RNA Sequencing Analysis

Sol Katzman

BME 235 May 6, 2015

Today's main topics

- RNA sequencing basics (Illumina)
- Aligning (with or without gene model)
- Gene by gene coverage
- Comparing conditions or tissues
- Alternative splicing

- Comparing gene to gene
- Alternative technologies (PacBio)
- Gene assembly (linking exons)
- Small rna (miRNA)
- CLIPseq (RNA binding proteins), Riboprof, …

DNAseq issues

- Data
  - How many reads do I need?
  - Quality of reads
  - PairedEnd vs SingleEnd sequencing
  - Length of reads
- Information
  - Mapping of reads to reference genome
  - Coverage of target region by reads
  - PCR duplicates

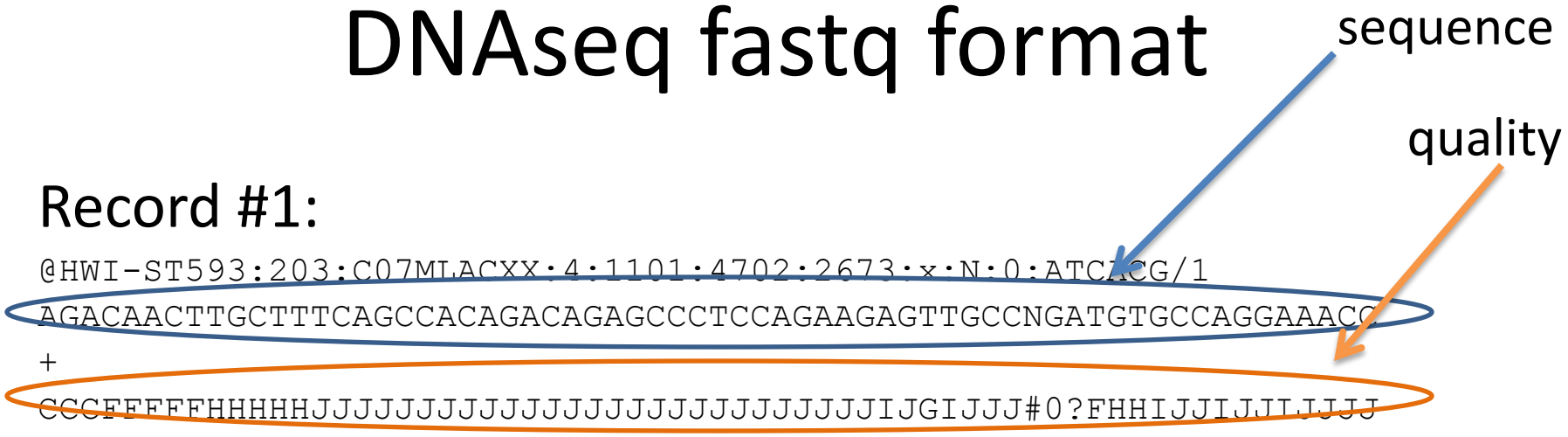
RNAseq issues

- All DNAseq issues, plus…
- Samples and library prep
  - Ribosomal RNA elimination
  - Strand specificity
  - Biological replicates
  - Infinite set of conditions, tissues to choose from
- Information
  - Mapping across splice junctions
  - RNA levels are not protein levels
  - 5’ or 3’ bias
  - Normalization among samples

RNAseq applications

- Annotated (model organism) genome
  - Relative gene expression
  - Alternative isoform expression
  - Differential gene or isoform expression
  - Novel (non-coding) gene expression
- Genome annotation
- Small rna (miRNA, etc.) expression
- Allele specific expression
- RNA editing

DNAseq fastq format



Record #1:

```
@HWI-ST593:203:C07MLACXX:4:1101:4702:2673:x:N:0:ATCACG/1
AGACAACTTGCTTTCAGCCACAGACAGAGCCCTCCAGAAGAGTTGCCNGATGTGCCAGGAAACC
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJIJGIJJJ#0?FHHIJJIJJIJJJJ
```

Quality is encoded:

C = Q34

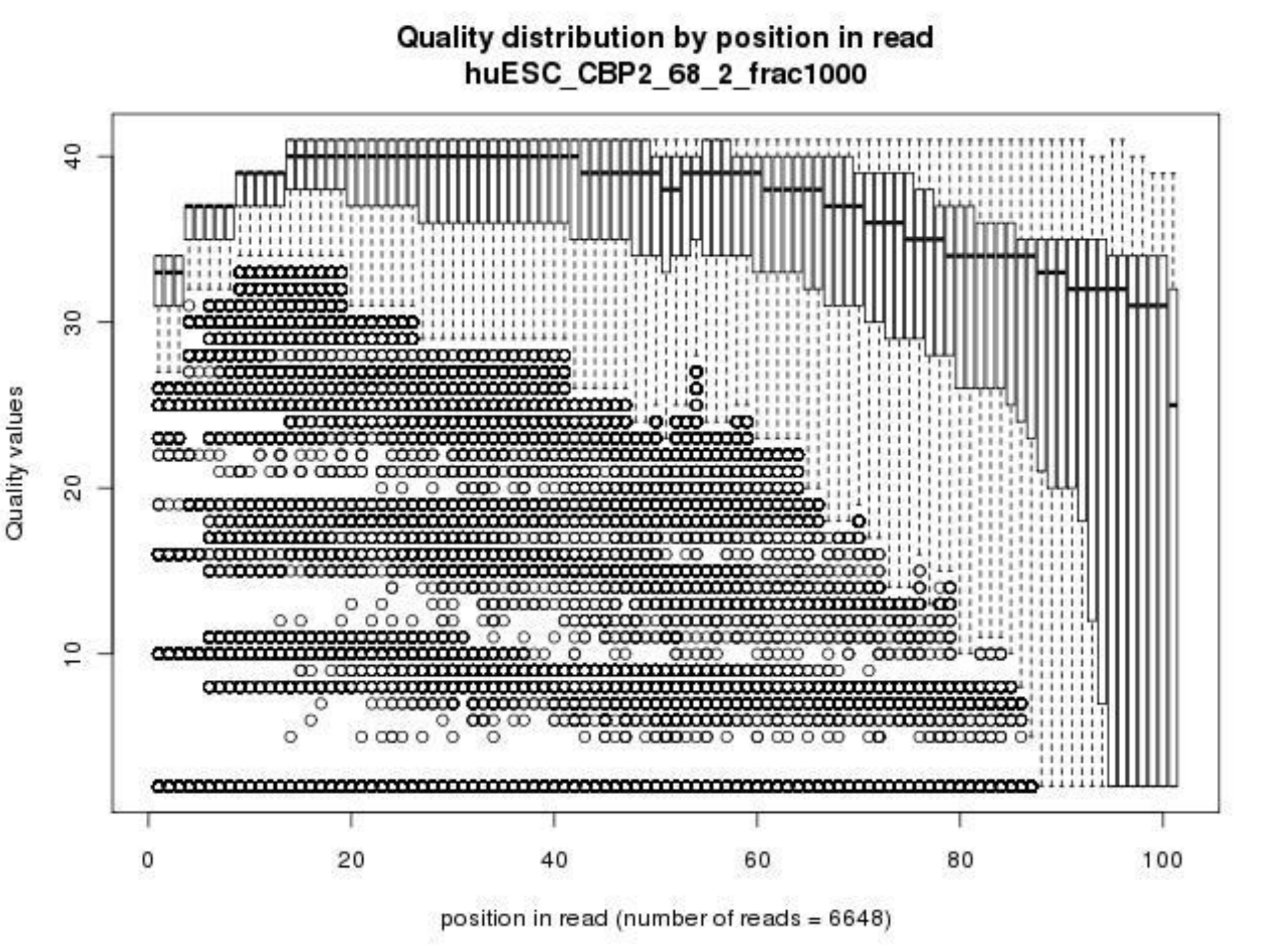
D = Q35

…

Q = -10 $\log_{10}$ (prob (error))

Q20: 1/100 prob error

Q30: 1/1000 prob error


Quality distribution by position in read
huESC_CBP2_68_2_frac1000
Quality values
10
20
30
40
0
20
40
60
80
100
position in read (number of reads = 6648)

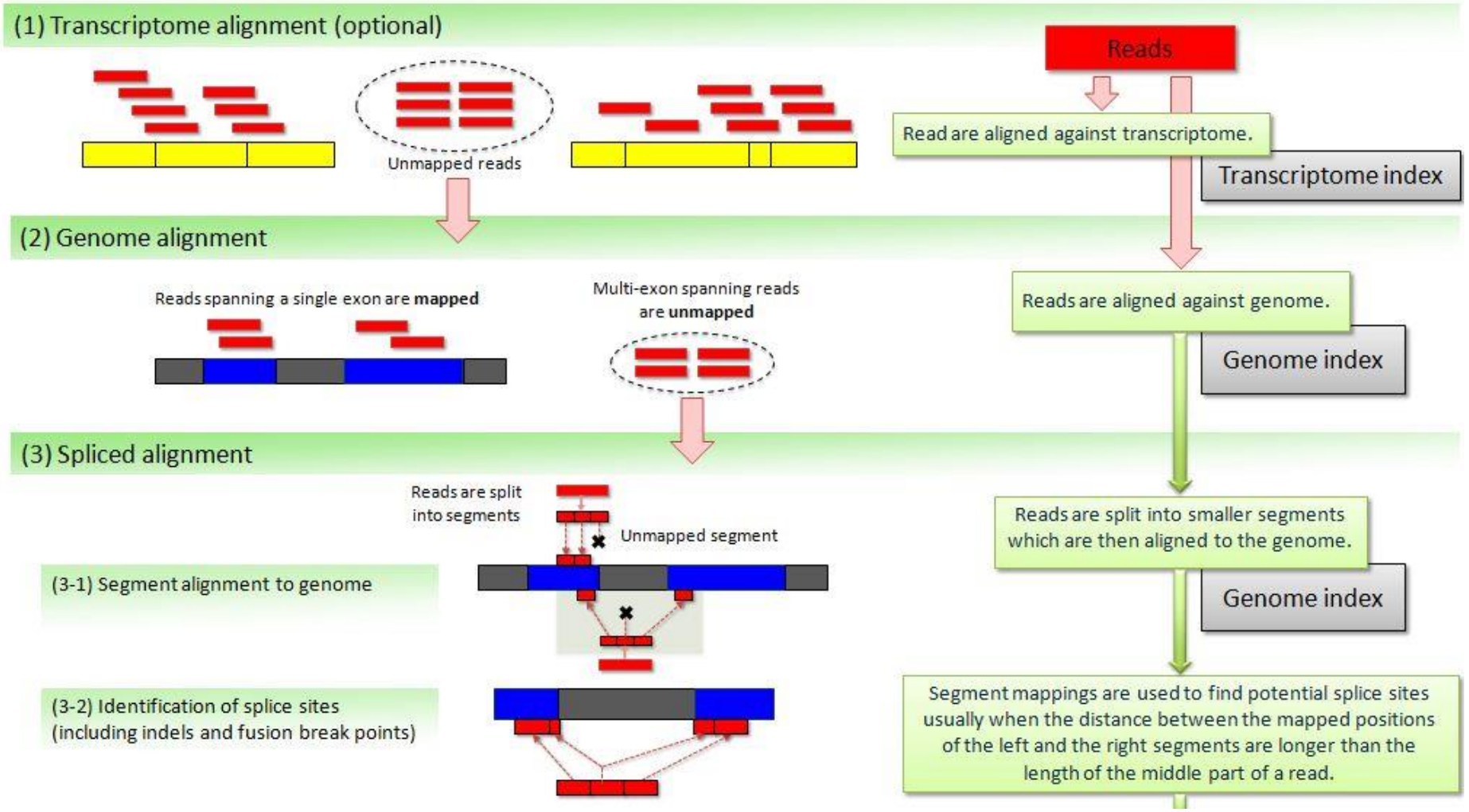
DNAseq mapping

- Mapping issues
  - Errors in the reads
  - Errors in the reference genome
  - Polymorphisms
  - Repetitive regions in the genome
- Mapping algorithms
  - Allow for mismatches, multiple matches, low quality positions in the read
  - Must be fast, parallelizable

RNAseq mapping

- All DNAseq mapping issues, plus
  - Splice junction mappings
  - High copy genes (rRNA, tRNA)
  - Paralogs
  - Pseudogenes
  - Degradation products (introns)
- Mapping algorithms
  - Map to genome, pre-split the reads (tophat1,2)
  - Map to transcriptome (gene model) (tophat2)
  - On-the-fly split reads during mapping (star)

Tophat2 (bowtie2)



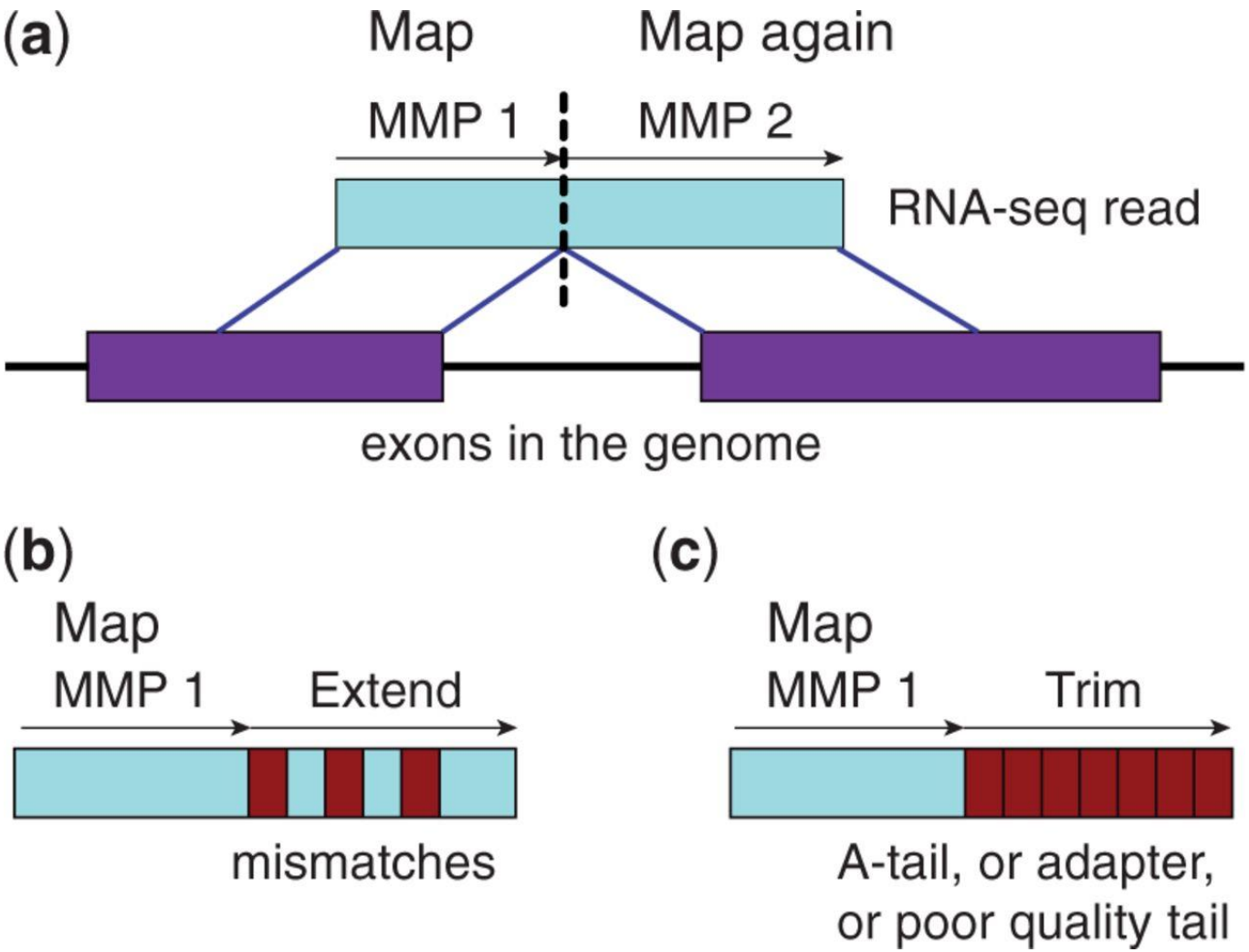
Kim et al Genome Biology 2013

tophat2

- Optional genome model for pre-mapping
- Reads split into thirds
- Need long enough reads to be split for splice junction finding (75bp)
- Where is the junction? Canonical splice sites (GU/AG) favored
- Coverage pileup method for finding splices too
- Iterative, more reads per run better (less parallelizable)
- Somewhat slow

STAR

Schematic representation of the Maximum Mappable Prefix search in the STAR algorithm for detecting (a) splice junctions, (b) mismatches and (c) tails.



Alexander Dobin et al. Bioinformatics 2013;29:15-21

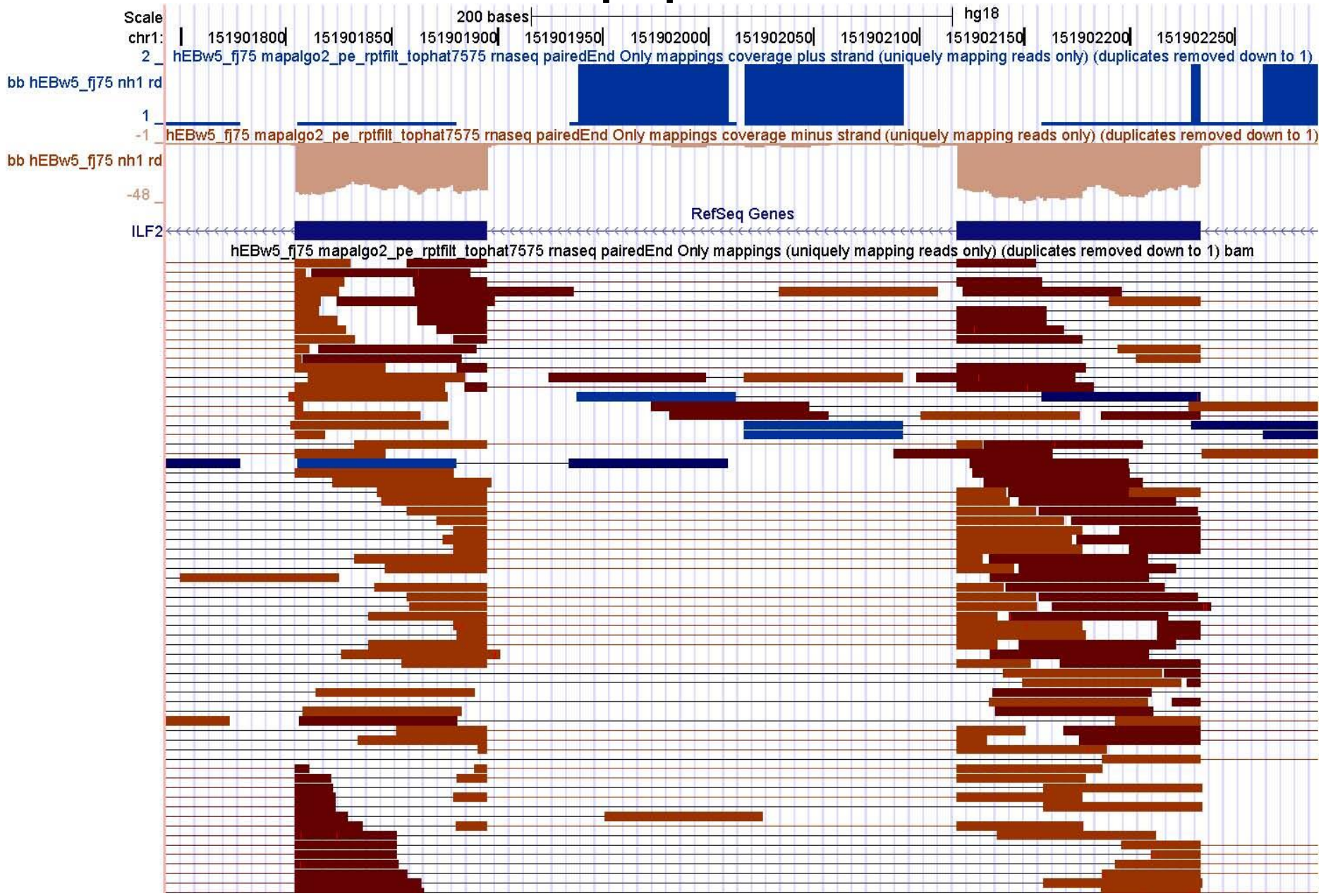
STAR

- Optional gene model included in mapping target (splice junction database)
- On-the-fly splicing
- Soft clipped mappings
- Post-filter based on splice sites (GU/AG…) (and reverse complement?)
- Can be run iteratively (internally or externally)
- Very fast (20X to 50X tophat)
- Separate output of “chimeric” mappings

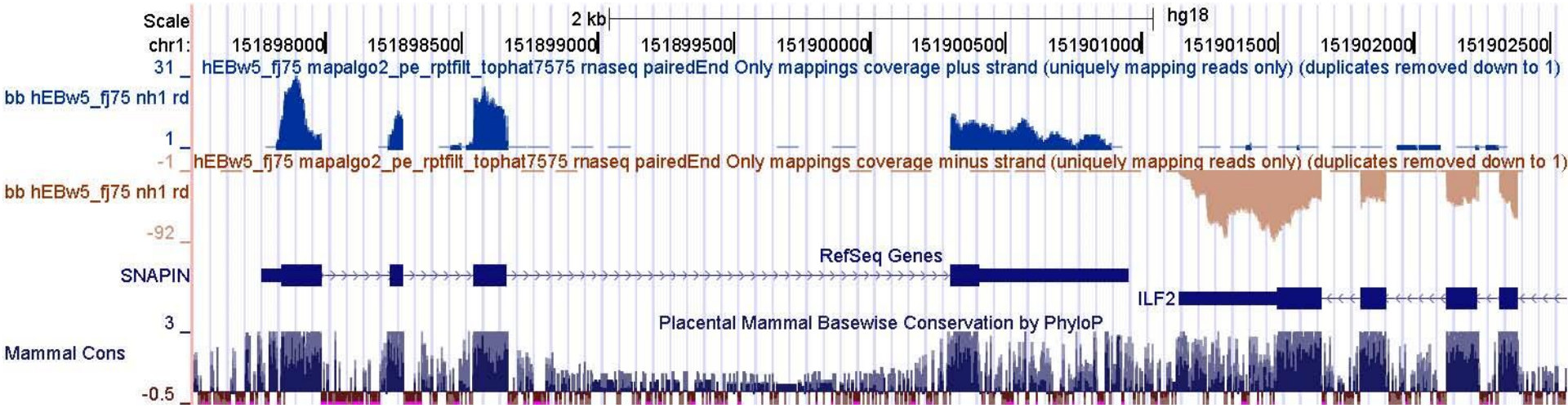
Other mappers

- BWA
- MapSplice (TCGA)
- SpliceMap
- Bowtie – “genomic” target only
  - for QC, for filtering
  - peLength estimation
  - mapping to transcriptome only (RSEM)
  - mapping to isoform alternatives (SpliceTrap)
  - it is the mapping engine for Tophat

RNAseq split reads



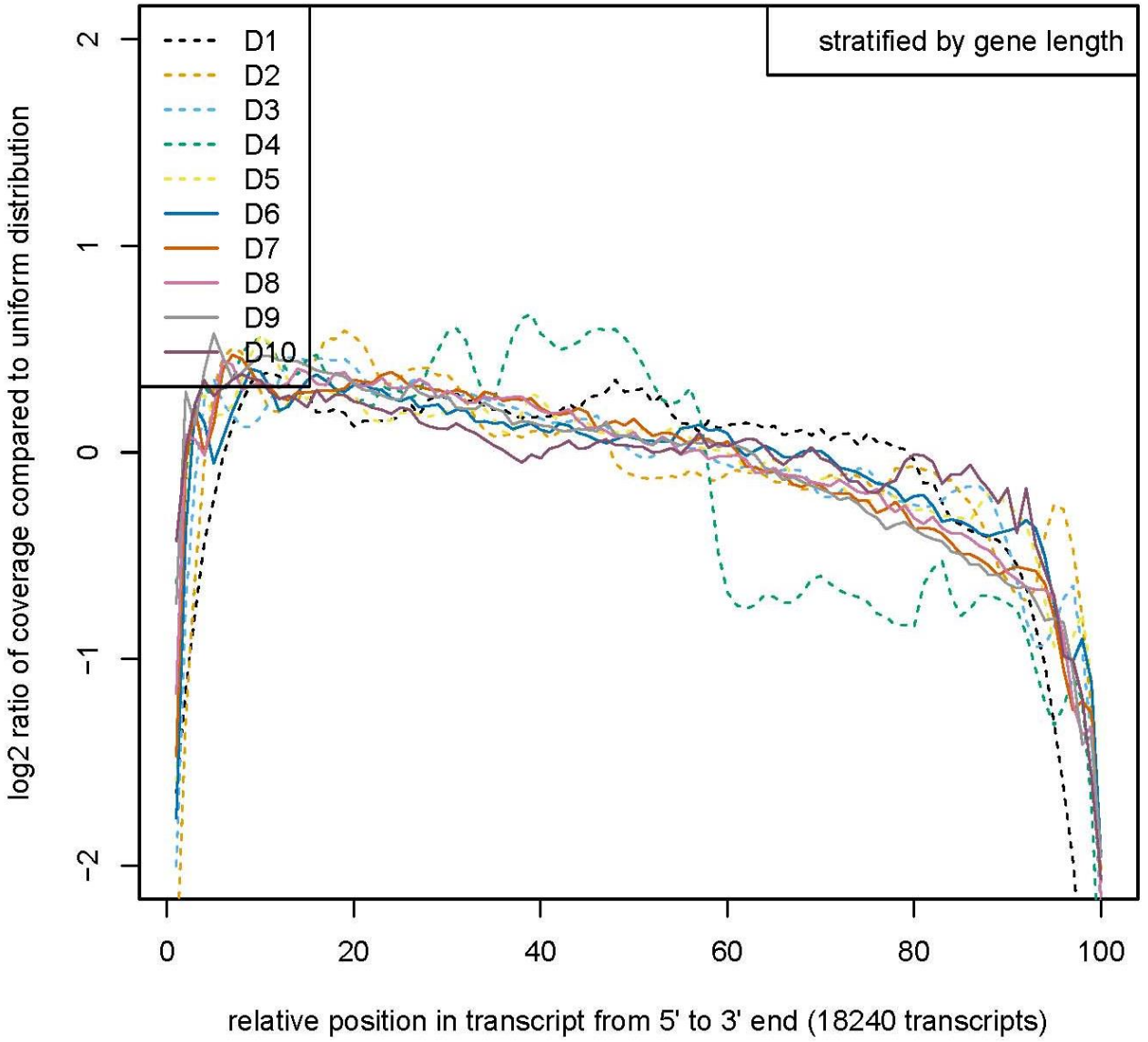
RNAseq exons coverage



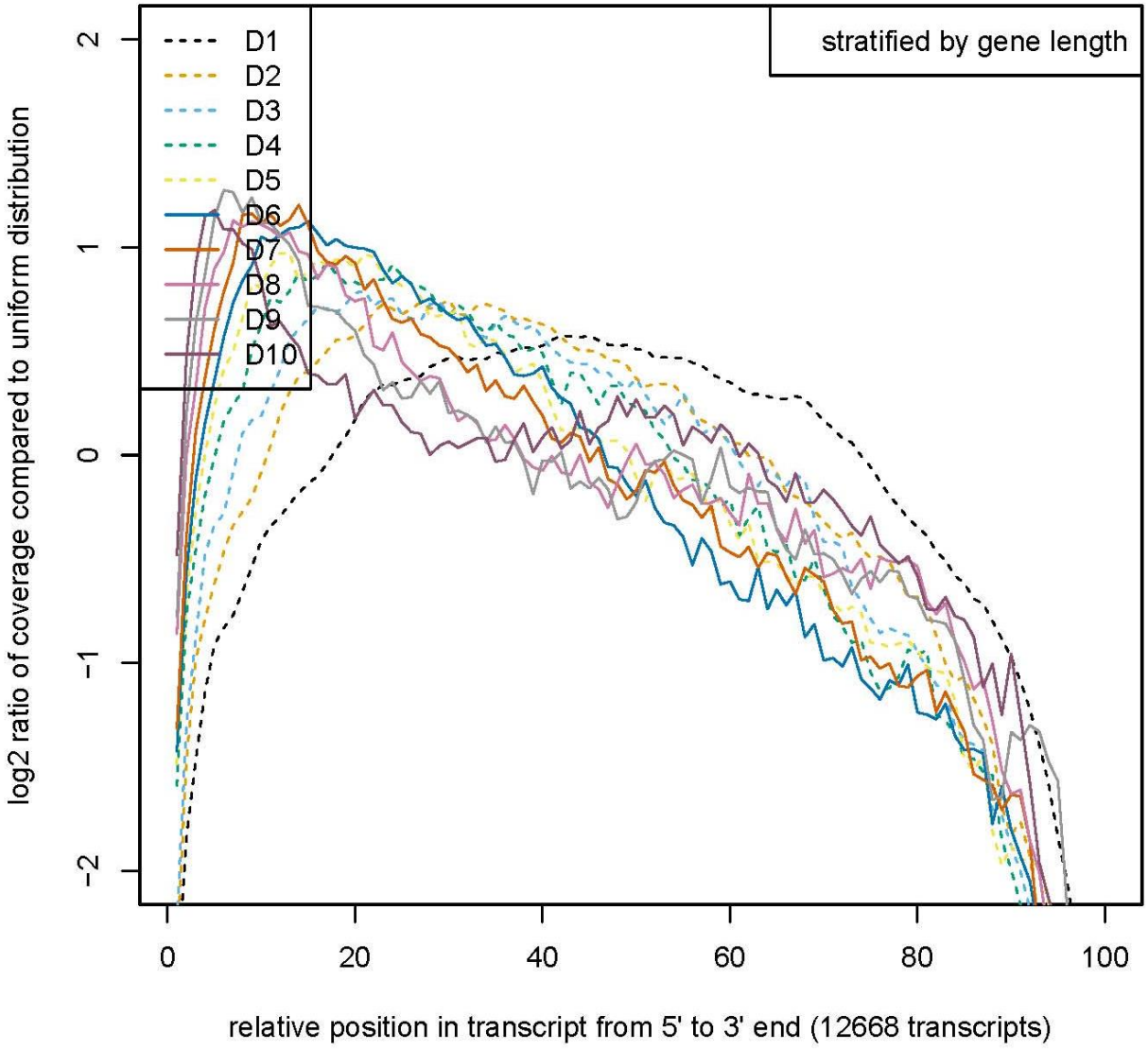
RNAseq issues (1)

- Ribosomal RNA elimination from library
  - Over 90% of RNA in the cell
  - polyA selection
    - 3’ bias
    - Pol II transcripts only
  - riboMinus, riboZero, (polyT priming)
    - 5’ bias?
    - pre-mRNA
  - QC: Filter reads against repetitive genomic elements before mapping
  - QC: 5’ to 3’ bias after mapping

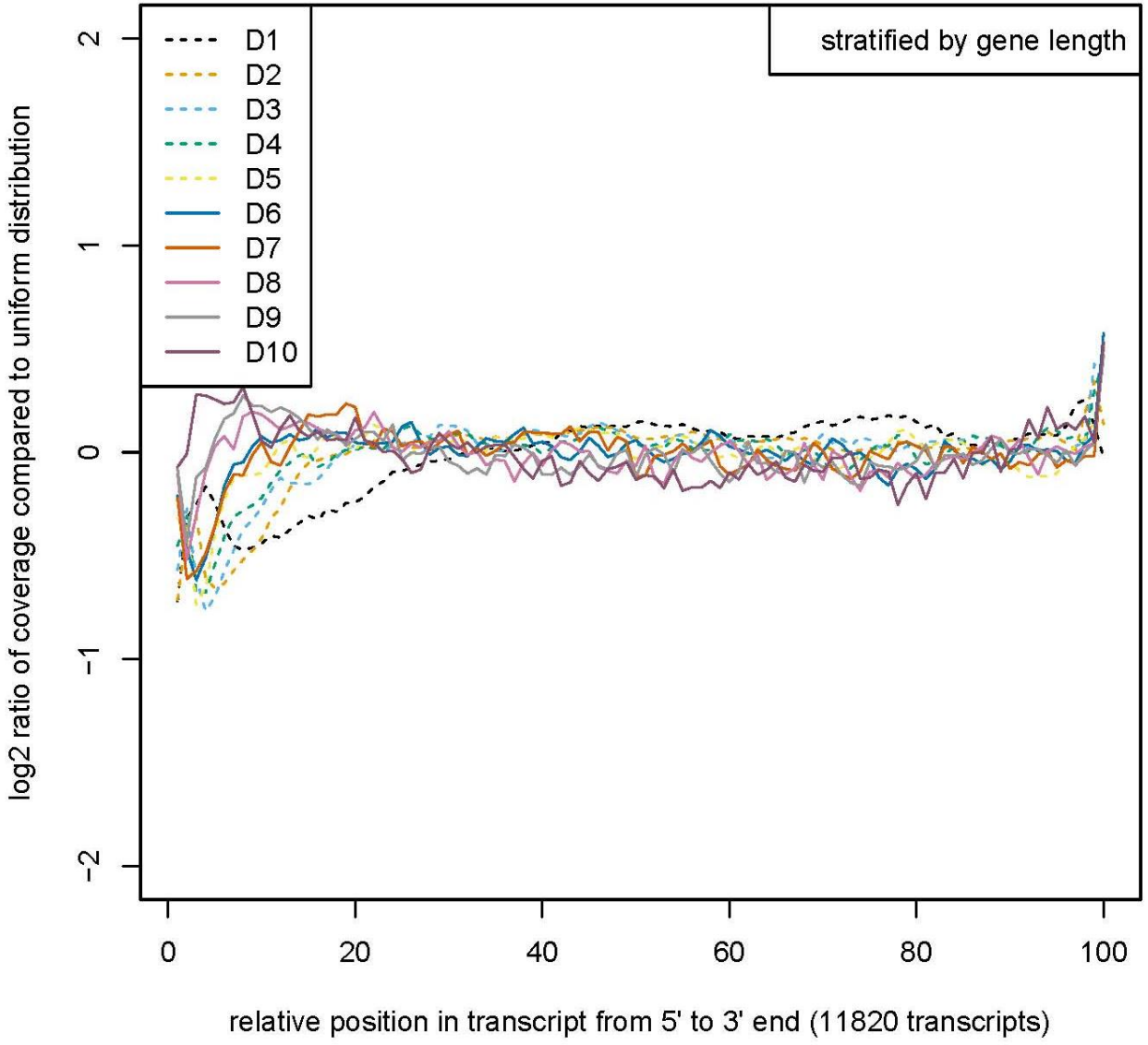
RiboZero positional bias



Ribo Prof Exons positional bias



Ribo Prof CDS positional bias



RNAseq issues (2)

- Strand specificity
  - Detect antisense transcription of known genes
  - Detect transcription of unknown genes, which strand?
  - Splice site bias (vs. revComp) for mapper

RNAseq issues (3)

- PCR duplicates
  - "PCR is not your friend"
  - Wastes sequencing $$
  - Wastes computational time, storage
  - 10-12 cycles max! But extracting RNA is hard. Extracting non-rRNA is harder still.
  - Detection and bioinformatic removal after mapping. Mitigate removal with:
    - Paired-end reads for an extra degree of freedom (if randomly digested)
    - Internal random barcodes in adaptors before PCR

RNAseq issues (other)

- Mitochondrial genes
  - Muscle cells
- “Uniquely” mapping reads only?
  - Determined by mapping parameters
  - Can lose common genes with paralogs (mouse GAPDH)
  - One best mapping (if scored) is reasonable
- Lengths of fragments vs. read length
  - Longer reads for more splice junctions (mapper must support multiple junctions per read)
  - Longer fragments for more mapping information (no overlap read1/read2)

Today's main topics

- RNA sequencing basics (Illumina)
- Aligning (with or without gene model)
- Gene by gene coverage
- Comparing conditions or tissues
- Alternative splicing

- Comparing genes
- Alternative technologies (PacBio)
- Gene assembly
- Small rna (miRNA)
- CLIPseq (RNA binding proteins), Riboprof, …

What is a Gene?

- Genes have many isoforms
- Canonical transcript?
- Any exon in any transcript?
  - Count reads? Read starts? Intron/exon reads?
  - Count total coverage, divide by read length to estimate number of reads in a “gene”

Today's main topics

- RNA sequencing basics (Illumina)
- Aligning (with or without gene model)
- Gene by gene coverage
- Comparing conditions or tissues
- Alternative splicing

- Comparing genes
- Alternative technologies (PacBio)
- Gene assembly
- Small rna (miRNA)
- CLIPseq (RNA binding proteins), Riboprof, …

Comparing expression

- FPKM: fragments per kilobase of gene length per million mapped reads

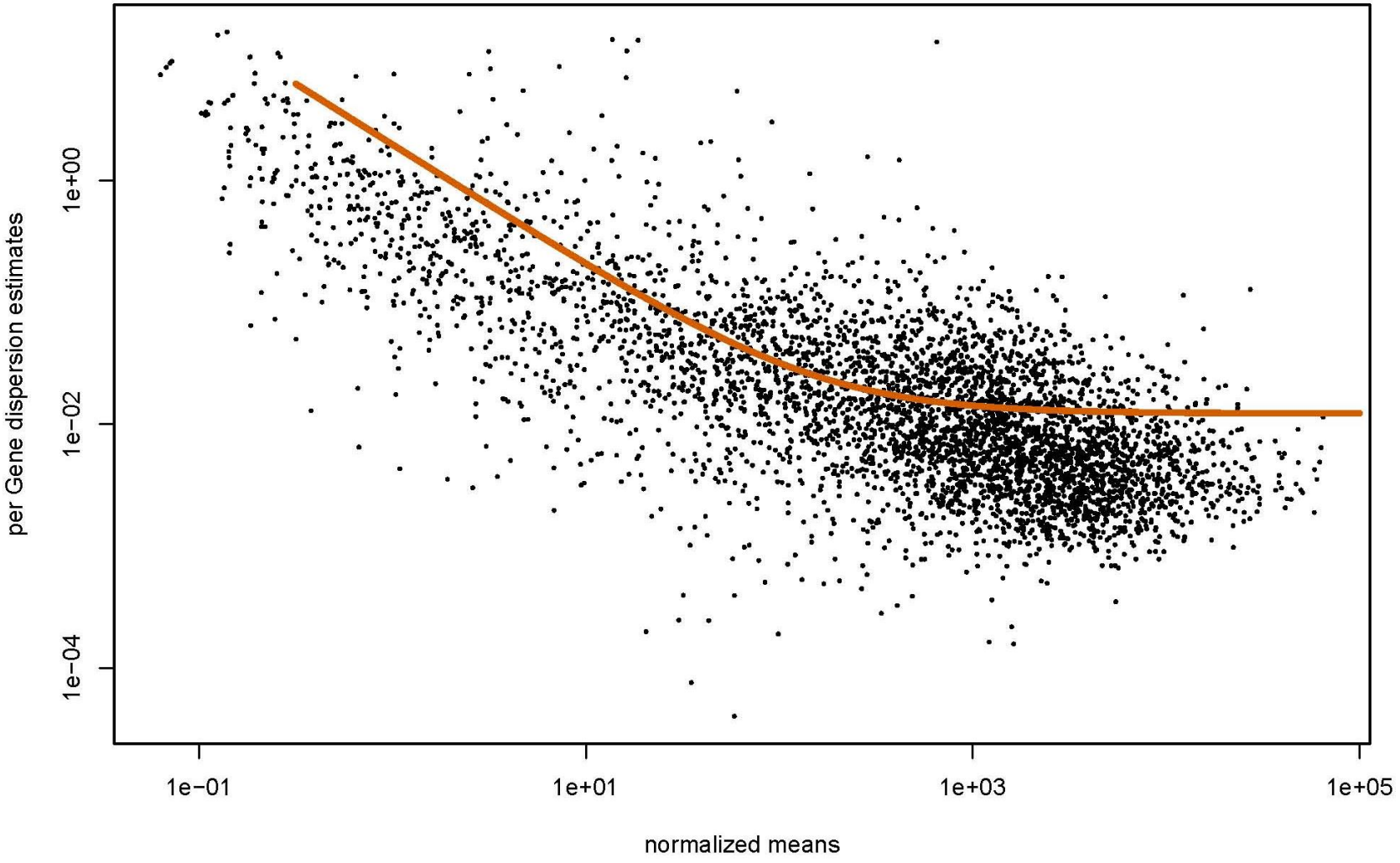
from Dillies et al. *Brief Bioinform.* 2013:

- Normalization of RNA-seq data in the context of differential analysis is essential in order to account for the presence of systematic variation between samples as well as differences in library composition.
- **The Total Count and RPKM normalization methods, both of which are still widely in use, are ineffective and should be definitively abandoned** in the context of differential analysis.
- Only the DESeq and TMM normalization methods are robust to the presence of different library sizes and widely different library compositions, both of which are typical of real RNA-seq data.

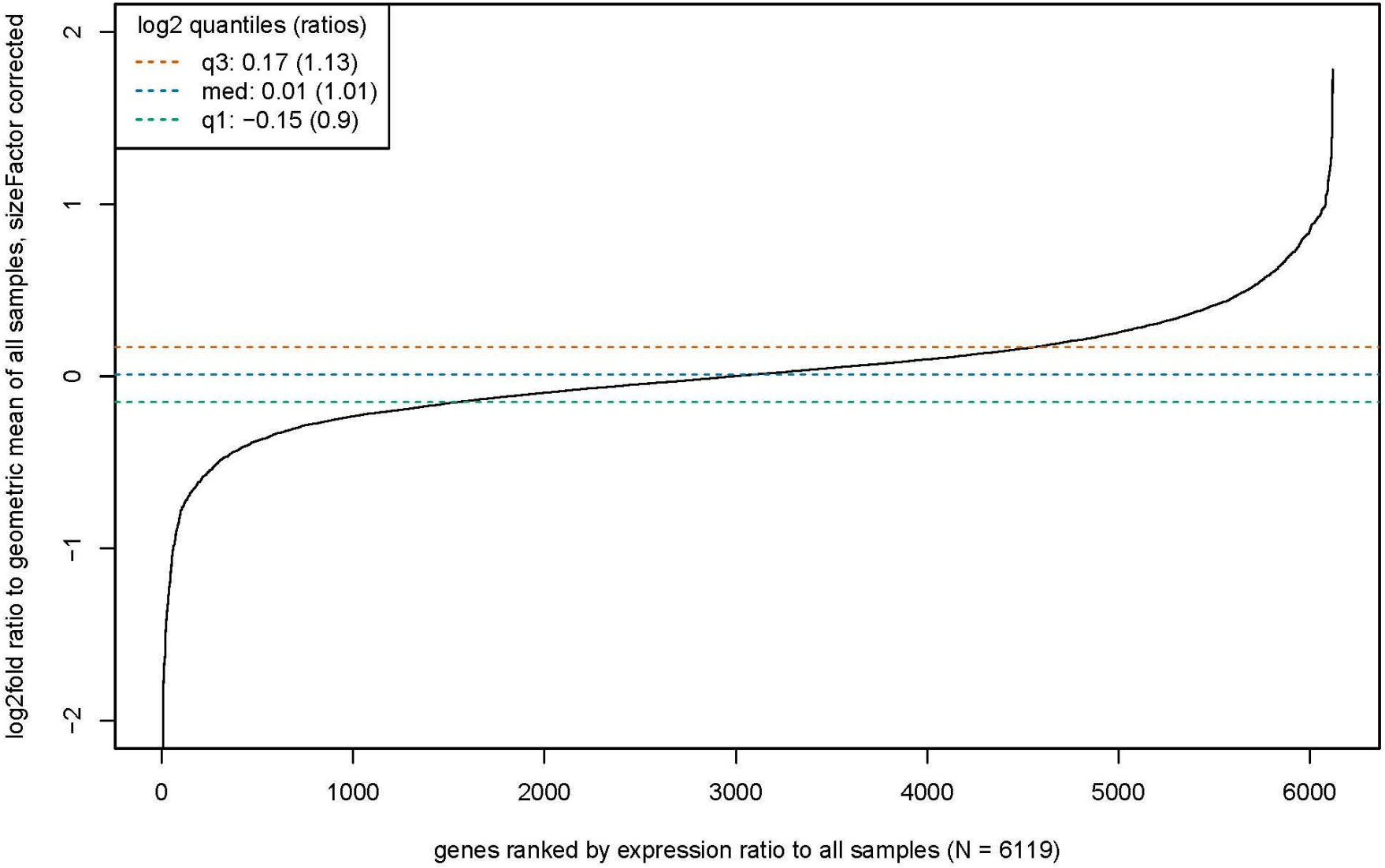
DESeq

- Uses replicates per “condition”
- Estimates dispersion as a function of mean expression for a gene
- Uses **median** relative expression to normalize sample gene counts
- Uses dispersion and normalized counts to determine **significance** of differential expression between conditions

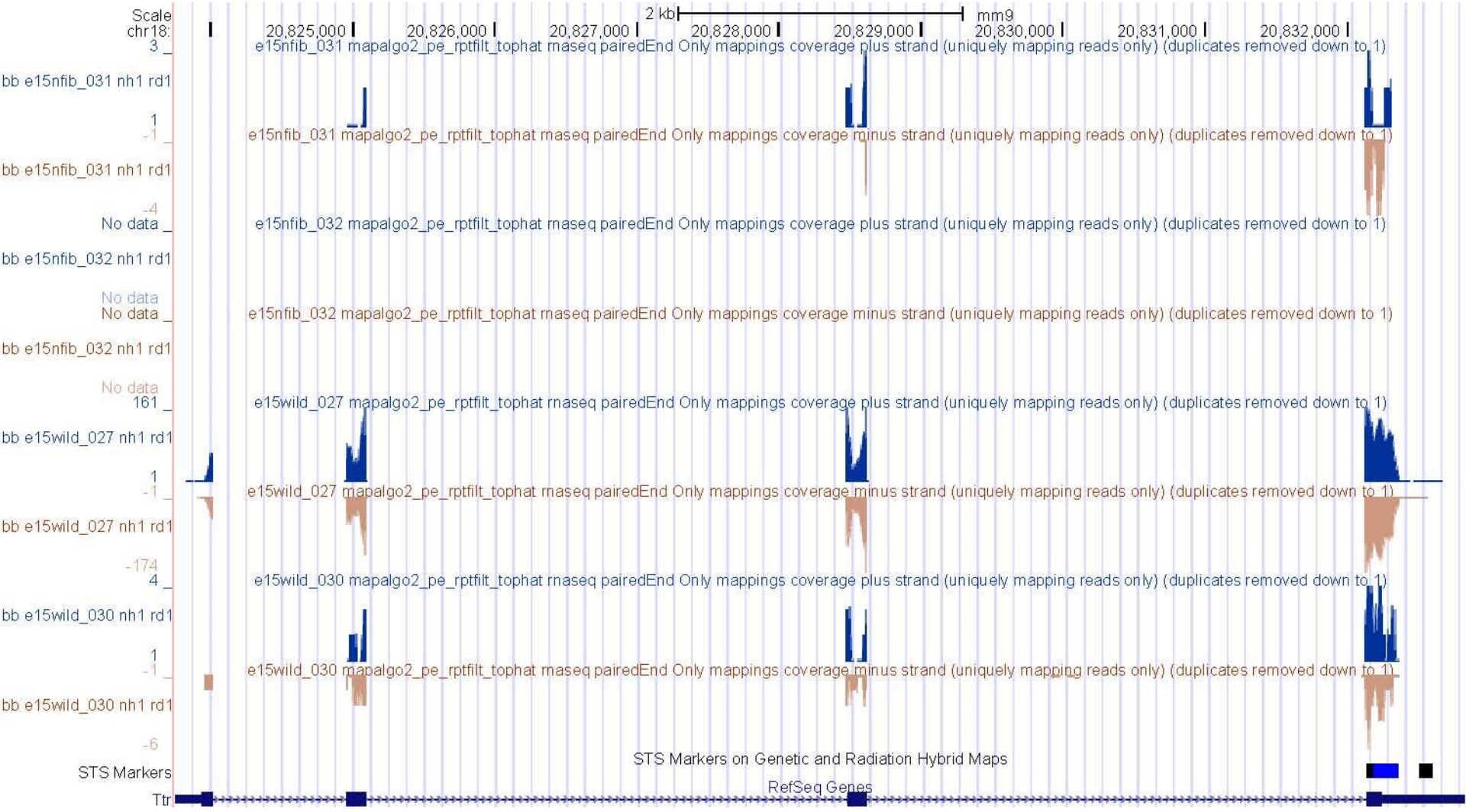
DESeq fitting dispersion estimates



DESeq normalization



RNAseq diffex with replicates



Today's main topics

- RNA sequencing basics (Illumina)
- Aligning (with or without gene model)
- Gene by gene coverage
- Comparing conditions or tissues
- Alternative splicing

- Comparing genes
- Alternative technologies (PacBio)
- Gene assembly
- Small rna (miRNA)
- CLIPseq (RNA binding proteins), Riboprof, …

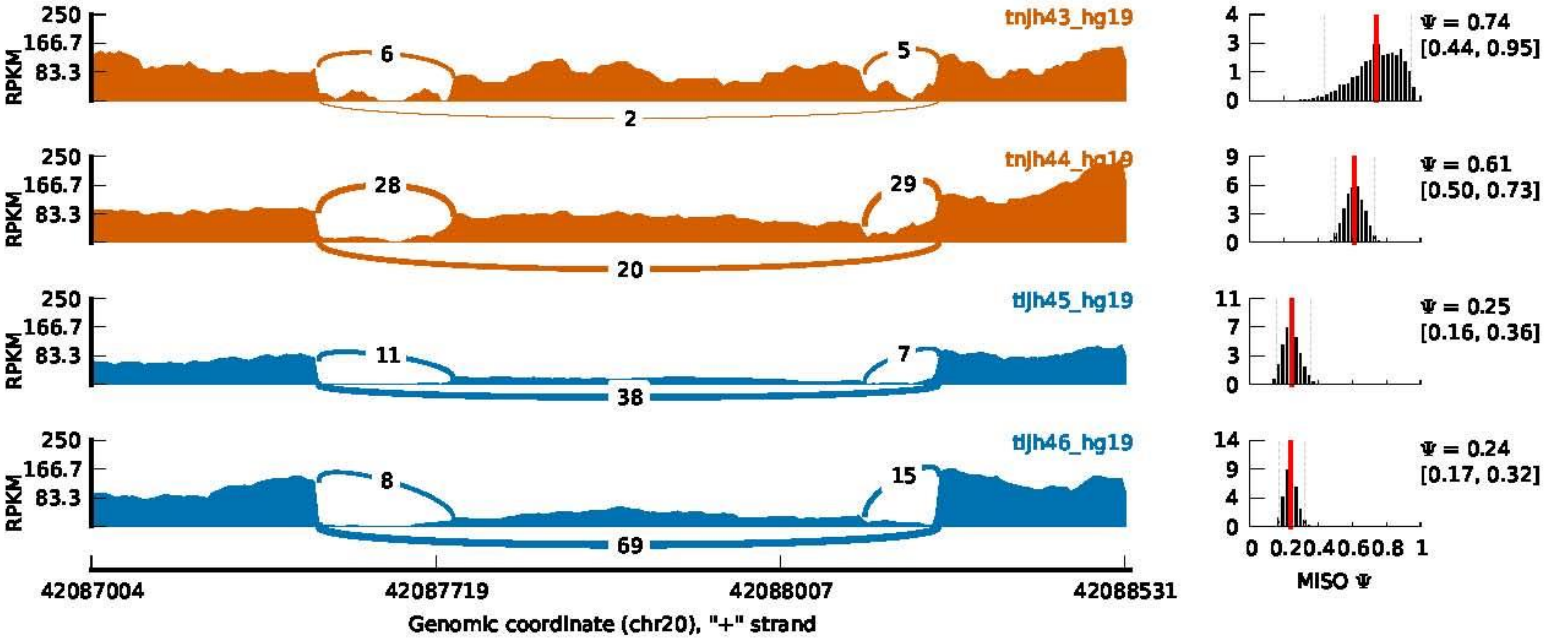
Alt Splicing tools

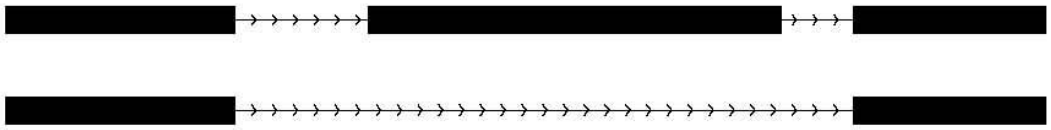
- MISO
- Splicetrap
- CuffDiff
- rMATs
- …

MISO

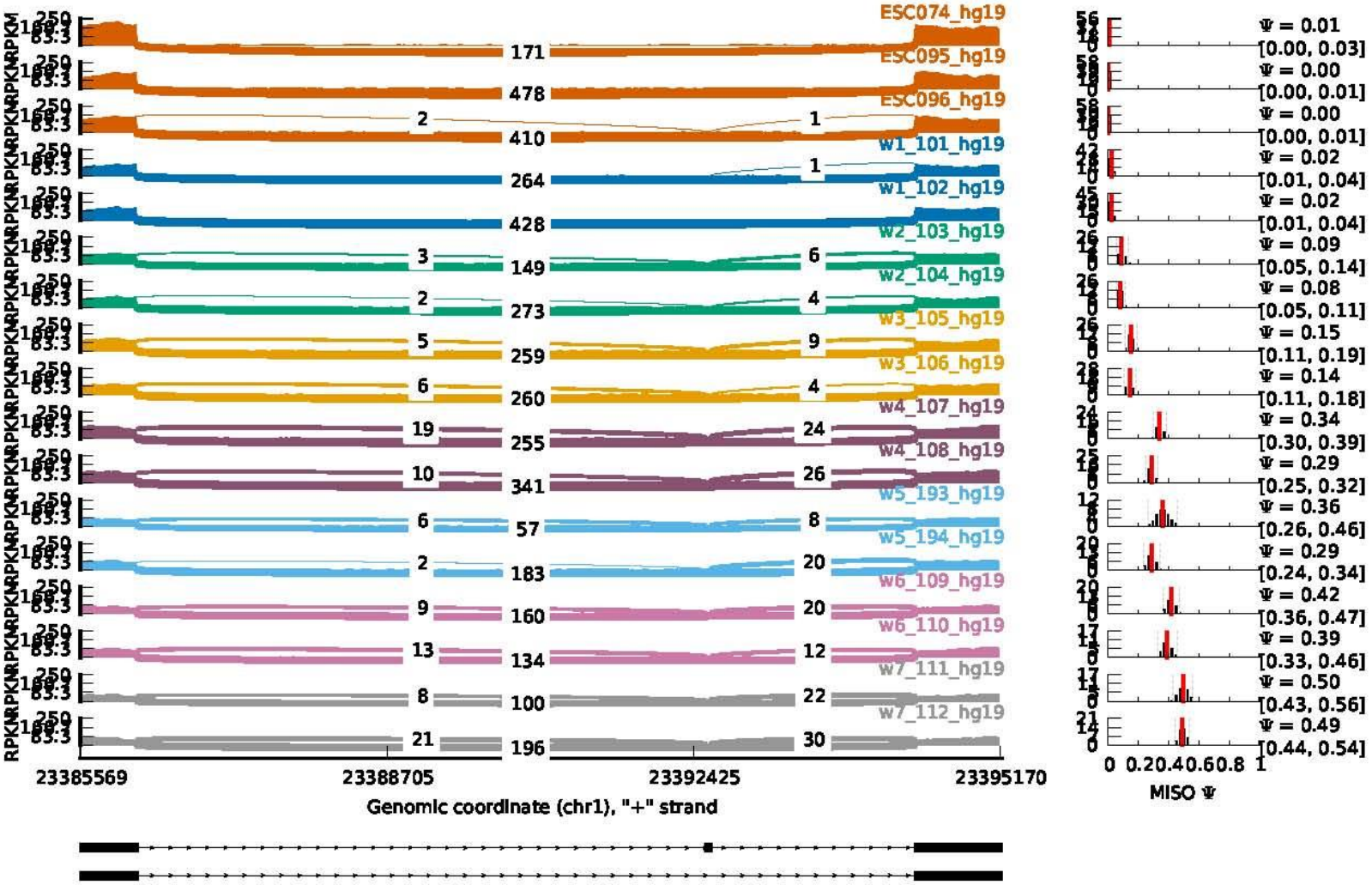
- Uses mapped reads as input
- Bayesian posterior estimate of “Percent Spliced In” (PSI) for a skipped exon
- Nice plotting tool

SE21101


tnJh43_hg19
tnJh44_hg19
tJh45_hg19
tJh46_hg19
RPKM
250
166.7
83.3
6
5
2
28
29
20
11
7
38
8
15
69
Ψ = 0.74
[0.44, 0.95]
Ψ = 0.61
[0.50, 0.73]
Ψ = 0.25
[0.16, 0.36]
Ψ = 0.24
[0.17, 0.32]
0 0.20.40.60.8 1
MISO Ψ
42087004
42087719
42088007
42088531
Genomic coordinate (chr20), "+" strand

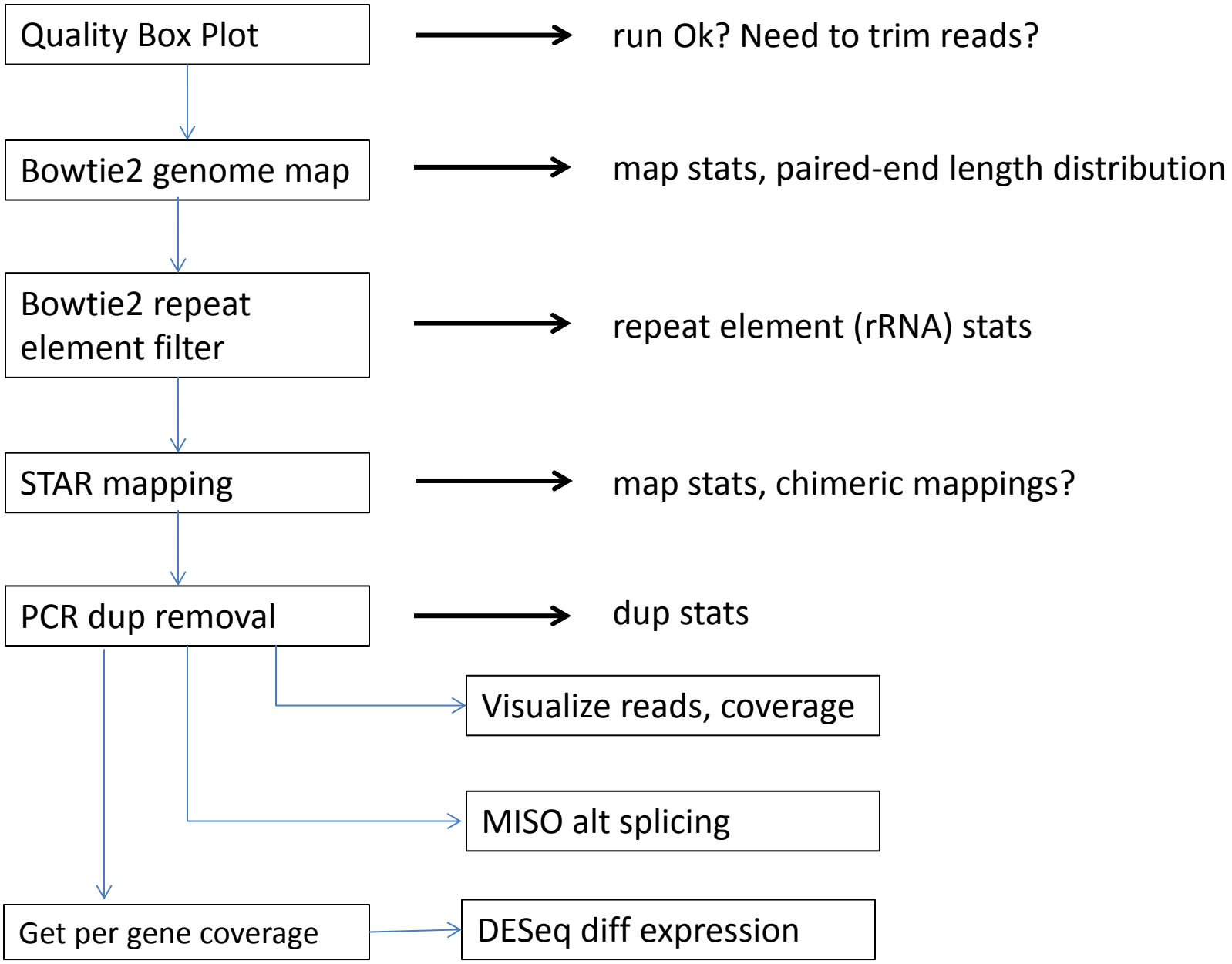


SE00574


ESC074_hg19
ESC095_hg19
ESC096_hg19
w1_101_hg19
w1_102_hg19
w2_103_hg19
w2_104_hg19
w3_105_hg19
w3_106_hg19
w4_107_hg19
w4_108_hg19
w5_193_hg19
w5_194_hg19
w6_109_hg19
w6_110_hg19
w7_111_hg19
w7_112_hg19
171
478
2
410
1
264
1
428
3
149
6
2
273
4
5
259
9
6
260
4
19
255
24
10
341
26
6
57
8
2
183
20
9
160
20
13
134
12
8
100
22
21
196
30
23385569
23388705
23392425
23395170
Genomic coordinate (chr1), "+" strand
Ψ = 0.01
[0.00, 0.03]
Ψ = 0.00
[0.00, 0.01]
Ψ = 0.00
[0.00, 0.01]
Ψ = 0.02
[0.01, 0.04]
Ψ = 0.02
[0.01, 0.04]
Ψ = 0.09
[0.05, 0.14]
Ψ = 0.08
[0.05, 0.11]
Ψ = 0.15
[0.11, 0.19]
Ψ = 0.14
[0.11, 0.18]
Ψ = 0.34
[0.30, 0.39]
Ψ = 0.29
[0.25, 0.32]
Ψ = 0.36
[0.26, 0.46]
Ψ = 0.29
[0.24, 0.34]
Ψ = 0.42
[0.36, 0.47]
Ψ = 0.39
[0.33, 0.46]
Ψ = 0.50
[0.43, 0.56]
Ψ = 0.49
[0.44, 0.54]
0 0.20.40.60.8 1
MISO Ψ

RNAseq SUMMARY

- PairedEnd, strand-specific library prep is preferred
- Biological replicates are a must
- Avoid PCR duplicates


Quality Box Plot
run Ok? Need to trim reads?
Bowtie2 genome map
map stats, paired-end length distribution
Bowtie2 repeat element filter
repeat element (rRNA) stats
STAR mapping
map stats, chimeric mappings?
PCR dup removal
dup stats
Visualize reads, coverage
MISO alt splicing
Get per gene coverage
DESeq diff expression