UMASS University of Massachusetts Medical School

RNA-Seq primer

Sequencing: applications

Counting applications

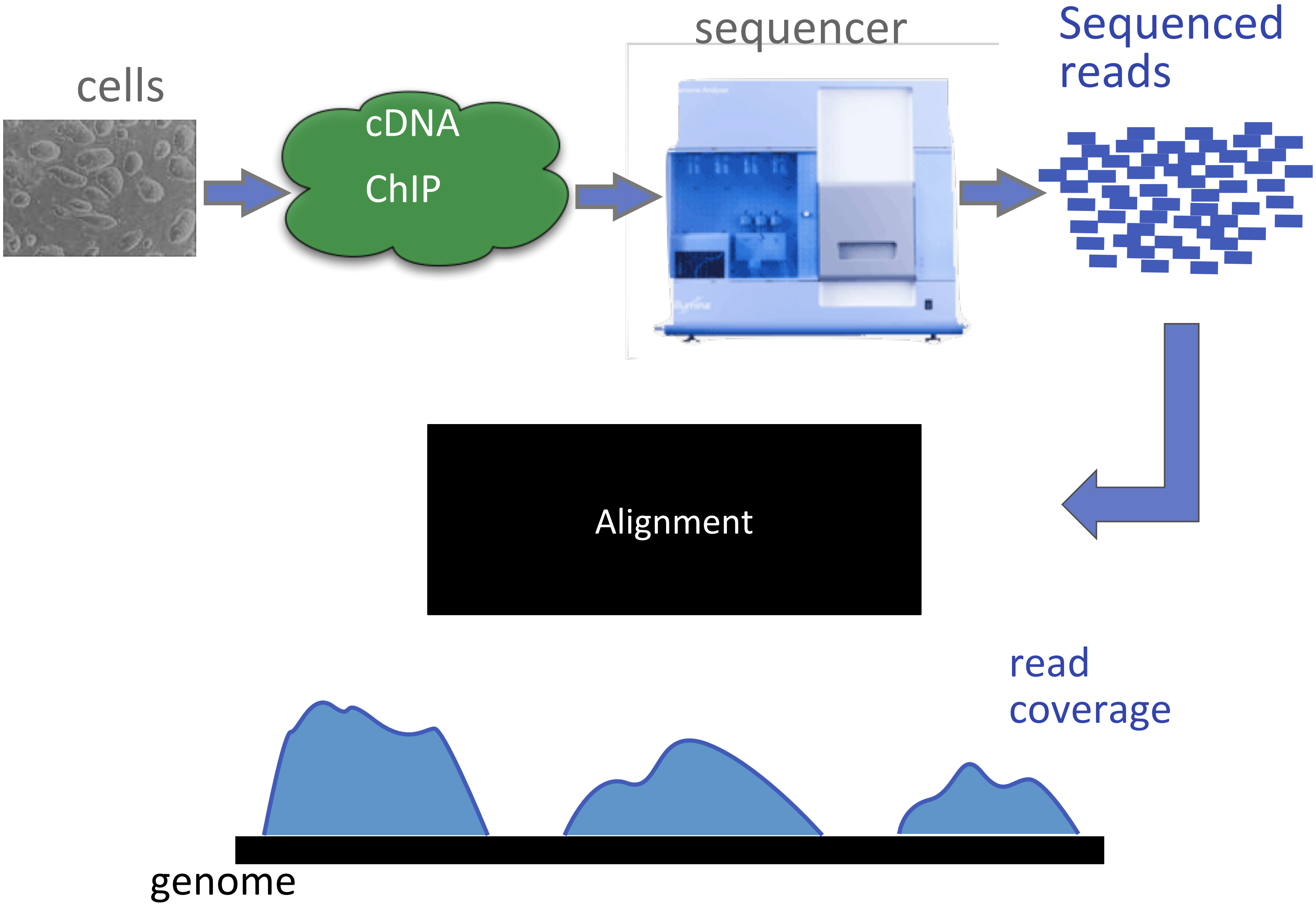
- Profiling
  - microRNAs
  - Immunogenomics
  - Transcriptomics
- Epigenomics
  - Map histone modifications
  - Map DNA methylation
  - 3D genome conformation
- Nucleic acid Interactions

Polymorphism/mutation discovery

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

- Cancer genomics
  - Map translocations, CNVs, structural changes
  - Profile somatic mutations
- Genome assembly
- Ancient DNA (Neanderthal)
- Pathogen discovery
- Metagenomics

Counting applications



cells

cDNA

ChIP

sequencer

Sequenced reads

Alignment

read coverage

genome

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

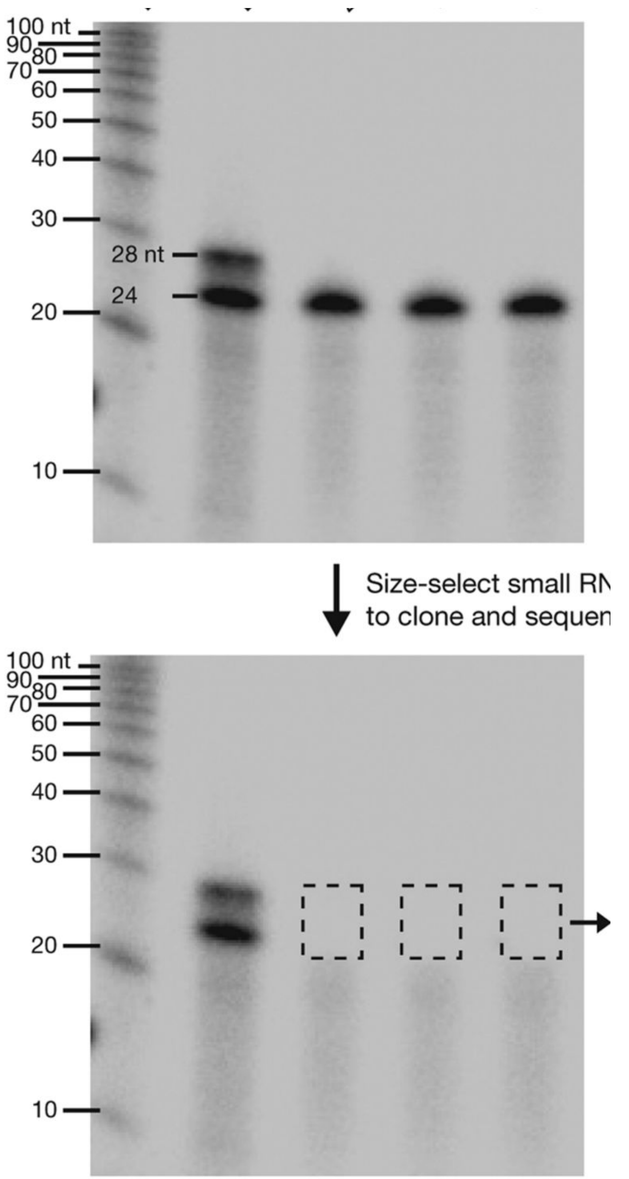
- “Source: intact, **high qual.** RNA (polyA selected or ribosomal depleted)
- RNA → cDNA → sequence
- Uses:
  - Annotation. Requires high depth, paired-end sequencing. ~50 mill
  - Gene expression. Requires low depth, single end sequence, ~ 5-10 mill
  - Differential Gene expression. Requires ~ 5-10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

RNA-Seq libraries III: Small RNA libraries

- Source: size selected RNA
- Uses: miRNA, piRNA annotation and quantification
  - Short single end 30-50 bp reads
  - Require "clipping"
  - Depth: 5-10 mill reads

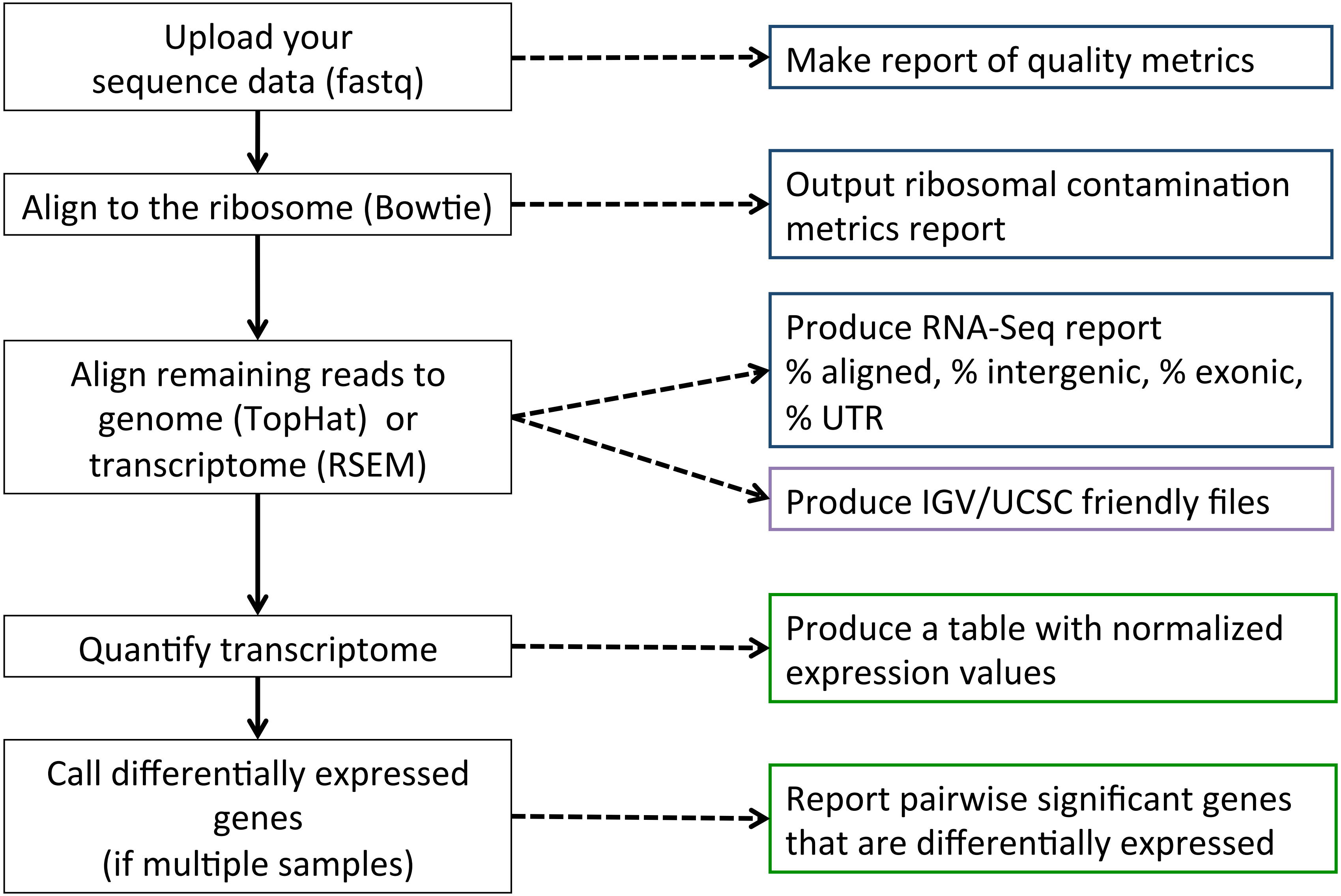


Malonne et al. CSHL protocols, 2011

When you need both annotation and quantification

- Attempt three replicates per condition
- Pool libraries to obtain ~15 mill reads per replicate
- Sequence using paired ends
- Analysis:
  - Merge replicate alignments for annotation
  - Split alignments for differential expression analysis

Our typical RNA quantification pipeline



Upload your sequence data (fastq)

Make report of quality metrics

Align to the ribosome (Bowtie)

Output ribosomal contamination metrics report

Align remaining reads to genome (TopHat) or transcriptome (RSEM)

Produce RNA-Seq report
% aligned, % intergenic, % exonic, % UTR

Produce IGV/UCSC friendly files

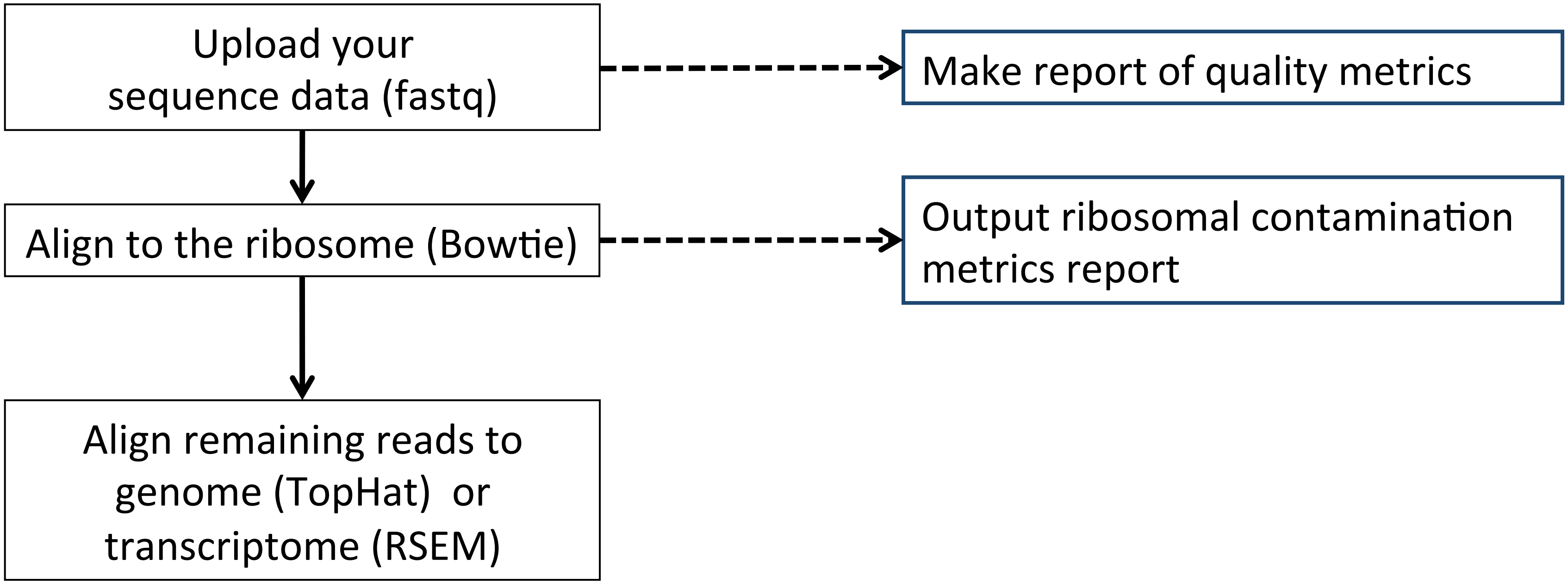
Quantify transcriptome

Produce a table with normalized expression values

Call differentially expressed genes
(if multiple samples)

Report pairwise significant genes that are differentially expressed

Alignment requires pre-processing



Upload your sequence data (fastq)

Make report of quality metrics

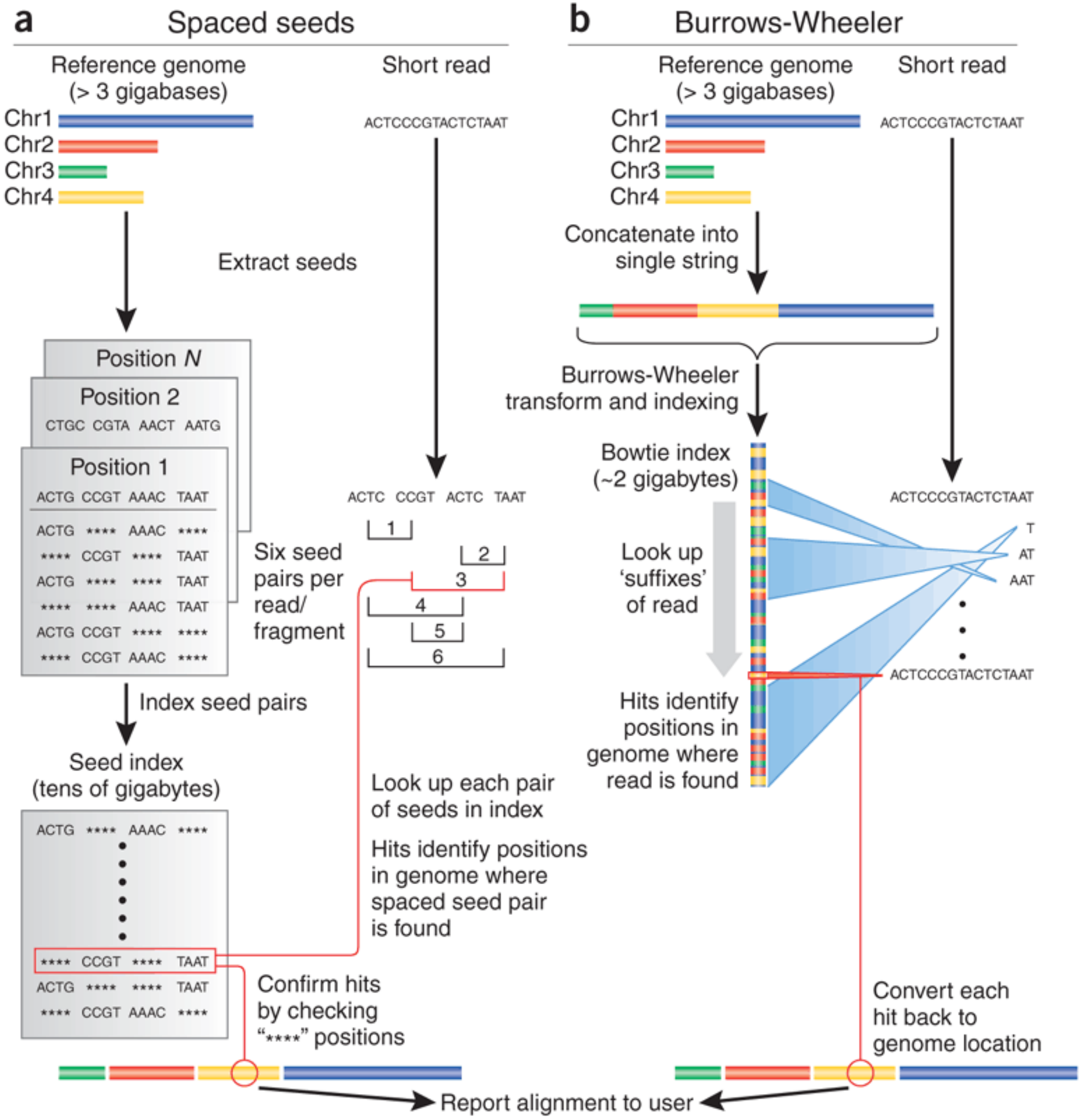
Align to the ribosome (Bowtie)

Output ribosomal contamination metrics report

Align remaining reads to genome (TopHat) or transcriptome (RSEM)

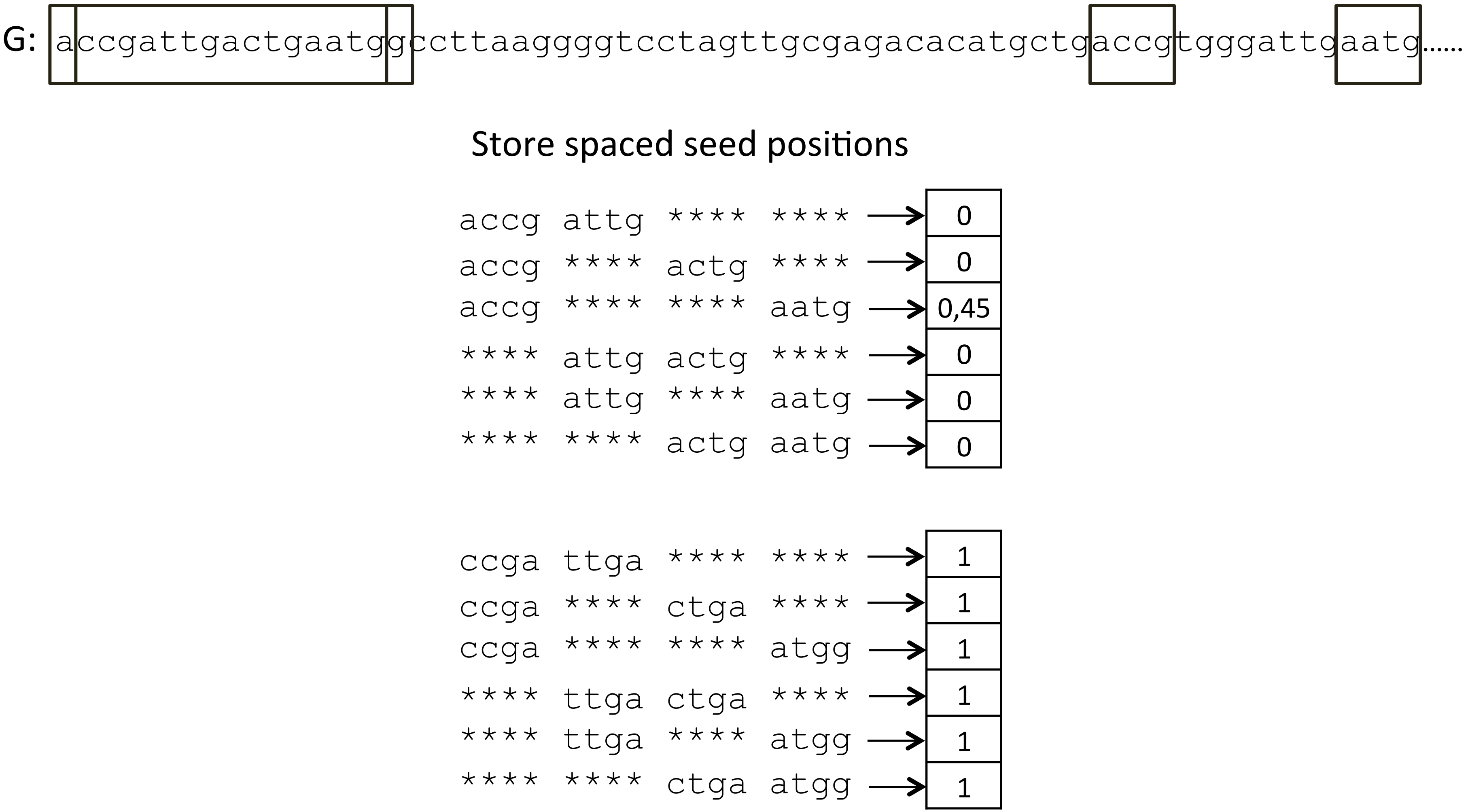
```
bowtie2-build -f mm10.fa mm10

rsem-prepare-reference \
--gtf ucsc.gtf --transcript-to-gene-map ucsc_into_genesymbol.rsem \
mm10.fa mm10.rsem
```



Trapnell, Salzberg, Nature Biotechnology 2009

Spaced seed alignment – Hashing the genome



```
G: accgattgactgaatggccttaaggggtcctagttgcgagacacatgctgaccgtgggattgaatg......
```

Store spaced seed positions

```
accg attg **** ****  → 0
accg **** actg ****  → 0
accg **** **** aatg  → 0,45
**** attg actg ****  → 0
**** attg **** aatg  → 0
**** **** actg aatg  → 0

ccga ttga **** ****  → 1
ccga **** ctga ****  → 1
ccga **** **** atgg  → 1
**** ttga ctga ****  → 1
**** ttga **** atgg  → 1
**** **** ctga atgg  → 1
```

Spaced seed alignment – Mapping reads

G: accgattgactgaatggccttaaggggtcctagttgcgagacacatgctgaccgtgggattgaatg......

accg	attg	****	****	0	✗
accg	****	actg	****	0	✗
accg	****	****	aatg	0,45	✓
****	attg	actg	****	0	✗
****	attg	****	aatg	0	✗
****	****	actg	aatg	0	✗

q: accg atag accg aatg

accgattgactgaatg

2 missmatches

accgtgggattgaatg

5 missmatches

ccga	ttga	****	****	1	✗
ccga	****	ctga	****	1	✗
ccga	****	****	atgg	1	✗
****	ttga	ctga	****	1	✗
****	ttga	****	atgg	1	✗
****	****	ctga	atgg	1	✗

Report position 0

But, how confidence are we in the placement?

$$q_{MS} = -10\log_{10} P(\text{read is wrongly mapped})$$

Mapping quality

What does $q_{MS} = -10\log_{10} P(\text{read is wrongly mapped})$ mean?

Lets compute the probability the read originated at genome position i

q: `accg atag accg aatg`

q_s: `30 40 25 30  30 20 10 20  40 30 20 30  40 40 30 25`

$q_s[k] = -10\log_{10} P(\text{sequencing error at base k})$, the PHRED score. Equivalently:

$$P(\text{sequencing error at base k}) = 10^{-\frac{q_s[k]}{10}}$$

So the probability that a read originates from a given genome position i is:

$$P(q\,|\,G,i) = \prod_{j\ \text{match}} P(q_j\text{good call}) \prod_{j\ \text{missmatch}} P(q_j\text{bad call}) \approx \prod_{j\ \text{missmatch}} P(q_j\text{bad call})$$

In our example

$$P(q\,|\,G,0) = \left[(1-10^{-3})^6(1-10^{-4})^4(1-10^{-2.5})^2(1-10^{-2})^2\right]\left[10^{-1}10^{-2}\right] = [0.97]*[0.001] \approx 0.001$$

Mapping quality

What we want to estimate is $q_{MS} = -10\log_{10} P(\text{read is wrongly mapped})$

That is, the posterior probability, the probability that the region starting at i was sequenced *given* that we observed the read *q*:

$$P(i \mid G,q) = \frac{P(q \mid G,i)P(i \mid G)}{P(q \mid G)} = \frac{P(q \mid G,i)P(i \mid G)}{\sum_j P(q \mid G,j)}$$

Fortunately, there are efficient ways to approximate this probability (see Li, H *genome Research* 2008, for example)

$$q_{MS} = -10\log_{10}(1 - P(i \mid G,q))$$

Considerations

- Trade-off between sensitivity, speed and memory
  - Smaller seeds allow for greater mismatches at the cost of more tries
  - Smaller seeds result in a smaller tables (table size is at most 4^k), larger seeds increase speed (less tries, but more seeds)

Short read mapping software

Seed-extend

	Short indels	Use base qual
Maq	**No**	**YES**
RMAP	Yes	**YES**
SeqMap	Yes	NO
SHRiMP	Yes	NO

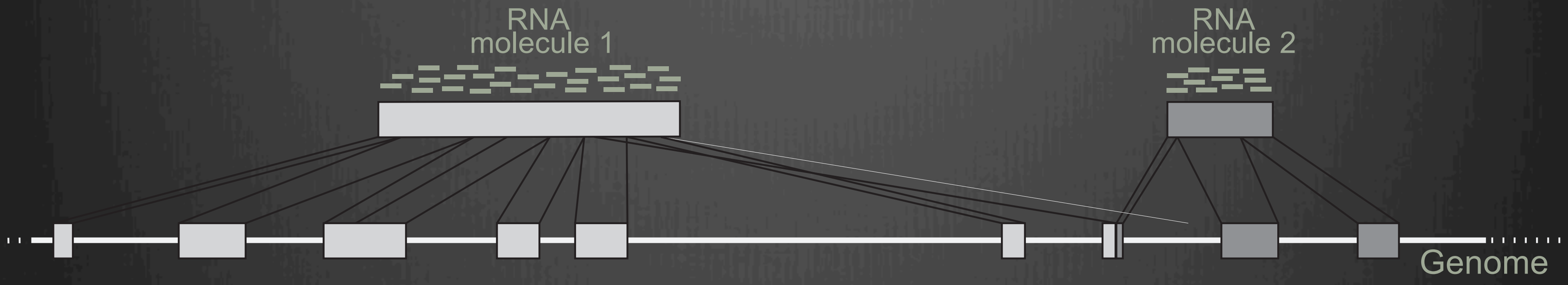
BWT

	Use Base qual
BWA	**YES**
Bowtie	NO
Stampy*	YES
Bowtie2*	(NO)

***Stampy is a hybrid approach which first uses BWA to map reads then uses seed-extend only to reads not mapped by BWA**

***Bowtie2 breaks reads into smaller pieces and maps these “seeds” using a BWT genome.**

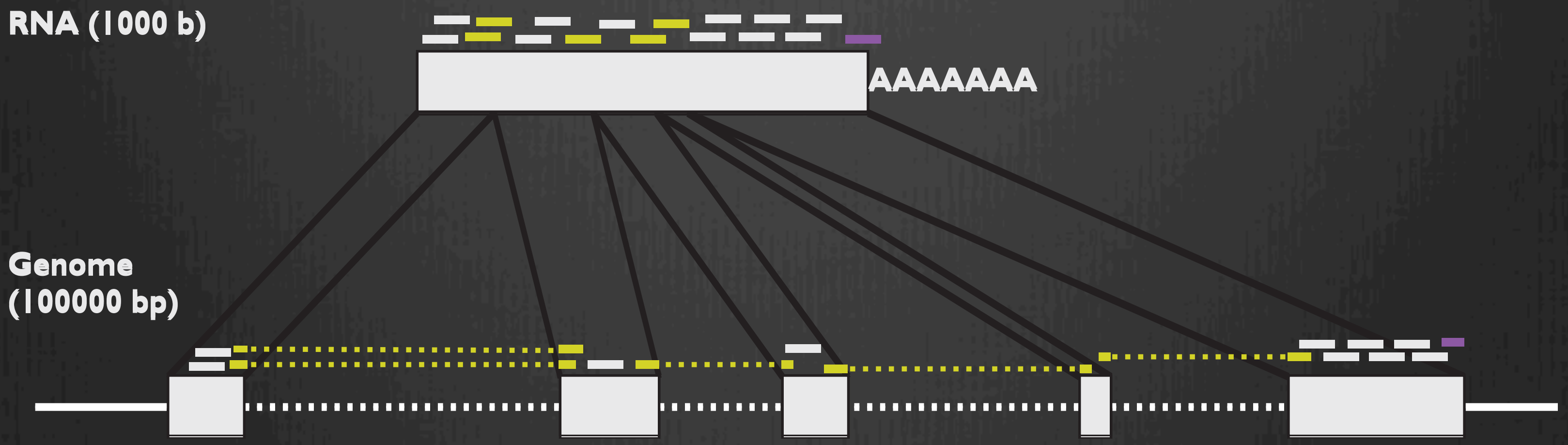
RNA-Seq Read mapping



RNA
molecule 1

RNA
molecule 2

Genome

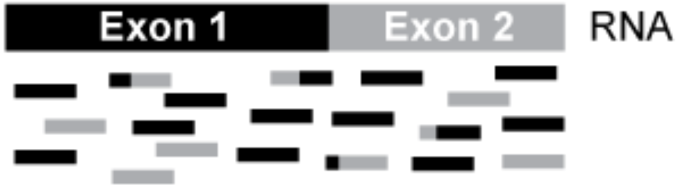


RNA (1000 b)

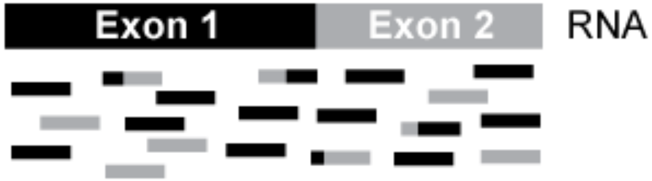
AAAAAAA

Genome
(100000 bp)

Mapping RNA-Seq reads: Seed-extend spliced alignment (e.g. GSNAP)



Mapping RNA-Seq reads: Exon-first spliced alignment (e.g. TopHat)



Short read mapping software for RNA-Seq

Seed-extend

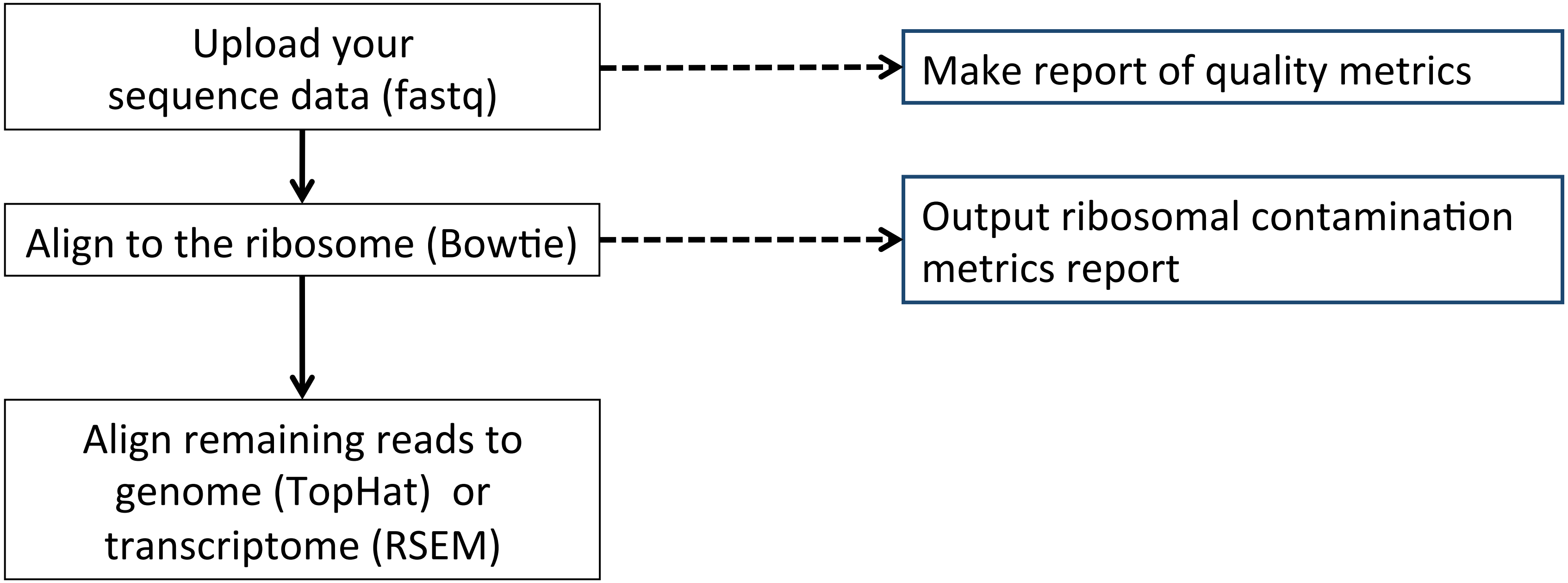
	Short indels	Use base qual
GSNAP	**Yes**	?
QPALMA	Yes	NO
BLAT	Yes	NO

Exon-first

	Use base qual
STAR	NO
TopHat	NO

Exon-first alignments will map contiguous first at the expense of spliced hits

Alignment requires pre-processing



Upload your sequence data (fastq)

Make report of quality metrics

Align to the ribosome (Bowtie)

Output ribosomal contamination metrics report

Align remaining reads to genome (TopHat) or transcriptome (RSEM)

```
tophat2 --library-type fr-firststrand --segment-length 20 \
-G  genome.quantification/ucsc.gtf -o  tophat/th.quant.ctrl1 \
genome.quantification/mm10 fastq.quantification/control_rep1.1.fq \
fastq.quantification/control_rep1.2.fq
```

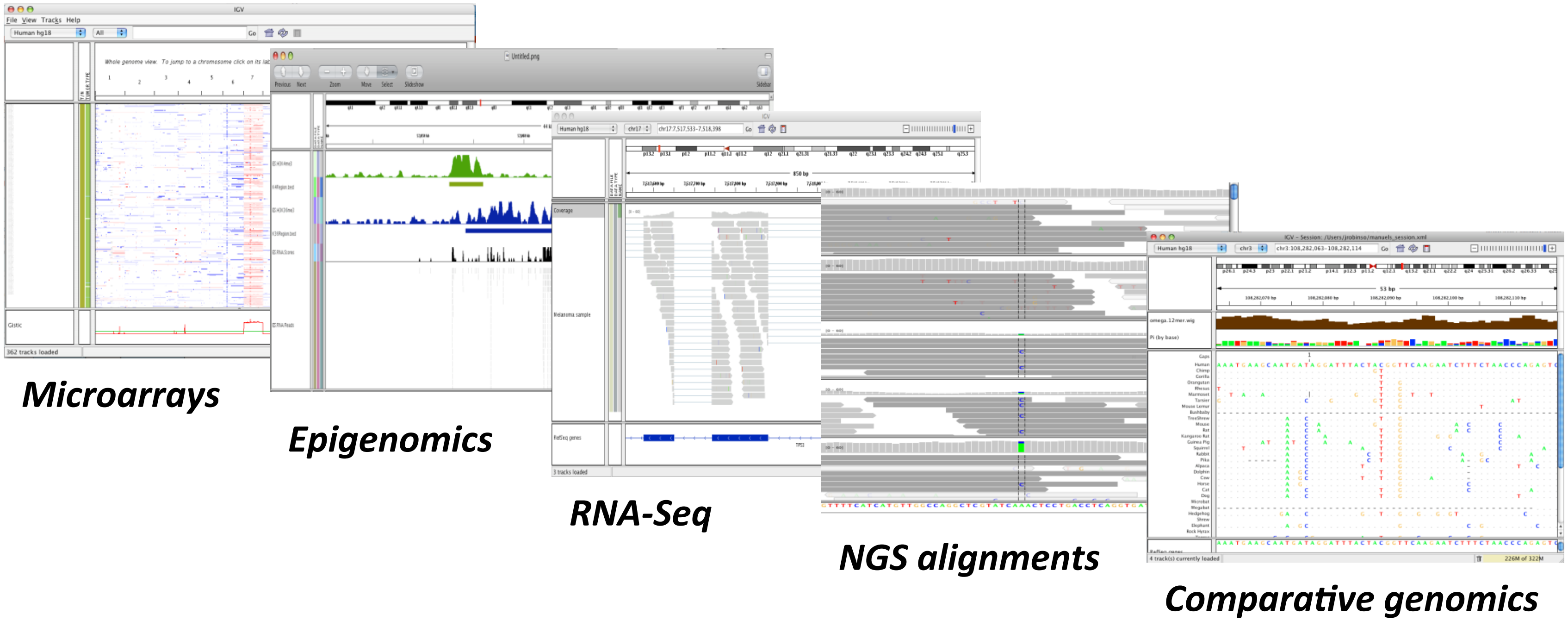
```
/project/umw_biocore/bin/igvtools.sh count -w 5 tophat/th.quant.ctrl1.bam \
tophat/th.quant.ctrl1.bam.tdf genome.quantification/mm10.fa
```

IGV: Integrative Genomics Viewer

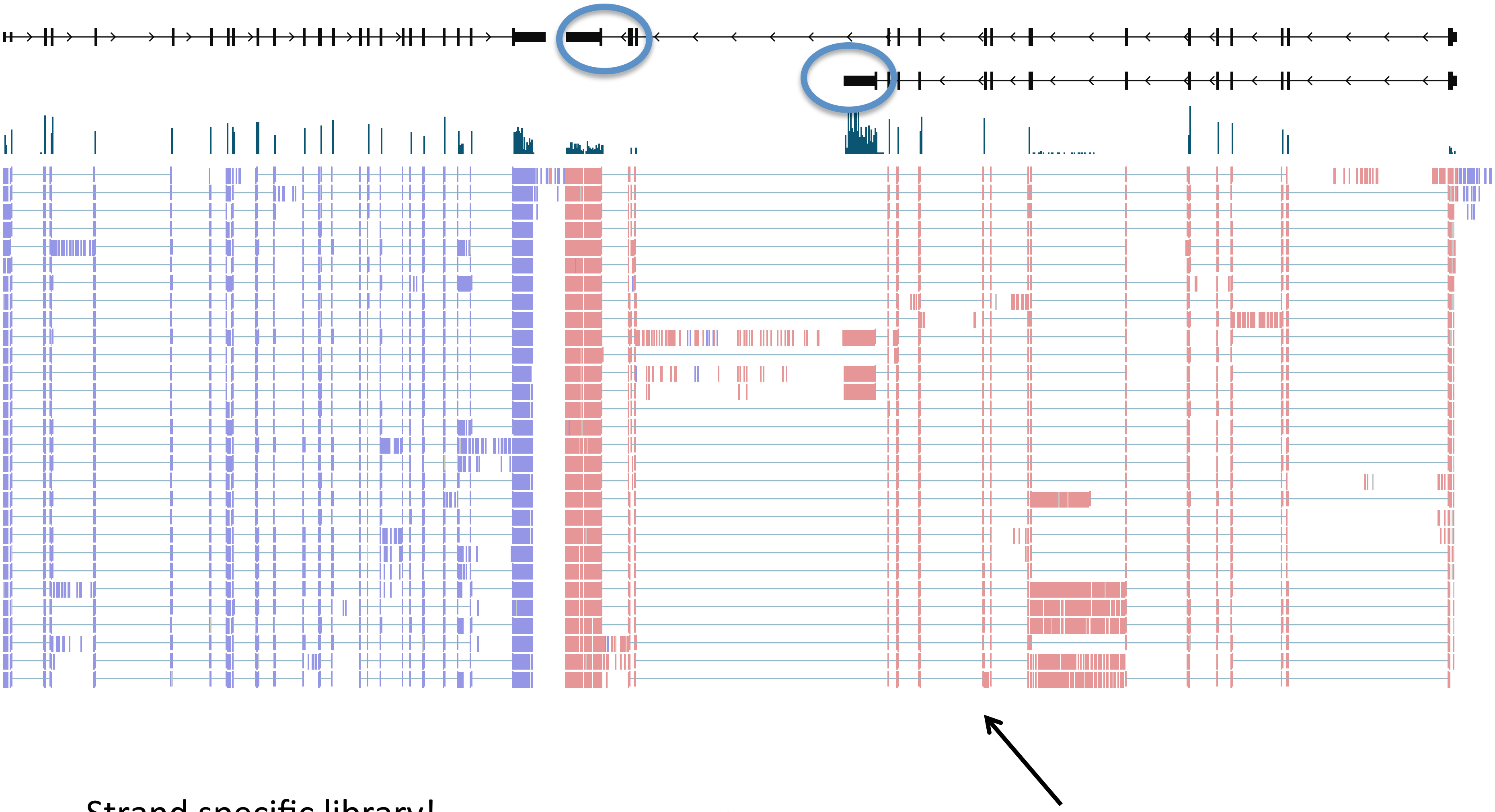
A desktop application

for the visualization and interactive exploration

of genomic data



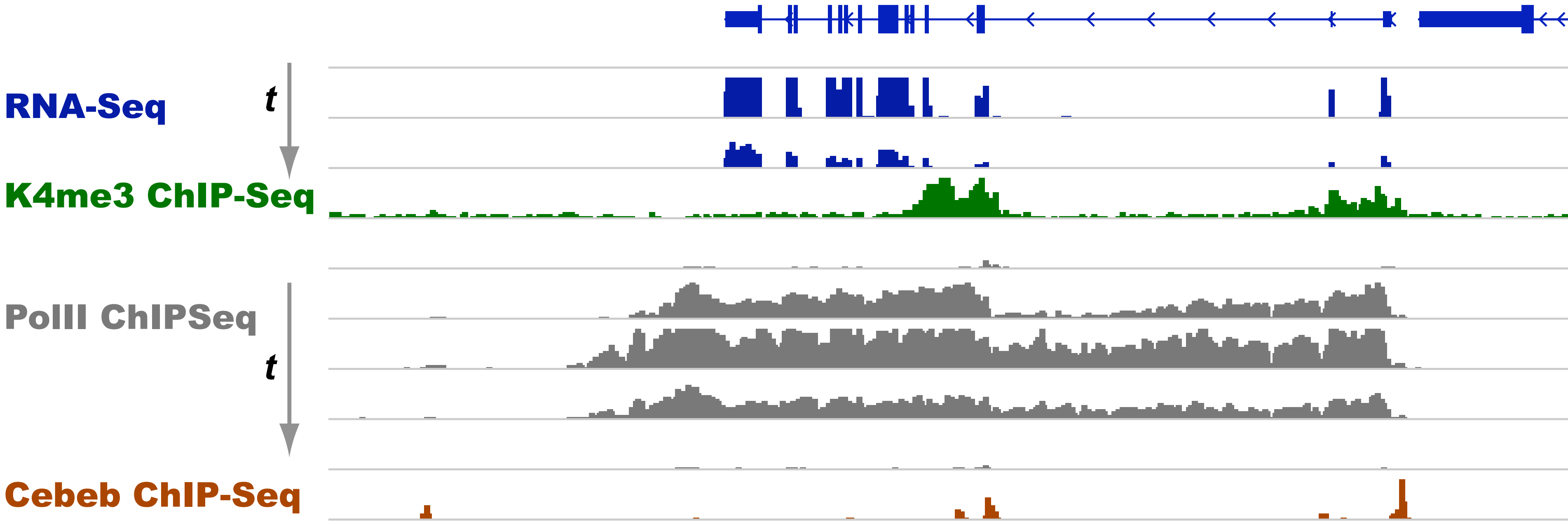
Visualizing read alignments with IGV — RNASeq



Strand specific library!

Gap between reads spanning exons

Visualizing read alignments with IGV — zooming out



RNA-Seq

K4me3 ChIP-Seq

PolII ChIPSeq

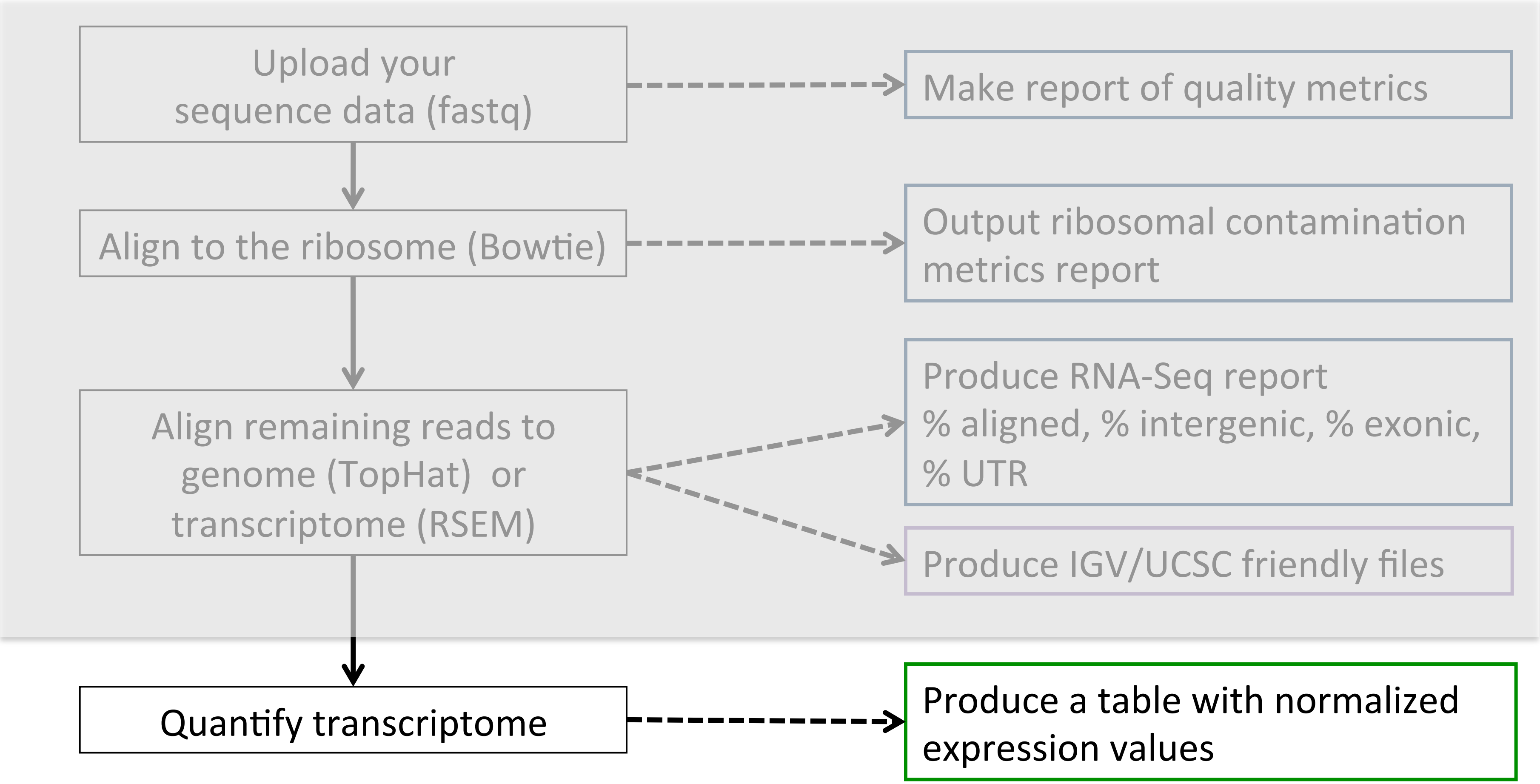
Cebeb ChIP-Seq

How do “short” read aligners responded to read increase?

- Break reads into seeds (e.g. 16nt every 10nt)
- Use BWT or HashTable to find candidate positions
- Prioritize candidates
- Extend top candidates using classical alignment techniques.

Aligner	Technique
TopHat2 (Bowtie2)	BWT
GSNAP	Hash Table
STAR	Suffix (similar to TopHat)

Computing gene expression



Upload your sequence data (fastq)

Make report of quality metrics

Align to the ribosome (Bowtie)

Output ribosomal contamination metrics report

Align remaining reads to genome (TopHat) or transcriptome (RSEM)

Produce RNA-Seq report
% aligned, % intergenic, % exonic, % UTR

Produce IGV/UCSC friendly files

Quantify transcriptome

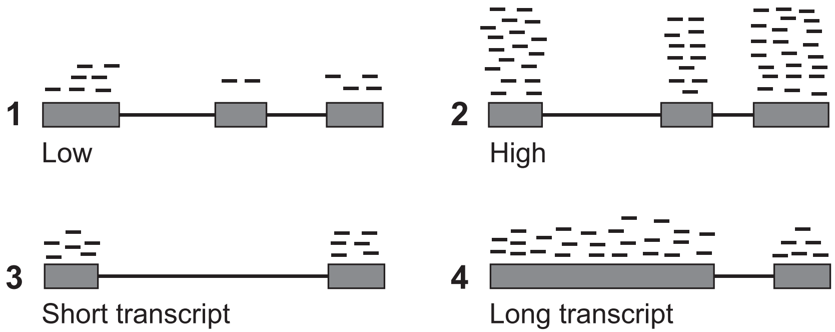
Produce a table with normalized expression values

```
rsem-calculate-expression --paired-end --strand-specific -p 2 \
 --output-genome-bam fastq.quantification/control_rep1.1.fq \
fastq.quantification/control_rep1.2.fq genome.quantification/mm10.rsem \
rsem/ctrl1.rsem
```

RNA-Seq quantification

- Is a given gene (or isoform) expressed?
- Is expression gene A > gene B?
- Is expression of gene A isoform a_1 > gene A isoform a_2?
- Given two samples is expression of gene A in sample 1 > gene A in sample 2?

Quantification: only one isoform



$$RPKM = 10^9 \frac{\#reads}{length \times TotalReads}$$

Reads per kilobase of exonic sequence per million mapped reads (*Mortazavi* et al Nature methods 2008)

•Fragmentation of transcripts results in length bias: longer transcripts have higher counts

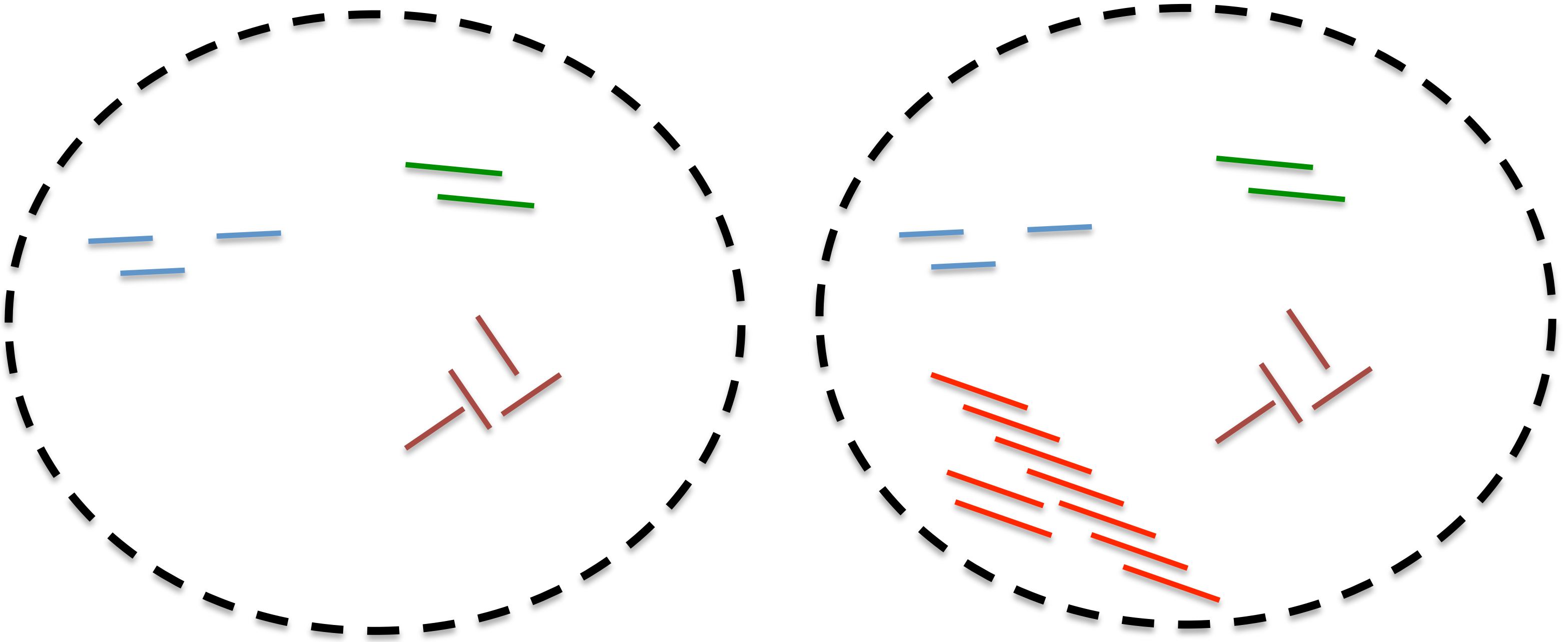
•Different experiments have different yields. Normalization is key for cross lane comparisons

Complexity increases when multiple isoforms exist

Normalization for comparing two different genes

- To compare within a sequence run (lane), RPKM accounts for length bias.
- RPKM is not optimal for cross experiment comparisons.
  - Different samples may have different compositions.
- FPKM superseded RPKM
- And later TPM = 10^6 x Fraction of transcript

Normalization for comparing a gene across samples



Cell type I

Cell type II

Normalizing by total reads does not work well for samples with very different RNA composition