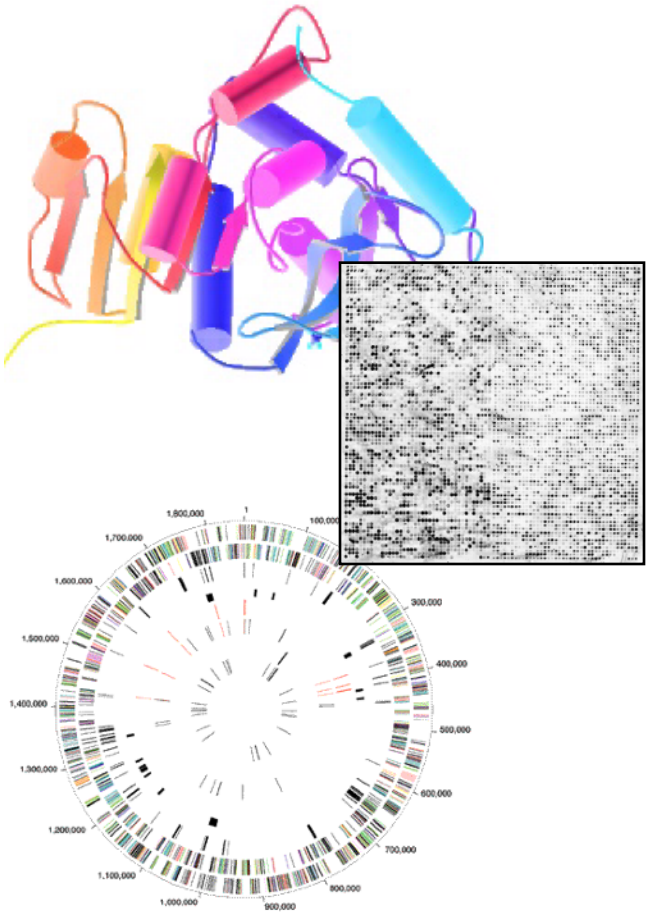
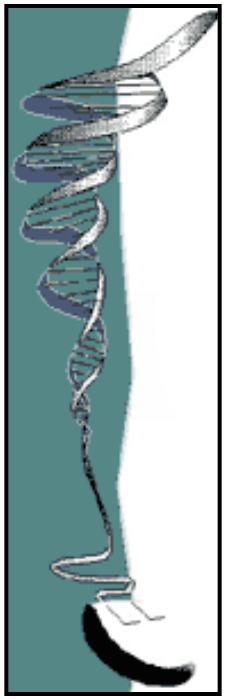
Biomed. Data Science:

# Basic RNA-seq & Chip-seq





Mark Gerstein, Yale University

gersteinlab.org/courses/452

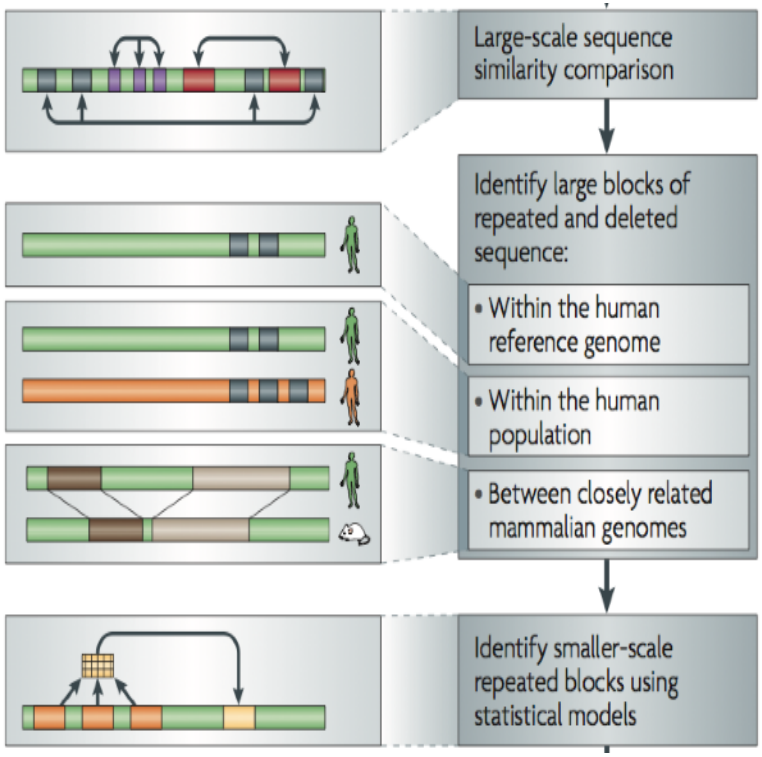
(last edit in spring ’19, pack #7)

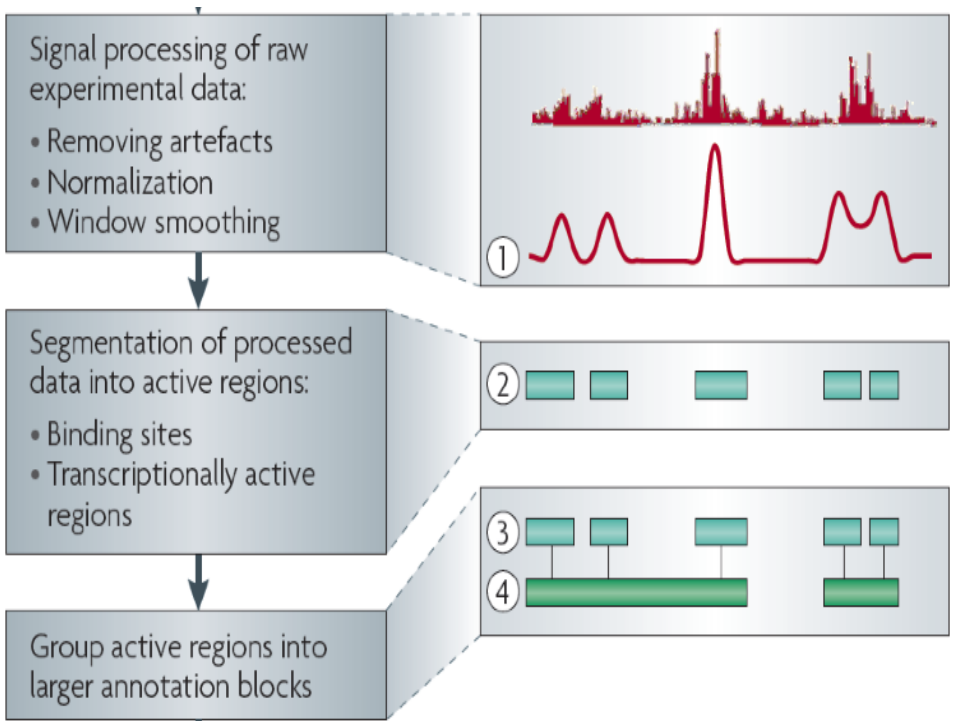
# Non-coding Annotations: Overview

Features are often present on multiple ”scale” (eg elements and connected networks)

Sequence features, incl. **Conservation**

**Functional Genomics**
**Chip-seq (Epigenome & seq. specific TF) and ncRNA & un-annotated transcription**

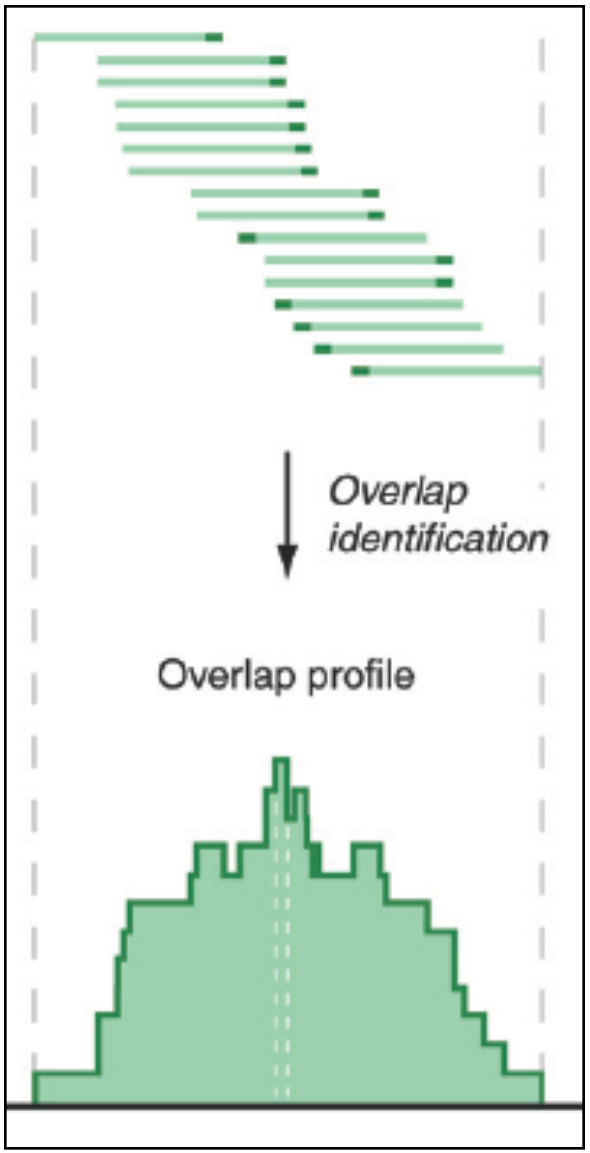



[*Nat. Rev. Genet.* (2010) 11: 559]

# Low-Level Data for RNA-seq & Chip-seq

```
@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhh]NYhhhhhhYIhhaZT[hYHNSPKXR
@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
+ILMN-GA001_3_208HWAAXX_1_1_111_879
hSWhRNJ\hFhLdhVOhAIB@NFKD@PAB?N?
```

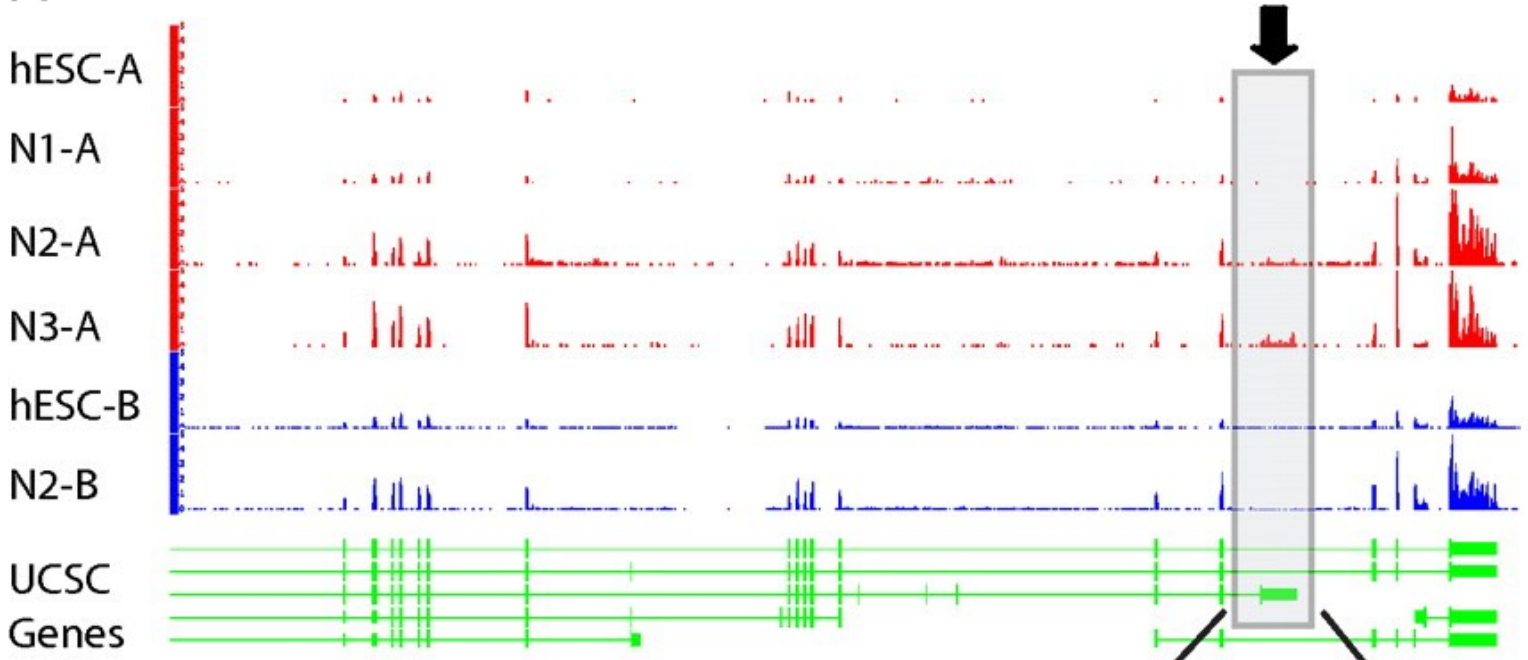


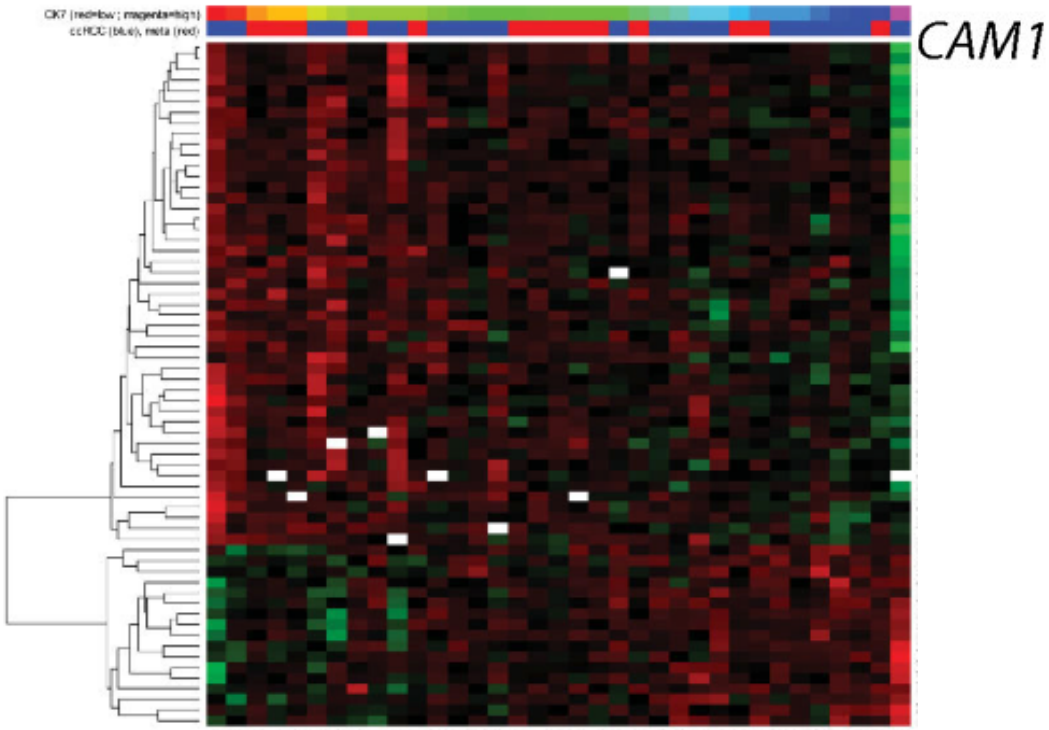

Reads (fasta)
+ quality scores (fastq)
+ mapping (BAM)

Reads => Signal (Intermediate file)

Accumulating @ >1 Pbp/yr (currently),
~20% of tot. HiSeq output

[*PLOS CB* 4:e1000158]




## Information from RNA-seq: Avg. signal at exons & TARs (RPKMs)

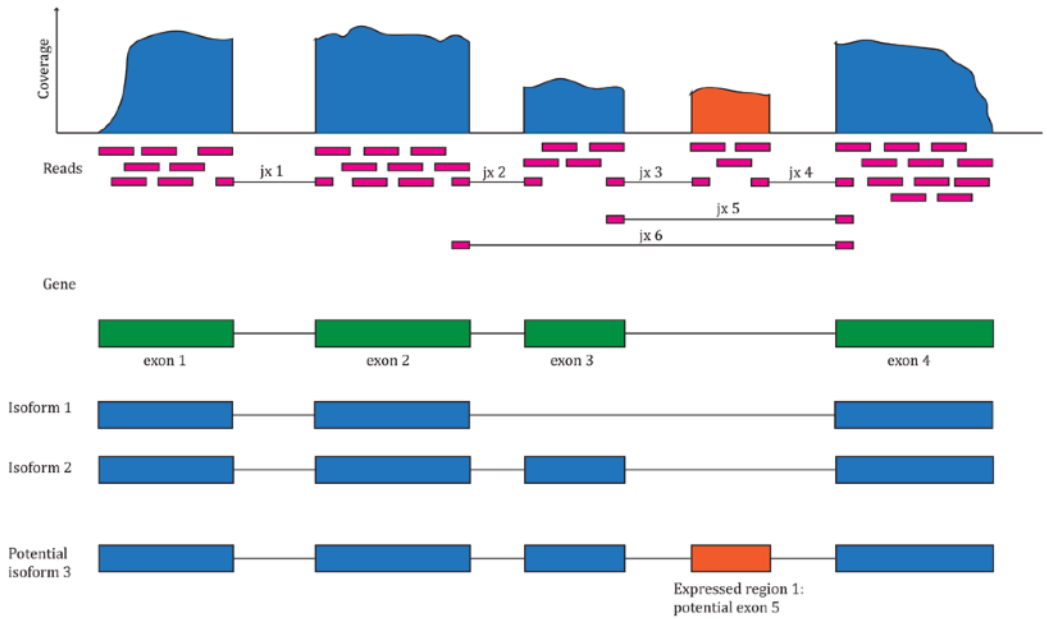
[*PNAS* 4:107: 5254 ; *IJC* 123:569]

## Activity Patterns

- RNA Seq. gives rise to activity patterns of genes & regions in the genome

# Transcript quantification

- Read counts (RC)
  - Summarized mapped reads to  gene, exon or isoform level




*F1000Research, 2017, 6:1558*

## Transcript quantification

The RC is roughly proportional to

- the effective length (exon) of the gene
- the total number of reads in the library

Question:

Gene A: RC=100

Gene B: RC=300

Expression of Gene A < Expression of Gene B?

# Transcript quantification

- FPKM /RPKM: (Reads/Fragments Per Kilobase Million)

$$\mathrm{FPKM}_i = \frac{X_i}{\left(\frac{\widetilde{l}_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\widetilde{l}_i N} \cdot 10^9$$

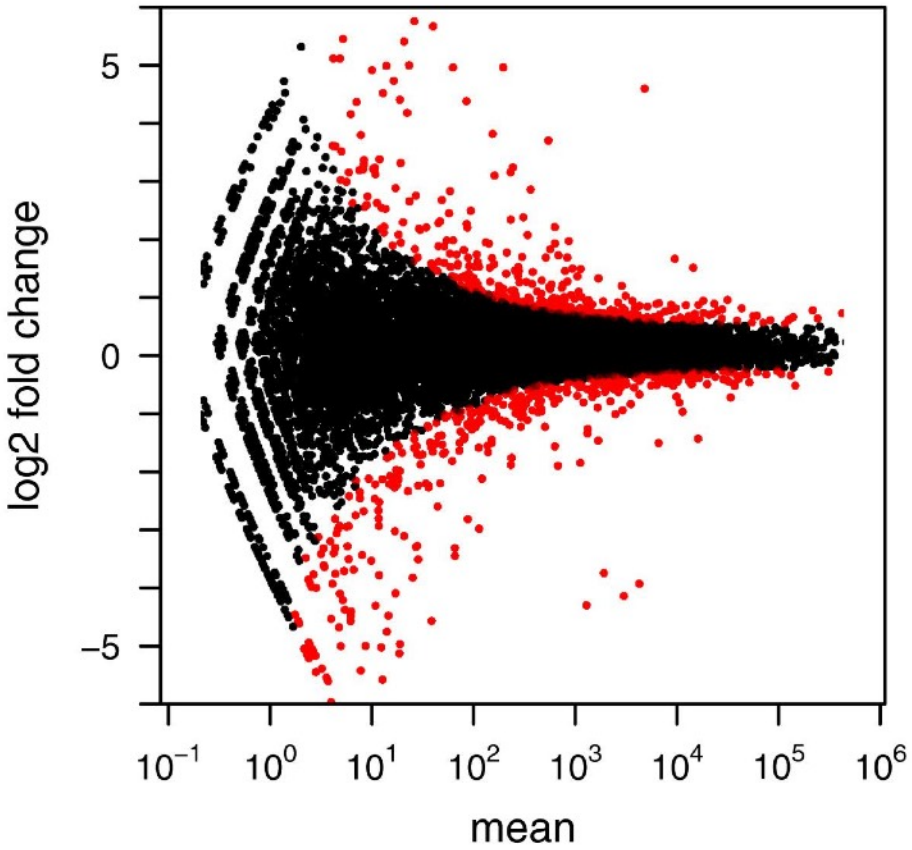
$i$.

$X_i$ : Counts of the mapped fragments (reads for RPKM) for gene (exon or isoform) $i$.

$\widetilde{l}_i$ : The effective length for gene (exon or isoform) $i$.

$N$ : Total mapped fragments (reads for RPKM).

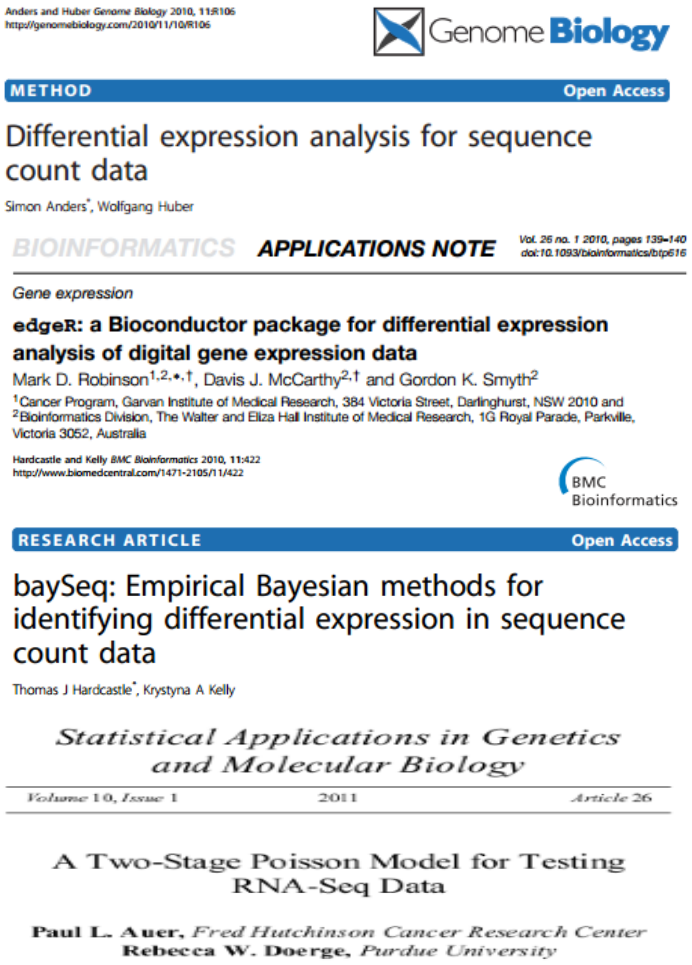
# Differential expression analysis




*Genome Biology, 2010 11:R106*

# Differential expression analysis: Count-based

1. **DESeq** -- based on negative binomial distribution
2. **edgeR** -- use an overdispersed Poisson model
3. **baySeq** -- use an empirical Bayes approach
4. **TSPM** -- use a two-stage poisson model



Anders and Huber *Genome Biology* 2010, 11:R106
http://genomebiology.com/2010/11/10/R106

Genome Biology

METHOD Open Access

Differential expression analysis for sequence count data

Simon Anders*, Wolfgang Huber

BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 1 2010, pages 139–140 doi:10.1093/bioinformatics/btp616

*Gene expression*

**edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**

Mark D. Robinson[1,2,*,†], Davis J. McCarthy[2,†] and Gordon K. Smyth[2]

[1]Cancer Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010 and [2]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia

Hardcastle and Kelly *BMC Bioinformatics* 2010, 11:422
http://www.biomedcentral.com/1471-2105/11/422

BMC Bioinformatics

RESEARCH ARTICLE Open Access

baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data

Thomas J Hardcastle*, Krystyna A Kelly

*Statistical Applications in Genetics and Molecular Biology*

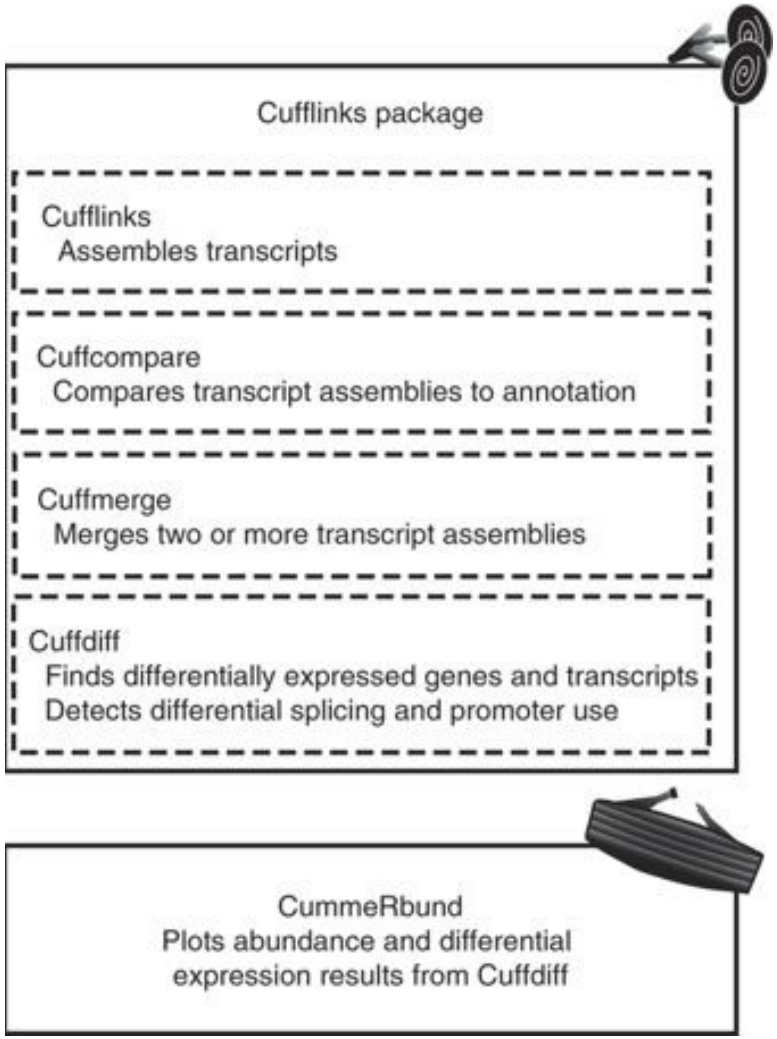
*Volume* 10, *Issue* 1 2011 *Article* 26

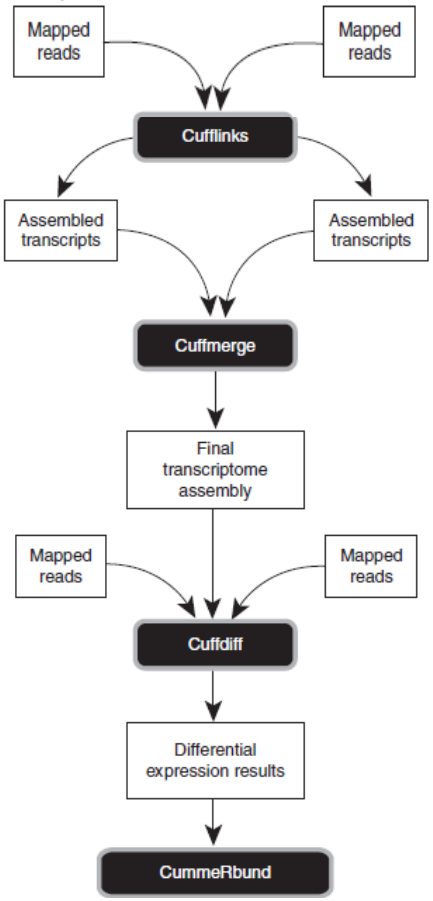
A Two-Stage Poisson Model for Testing RNA-Seq Data

**Paul L. Auer**, *Fred Hutchinson Cancer Research Center*
**Rebecca W. Doerge**, *Purdue University*

# Differential expression analysis: RPKM/FPKM-based

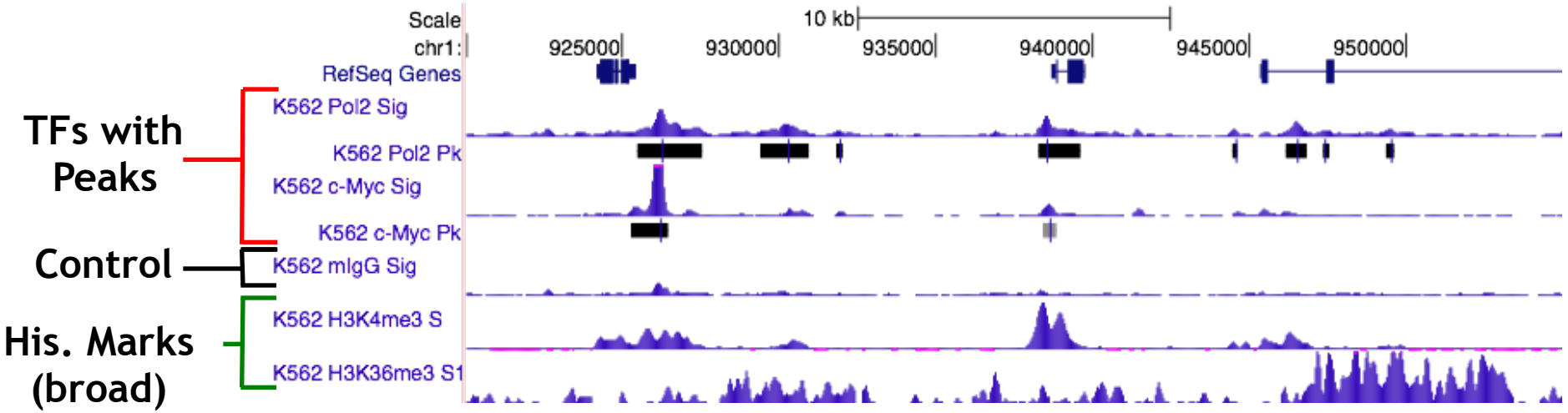
## Cufflinks & Cuffdiff




*Nature Protocols 7, 562-578 (2012)*

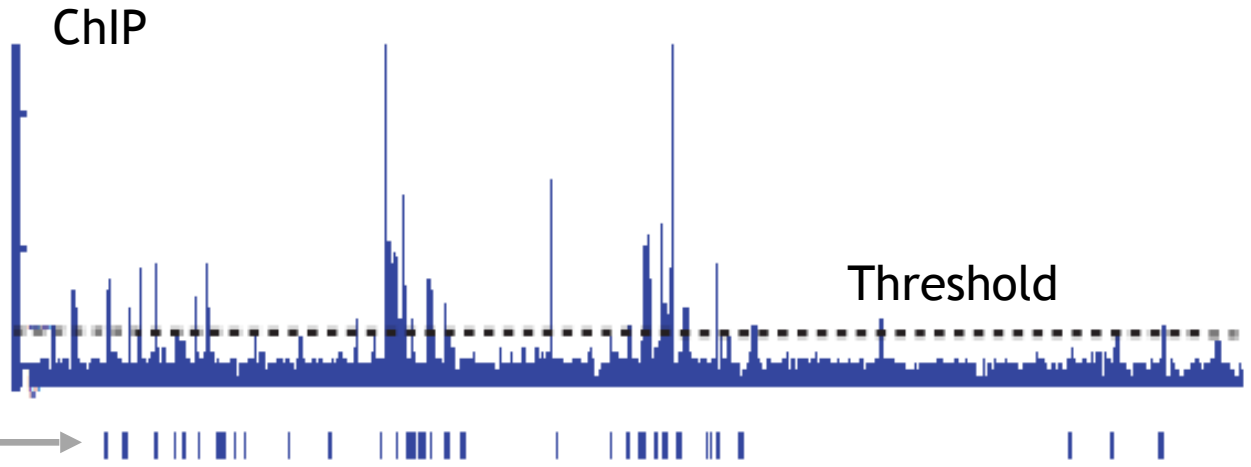
# Information from Chip-seq



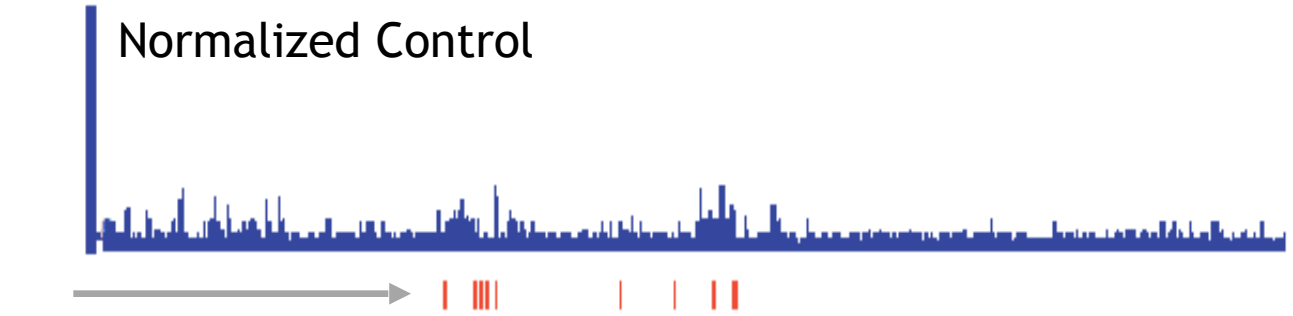

[*Science* 330: 1775
+ ENCODE Data Sources
TFs & Control: Yale
HMs: UW & Broad ]

# Summarizing the Signal: "Traditional" ChipSeq Peak Calling

- Generate & threshold the signal profile to identify candidate target regions
  - Simulation (PeakSeq),
  - Local window based Poisson (MACS),
  - Fold change statistics (SPP)



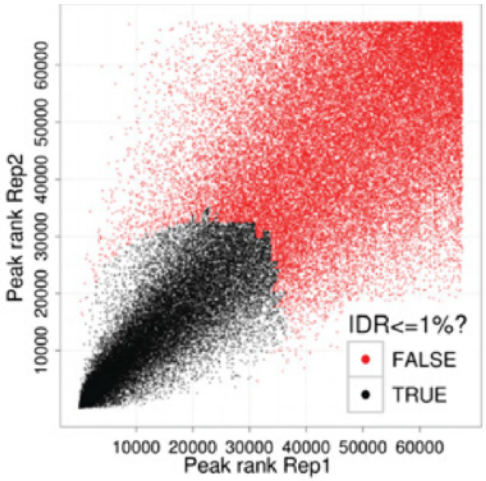

- Score against the control

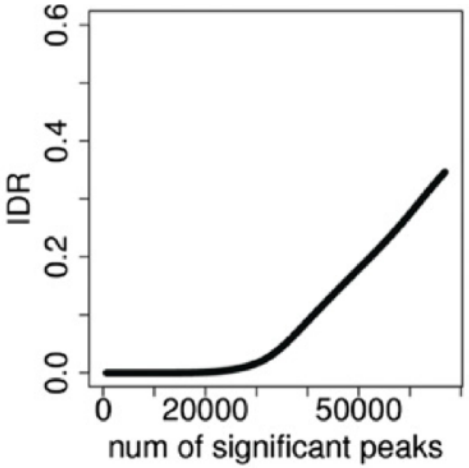



[Rozowsky et al. ('09) *Nat Biote*

# The irreproducible discovery rate (IDR)

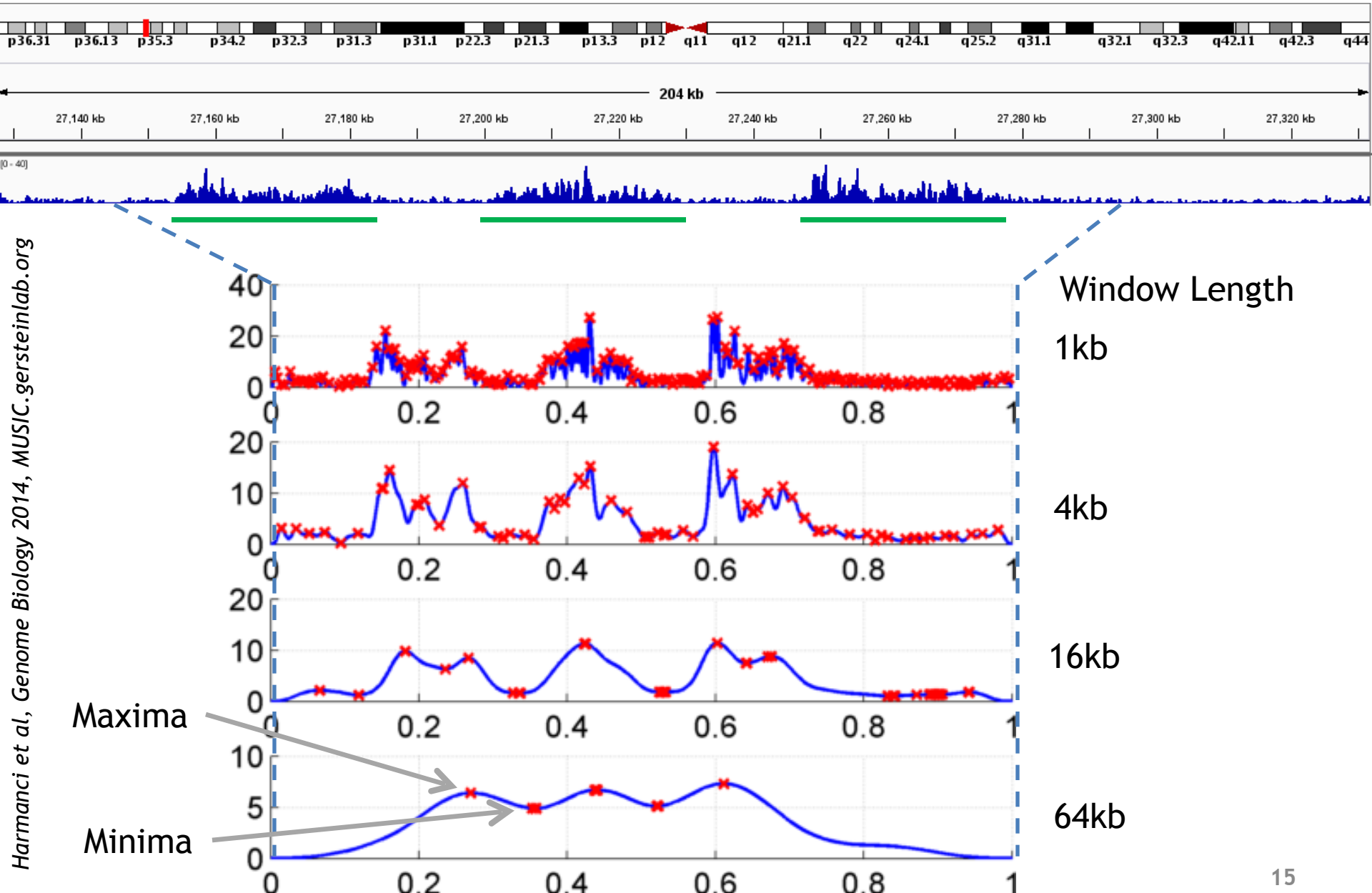
- Unified approach to measure the reproducibility of findings identified from replicate high-throughput experiments.
- Idea : call peaks with low cutoff and classify peaks as reproducible or not (bivariate rank distributions) based on overlap of ranked peaks (consistency)



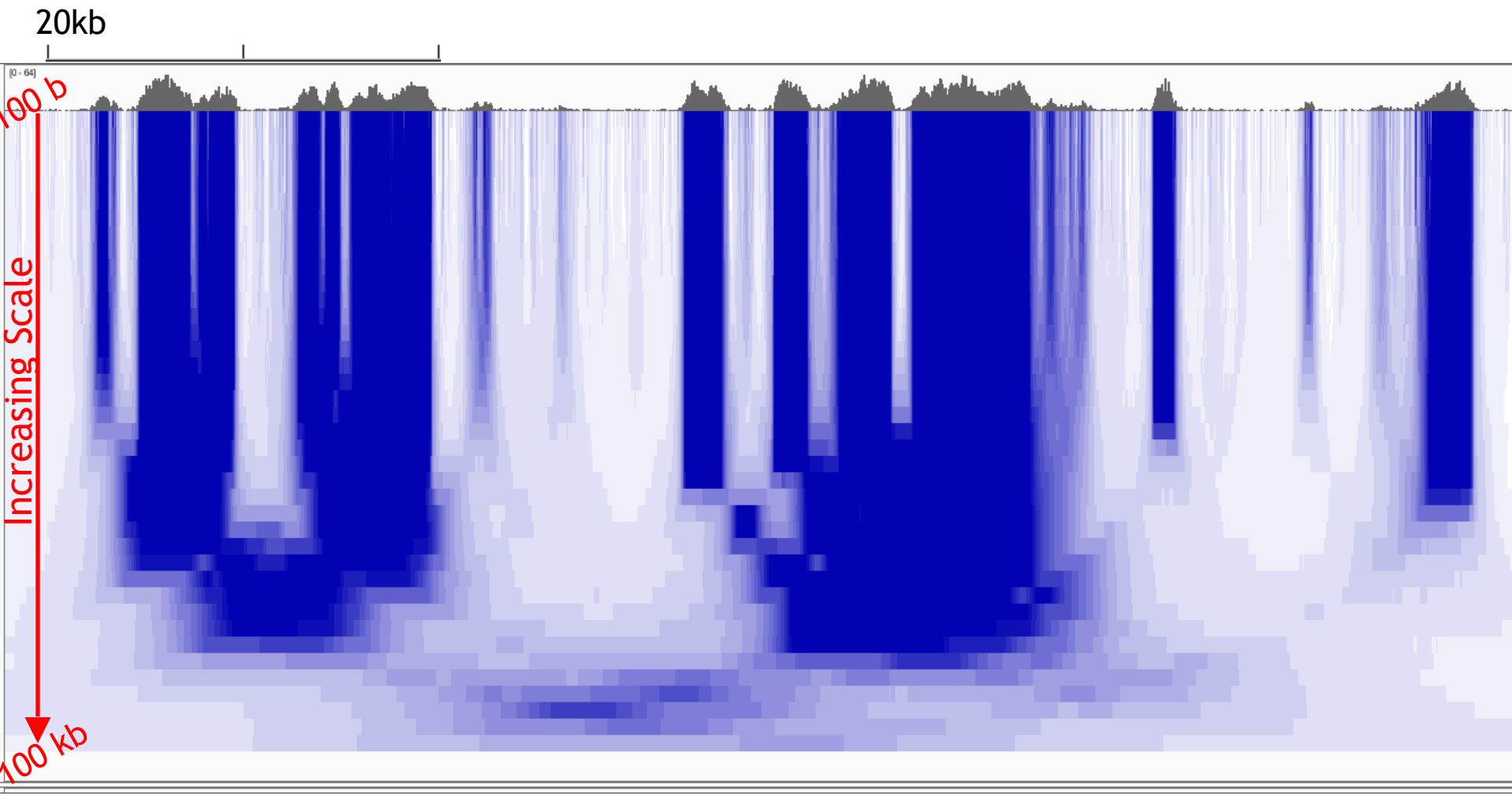




*Genome Research 2012, 22:1813-1831*

# Multiscale Analysis, Minima/Maxima based Coarse Segmentation

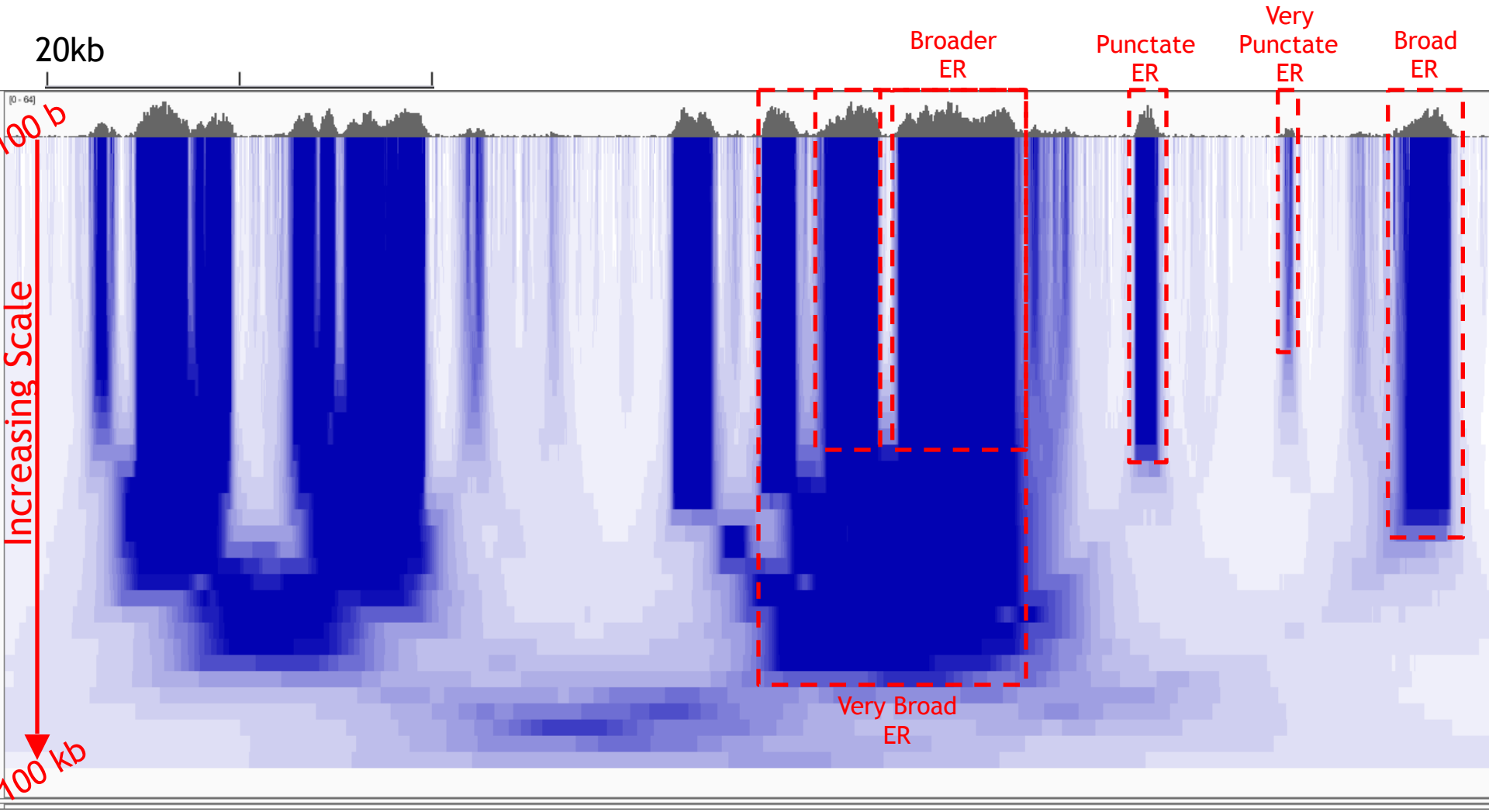


# Multiscale Decomposition



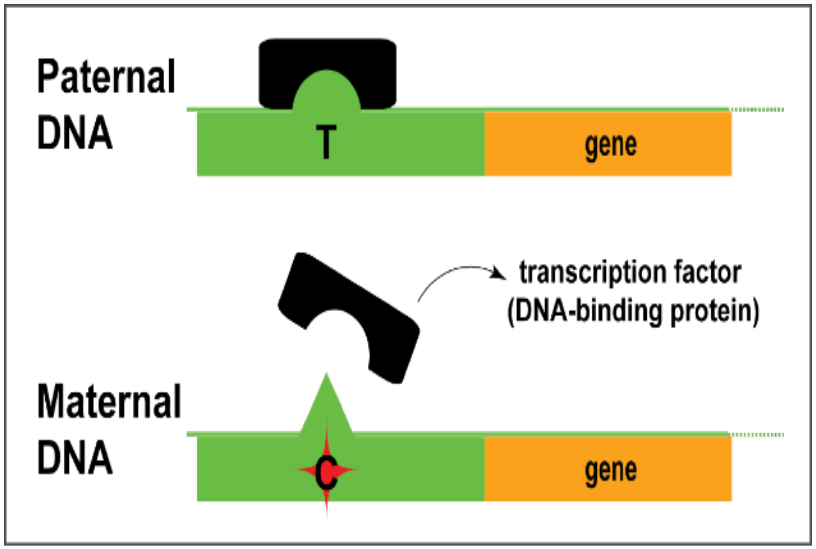

[Harmanci *et al*, *Genome Biol*. ('14)]

# Multiscale Decomposition

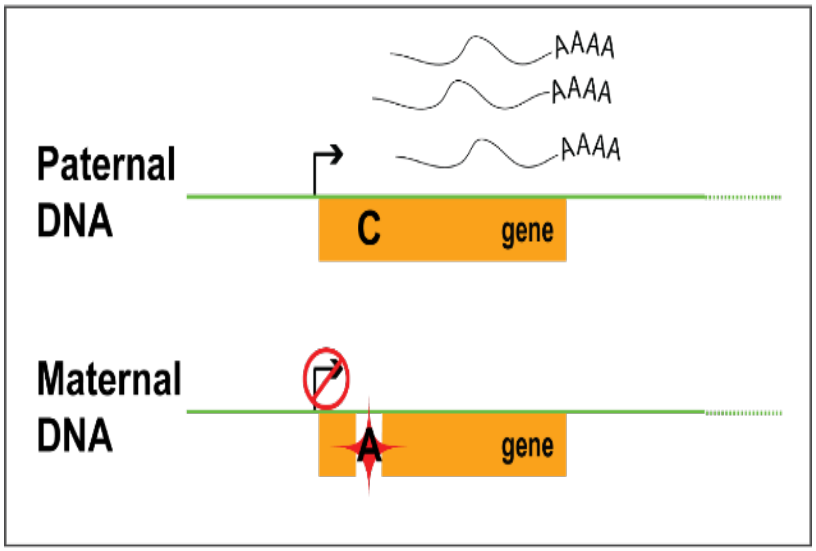



[Harmanci *et al*, *Genome Biol*. ('14)]

# Allele-specific binding and expression




Genomic variants
affecting allele-specific behavior
e.g. allele-specific binding
(ASB)



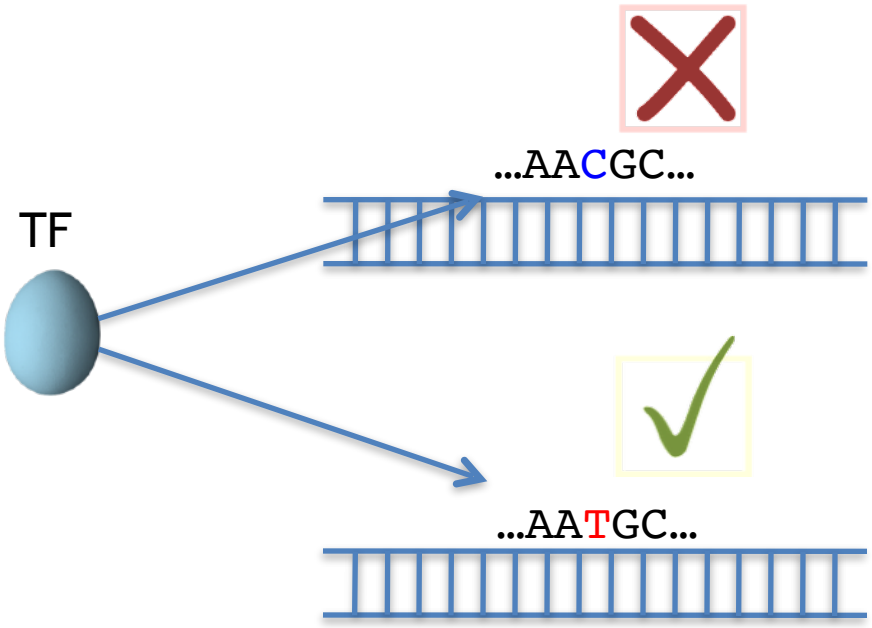

e.g. allele-specific expression
(ASE)

# Inferring Allele Specific Binding/Expression using Sequence Reads

**RNA/ChIP-Seq Reads**

```
ACTTTGATAGCGTCAATG
 CTTTGATAGCGTCAATGC
 CTTTGATAGCGTCAACGC
   TTGACAGCGTCAATGCAC
    TGATAGCGTCAATGCACG
      ATAGCGTCAATGCACGTC
       TAGCGTCAATGCACGTCG
          CGTCAACGCACGTCGGGA
           GTCAATGCACGTCGAGAG
             CAATGCACGTCGGGAGTT
              AATGCACGTCGGGAGTTG
                 TGCACGTTGGGAGTTGGC
```

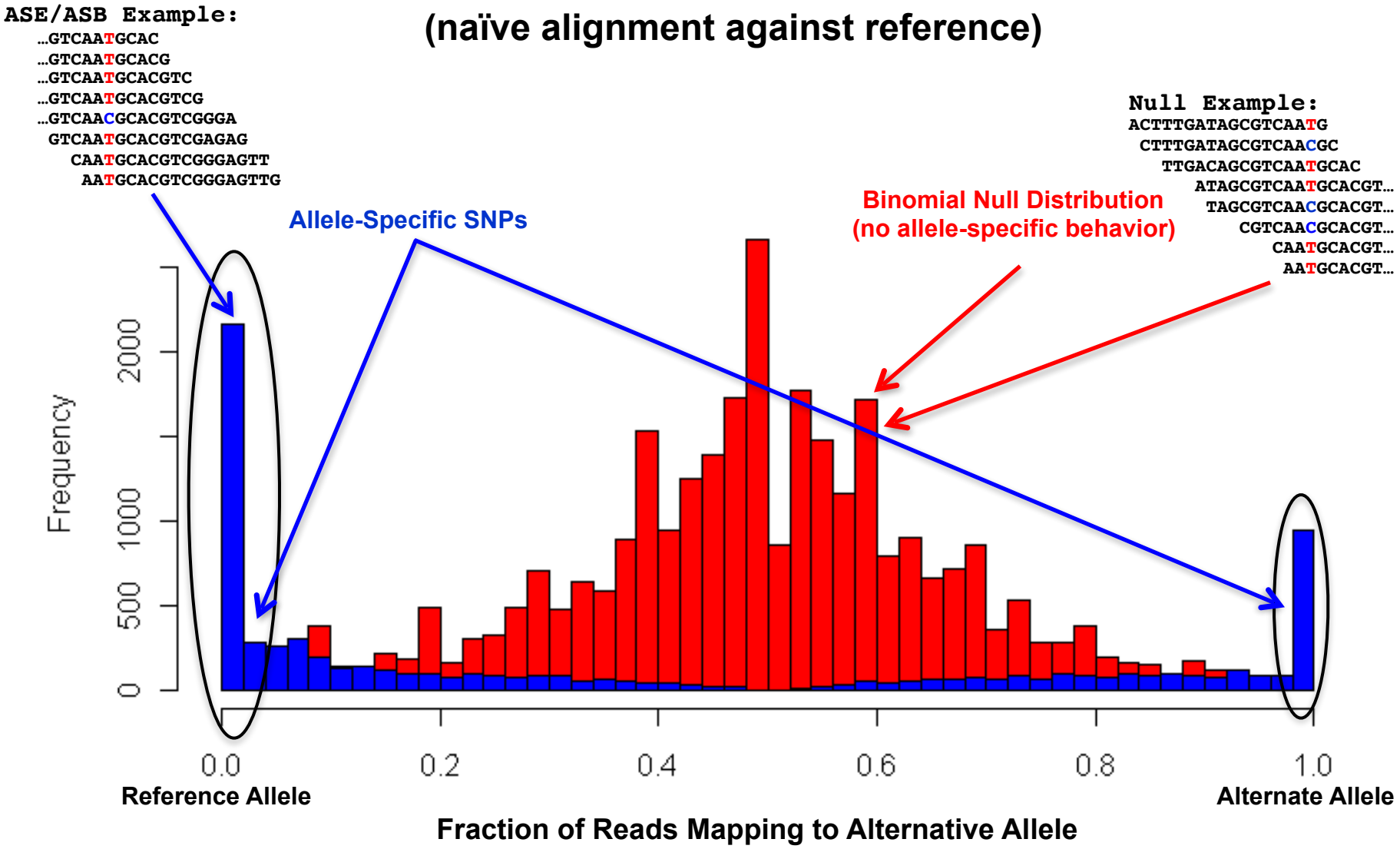
```
10 x T
 2 x C
```




**Haplotypes with a Heterozygous Polymorphism**

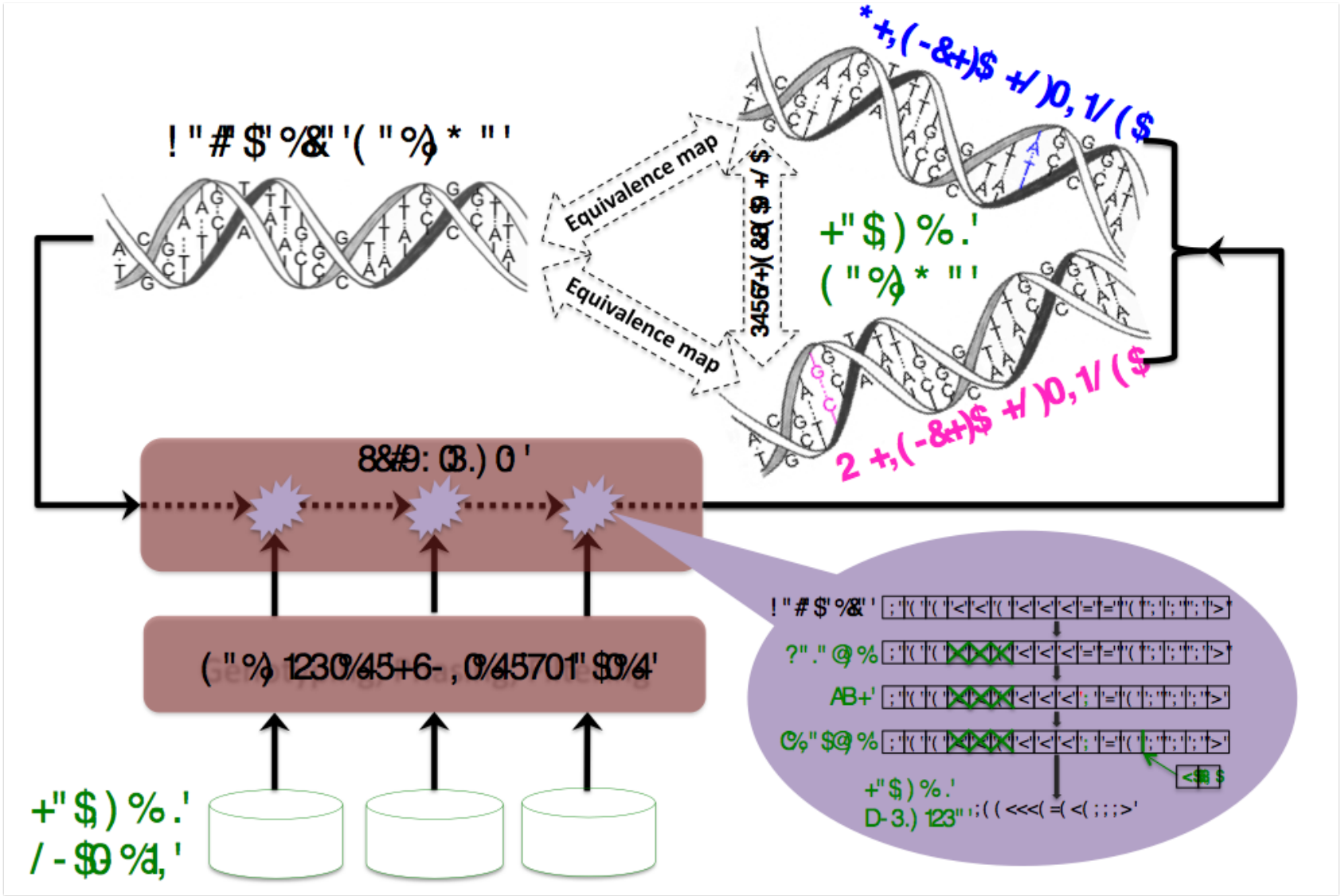
Interplay of the annotation and individual sequence variants

# Many Technical Issues in Determining ASE/ASB: Reference Bias (naïve alignment against reference)




[Rozowsky et al., MSB ('11)]

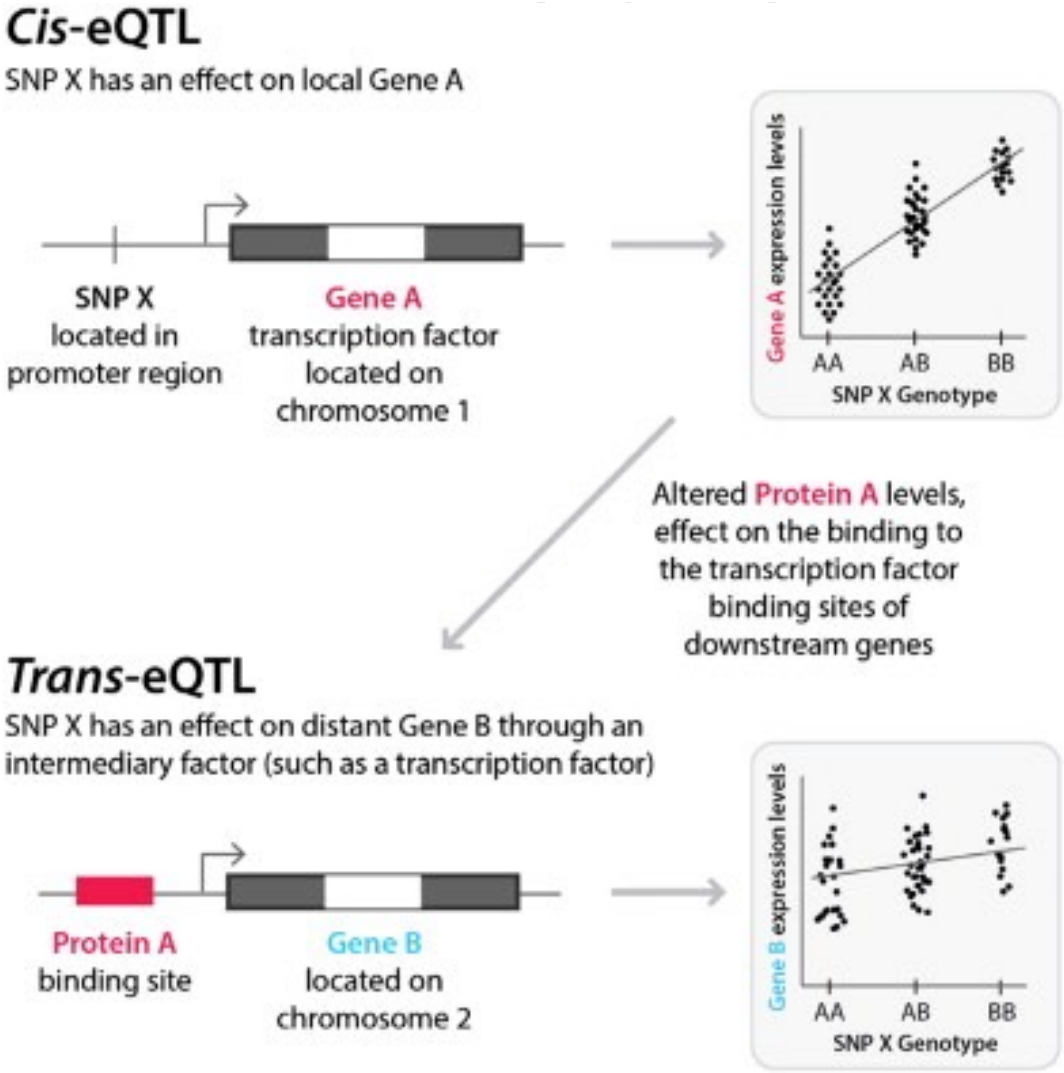
# Construction of a Personal Diploid Genome & Transcriptome



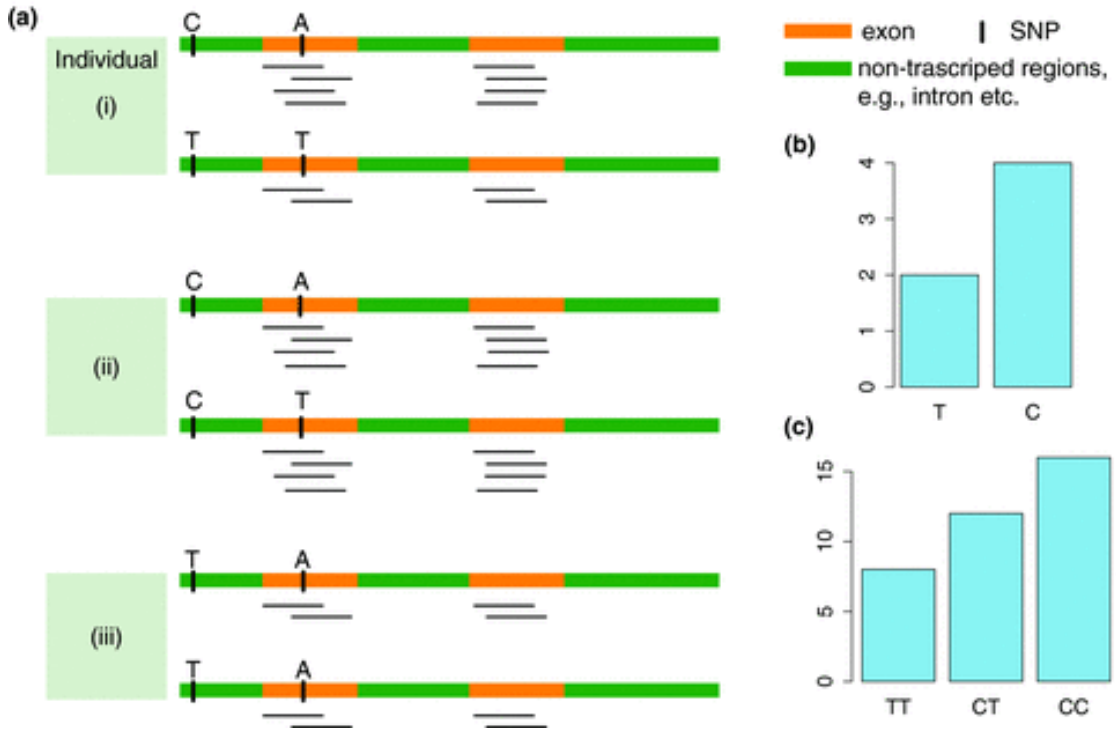

[alleleseq.gersteinlab.org]

[Rozowsky et al., MSB ('11)]

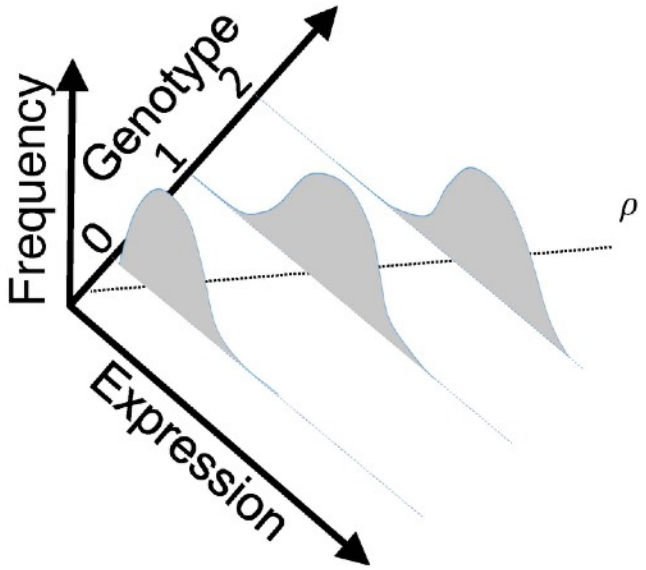
# Expression quantitative trait




*Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 2014, 10:1896-1902




[*Biometrics 68(1) 1–11*]




# eQTL Mapping Using RNA-Seq Data

- eQTLs are genomic loci that contribute to variation in mRNA expression levels
- eQTLs provide insights on transcription regulation, and the molecular basis of phenotypic outcomes
- eQTL mapping can be done with RNA-Seq data