

RNA-Seq primer

Sequencing: applications

Counting applications

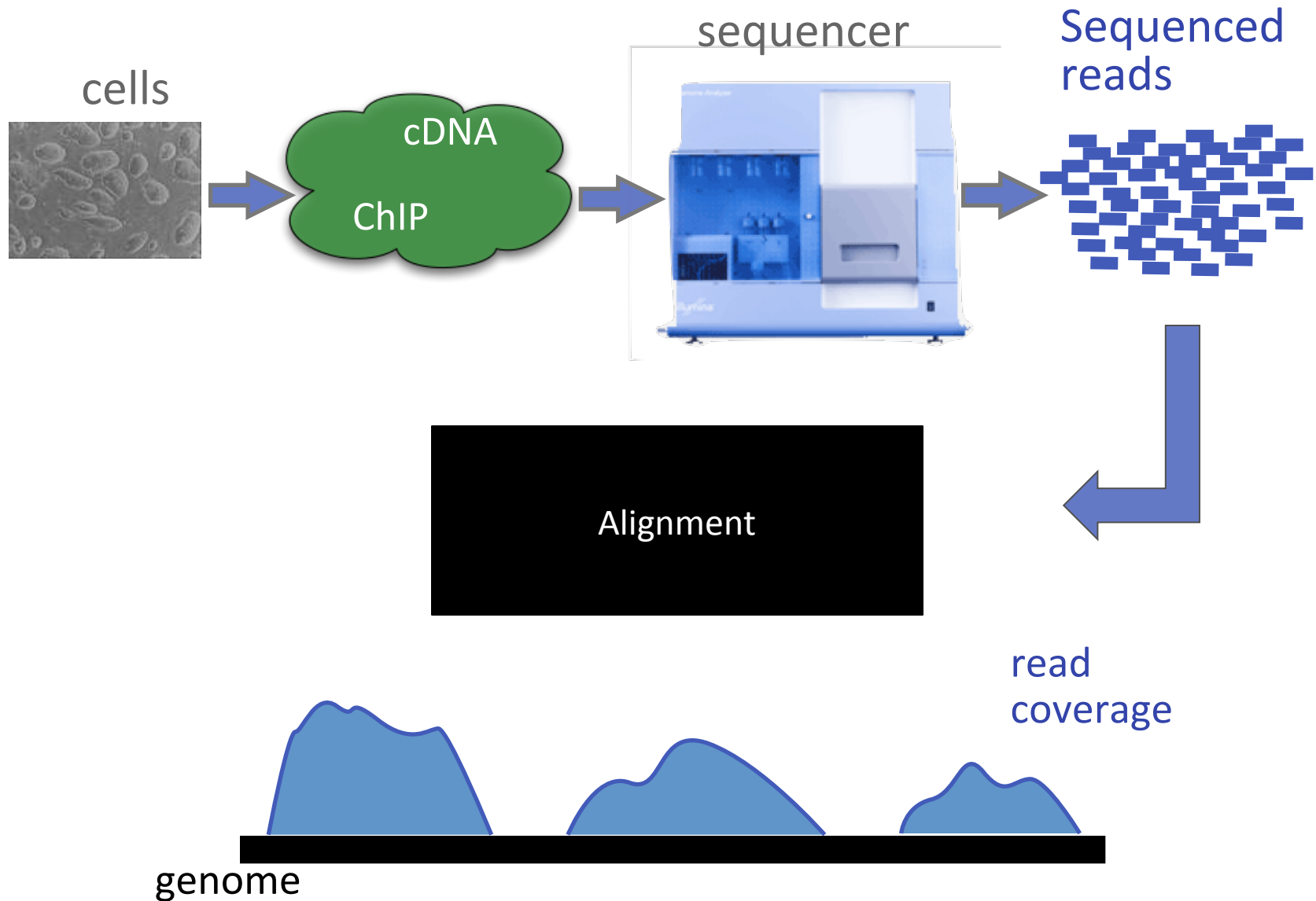
- Profiling
 - microRNAs
 - Immunogenomics
 - Transcriptomics
 - Epigenomics
 - Map histone modifications
 - Map DNA methylation
 - 3D genome conformation
 - Nucleic acid Interactions
- Cancer genomics
 - Map translocations, CNVs, structural changes
 - Profile somatic mutations
 - Genome assembly
 - Ancient DNA (Neanderthal)
 - Pathogen discovery
 - Metagenomics

Polymorphism/mutation discovery

- Bacteria
- Genome dynamics
- Exon (and other target) sequencing
- Disease gene sequencing
- Variation and association studies
- Genetics and gene discovery



Counting applications



Sequencing libraries to probe the genome

- RNA-Seq
 - Transcriptional output
 - Annotation
 - miRNA
 - Ribosomal profiling
- ChIP-Seq
 - Nucleosome positioning
 - Open/closed chromatin
 - Transcription factor binding
- CLIP-Seq
 - Protein-RNA interactions
- Hi-C
 - 3D genome conformation


RNA-Seq libraries I: “Standard” full-length

- “Source: intact, **high qual.** RNA (polyA selected or ribosomal depleted)
- RNA → cDNA → sequence
- Uses:
 - Annotation. Requires high depth, paired-end sequencing. ~50 mill
 - Gene expression. Requires low depth, single end sequence, ~ 5-10 mill
 - Differential Gene expression. Requires ~ 5-10 mill, at least 3 replicates, single end

RNA-Seq libraries II: End-sequence libraries

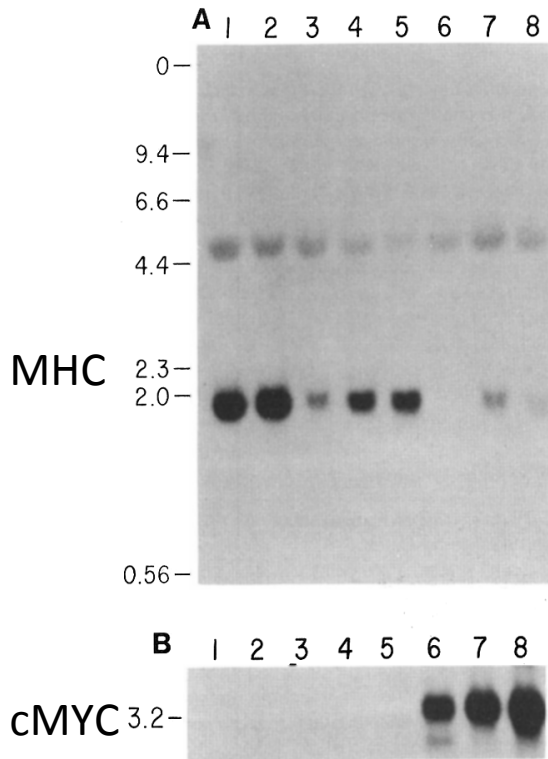
- Target the start or end of transcripts.
- Source: End-enriched RNA
 - Fragmented then selected
 - Fragmented then enzymatically purified
- Uses:
 - Annotation of transcriptional start sites
 - Annotation of 3' UTRs
 - Quantification and gene expression
 - Depth required 3-8 mill reads
 - Low quality/quantity (single cell) RNA samples

Analysis of counting data requires 3 broad tasks

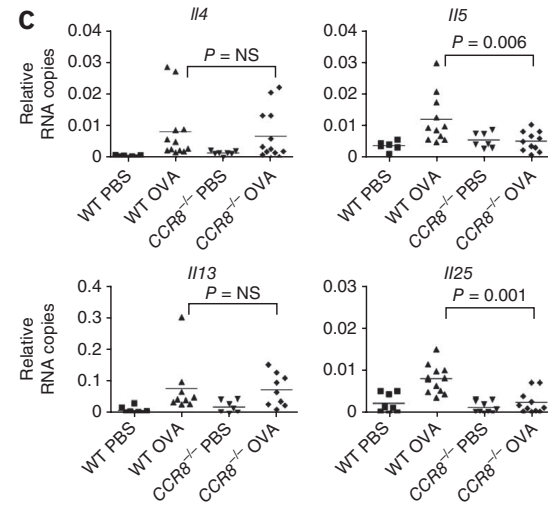
- Read mapping (alignment) or assembly: Finding what generated the reads
 - **Computationally intensive** 
- Quantification:
 - Transcript relative abundance estimation
 - Determining whether a gene is expressed
 - Normalization
 - Finding genes/transcripts that are differentially represented between two or more samples
 - **Computationally and statistically intensive**
- Data analysis:
 - Cross sample comparison
 - Feature enrichment
 - **Statistically intensive**

Where are we?

80s

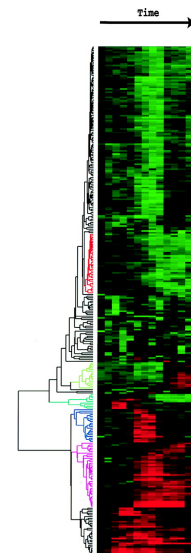


90s



qPCR

Islan et al Nat. Imm. 2011

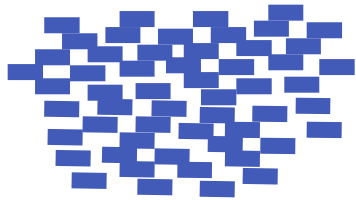


microarrays

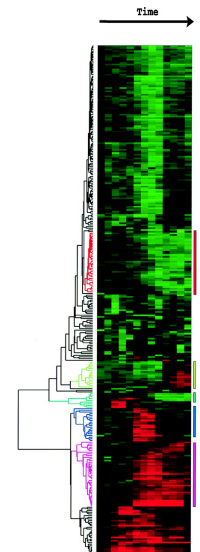
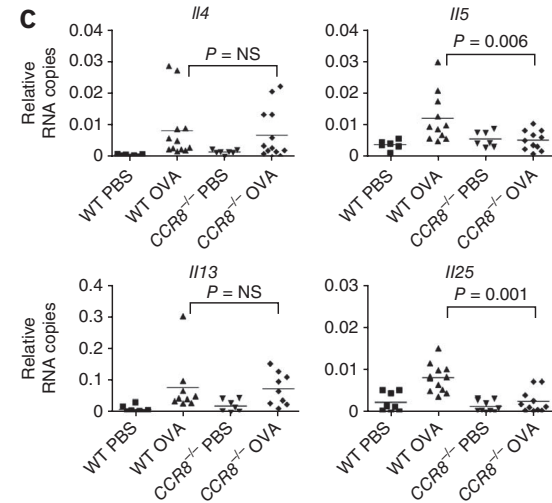
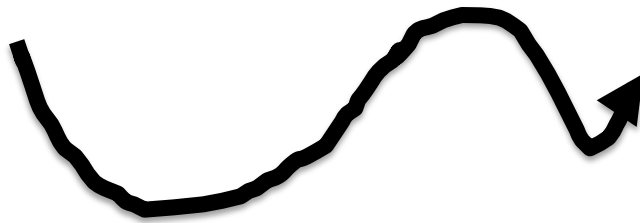
Biology slowly becoming a “big data” science

2010s

Sequenced
reads



Millions-billions



Statistical methods are deeply embedded – two concepts

Multiple testing problems
Modeling count data

“One” slide probability review – Experimental data

- An experimental design consists of a choice of populations and a measurable property
 - Two or more populations:
 - Cell types
 - Developmental times
 - Affected / not affected individuals
 - WT / KO / KD
 - Measurements
 - # of cells
 - Gene expression
 - Fluorescence

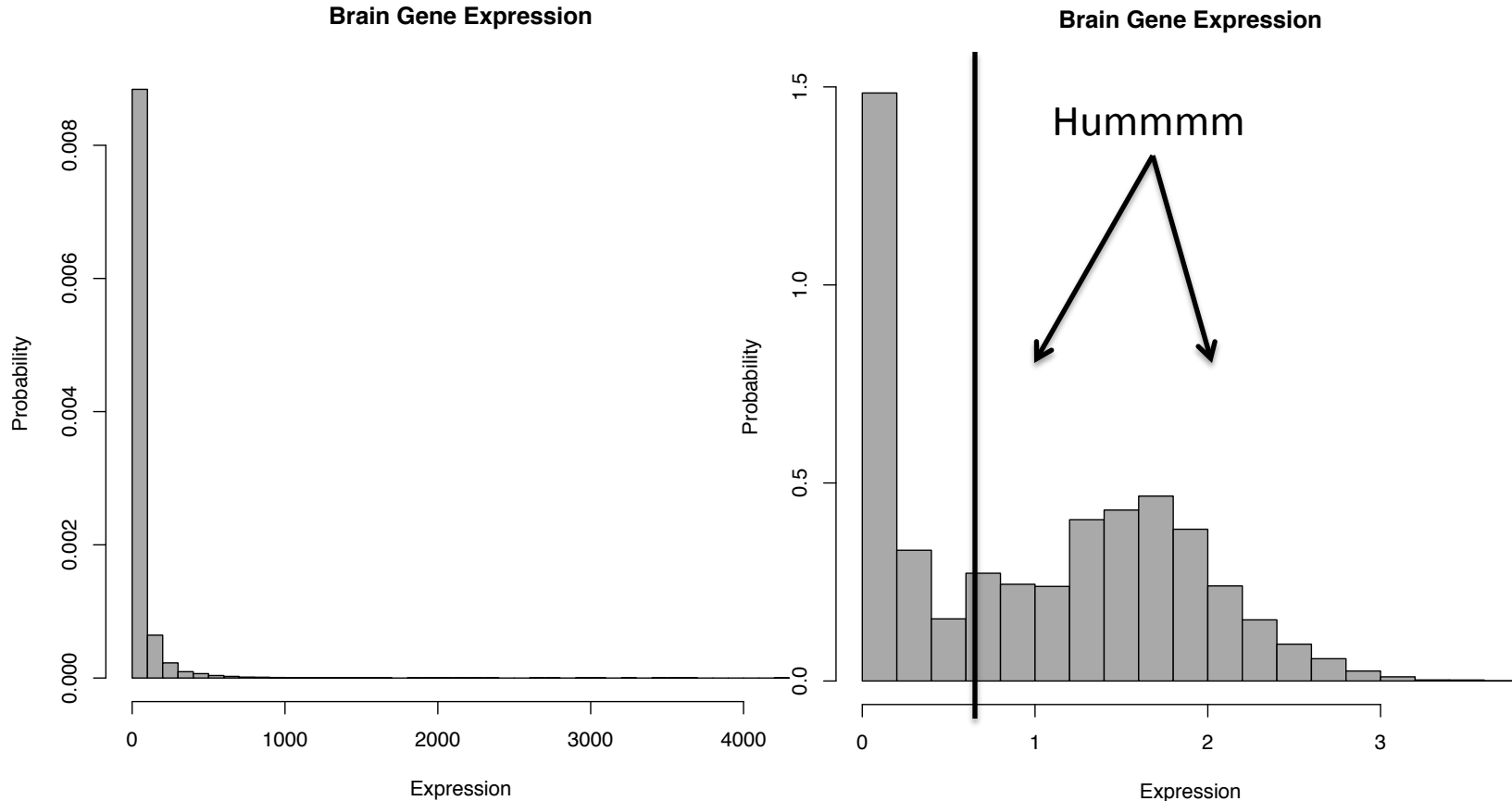
“One” slide probability review – Measurements

- Measurements are what we do statistics on. We can usually look for “outliers”. Things out of the ordinary
- When we do comparisons we use “test statistics” built on the original measurements, similarly we look for values of the test statistics that are “out of the ordinary”
- What is an “outlier”? When is a measurement or a test statistic “out of the ordinary”?

“One” slide probability review – Distributions

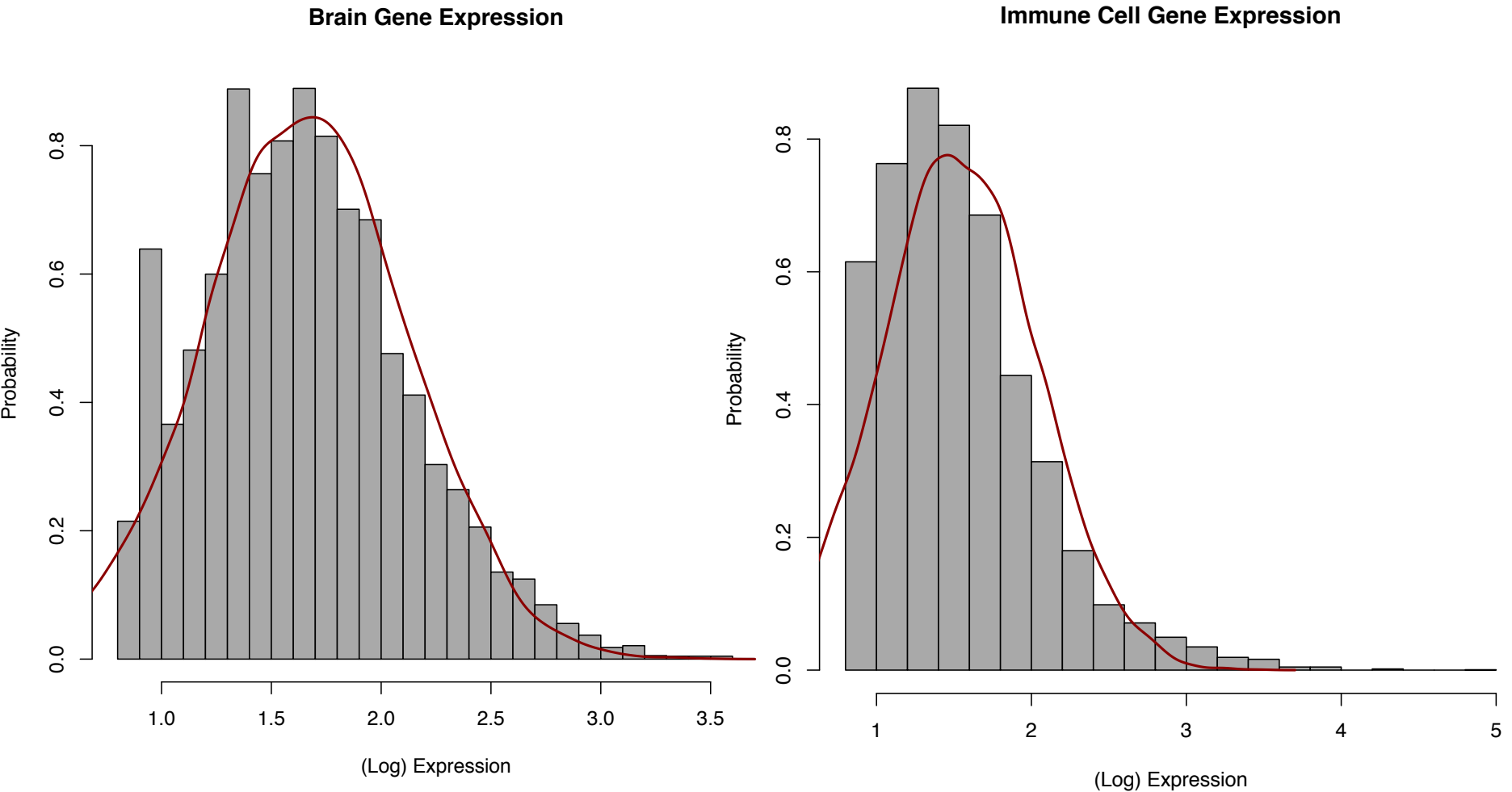
- Measurements values follow a “distribution”, that is, some values are more likely than others. The probability of observing a value is described by the measurement “distribution”
- The number of mRNAs per cell for each gene ranges from a handful to thousands of copies. Overall, the “distribution” of expression is similar between cells
- “Shotgun” sequencing a genome results in a roughly uniform coverage of the original genome. In many cases we care about the “number of reads that land on a given region” the distribution of this counts have important properties

“One” slide probability review – Distributions (cont.)



Gene expression is generally “bi-modal” if you measure expression of ALL genes in the genomes many genes are expressed, and many are not

“One” slide probability review – Distributions (cont.)

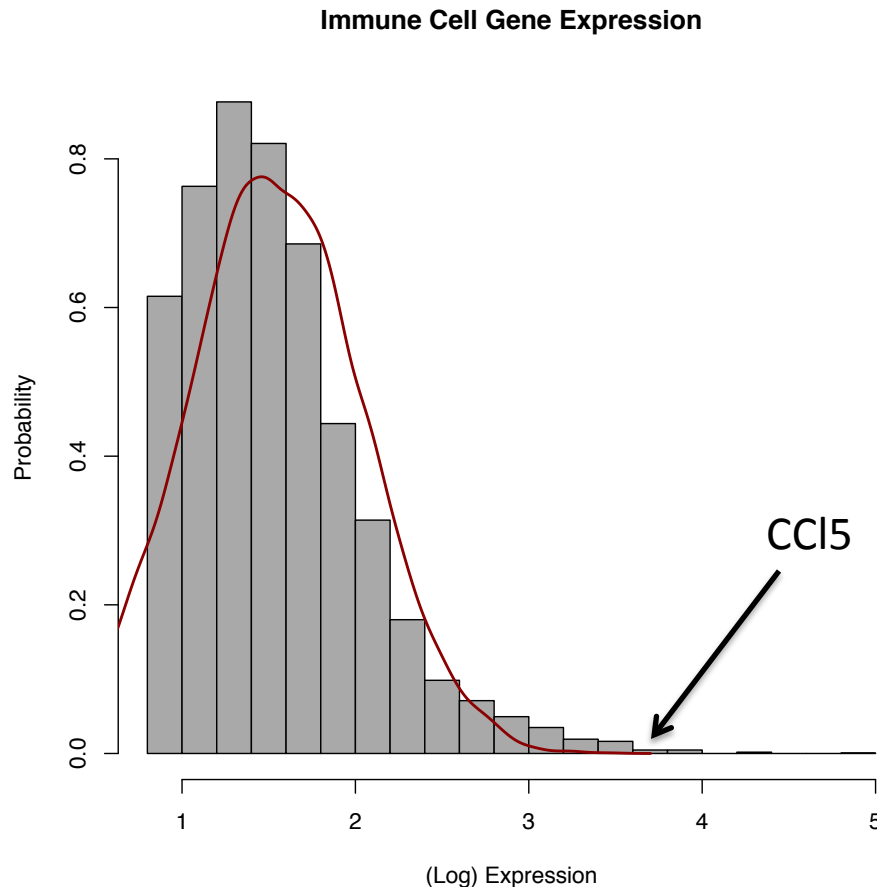


— Normal distribution with mean and variance estimated from gene exp. data

Gene expression distributes roughly log-normal

“One” slide probability review – p-values

You find that CCL5 is a critical chemokine, it recruits monocytes and T-Cells to sites of inflammation. Given that it is secreted by non-abundant cells. You hypothesize that it must be one of the most expressed proteins in the cell you found it.



- You eagerly purify your favorite cells and isolate and sequence its RNA.
- Is CCL5 a top expressed gene?
- CCL5 is expressed at 3,500 units!
- Can I say that my hypothesis is correct and even compute a significance for it?
- The null hypothesis: CCL5 is an averagely expressed gene. If I pick a gene randomly, what is the chance that it has an expression level higher than CCL5?

$$P(g > CCL5) = \frac{\# \text{ genes} > CCL5}{\text{all other genes}} = \frac{27}{8300} \approx 0.003$$

So I can reject my null hypothesis with confidence! CCL5 is highly expressed (no kidding!)

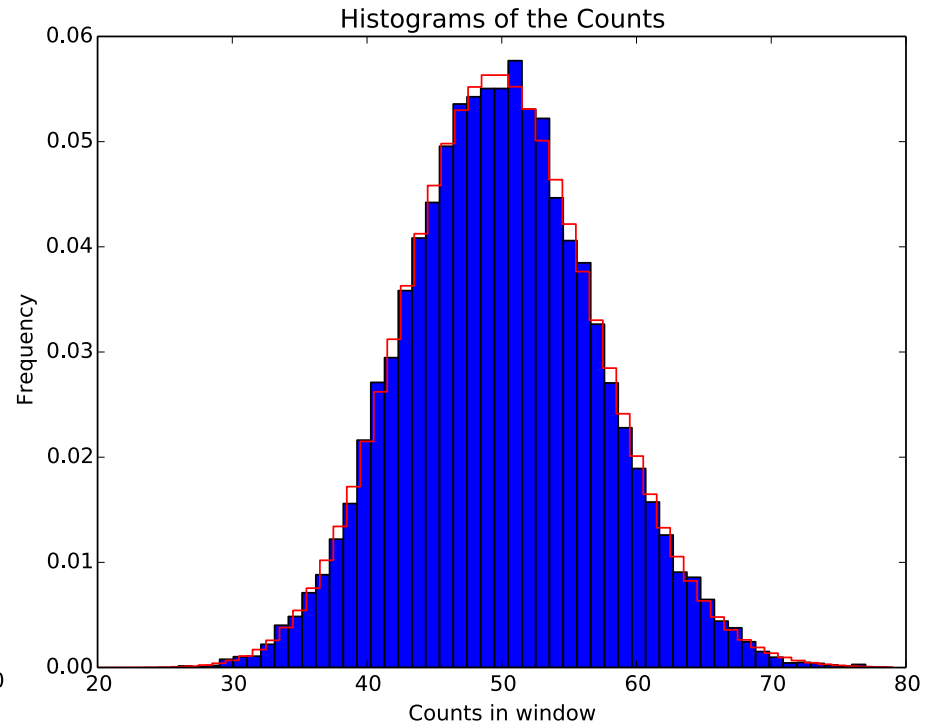
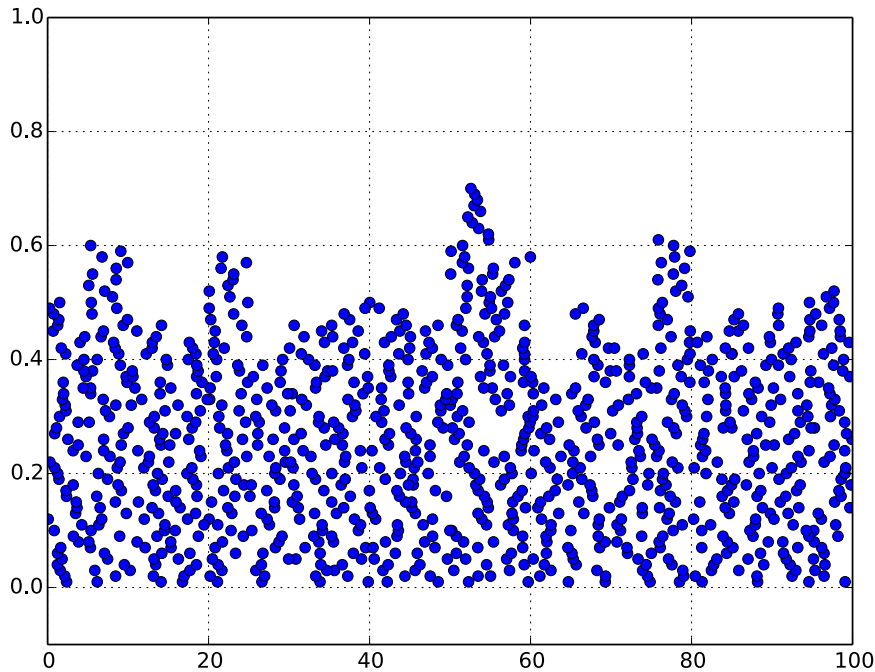
More interesting example: Modeling sequencing data

Lets try a simple simulation

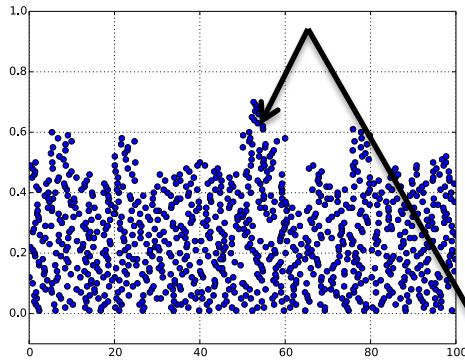
More interesting example: Modeling sequencing data

Suppose we are working on a genome that has 100 bp. We sequence deeply, how do reads map to the genome?

Now usually (e.g. ChIP-Seq) we are interested in knowing the **NUMBER** of reads that mapped to a specific location.

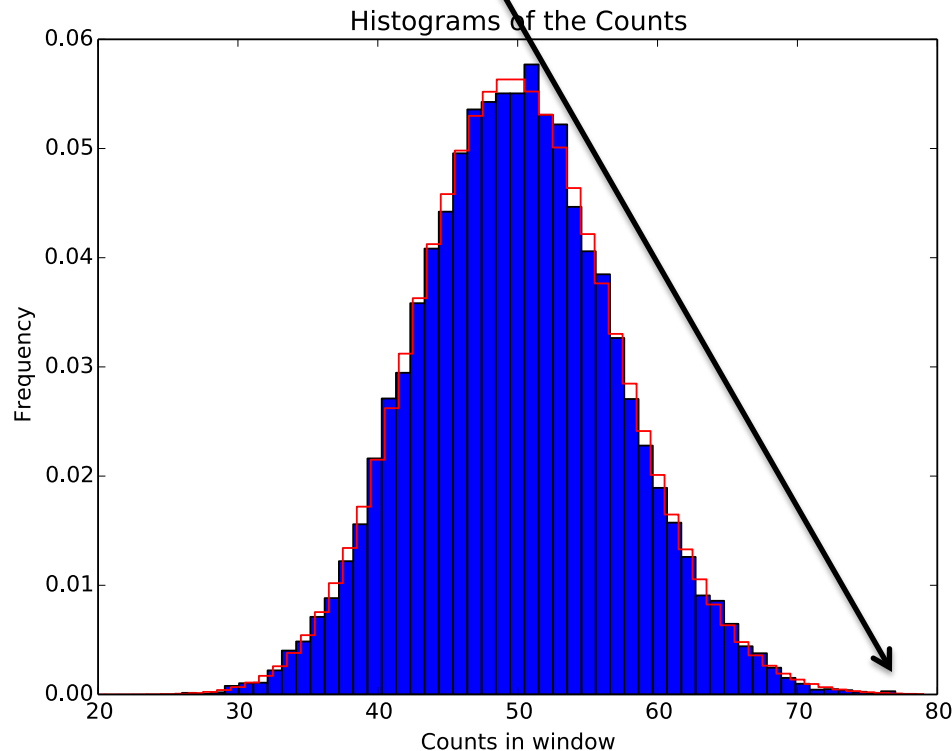


Two very different questions



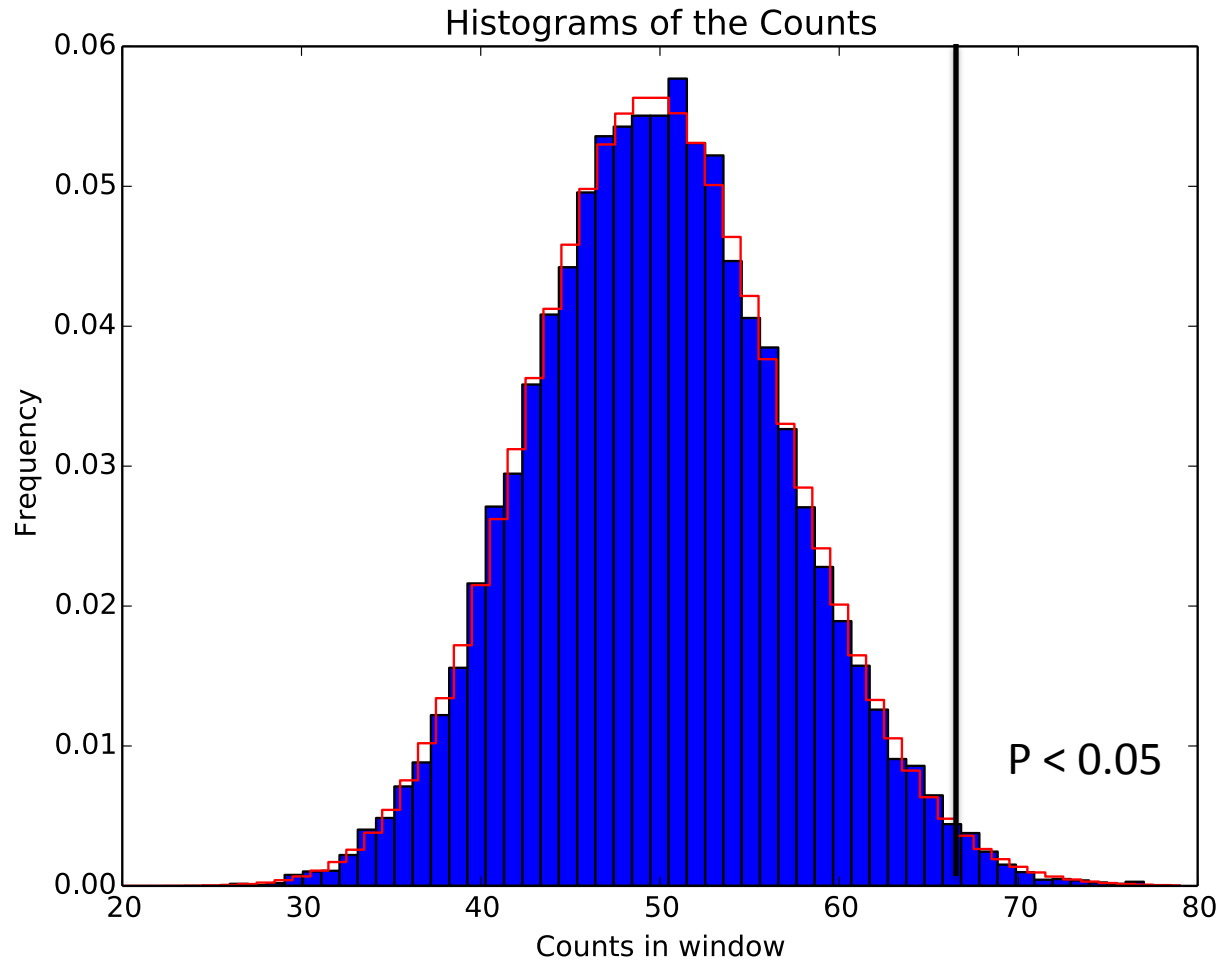
Promoter of myGene. Null hypothesis coverage of the gene is average, if I were to choose any other window how likely it is that I get this value?

$$P(\text{window with count} > \text{myGene promoter}) = 0$$



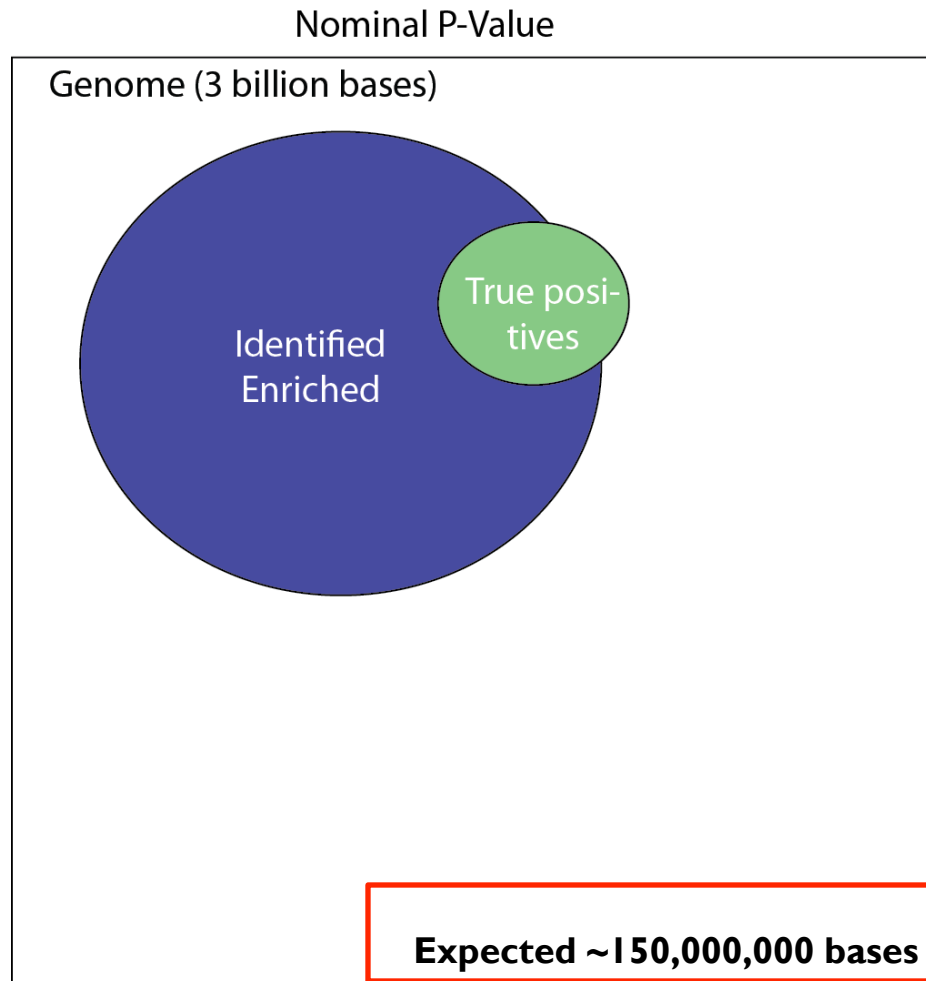
This is not the usual question. In an unbiased approach we want to DISCOVER windows that have greater coverage than expected under the null hypothesis. That is we want to identify OUTLIERS.

We can't use a nominal p-value any longer



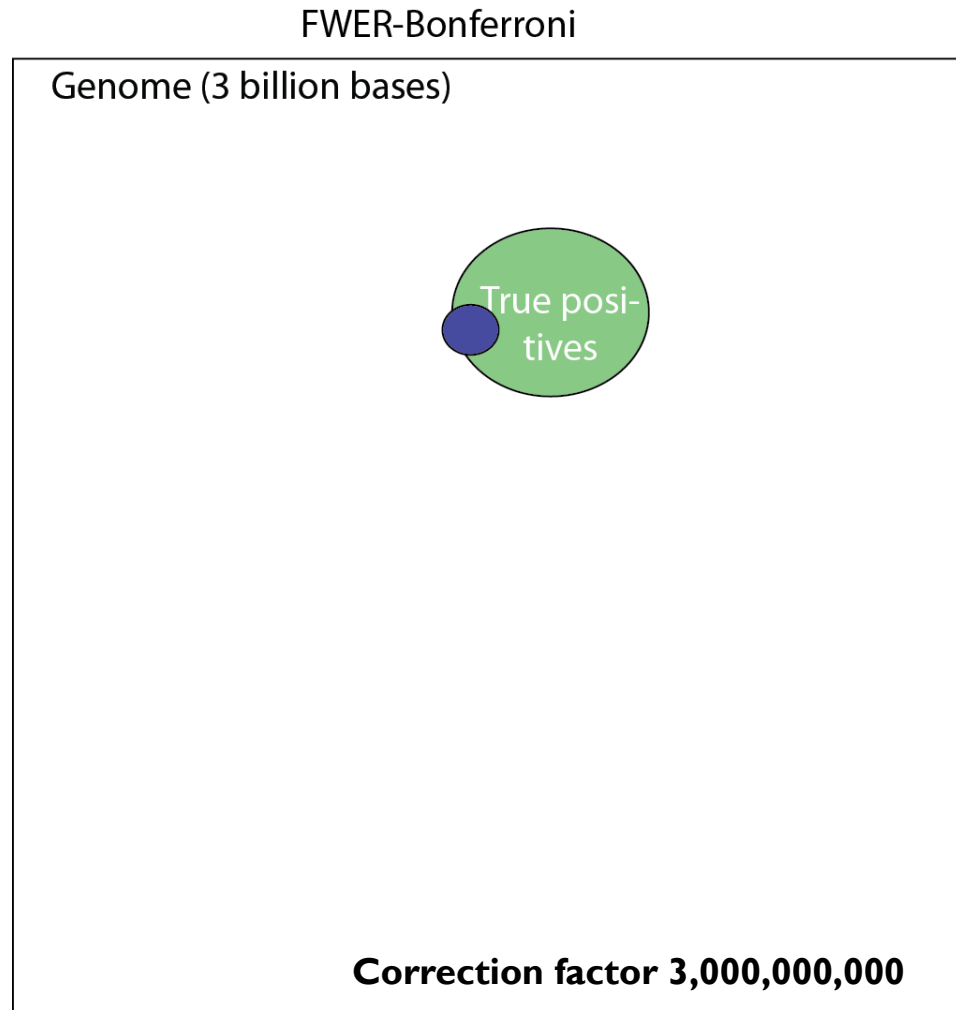
All will be noise!

The genome is large, many things happen by chance



We need to correct for multiple hypothesis testing

Bonferroni correction is way to conservative



Bonferroni corrects the number of hits but misses many true hits because its too conservative – How do we get more power?

How do we compute significance when we have this much data?

J. R. Statist. Soc. B (1995)
57, No. 1, pp. 289–300

Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

By YOAV BENJAMINI† and YOSEF HOCHBERG

Tel Aviv University, Israel

[Received January 1993. Revised March 1994]

Downloadable from: <http://garberlab.umassmed.edu/bootcamp.2015/BH.pdf>