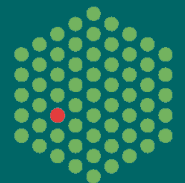


Processing data from high-throughput sequencing experiments II

Simon Anders

EMBL



Alignment

Short-read algorithms: Seed matches

Maq claims to find all alignments with up to 2 mismatches and may find alignments with more than two mismatches.

How does it work?

Aligning hashed reads

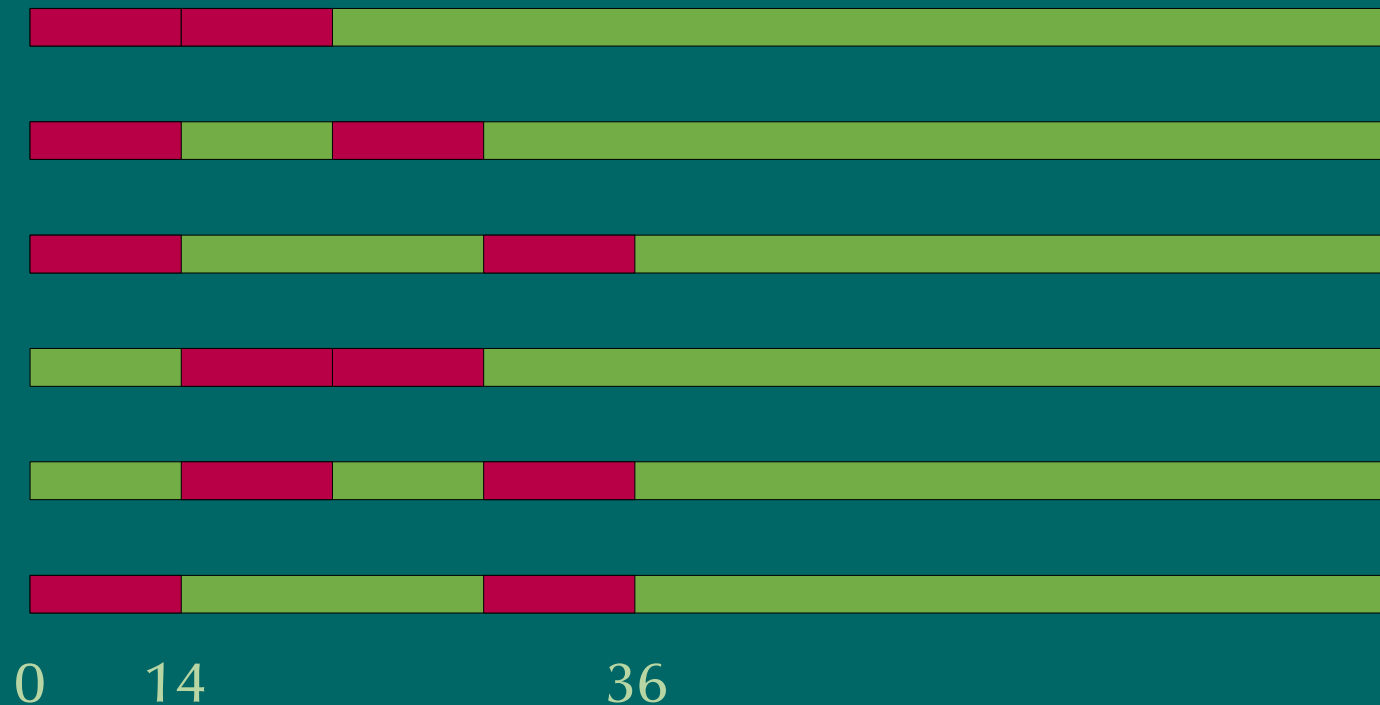
Naive algorithm:

- Make a hash table of the first 28mers of each read, so that for each 28mer, we can look up quickly which reads start with it.
- Then, go through the genome, base for base. For each 28mer, look up in the hash table whether reads start with it, and if so, add a note of the current genome position to these reads.

Problem: What if there are read errors in the first 28 base pairs?

Spaced seeds

Maq prepares six hash table, each indexing 28 of the first 36 bases of the reads, selected as follows:



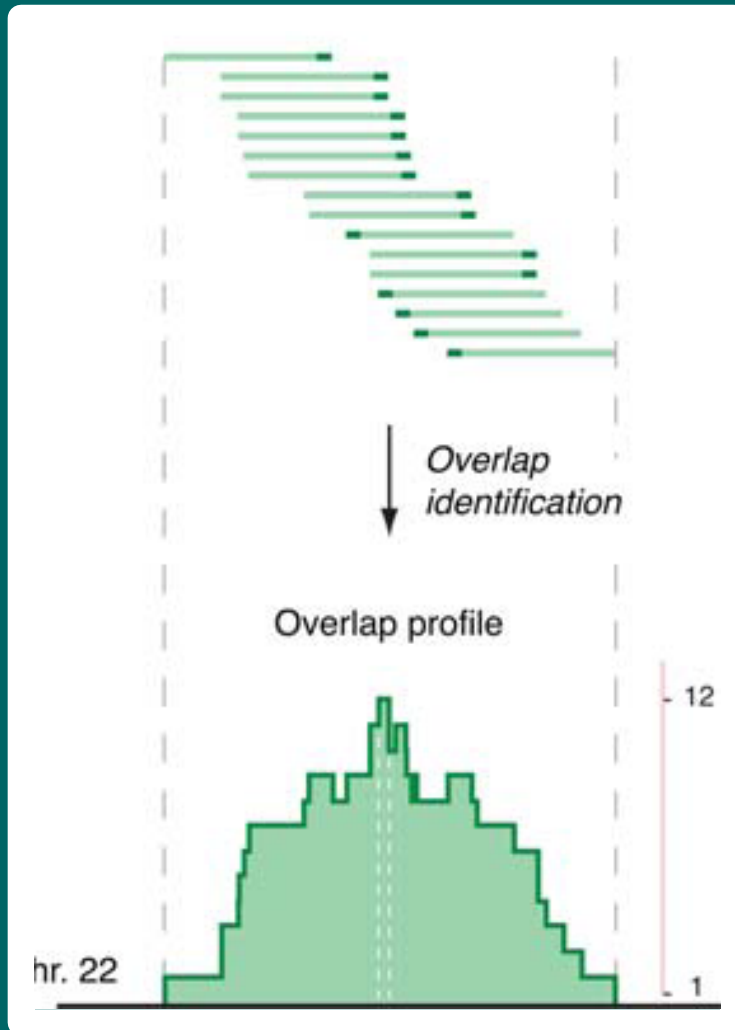
Hence, Maq finds all alignments with at most 2 mismatches in the first 36 bases.

Coverage vectors

Coverage

- In resequencing, we hope to sequence uniformly, i.e., see each part of the genome represented by the same amount of reads.
- Due to the random nature of shotgun sequencing, we need to “cover the genome several times” in order to see each position at least once.
- In other techniques (ChIP-Seq, RNA-Seq, Tag-Seq, CNV-Seq, etc.), the local coverage is what we are interested in.

Coverage vectors



<-- Solexa reads,
aligned to genome

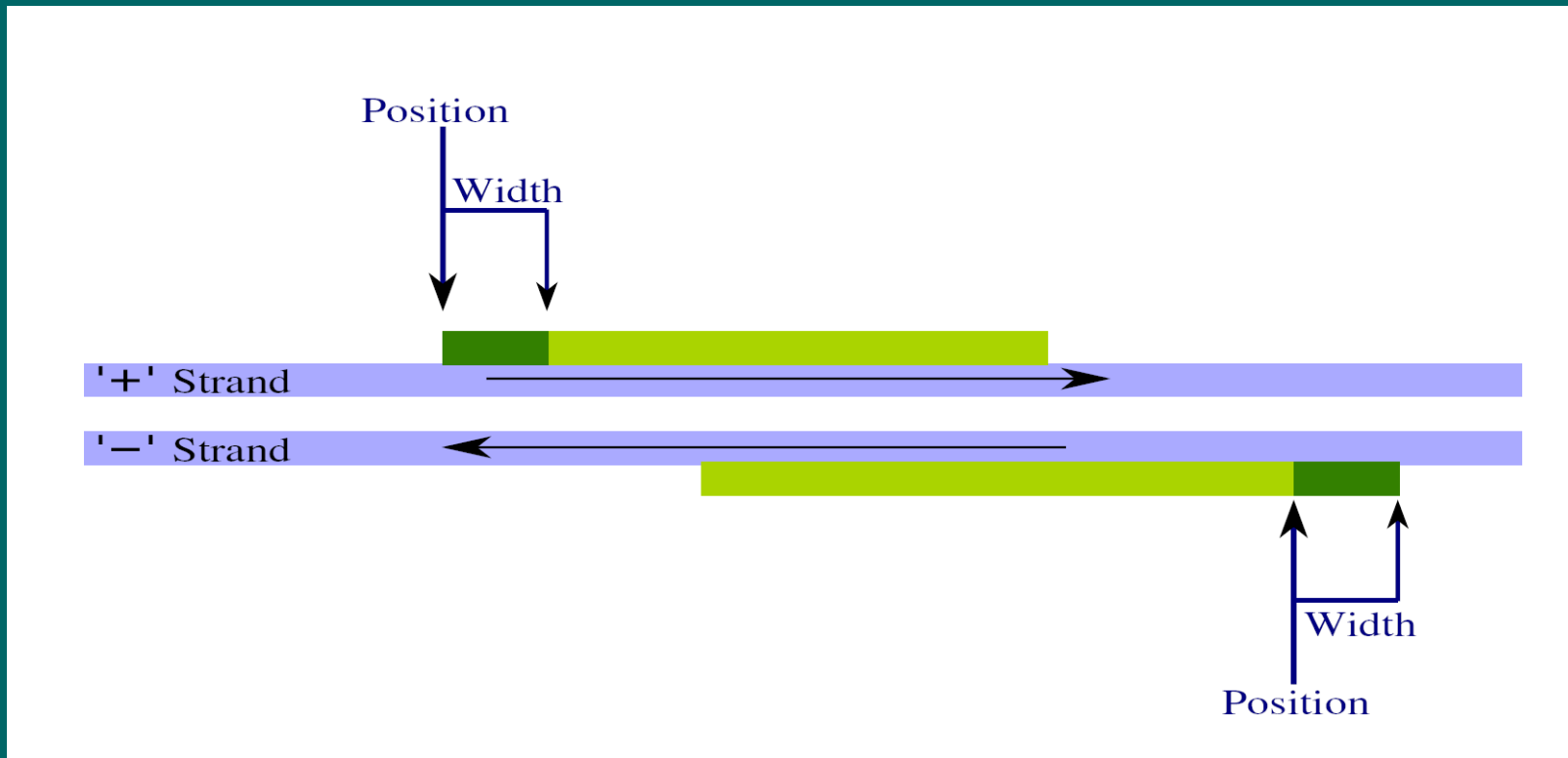
<-- coverage vector

Coverage vectors

- A coverage (or: “pile-up”) vector is an integer vector with one element per base pair in a chromosome, tallying the number of reads (or fragments) mapping onto each base pair.
- It is the essential intermediate data type in assays like ChIP-Seq or RNA-Seq
- One may ever count the coverage by the reads themselves, or extend to the length of the fragments

Calculating coverage vectors

Extending reads to fragments:



Chip-Seq coverage: examples

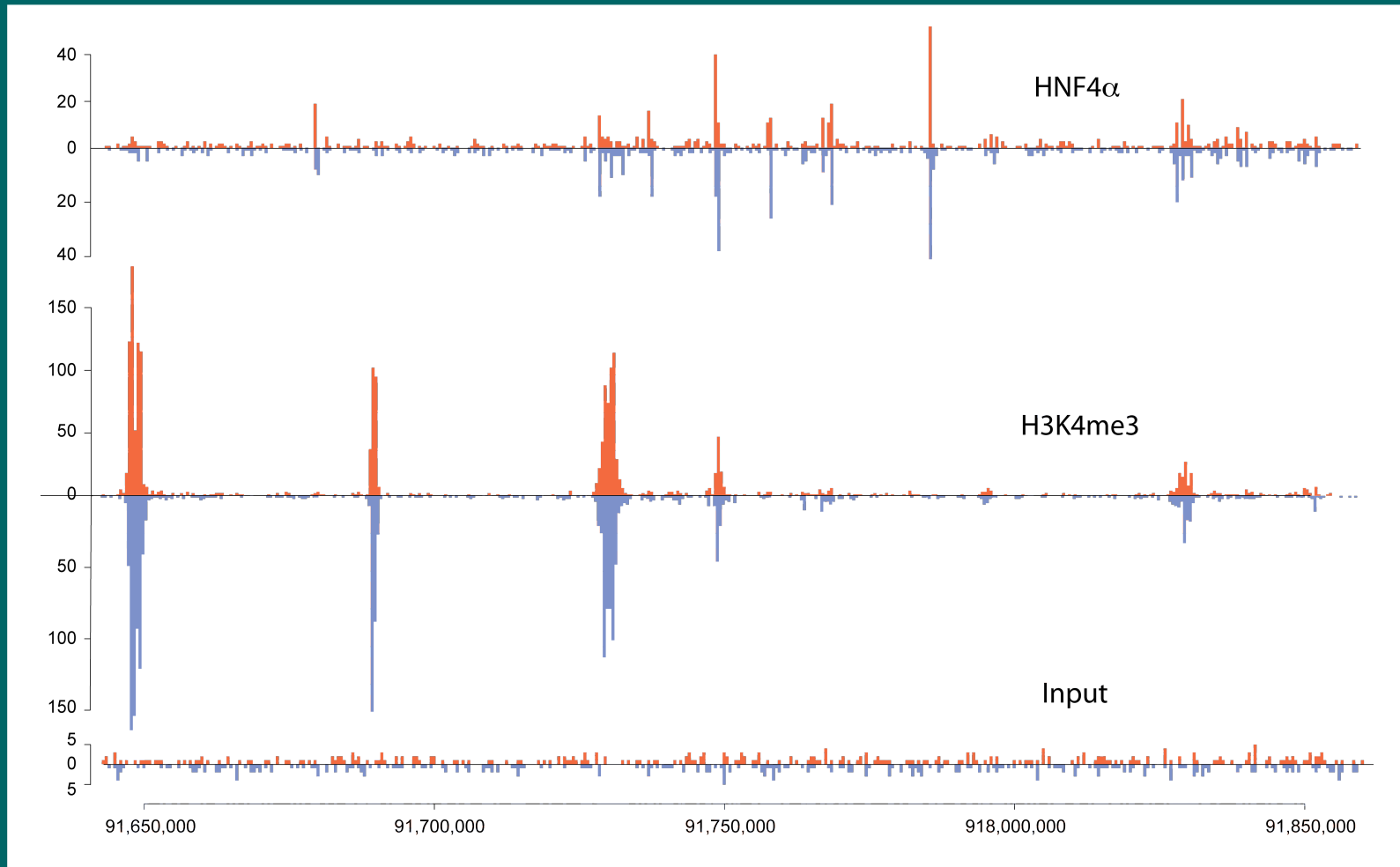
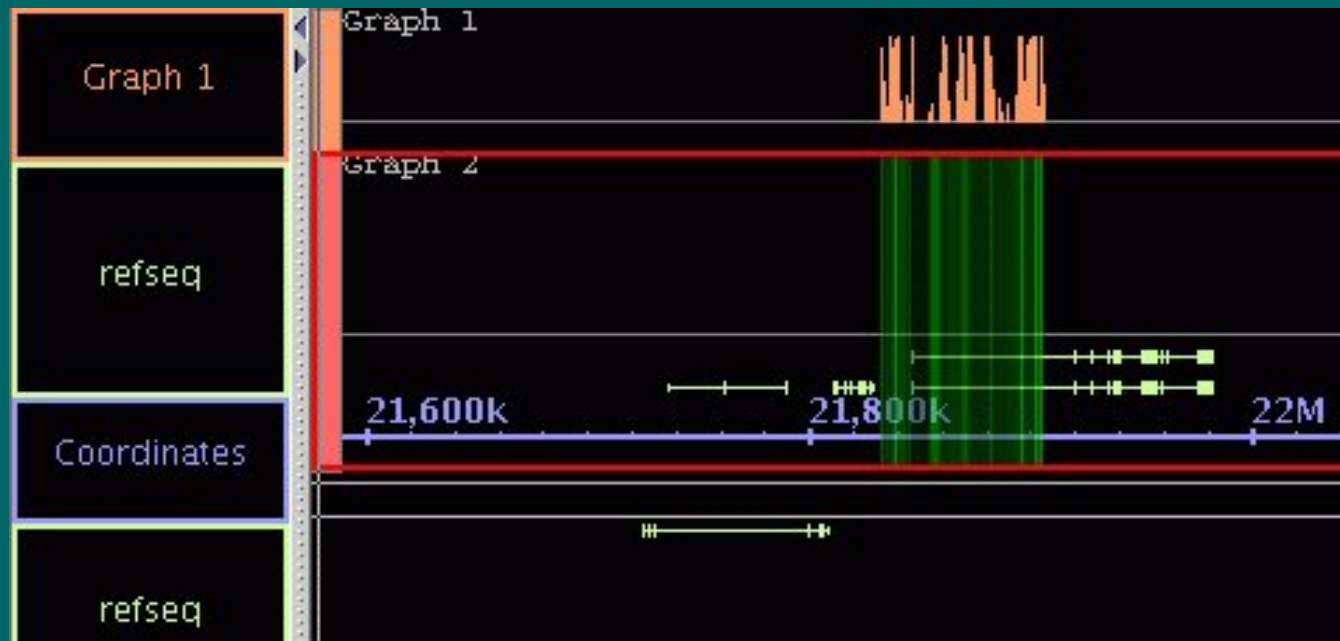


Figure courtesy of Christiana Spyrou (CR UK)

Genome browsers

- Genoviz Integrated Genome Browser (IGB)
- ...



File formats

- Sequences, reads:
 - FASTA
 - FASTQ
- Alignments:
 - SAM
 - ...
- Features, annotations, scores:
 - GFF, GTF
 - BED
 - Wiggle

The issue with multiple reads

If one finds several reads with the exact same sequence, does this mean

- that many fragments from this locus were precipitated and often got cut at the exact same place, or
- that there was only a single fragment, but it was amplified more efficiently than fragments from other loci in the PCR (or more efficiently transcribed to cDNA)?
 - If you consider the latter more likely, you should count these reads only once. However, this dramatically compresses your dynamic range.

Ambiguous matches and mappability

- If a read matches at several places in the reference, the best match should be used.
- If there are several equally good matches, an aligner may
 - chose an alignment at random
 - discard the read
 - report all alignments and delay the choice to downstream analysis
- It is useful to know which regions in the genome are repetitive on the scale of the read length and hence give rise to alignment ambiguities.

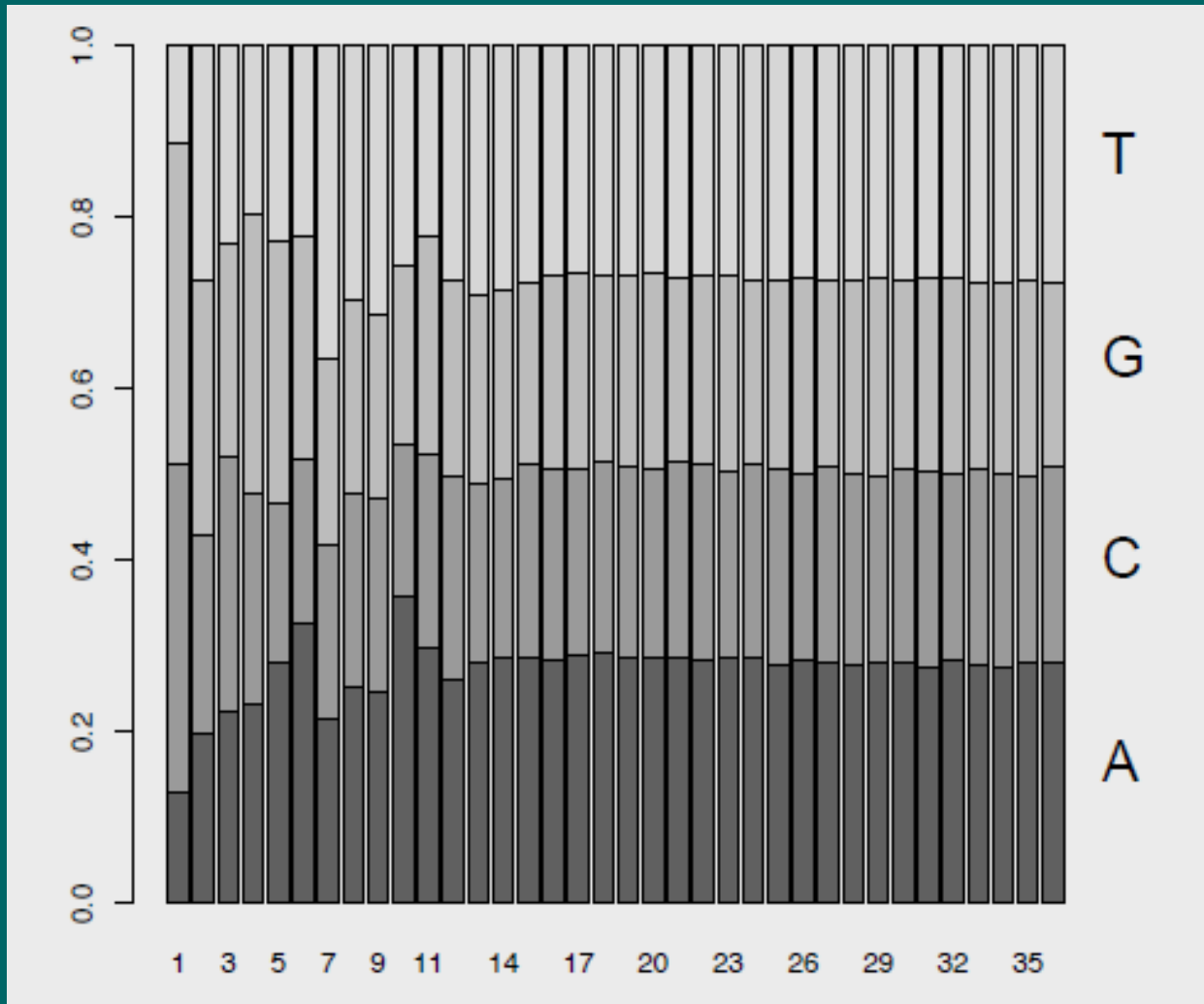
Biases in RNA-Seq

Coverage in RNA-Seq

- When sequencing genomic DNA, the coverage seems reasonably even.
- In RNA-Seq, this quite different

RNA-Seq: Base calls by position in read

(Illumina's standard RNA-Seq protocol)

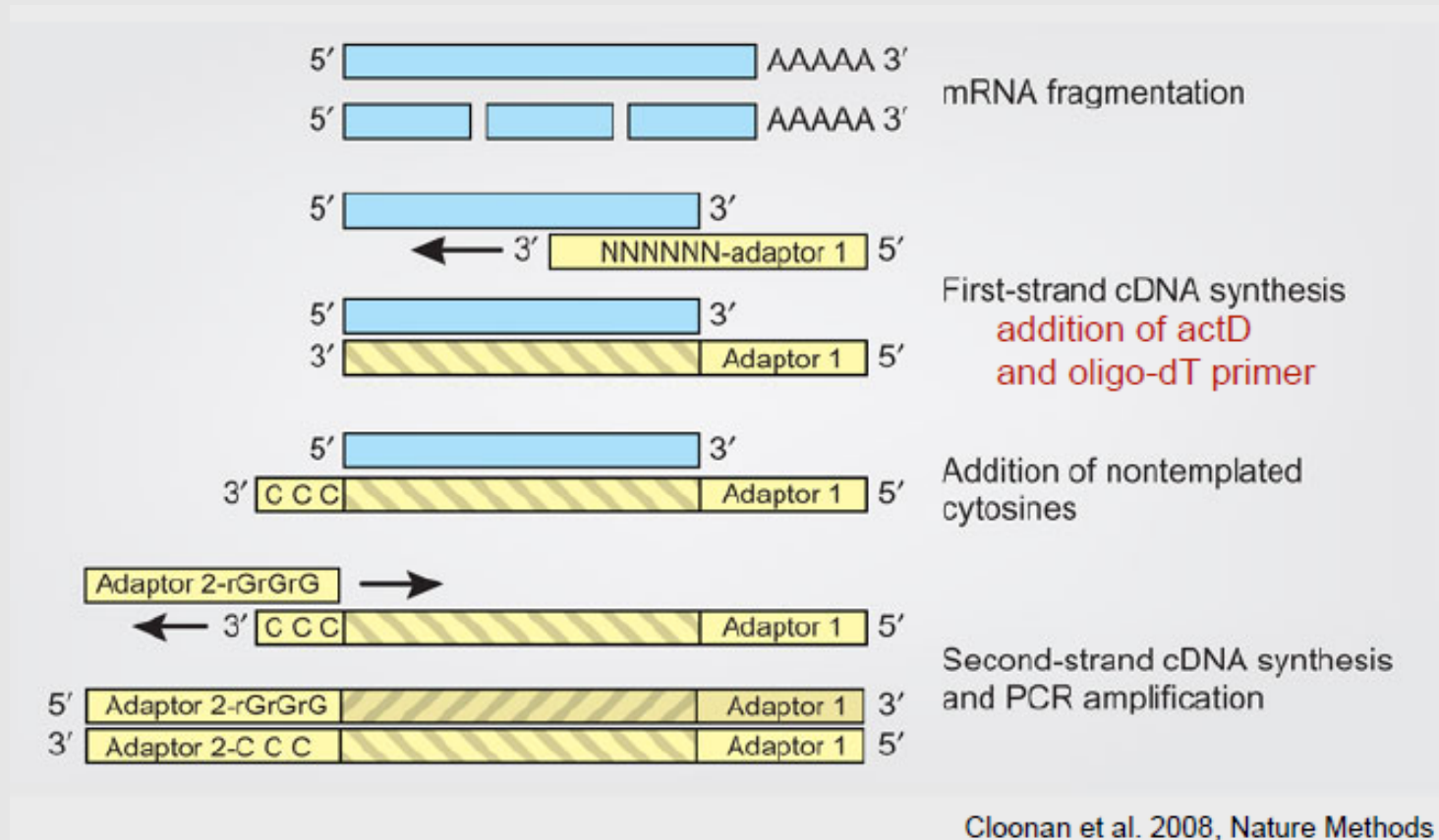


Solexa standard protocol for RNA-Seq

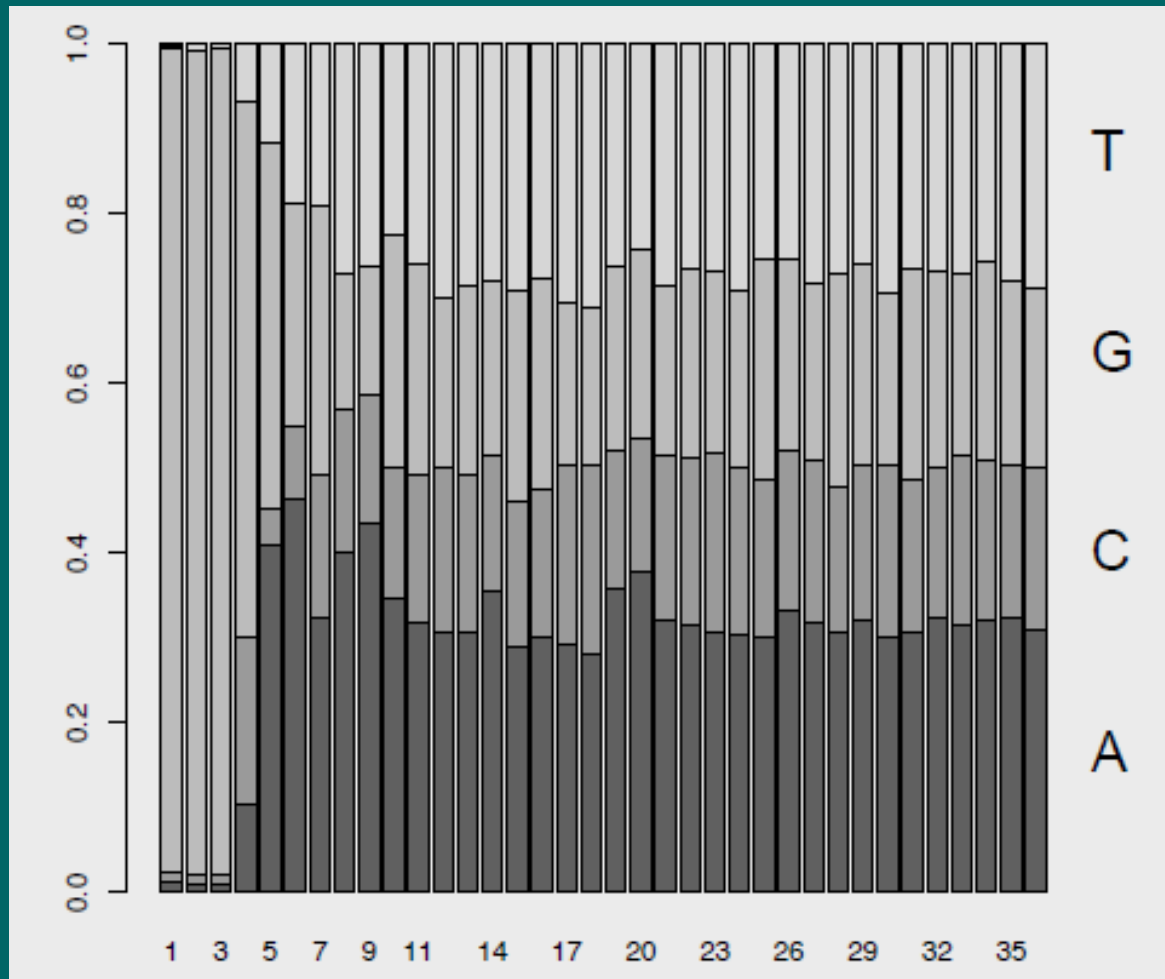


Mortazavi et al. 2008, Nature Methods

Strand-specific RNA-Seq with random hexamer priming



Strand-specific RNA-Seq with random hexamer priming



Are the random hexamers at fault?

Nucleic Acids Research Advance Access published April 14, 2010

Nucleic Acids Research, 2010, 1–7
doi:10.1093/nar/gkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen^{1,*}, Steven E. Brenner² and Sandrine Dudoit^{1,3}

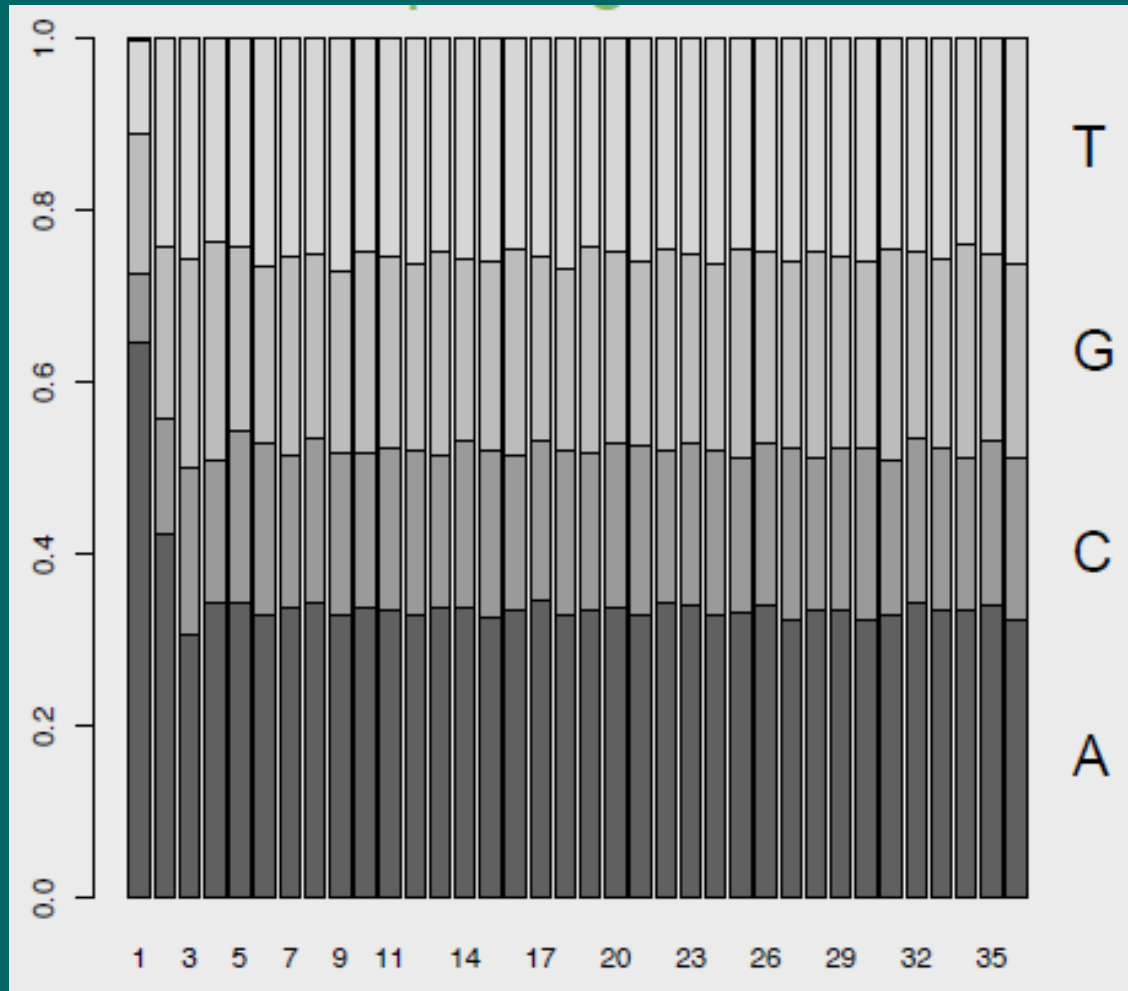
Not all protocols use random hexamers, though.

Strand-specific RNA-Seq with adapter ligation



Lister et al. 2008, Cell

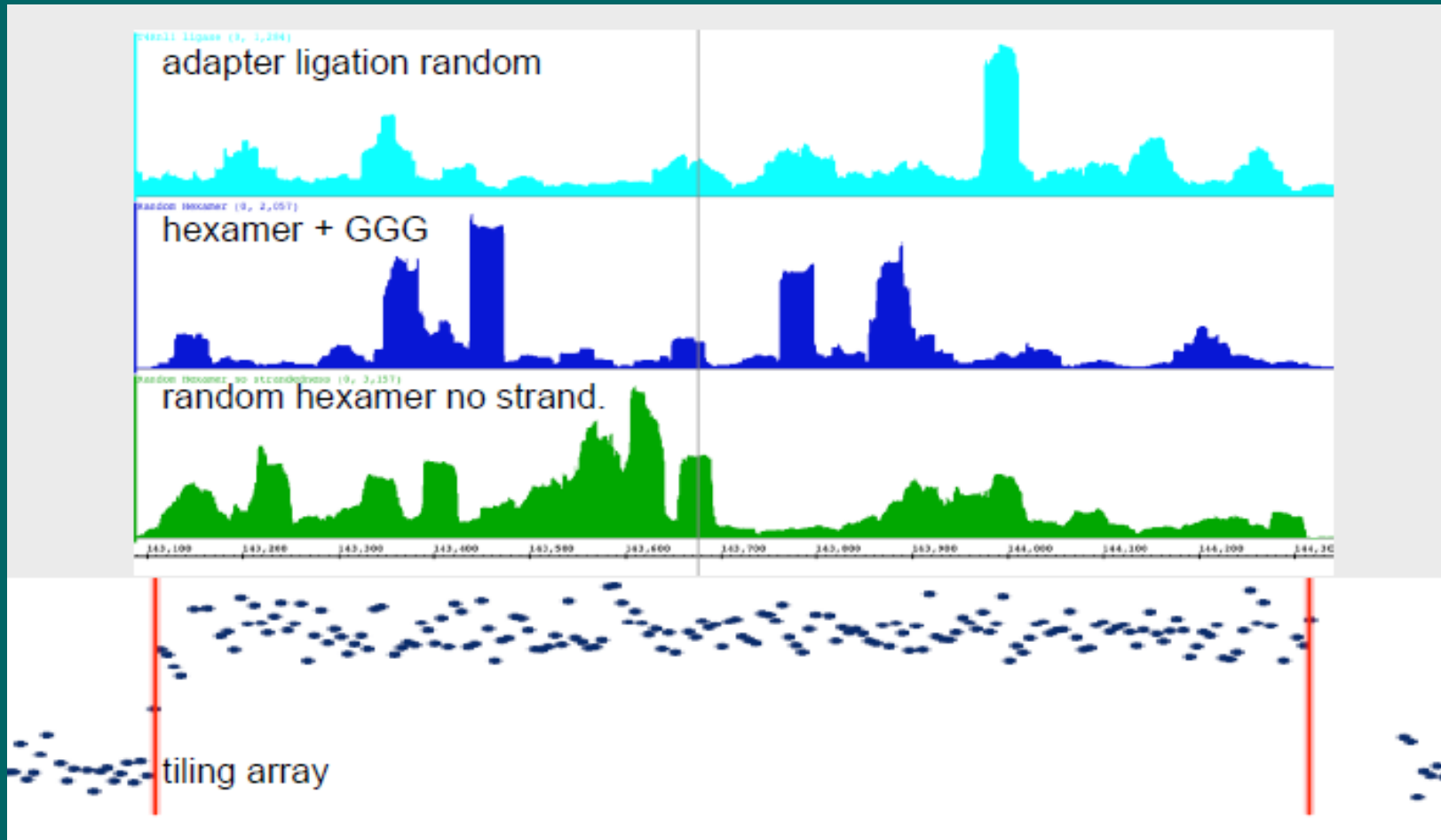
Strand-specific RNA-Seq with random hexamer priming



Problem solved?

Not so fast ...

Coverage of a single-exon gene (PGK1 in yeast)



Coverage in RNA-Seq

- Coverage in RNA-Seq is highly non-uniform
- Within a single exon, there are regions with high coverage and regions with zero coverage.
- These patterns are reproducible.
- They change when the library preparation protocol is changed.
- The binding preferences of random hexamer primers explain them only partially.
- So far, we simply hope that this averages out over the whole transcript.

RNA-Seq: Other biases

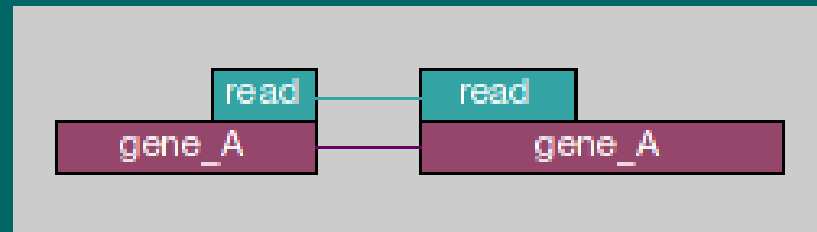
- Depending on transcript and fragment length, and on the protocol, coverage also depends on distance to the ends.

[show slides]

Back to RNA-Seq alignment

Spliced alignment

- When aligning RNA-Seq data to the genome, a read might straddle an intron.



- Most aligner will not be able to align this properly.
- Hence, special tools for RNA-Seq have been developed.

RNA-Seq alignments: Basic strategy (as used e.g. by TopHat/Cufflinks)

- Try to align all reads to the genomes.
- Split those reads that could not be aligned into pieces (of ~ 25 bp); try to align each piece separately.
- For pieces that are still unaligned, look for gapped alignments next to their neighbours.
- Use coverage gaps and intron-straddling reads to infer a parsimonious set of gene models (Dilworth's theorem).
- Infer isoform abundance ratios by likelihood maximization.

The TopHat tool chain

FASTQ file

- **TopHat** (which internally calls **Bowtie**)
- SAM file with mapped reads,
GFF file with inferred splice junctions,
Wiggle file with coverage vector
- **Cufflinks**
- GFF file with gene models and
isoform abundances and uncertainties

Other tools for splica-aware RNA-Seq alignment

- ERange
- TopHat

- SpliceMap
- GSNAP

- SPA
- QPALMA / PALMapper
- MapNext

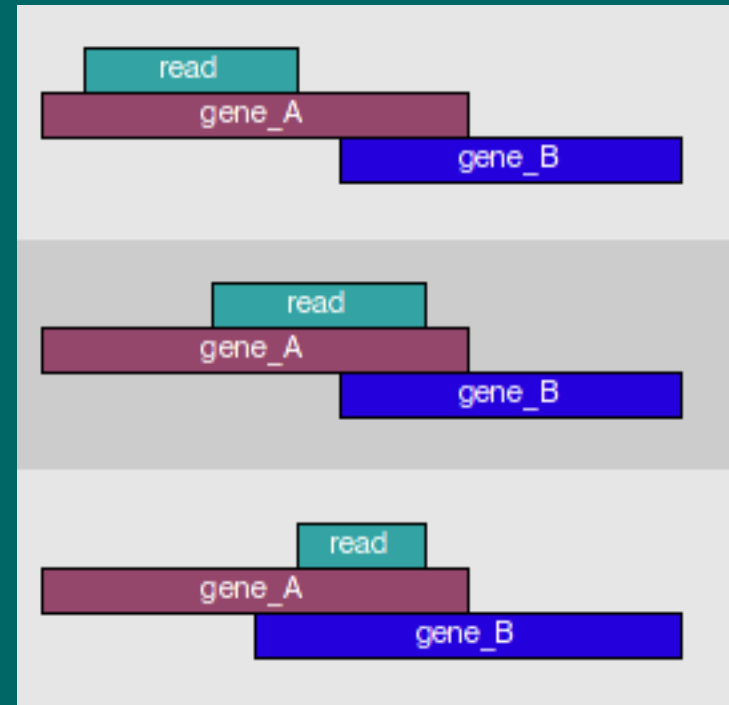
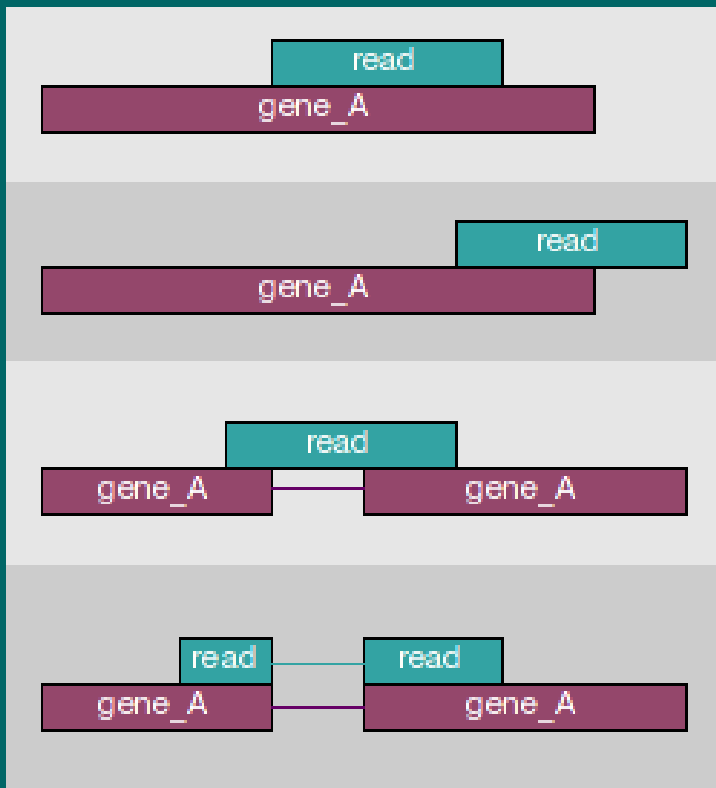
RPKM, FPKM, raw counts

- Mortazavi et al. Suggested to do state mRNA abundances as *RPKM*:
Reads per kilobase of transcript length per one million mapped reads
- The Cufflinks author criticize that the transcript length depends on isoform inference. To emphasize, they call their measure *FPKM*:
Fragments per ...
- Raw counts side-step the issues.

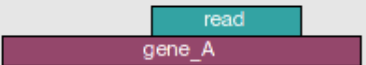
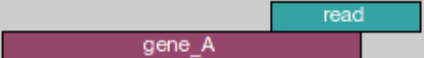


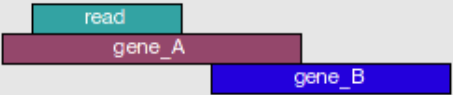

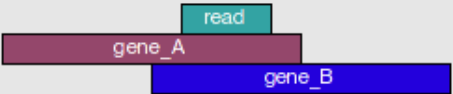
Raw counts

- Raw counts: Simply count how many reads were mapped onto any of the exons of a given gene.
- This side-steps the transcript-length issue.
- However, expression of different genes is no longer comparable.
- As advantage, raw counts allow to quantify shot noise.

Ambiguities with raw counts



Ambiguities in counting

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

RNA-Seq count table

Gene	GLiNS1	G144	G166	G179	CB541	CB660
13CDNA73	4	0	6	1	0	5
A2BP1	19	18	20	7	1	8
A2M	2724	2209	13	49	193	548
A4GALT	0	0	48	0	0	0
AAAS	57	29	224	49	202	92
AACS	1904	1294	5073	5365	3737	3511
AADACL1	3	13	239	683	158	40
[...]						

*