Functional Genomics with Deep Sequencing

Gunnar Rätsch[1] **and Ali Mortazavi**[2]

[1] FML, Max Planck Society, Tübingen, Germany

[2] Caltech, Pasadena, California, USA

ISMB Tutorial, July 10, 2010

Roadmap

Introduction
- Quantitation in biology
- The Rise of the Sequencers
- Sequence-counting Assays
- Important Considerations

Transcriptome Analysis
- Motivation
- Common Steps and Challenges
- RNA-seq Read Mapping
- Quantitation & Diff. Expression
- Transcript Reconstruction
- Transcript Assembly

ChIP-seq
- Introduction
- Mapping the reads
- Taking Advantage of Directionality
- Shifting Reads
- Motif Finding

Integration
- The Combinatorial Problem
- Analyzing Chip-seq Datasets Jointly
- Segmenting the Genome
- Learning Regulatory States
- Joint Analysis of Chip- and RNA-Seq

Introduction



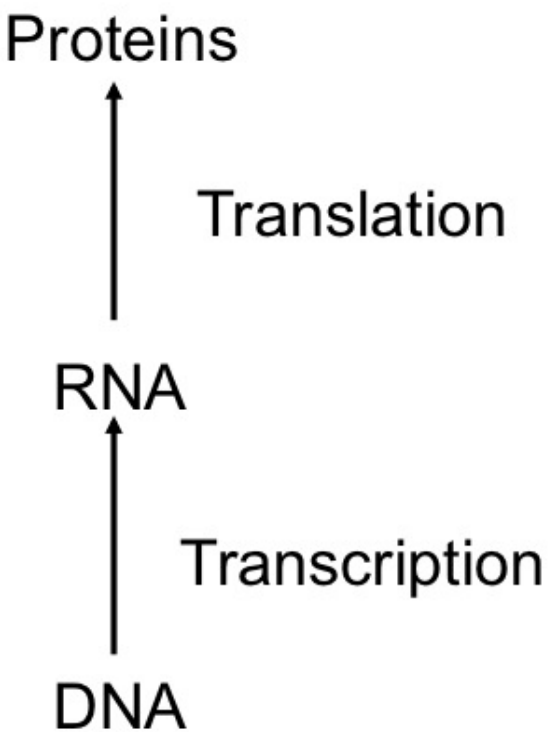
- Quantitation in Biology
- The Rise of the Sequencers
- Sequence-Counting Assays
  - ChIP-seq
  - RNA-seq
- Important Considerations:
  - Read Length
  - Error Rate
  - Mapping Strategy

Quantitation in Biology



- Biology has a rich tradition of quantitative analysis
  - Biostatistics for ecology and genetics
  - Biochemistry & X-ray crystallography
- But the rise of molecular biology in the 1970s and 1980s led to a more qualitative approach:
  - "I see a band on this gel at the right location"
  - "We have cloned gene X, which is related to gene Z."

The Central Dogma and Transcriptional Regulation

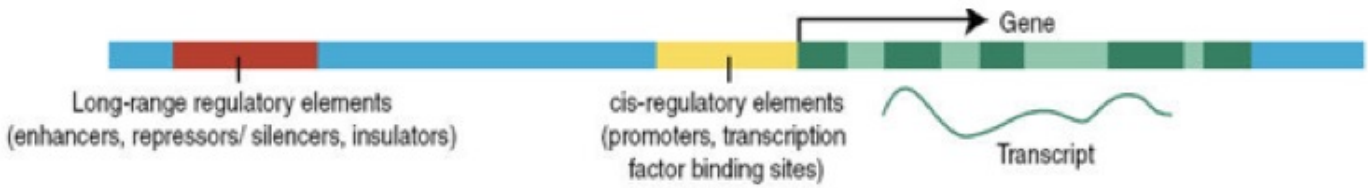


- Transcription is tightly regulated as part of development
  - Also easier to measure than protein levels
- Critical questions:
  - Which genes are turned on and off in a given cell at a given time?
  - What is the expression level of these genes?
  - How is this all encoded in the DNA?

Generalizing by Going Genome-wide



- The cloning of individual genes during the last quarter of the 20th century revealed tantalizing hints to the structure of eukaryotic genes.



- But a comprehensive picture of gene regulation needs to include the entire gene collection for a given organism as well as its intergenic regions, collectively called “genome”.

Sequencing Genomes



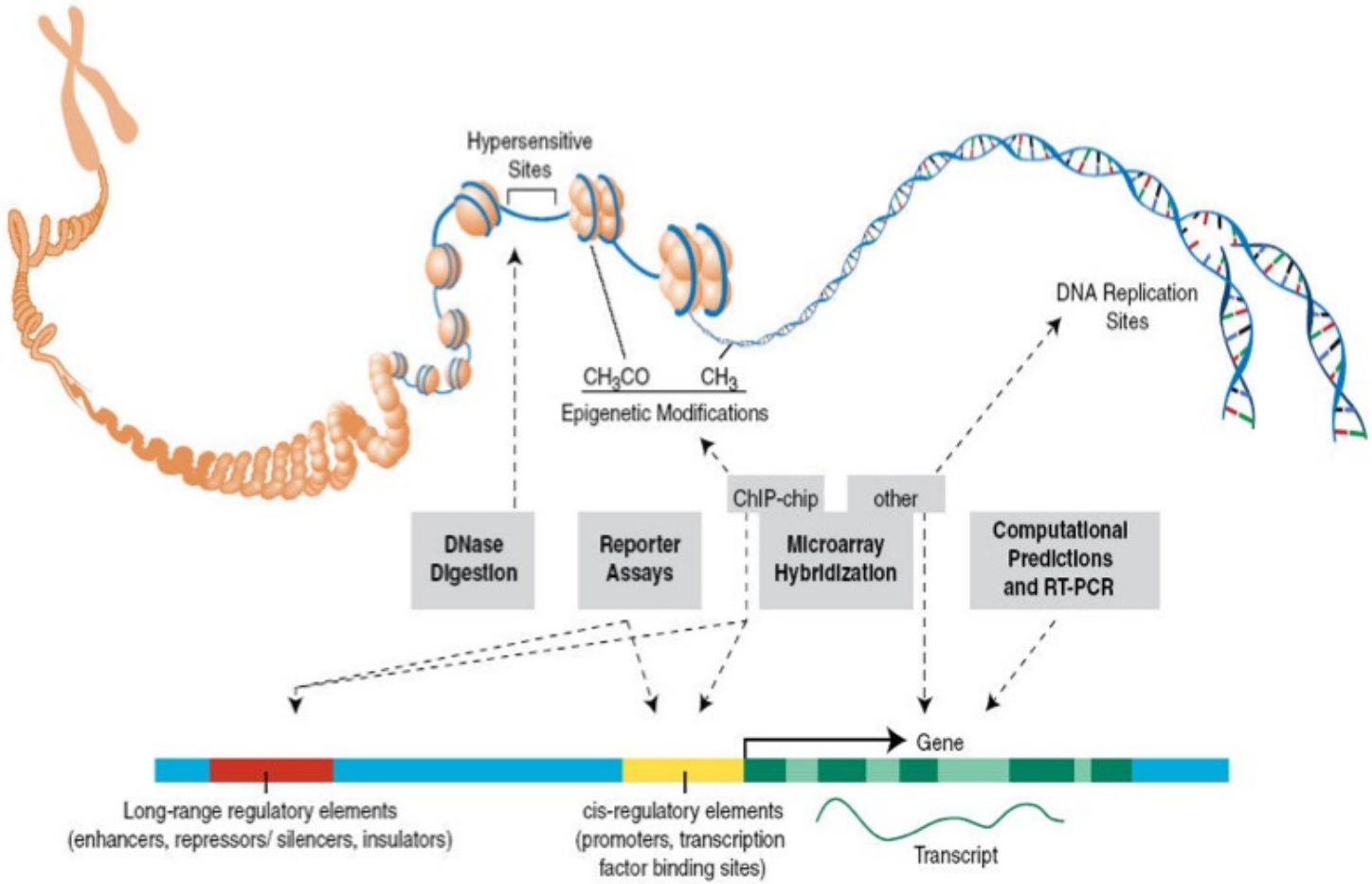
- Manual sequencing using gels was quickly automated by the mid-1980s.
  - Applied Biosystems
  - Capillary sequencing
  - 150-200 bp at first, paired 600-700 bp now
- A concerted effort from the NIH to sequence genomes of model organisms:
  - E coli (bacteria) 4.5 Mb (1997)
  - S cerevisiae (yeast) 6.0 Mb (1997)
  - C elegans (nematode worm) 98 Mb (1998)
  - Human 3 Gb (2000)
    - Estimated cost: $2.7 billion in 1991 dollars
    - Estimated time in 1990: 15 years

The Encyclopedia of DNA Elements



- Once the genome was sequenced, the next question became how to make sense of it.
  - Which nucleotides are functional ?
  - What is their function
- The National Human Genome Research Institute (NHGRI) started the (mod)ENCODE projects to annotate the human and model organism genomes:
  - 2004: Human 1%
  - 2007: Human whole-genome
  - 2008: Drosophila and C. elegans
  - 2010: Mouse

Originally, microarrays were used to read out genome-wide functional assays



ENCODE Project Consortium (2004). Science 306: 636.

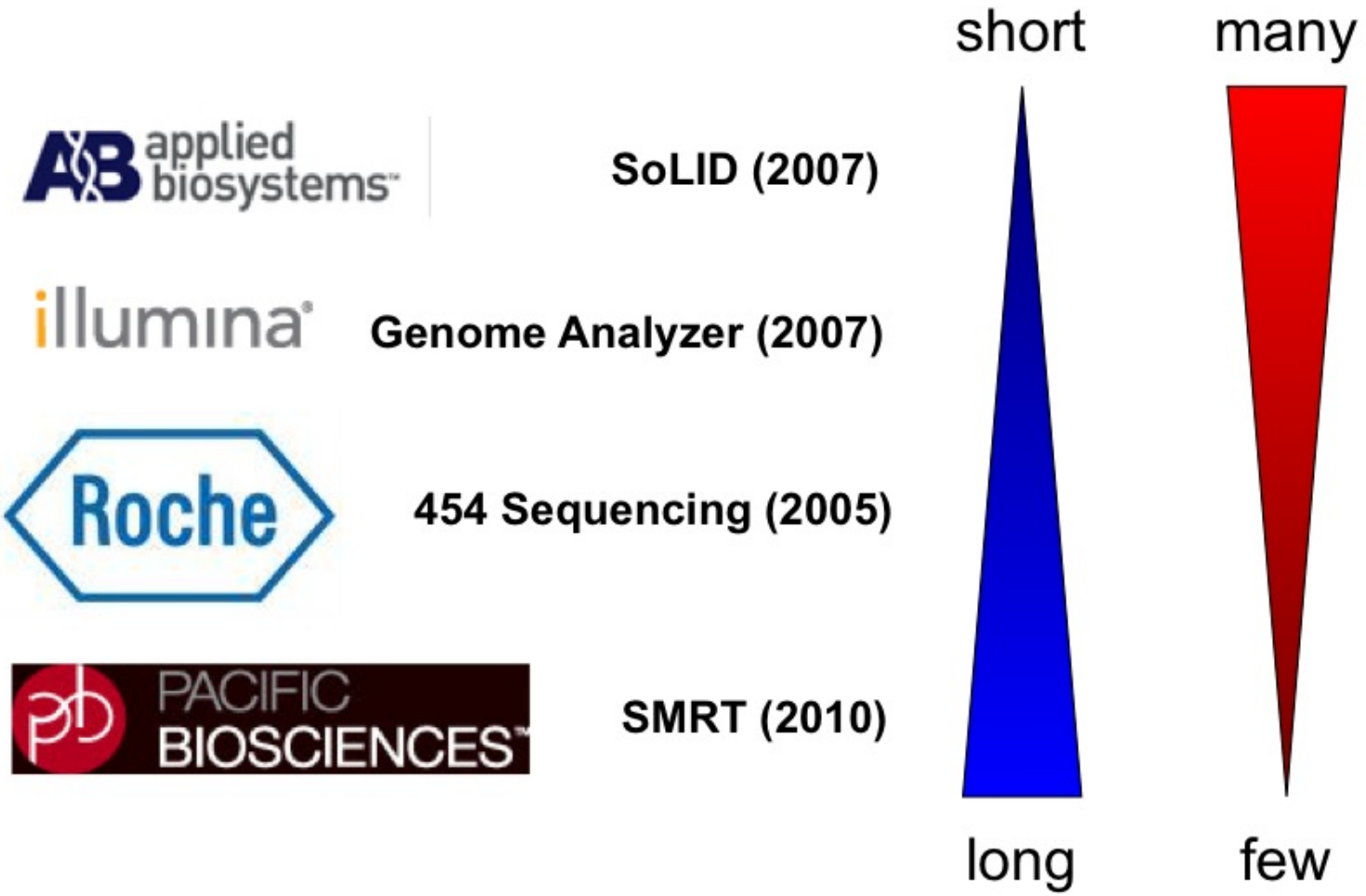
An Investment: the $1000 human genome

- The situation by the turn of the century:
  - Cost of a human-size draft genome (8x) in 2003: $50M
  - 4 main publicly supported genome centers in the US received the bulk of the money set aside for sequencing:
    - MIT (Broad)
    - Washington University
    - Baylor
    - DOE
- In 2003, NHGRI committed to develop next-generation sequencing technologies to lower the cost of 30x a human genome (~100 Gbp):
  - $100,000 genome
  - $1,000 genome
- Originally targeted for *de novo* sequencing, and resequencing for population genetics.



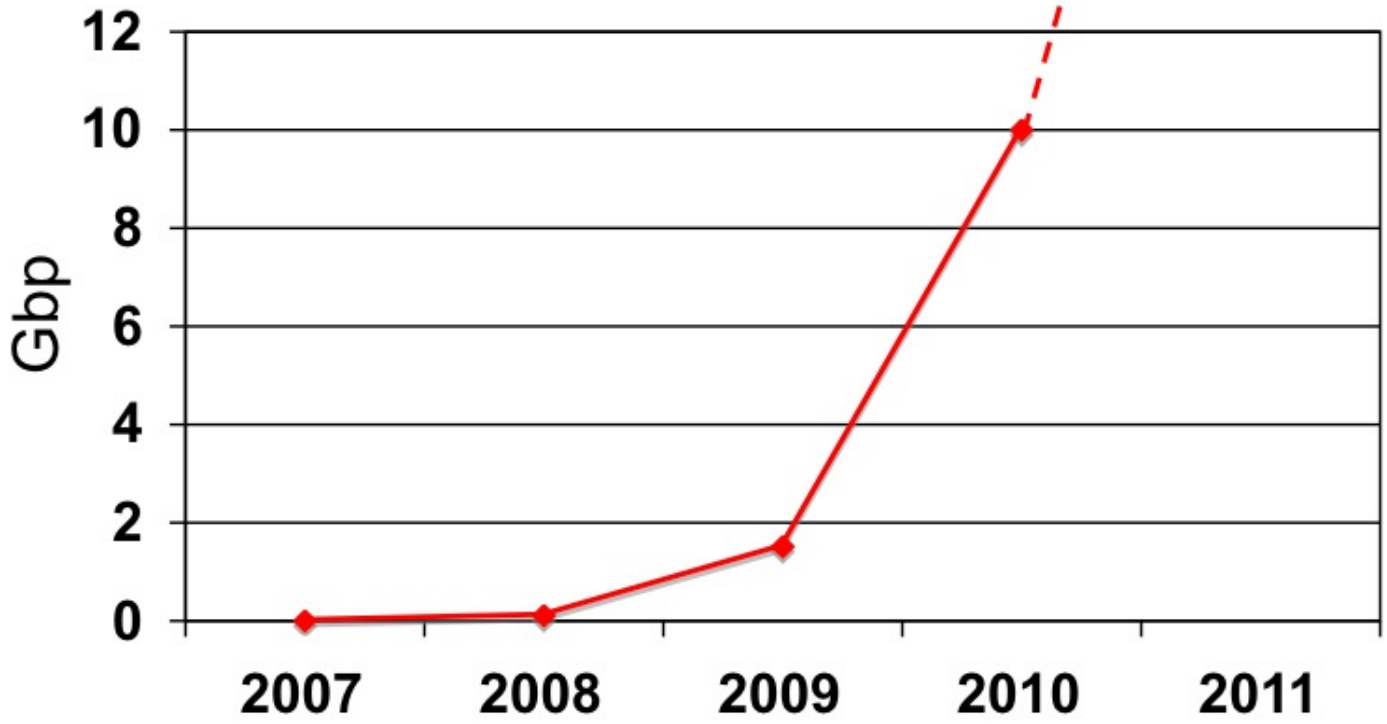
A variety of current technologies are available with different tradeoffs





Exponential growth of Illumina mapped sequence / lane throughput





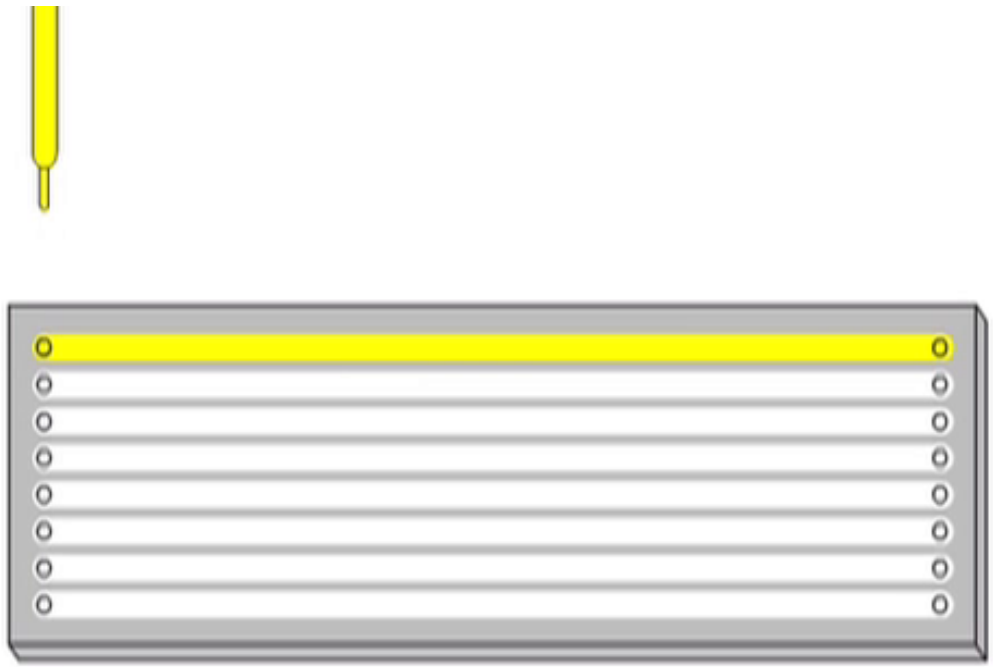
	Jan-07	Jan-08	Jan-09	Jan-10	Jan-11
Gbp / lane	0.025	0.128	1.5	10	30-50 ?
Read type:	1x25	1x36	2x75	2x100	2x150

Cost/lane is relatively stable at $600 to $1,200

Illumina and SOLiD Sequencing Technologies

Illumina Sequencing

- *Solexa* released a sequencing machine in 2006
- Fragment sizes from 28 – 100
- Probes are fixed to a glass plate "flow cell"
- Reagents are directed through flow cell

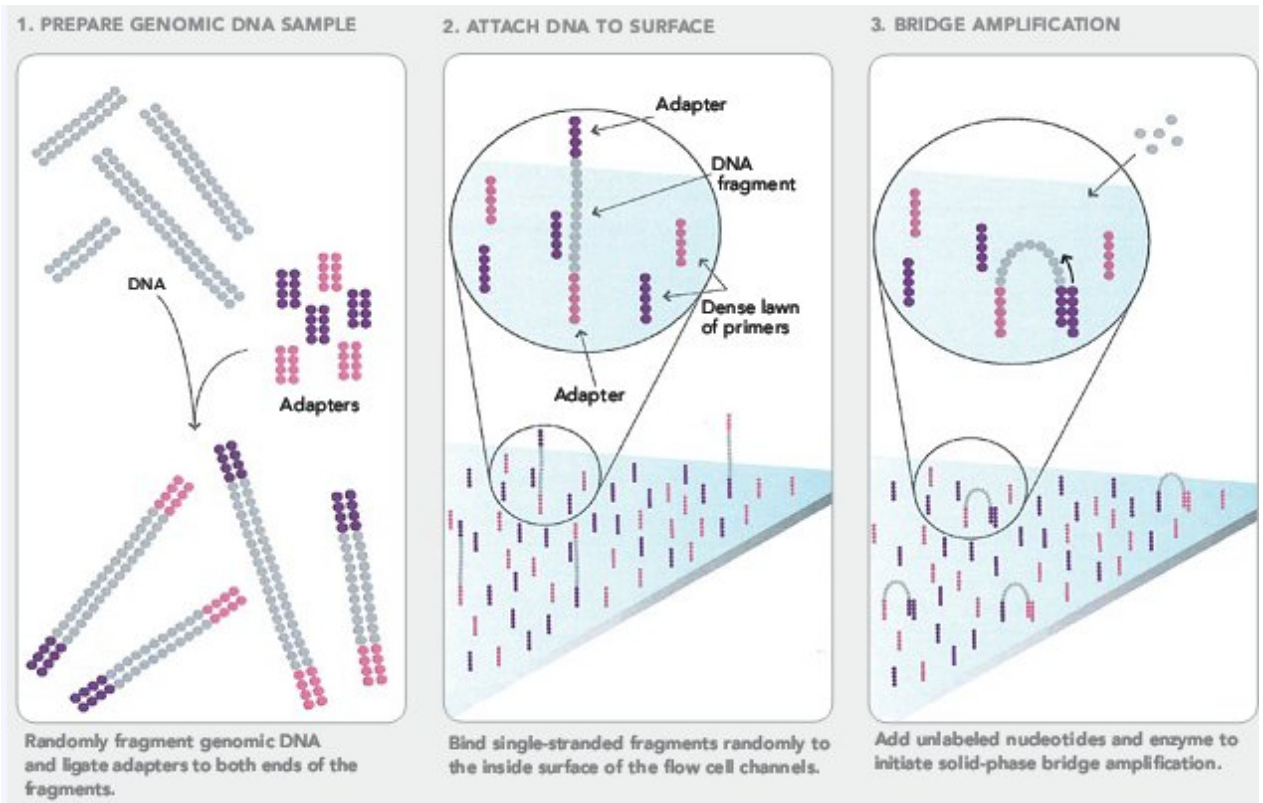


Watch a movie describing the sequencing process can be found here:
`http://www.youtube.com/watch?v=HtuUFUnYB9Y`

Illumina Sequencing

Step 1 - 3

- sample preparation: randomly fragment genomic DNA and ligate adaptors to both ends of the fragments
- attach DNA (probes) to surface
- bridge amplification

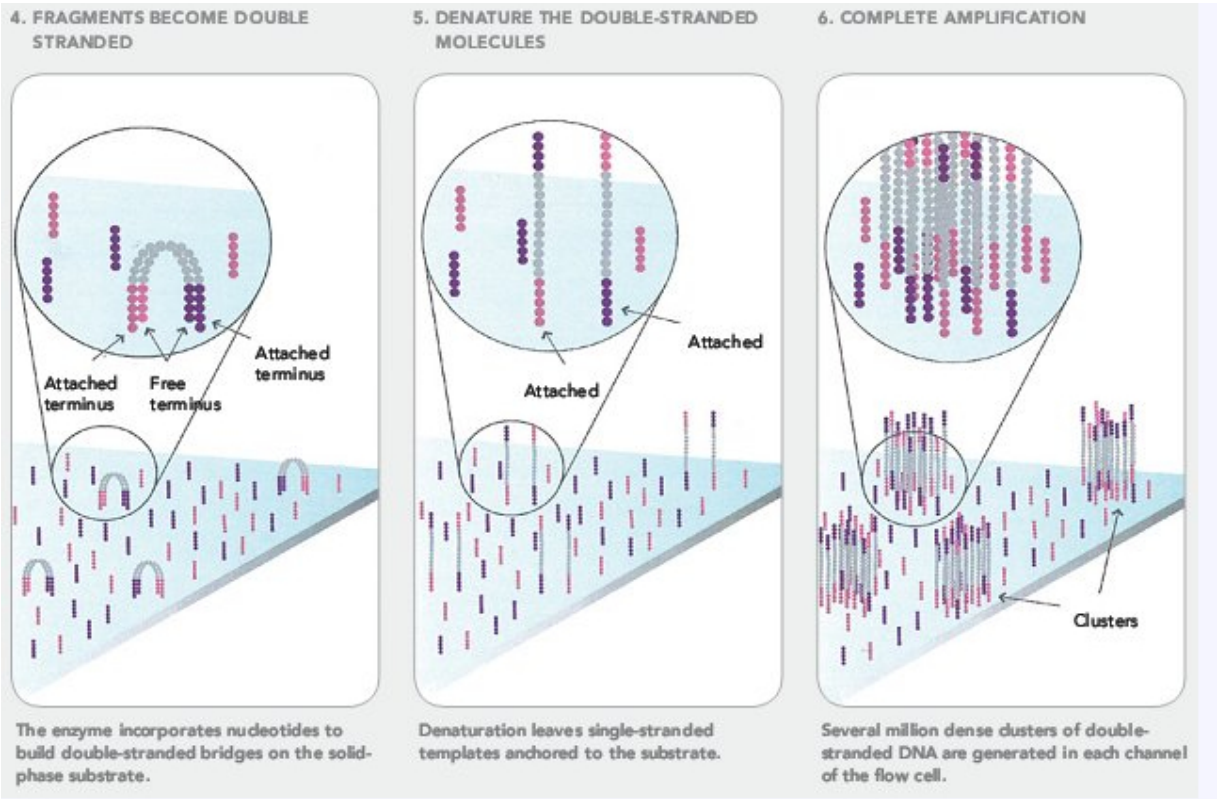


[Illumina]

Illumina sequencing

Step 4 - 6

- synthesize second strands, fragments become double-stranded
- denature single-stranded samples
- after several cycles clusters are ready for sequencing

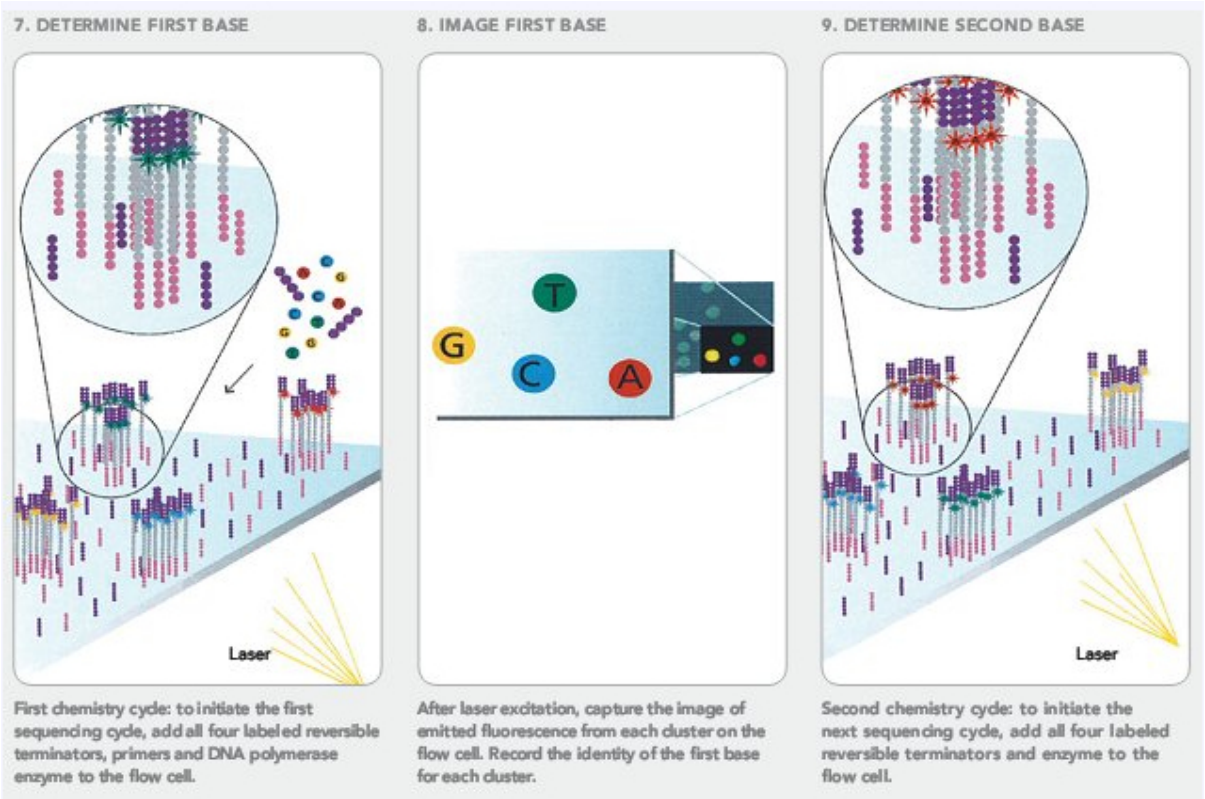


[Illumina]

Illumina sequencing

Step 7 - 9

- determine first base
- image first base
- determine second base

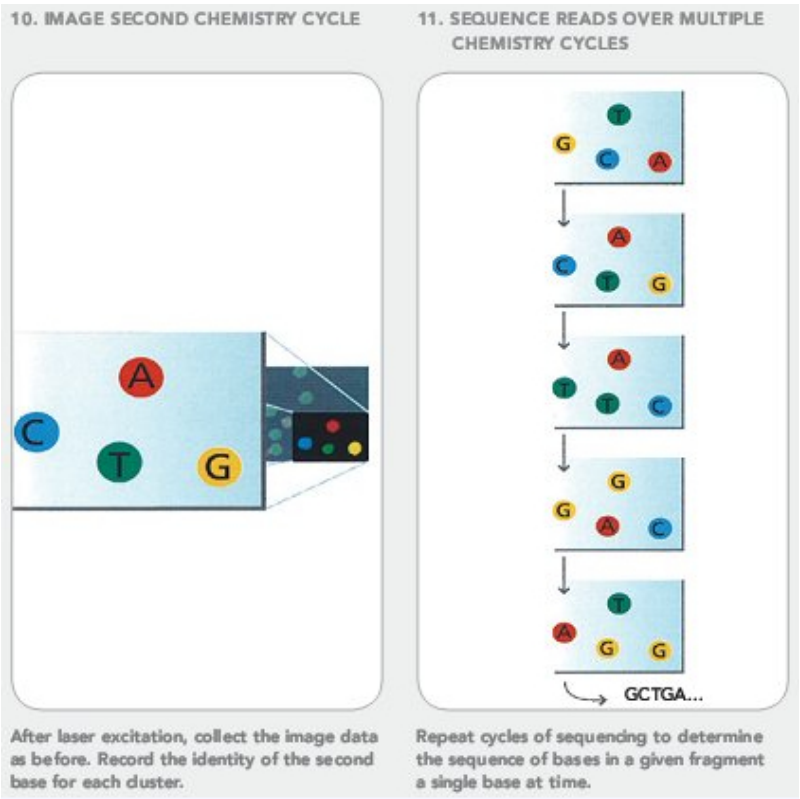


[Illumina]

Illumina sequencing

Step 10 and 11

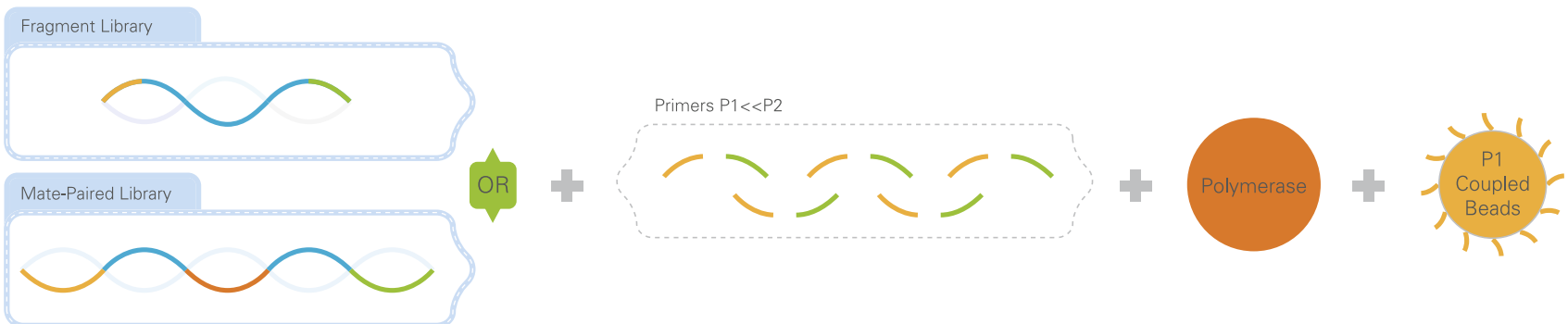
- image second chemistry cycle
- sequence reads over multiple chemistry cycles



[Illumina]

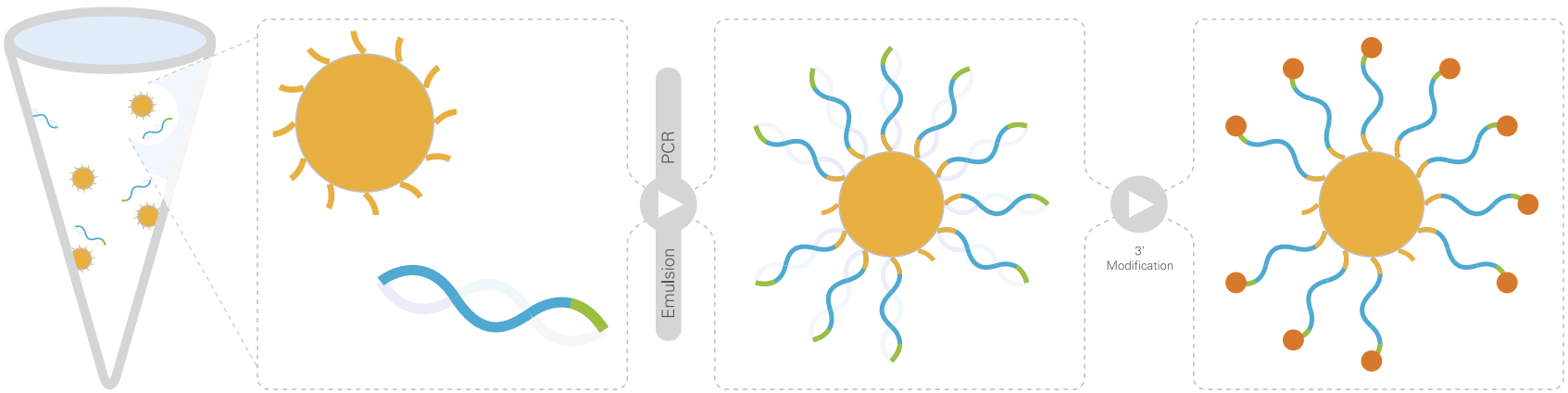
SOLiD Sequencing

- Sequencing by ligation: Fragments ligated to "beads"
- PCR, beads enriched with fragments, ends of the templates modified to allow for an attachment to the slide
- Beads are deposited onto a glass slide
- Di-base probes compete for ligation to the sequencing primer



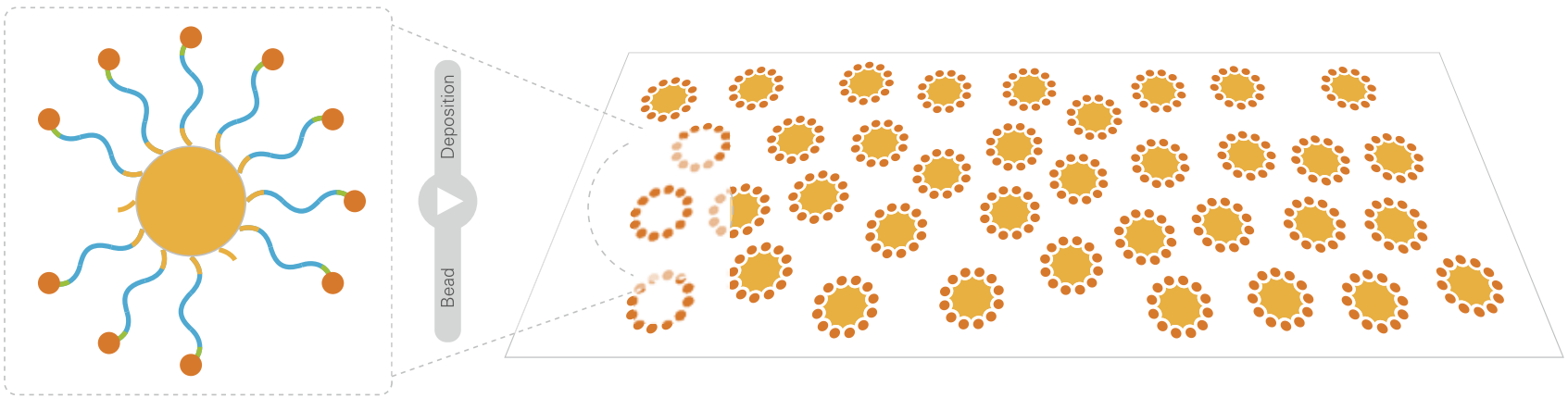
SOLiD Sequencing

- Sequencing by ligation: Fragments ligated to "beads"
- PCR, beads enriched with fragments, ends of the templates modified to allow for an attachment to the slide
- Beads are deposited onto a glass slide
- Di-base probes compete for ligation to the sequencing primer



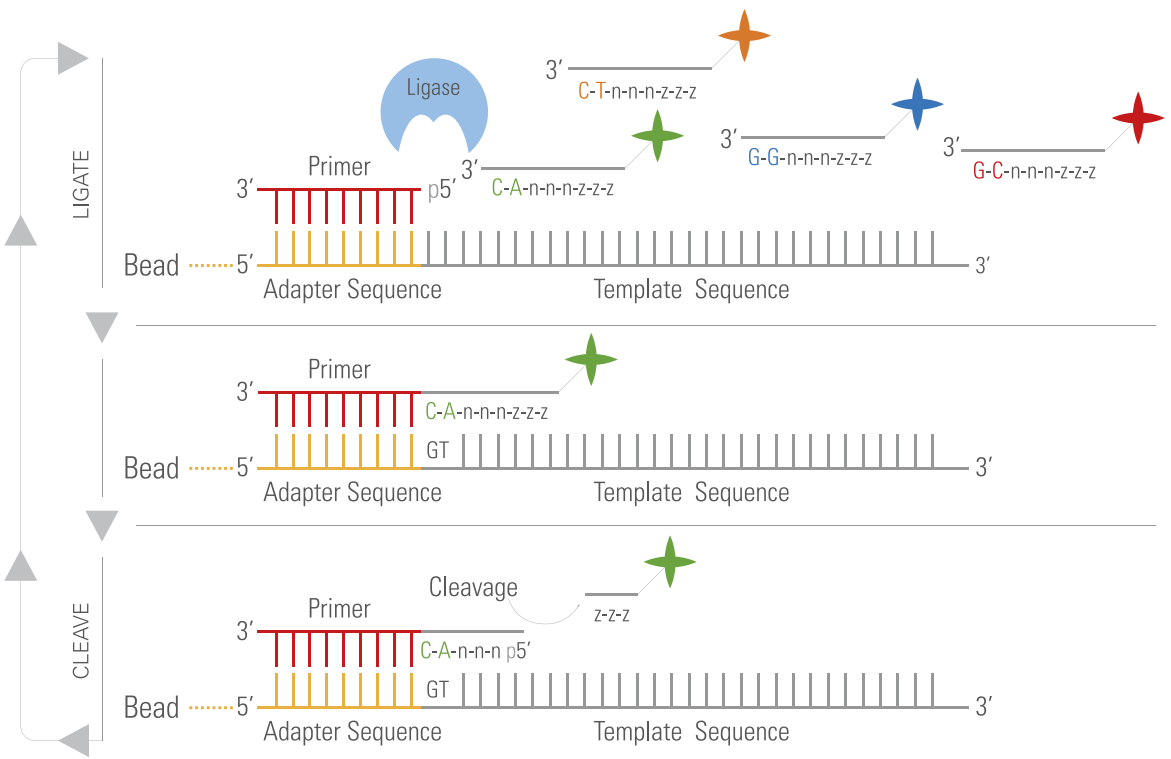
SOLiD Sequencing

- Sequencing by ligation: Fragments ligated to "beads"
- PCR, beads enriched with fragments, ends of the templates modified to allow for an attachment to the slide
- Beads are deposited onto a glass slide
- Di-base probes compete for ligation to the sequencing primer



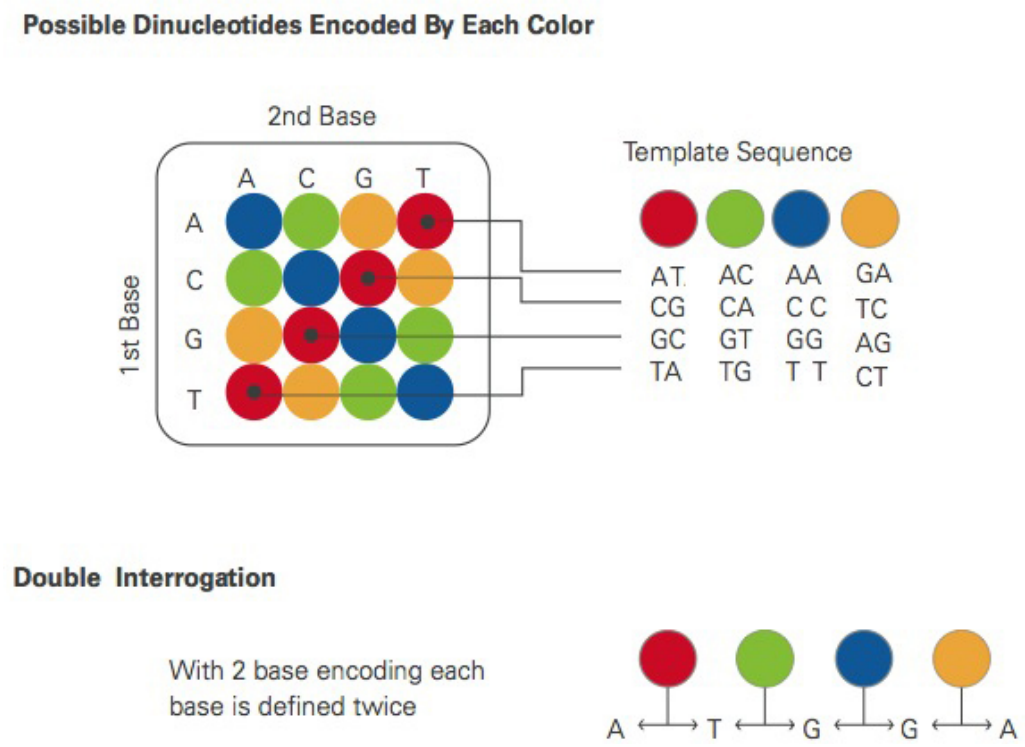
SOLiD Sequencing

- Sequencing by ligation: Fragments ligated to "beads"
- PCR, beads enriched with fragments, ends of the templates modified to allow for an attachment to the slide
- Beads are deposited onto a glass slide
- Di-base probes compete for ligation to the sequencing primer



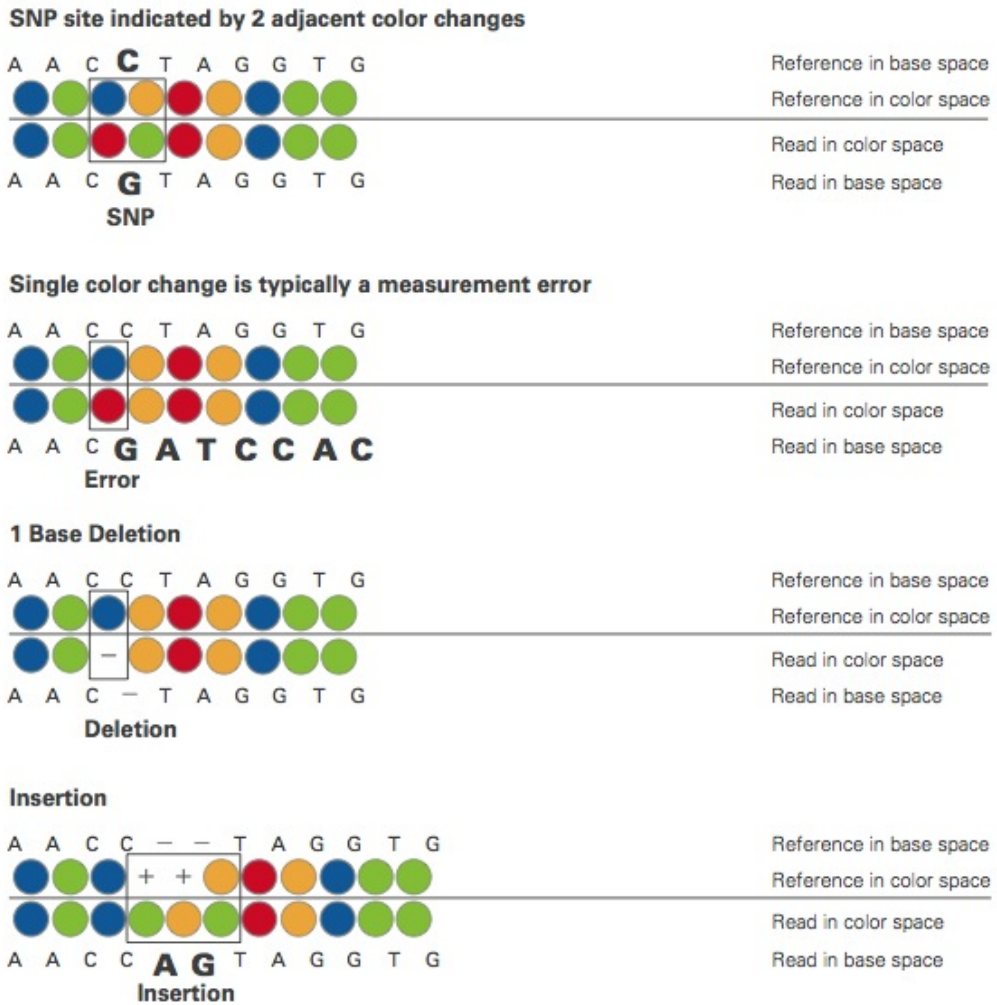
SOLiD Sequencing - Color Space

- 4 fluorescent dyes for 16 possible 2-mers
- Reverse, complement and reverse complement are always of same color



SOLiD Sequencing - Color Space

- 4 fluorescent dyes for 16 possible 2-mers
- Reverse, complement and reverse complement are always of same color



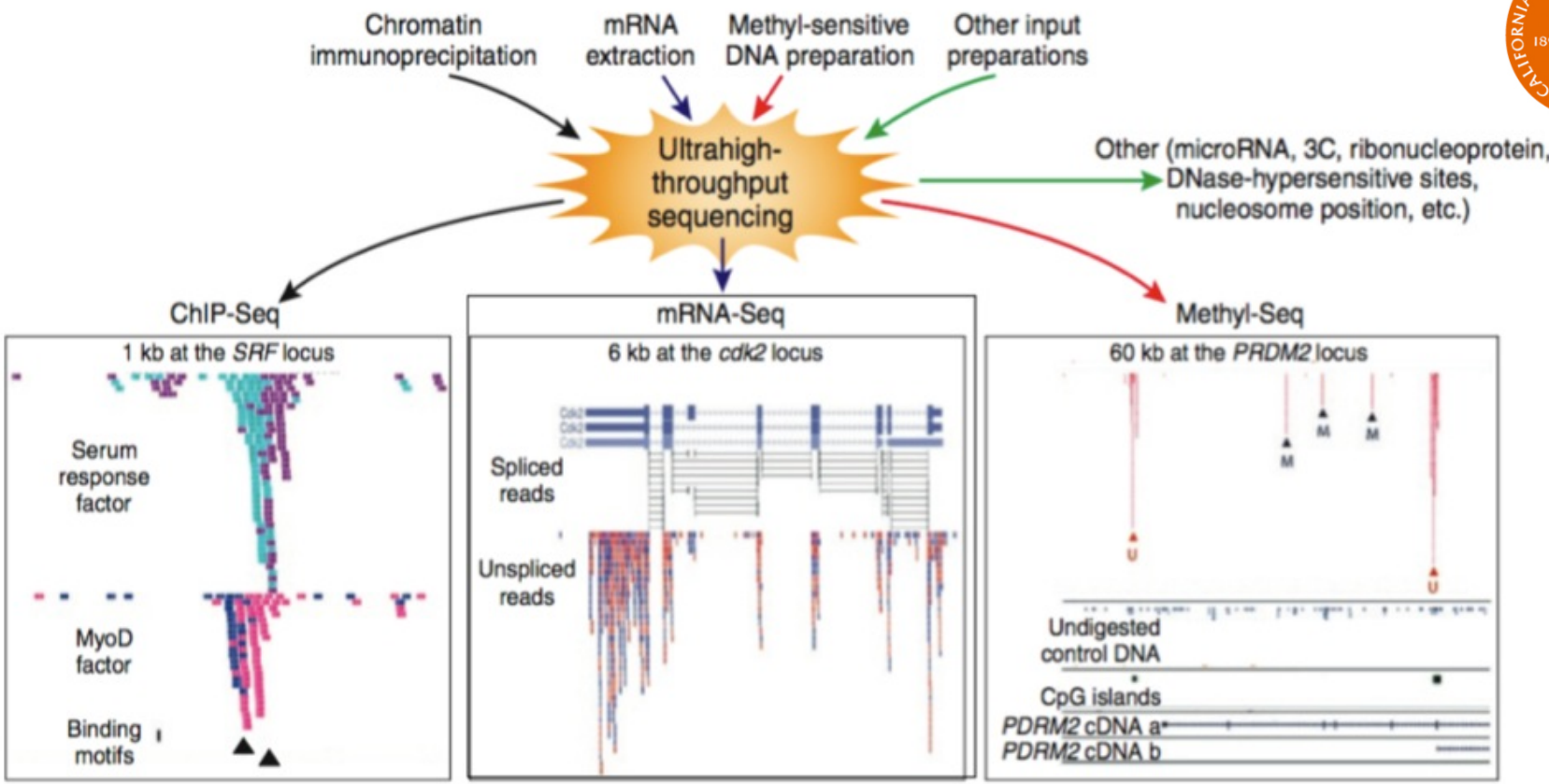
Genomics: a maturing field



Global	General	Specific
	Counting assays	
Genomes	Transcriptome discovery	Transcription factor interactome
Comparative Genomics	Transcriptional regulation	Transcriptome quantitation
		Variant discovery / resequencing

Pure → Applied

Longer reads — Shorter reads



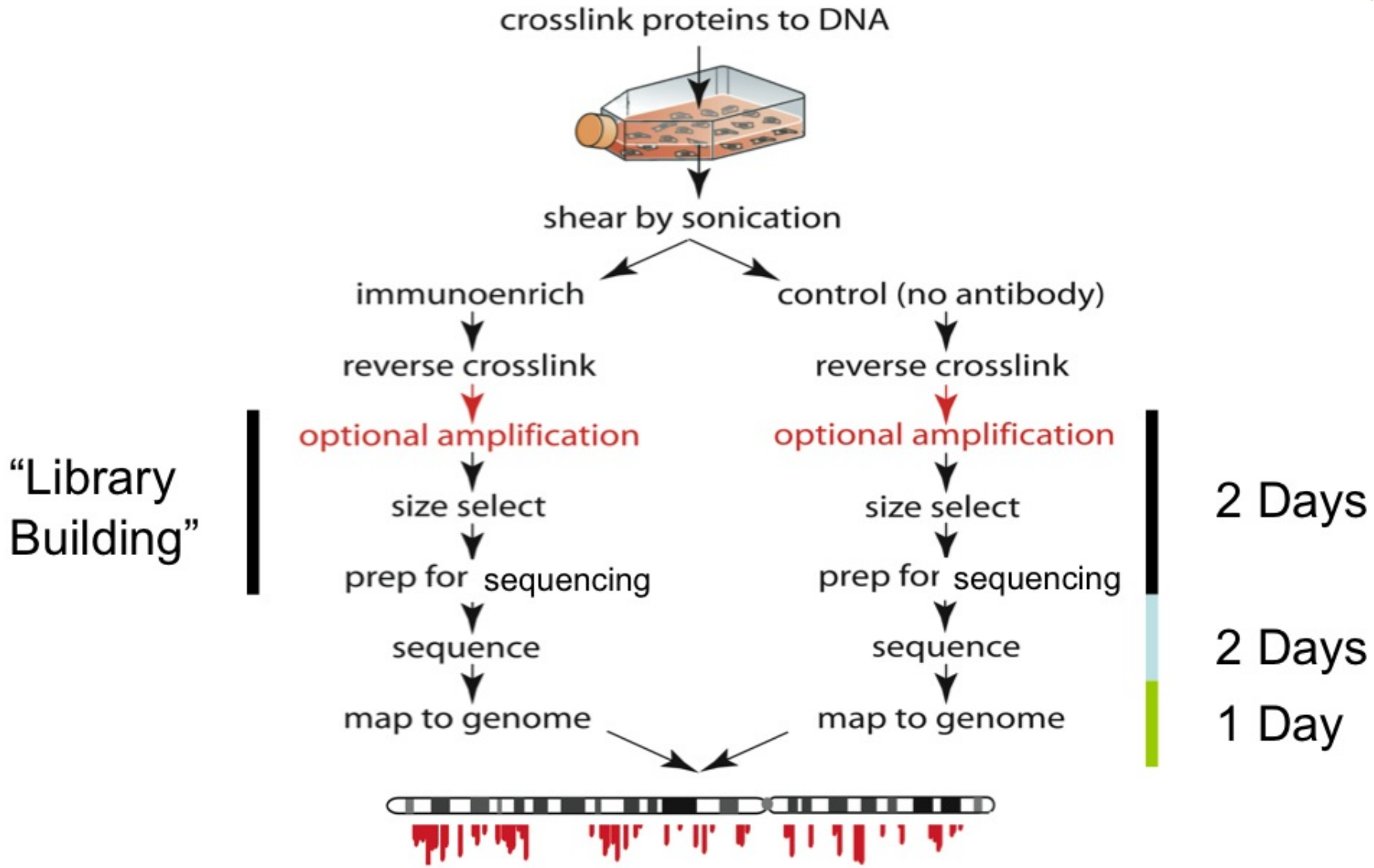
Wold & Myers
Nature Methods, 2007

For all sequence-counting assays, the more reads, the better

About half of the worldwide current generation of sequencing capacity is dedicated to these assays.

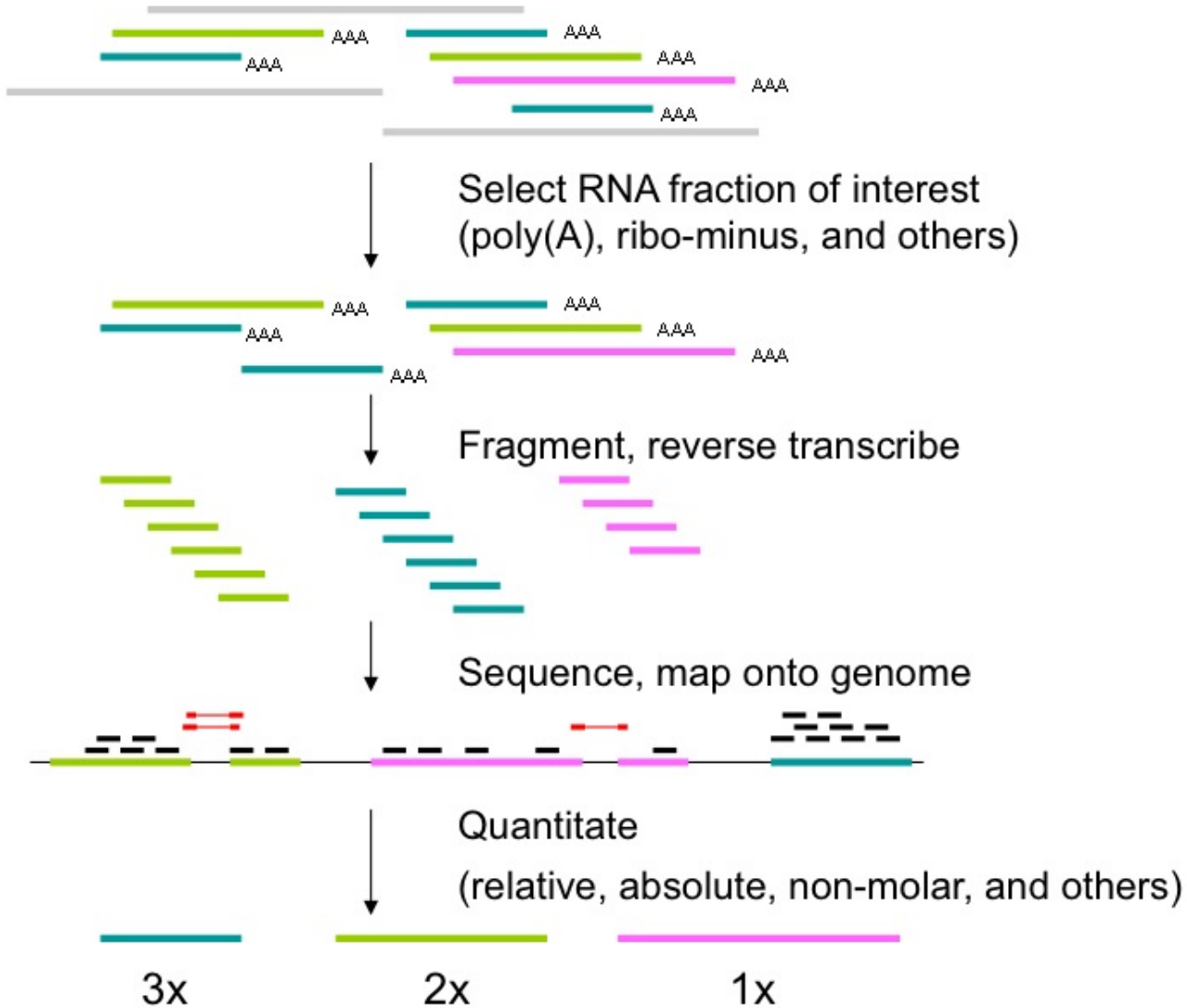
ChIP-seq

A digital counting method to score site occupancy by DNA binding proteins



RNA-seq

A digital counting method for transcriptome discovery and quantification



RNA-seq

Splice mapping strategies



de novo assembly of the transcriptome

highly expressed gene



lowly expressed gene



Read coverage must be high enough to build EST contigs (green solid bar)

Map onto the genome

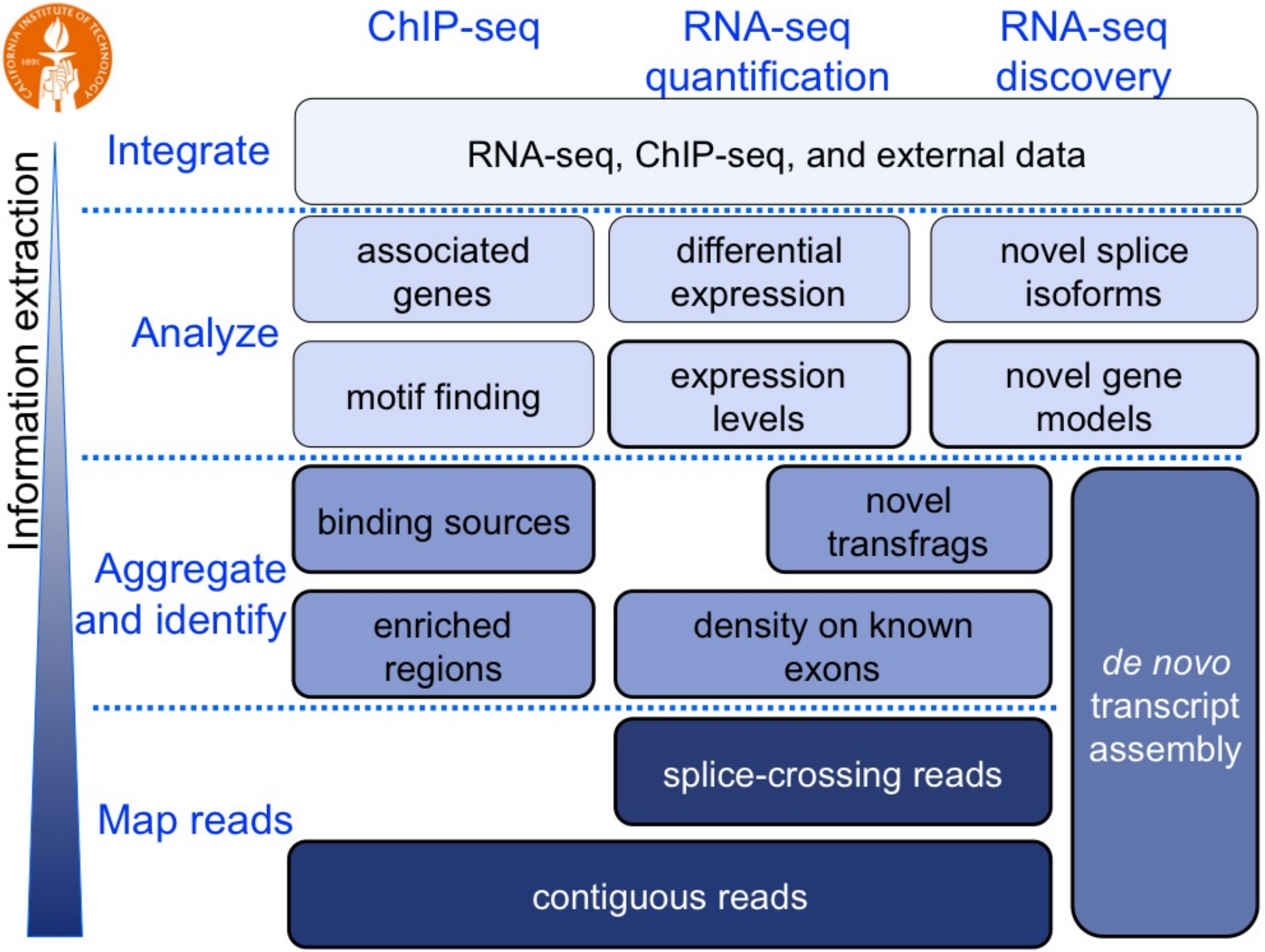


Read mapper must support splitting reads to record splices

Map onto the genome and splice junctions



Splice junctions sequences from either annotations or inferred


ChIP-seq
RNA-seq quantification
RNA-seq discovery
Integrate
RNA-seq, ChIP-seq, and external data
Analyze
associated genes
differential expression
novel splice isoforms
motif finding
expression levels
novel gene models
Aggregate and identify
binding sources
novel transfrags
enriched regions
density on known exons
Map reads
splice-crossing reads
contiguous reads
de novo transcript assembly
Information extraction

Sequencing considerations



- Amount of sequence:
  - 6-15 million reads for ChIP-seq of sequence specific factor in mammals
  - 20-40 million reads for RNA-seq quantitation, more for transcript assembly
- Type of sequence:
  - 1x36bp for ChIP-seq, RNA-seq quantitation
  - 2x75bp or more for RNA-seq transcript assembly
- For RNA-seq, longer reads translate into more splice events observed

Sequence mapping considerations



- Quality of sequence:
  - "Filtering by mapping" for quantitation
  - Mapping using quality score for more sensitivity and for SNP-calling
- How to handle and weight multireads, i.e. reads that map to more than one place in the genome
  - Ignore them
  - Take reads that map to a small number of places
  - Take all
- How many mismatches to allow ? Will this be used for SNP calling ?
  - A finite number (e.g. 5 / 100)
  - A finite amount of quality budget (e.g. up to 200 quality values, distributed across as many positions)
  - As many mismatches as still uniquely mappable.

Roadmap

Introduction
- Quantitation in biology
- The Rise of the Sequencers
- Sequence-counting Assays
- Important Considerations

Transcriptome Analysis
- Motivation
- Common Steps and Challenges
- RNA-seq Read Mapping
- Quantitation & Diff. Expression
- Transcript Reconstruction
- Transcript Assembly

ChIP-seq
- Introduction
- Mapping the reads
- Taking Advantage of Directionality
- Shifting Reads
- Motif Finding

Integration
- The Combinatorial Problem
- Analyzing Chip-seq Datasets Jointly
- Segmenting the Genome
- Learning Regulatory States
- Joint Analysis of Chip- and RNA-Seq

Transcription & RNA Processing in Eucaryotes

- Newly synthesized pre-mRNA is capped.

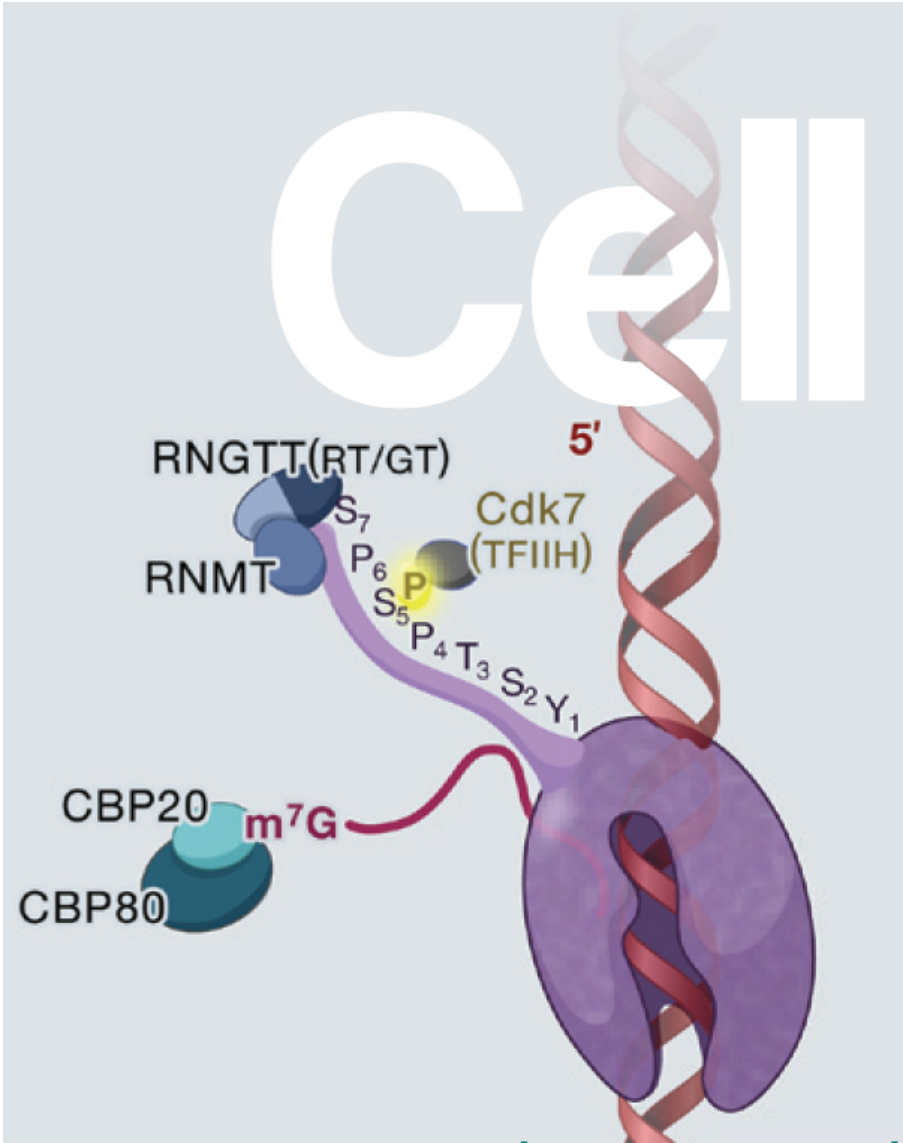
  [CBP20 & CBP80: cap-binding proteins]

- Introns are spliced from pre-mRNA.

  [U1, U2, U4-6: spliceosome

  SF1, U2AF, SR proteins: splicing factors]

- A polyA-tail is added to the 3' terminus of pre-mRNA.



[Bergkessel et al., 2009]

Transcription & RNA Processing in Eucaryotes

- Newly synthesized pre-mRNA is capped.

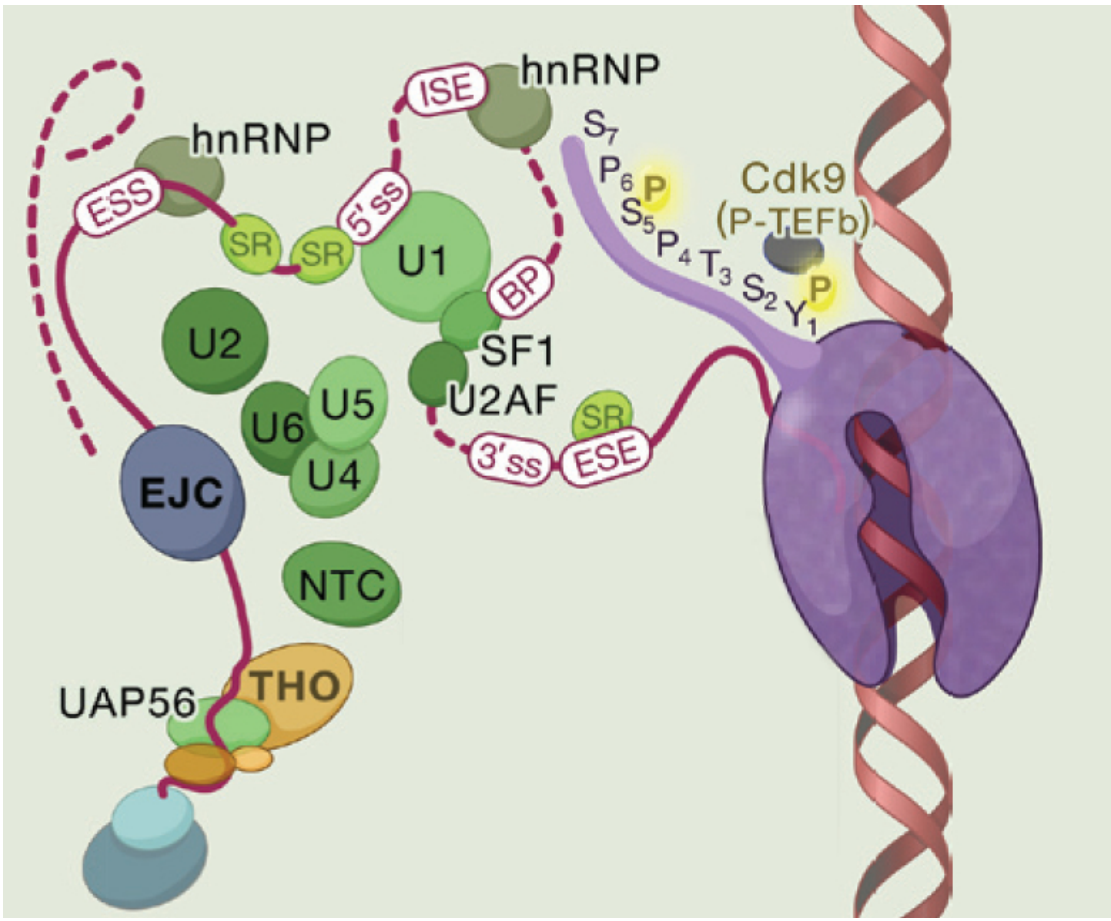
  [CBP20 & CBP80: cap-binding proteins]

- Introns are spliced from pre-mRNA.

  [U1, U2, U4-6: spliceosome

  SF1, U2AF, SR proteins: splicing factors]

- A polyA-tail is added to the 3' terminus of pre-mRNA.



[Bergkessel et al., 2009]

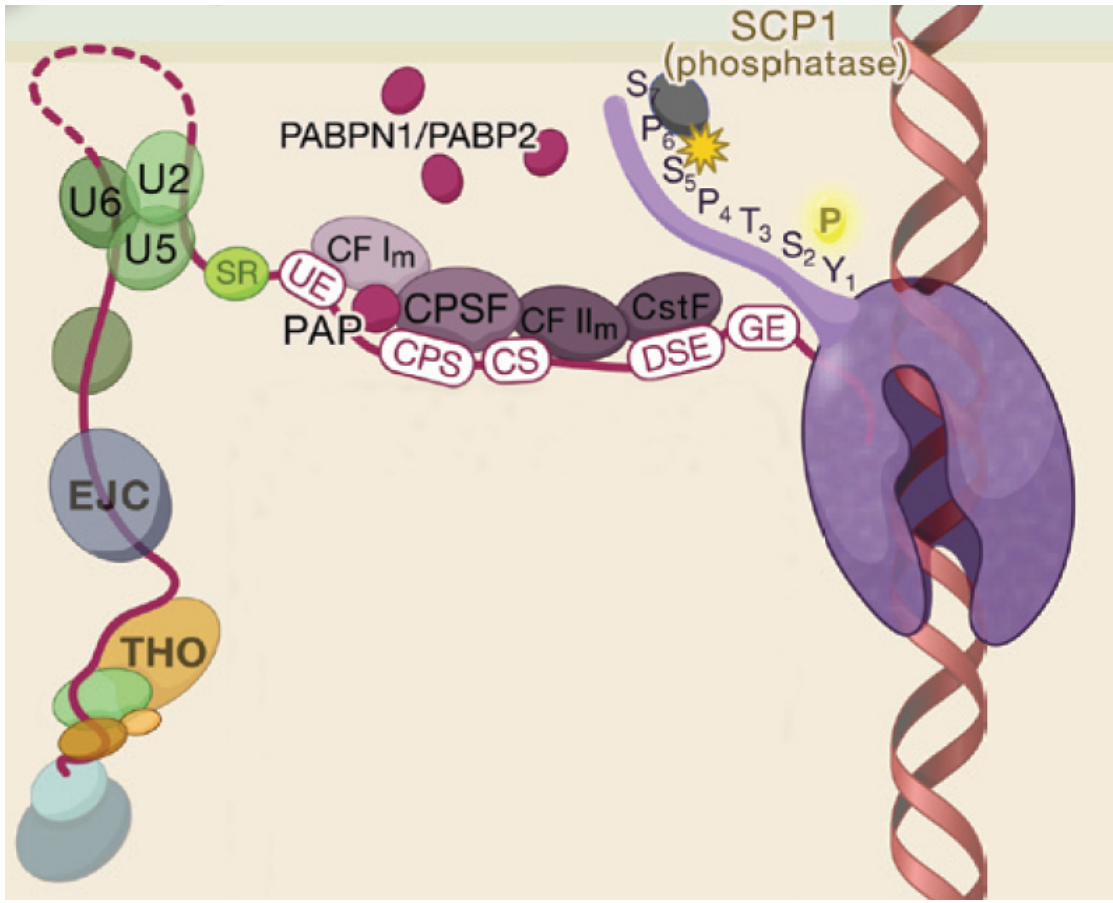
Transcription & RNA Processing in Eucaryotes

- Newly synthesized pre-mRNA is capped.

  [CBP20 & CBP80: cap-binding proteins]
- Introns are spliced from pre-mRNA.

  [U1, U2, U4-6: spliceosome

  SF1, U2AF, SR proteins: splicing factors]
- A polyA-tail is added to the 3' terminus of pre-mRNA.



[Bergkessel et al., 2009]

Key Research Questions



- Characterize an organism's full complement of genes
  - ⇒ Find new (possibly noncoding) genes
  - ⇒ Compare genes among organisms
- Characterize transcript isoforms
  - ⇒ Find new alternative splice forms / transcript ends
- Monitor transcriptome changes between tissues or in response to environmental changes (e.g. stress)
  - ⇒ Identify significant expression changes
- Understand transcriptome regulation
  - ⇒ Knock-out / knock-down analysis of regulators
    - ⇝ Identify regulated targets with significant expression changes
  - ⇒ Identify binding sites used in regulation (e.g. ChIP-seq)
  - ⇒ QTL-Mapping of transcriptome-phenotypes

RNA Transcripts and Library Preparation



There are many different kinds of RNAs:

- Protein-coding mRNAs
- Noncoding RNAs
  - Structural RNAs (e.g. rRNAs, tRNAs, ...)
  - Small RNAs (e.g. miRNAs, endogenous siRNAs, ...)
  - Antisense / promoter-associated transcripts
  - ...

Analysis of biological sample starts with sample/library preparation.

Depending on which RNAs should be targeted, different preparation strategies have to be used.

Sample/Library Preparation Choices



Directly sequencing total-RNA is suboptimal in most cases:

- rRNA, tRNAs constitute the largest fraction of RNA ($> 90\%$)

Sample preparation choices:

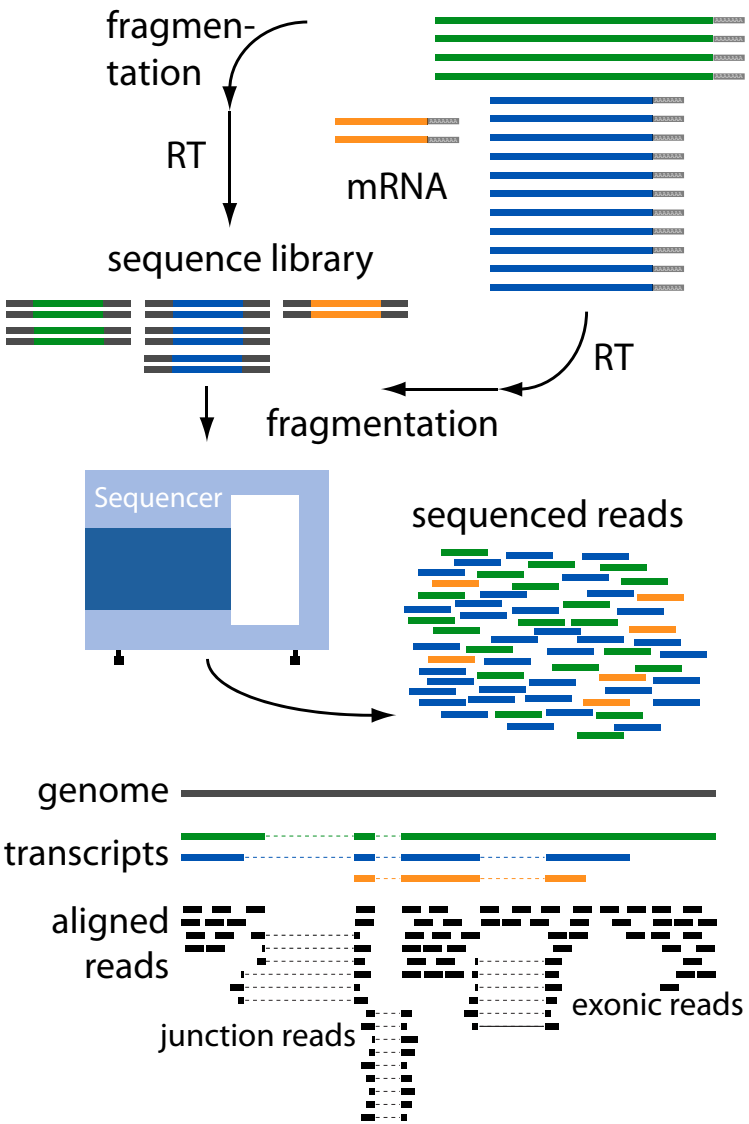
- ribo-minus (rRNA depletion, if it works)
- oligo-dT (selection of poly-adenylated transcripts),
- exonuclease treatment (degrade 5'-P RNAs)

Library preparation options (depend on sequencing technology)

- Strand information
- paired end, mate-pair sequencing

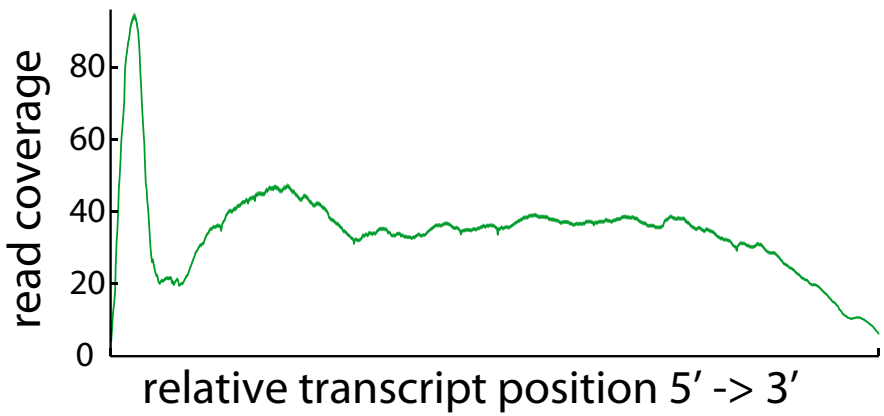
Most of these steps distort RNA transcript concentrations.

RNA-Seq Biases



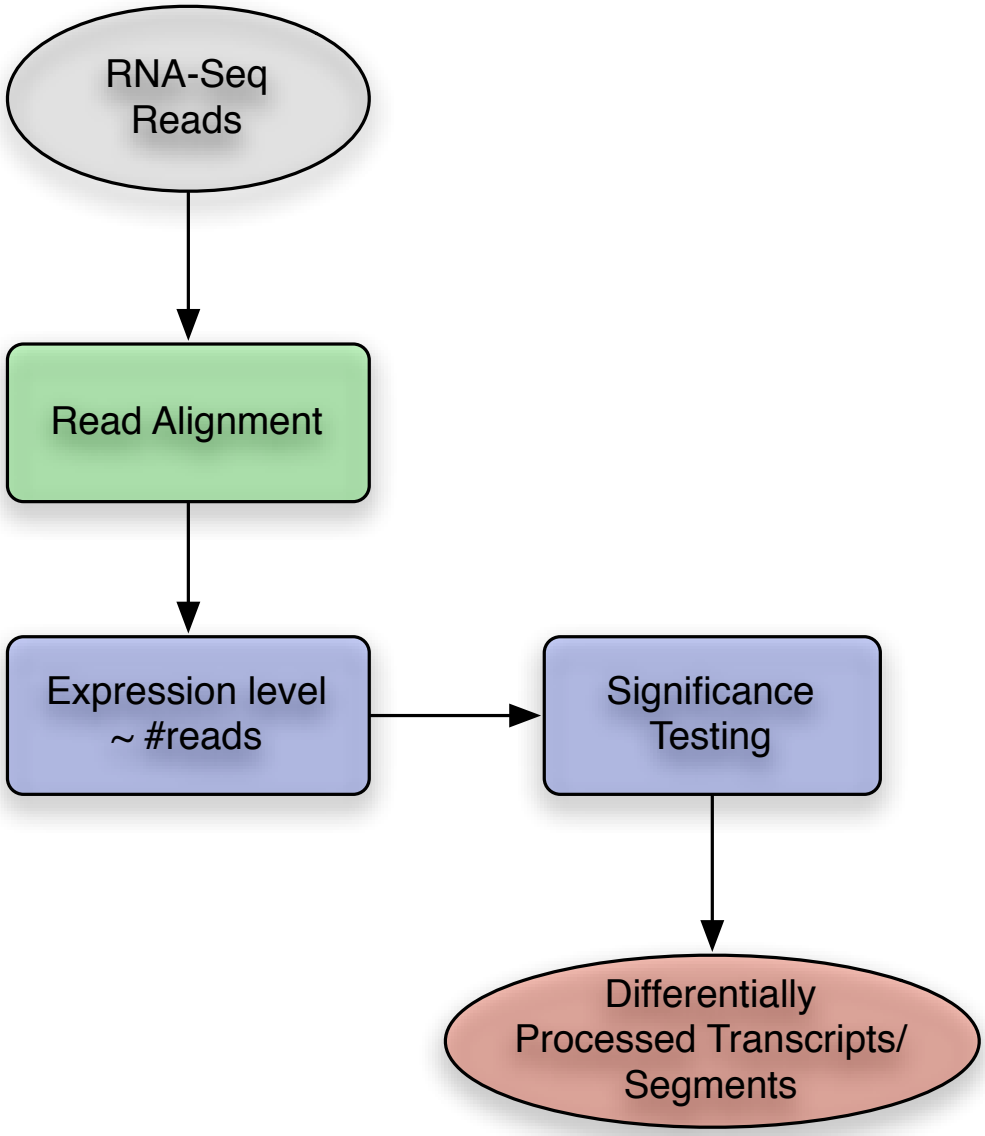
Biases due to ...

- cDNA library construction
- Sequencing
- Read mapping

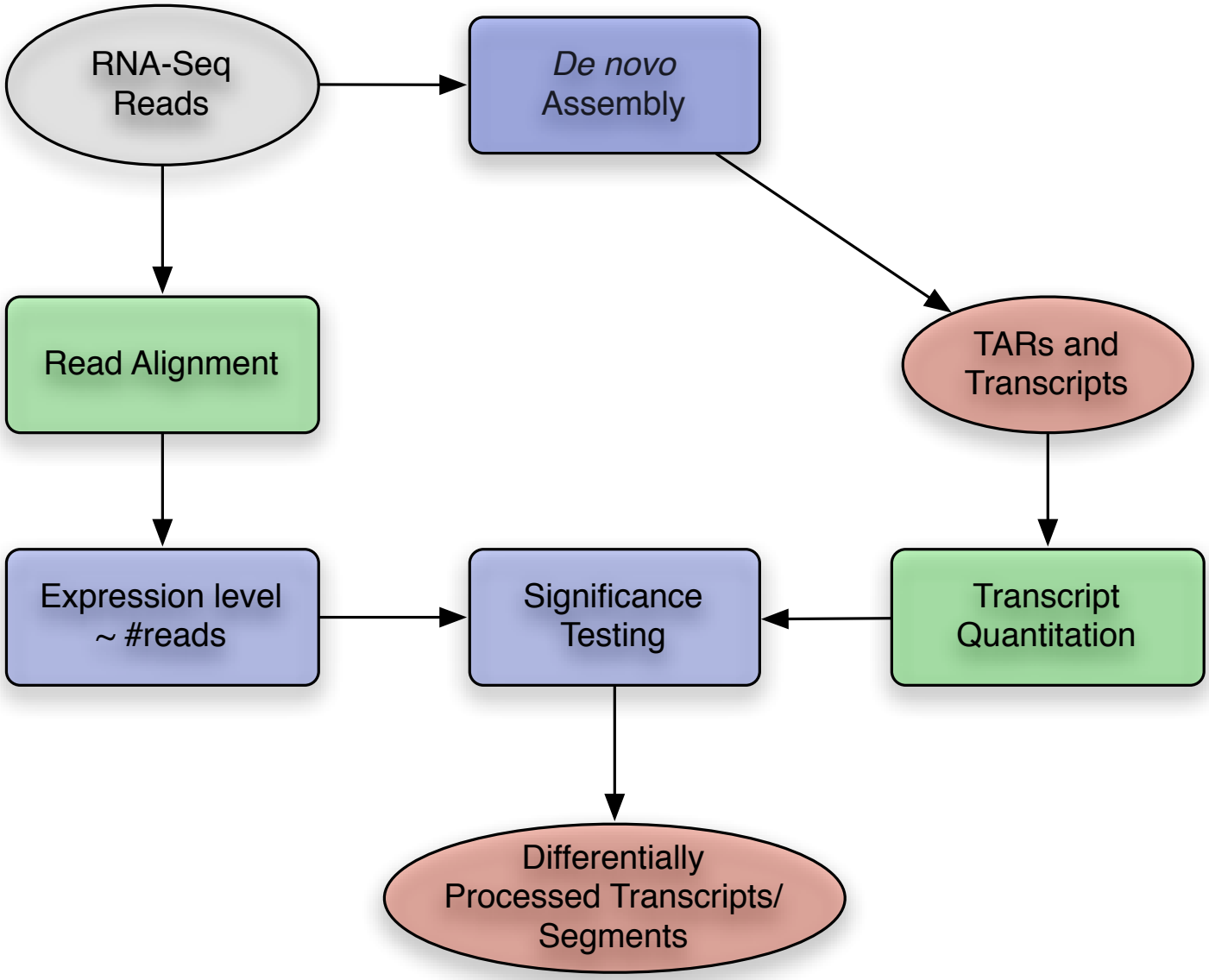


(average over annotated transcripts of length ≈1kb for the *C. elegans* SRX001872 dataset)

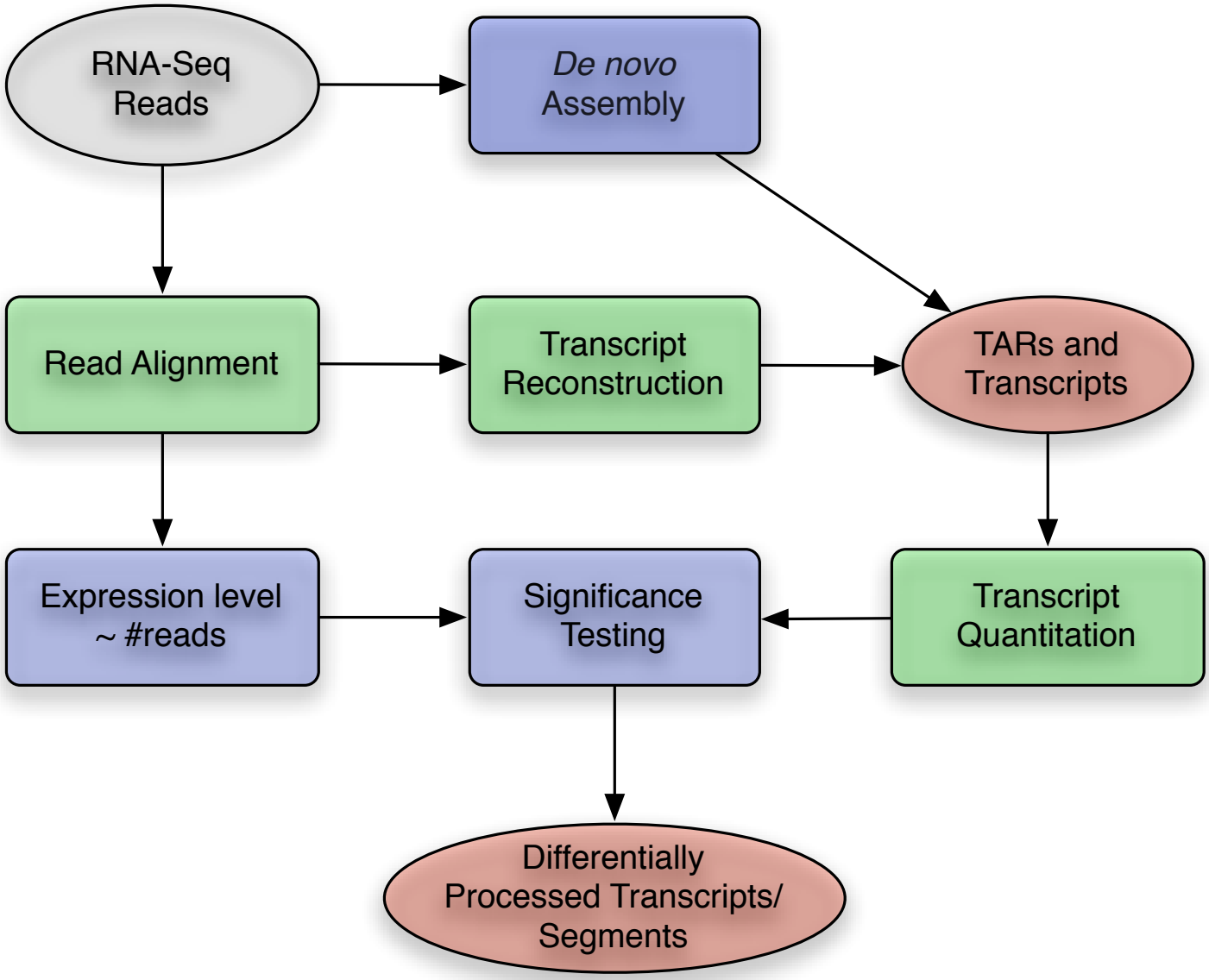
Common RNA-Seq Analysis Steps



Common RNA-Seq Analysis Steps



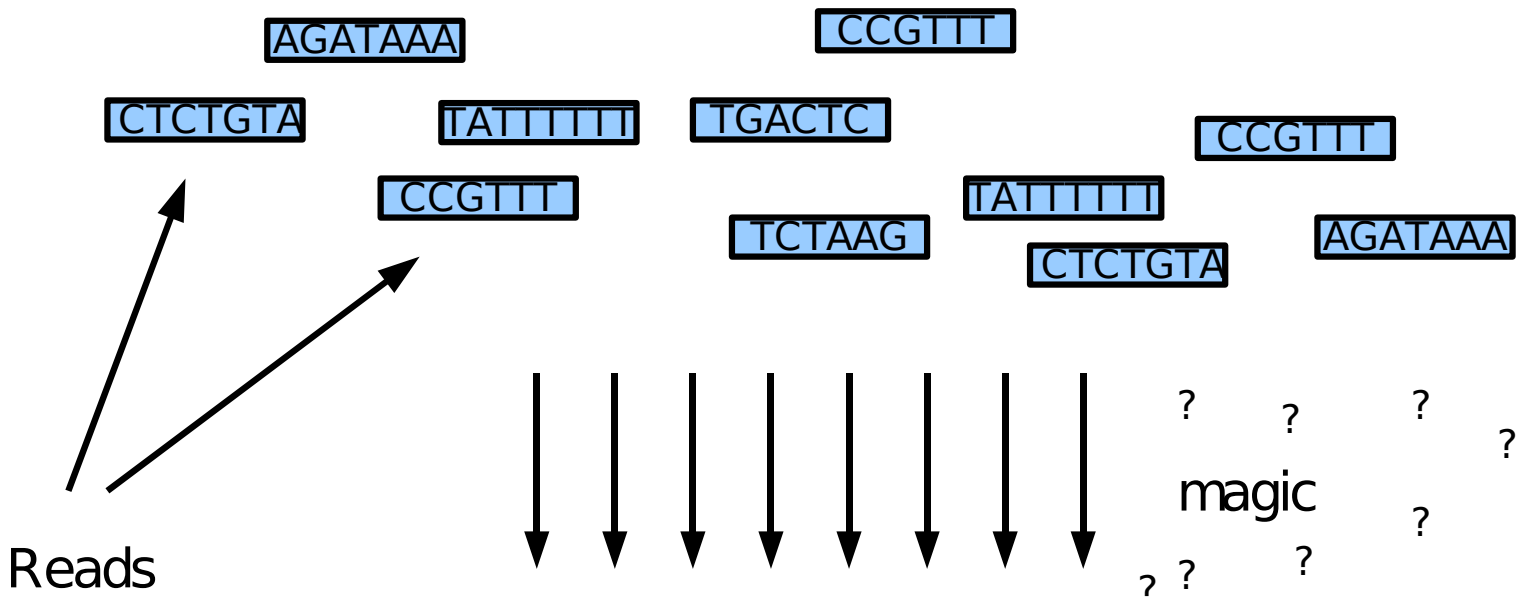
Common RNA-Seq Analysis Steps



RNA-seq Read Analysis I

Given read data, the following analysis steps are possible:

- Assembly
  ⇒ generate contigs
- Mapping/Alignments
  ⇒ map/align reads back to a known genome
- Quantification of transcripts
  ⇒ Estimate abundances for alternative splice forms



RNA-seq Read Analysis I

Given read data, the following analysis steps are possible:

- Assembly
  ⇒ generate contigs
- Mapping/Alignments
  ⇒ map/align reads back to a known genome
- Quantification of transcripts
  ⇒ Estimate abundances for alternative splice forms

```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT...
   GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
    CAAACCAGTGACCTGACTACTACGTCGTAACGTACA
     AAACCAGTGACCTGACTACTACGTCGTAACGTACAC
      AACCAGTGACCTGACTACTACGTCGTAACGTACACG
       ACCAGTGACCTGACTACTACGTCGTAACGTACACG
```

RNA-seq Read Analysis I



Given read data, the following analysis steps are possible:

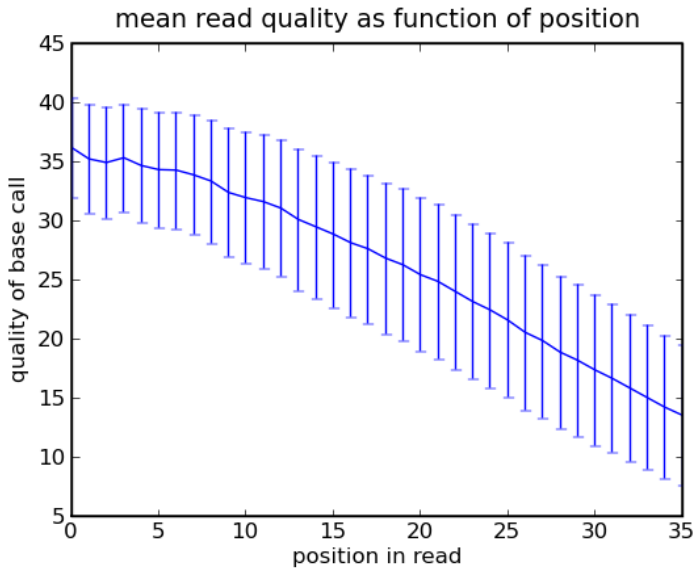
- Assembly
  $\Rightarrow$ generate contigs
- Mapping/Alignments
  $\Rightarrow$ map/align reads back to a known genome
- Quantification of transcripts
  $\Rightarrow$ Estimate abundances for alternative splice forms

Problem: hundreds of millions of reads of short length

$\Rightarrow$ Big computational challenge

RNA-seq Read Analysis II

- Experiment leads to millions of reads $36 - 100$nt
- Reads may have a position-wise varying quality



- Quality corresponds to error probability:

$$q = -10 \cdot log_{10}(\frac{p}{1-p})$$

- Example: If we have an error probability of 10^{-3} per base then the quality is 30

RNA-seq Reads Analysis - Mapping



Reads mapping problem

For each read find its target regions on the reference genome such that are at most k mismatches between read and target.

- Global/local alignment of all reads prohibitive
- A read stems from a certain small region
- Find this region and then do an alignment
  - (Spaced) seeds
  - Suffix trees/arrays
  - Burrows-Wheeler
- Common tools: bowtie [Langmead et al., 2009], bwa [Li and Durbin, 2009, 2010], *GenomeMapper* [Schneeberger et al., 2009a], *Shrimp* [Rumble et al., 2009], *SOAP(2)* [Li et al., 2009], *VMATCH*, MAQ [Li et al., 2008], ELAND, segemehl [Hoffmann et al., 2009], . . . (≈50 more)
- Main issues:
  - Accuracy
  - Speed
  - Memory Consumption

Mapping via Spaced Seeds

- Blast-like searches suffer from two problems:
  - longer seeds lose distant homologies
  - shorter seeds create too many hits
- Idea: Create seeds that have a higher probability of a hit in a homologous region while lower expectation of random hits
  $\Rightarrow$ **Spaced seeds**

```
..GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT...
  GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
                  11111111
```

Mapping via Spaced Seeds

- Blast-like searches suffer from two problems:
  - longer seeds lose distant homologies
  - shorter seeds create too many hits
- Idea: Create seeds that have a higher probability of a hit in a homologous region while lower expectation of random hits
  $\Rightarrow$ **Spaced seeds**

```
..GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT...
  GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
                       11111111
```

Mapping via Spaced Seeds

- Blast-like searches suffer from two problems:
  - longer seeds lose distant homologies
  - shorter seeds create too many hits
- Idea: Create seeds that have a higher probability of a hit in a homologous region while lower expectation of random hits
  $\Rightarrow$ **Spaced seeds**

```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT...
   GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
          1000010011010001000000011
```

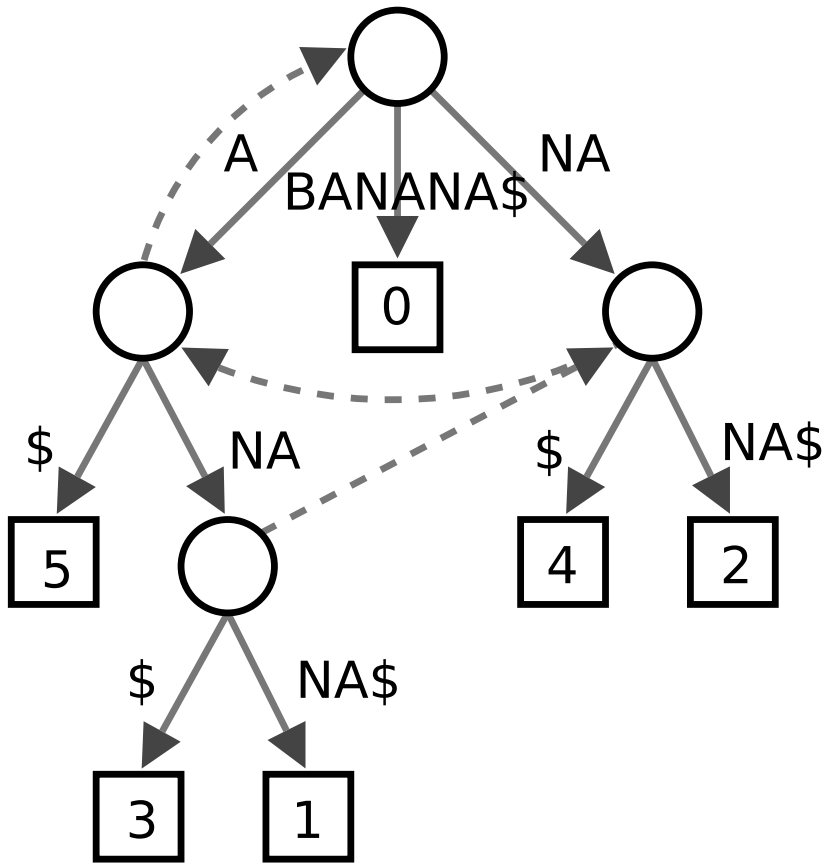
Mapping via Spaced Seeds

- Blast-like searches suffer from two problems:
  - longer seeds lose distant homologies
  - shorter seeds create too many hits
- Idea: Create seeds that have a higher probability of a hit in a homologous region while lower expectation of random hits
  $\Rightarrow$ **Spaced seeds**

```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT...
   GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
          1000010011010001000000011
          0111100000100101010000000
          1100001100000000001111000
```

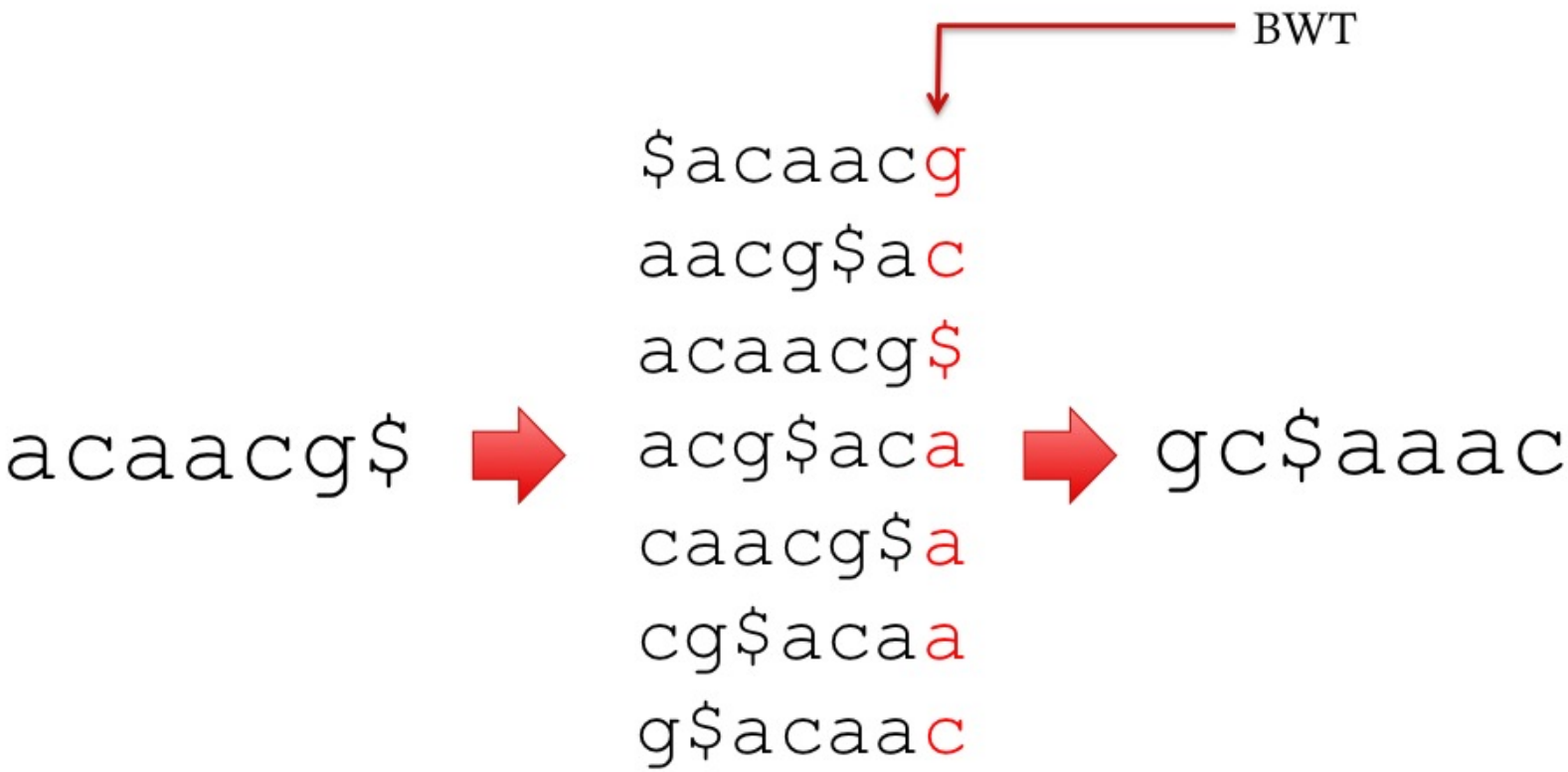
Mapping via Suffix Arrays

- Given a long fixed string of length n and smaller patterns of lengths m to be searched for
- Construction in $\mathcal{O}(n)$, Patterns can be detected in $\mathcal{O}(m)$



[Wikipedia]

Burrows-Wheeler Transform (BWT)



Burrows-Wheeler Matrix (BWM)

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

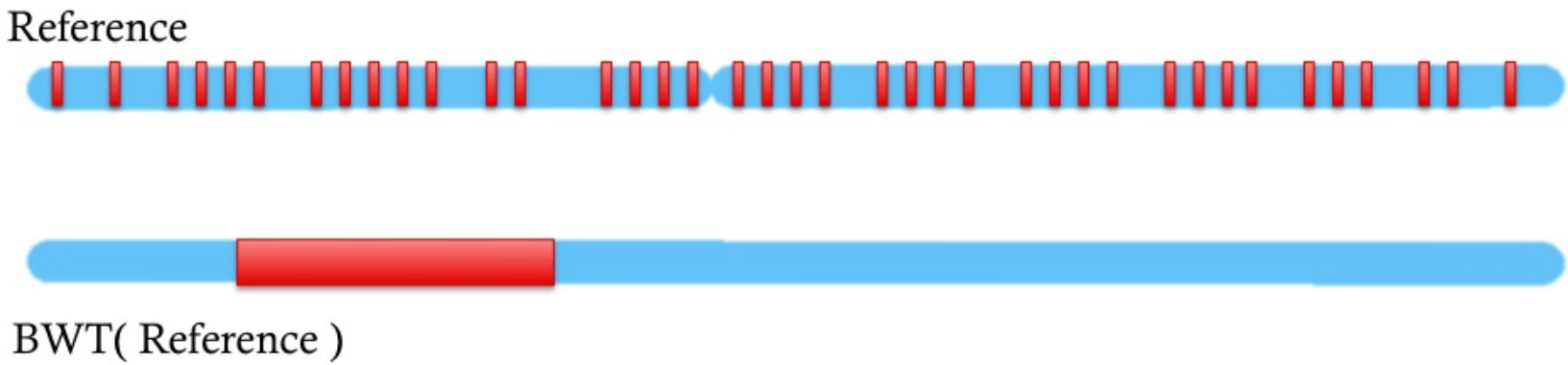
Bowtie algorithm



@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Bowtie algorithm



@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Bowtie algorithm



@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Bowtie index

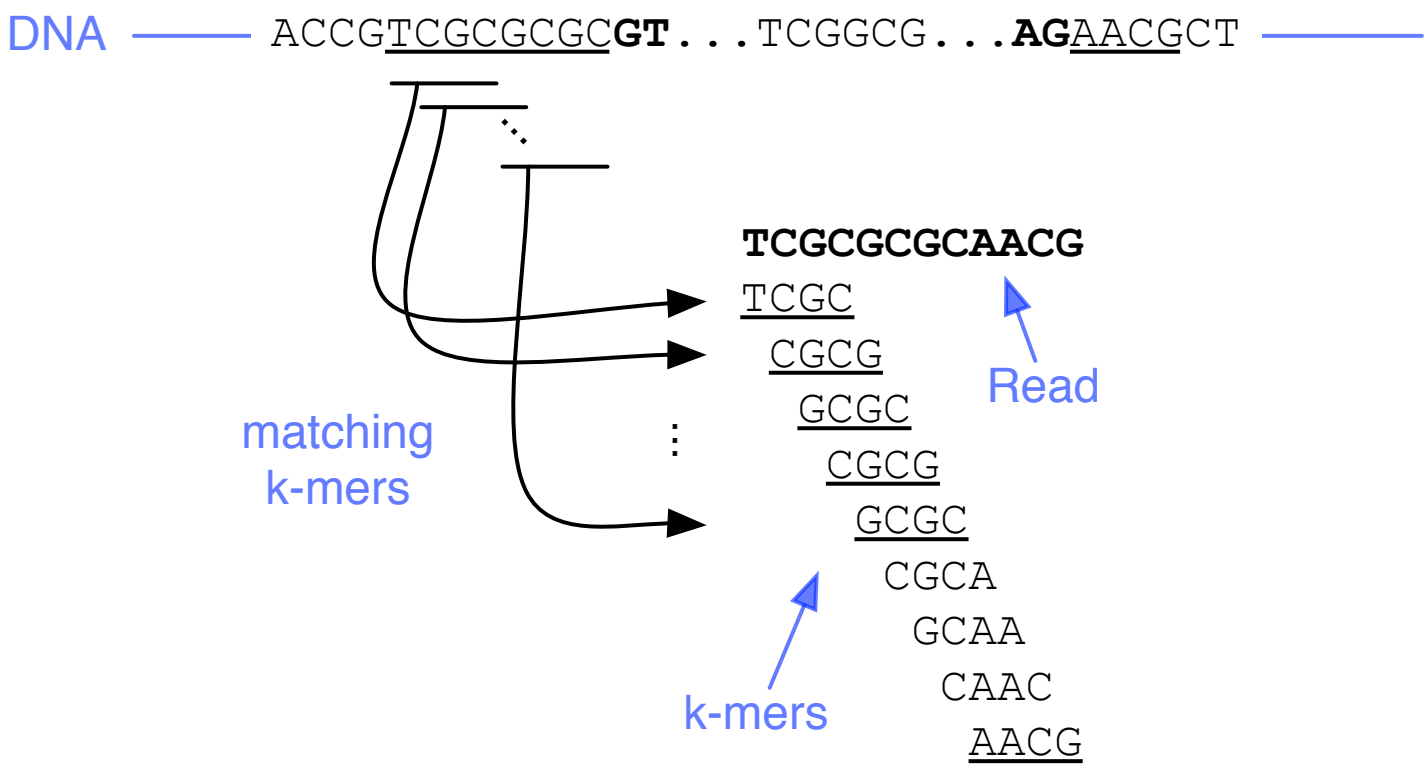
- The index is a sampled, enhanced BWT
- `bowtie-build` implements a difference-cover suffix sorting algorithm from:
  - Kärkkäinen, J. Fast BWT in small space by blockwise suffix sorting. *Theor. Comput. Sci.* **387**, 249-257 (2007).
- Can index the human genome in less than 1 day
- Or download from http://tinyurl.com/bowtiesrc

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

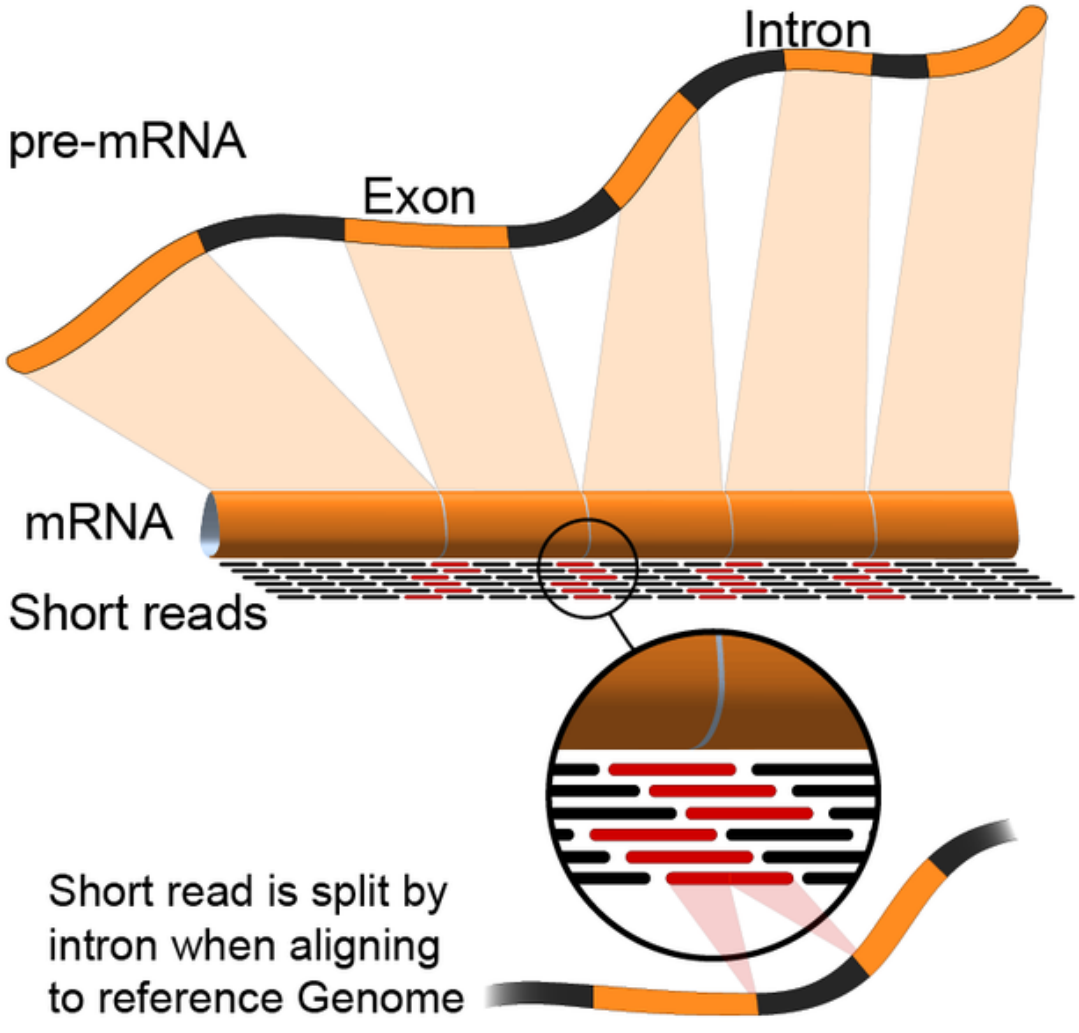
Read Alignment with GenomeMapper

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009b]
- *k*-mer based index, well suited for smaller genomes

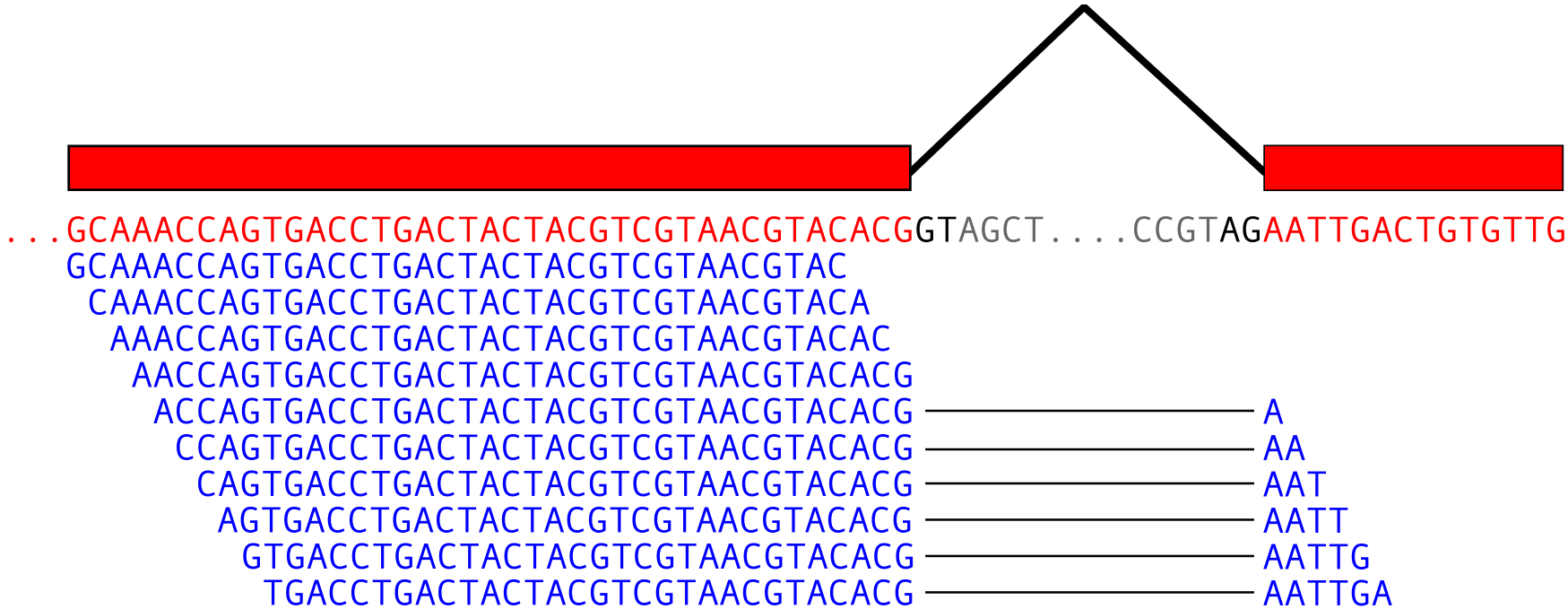


Spliced vs. Unspliced Alignments



[Wikipedia]

Spliced vs. Unspliced Alignments



- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads

Tools for Spliced Read Alignments



Traditional ones developed for cDNA sequence alignment:

- *blast* [Altschul et al., 1990], *spliced alignments* [Gelfand et al., 1996], *sim4* [Florea et al., 1998], *GeneSeqer* [Usuka et al., 2000], *Spidey* [Wheelan SJ, 2001], *blat* [Kent, 2002], *exalin* [Zhang and Gish, 2006], *Palma* [Schulze et al., 2007]

⇒**Too slow for RNA-seq read alignment**Variety of new tools specific for spliced NGS read alignment:

- Erange [Mortazavi et al., 2008], GEM [Ribeca], MapNext [Bao et al., 2009], MapSplice [Prins], PALMapper [Rätsch et al., 2010] (=GenomeMapper/QPALMA [Schneeberger et al., 2009b, De Bona et al., 2008]), PASS [Campagna et al., 2009], Star (Dobin), TopHat [Trapnell et al., 2009], ...
- Issues:
  - Assumptions on splice consensus
  - Accuracy of intron predictions
  - Speed (often higher than for unspliced alignments)
  - Memory consumption (similar to unspliced mappers)

Tools for Spliced Read Alignments



Traditional ones developed for cDNA sequence alignment:

- *blast* [Altschul et al., 1990], *spliced alignments* [Gelfand et al., 1996], *sim4* [Florea et al., 1998], *GeneSeqer* [Usuka et al., 2000], *Spidey* [Wheelan SJ, 2001], *blat* [Kent, 2002], *exalin* [Zhang and Gish, 2006], *Palma* [Schulze et al., 2007]

$\Rightarrow$**Too slow for RNA-seq read alignment**Variety of new tools

specific for spliced NGS read alignment:

- Erange [Mortazavi et al., 2008], GEM [Ribeca], MapNext [Bao et al., 2009], MapSplice [Prins], PALMapper [Rätsch et al., 2010] (=GenomeMapper/QPALMA [Schneeberger et al., 2009b, De Bona et al., 2008]), PASS [Campagna et al., 2009], Star (Dobin), TopHat [Trapnell et al., 2009], ...
- Issues:
  - Assumptions on splice consensus
  - Accuracy of intron predictions
  - Speed (often higher than for unspliced alignments)
  - Memory consumption (similar to unspliced mappers)

Tools for Spliced Read Alignments



Traditional ones developed for cDNA sequence alignment:

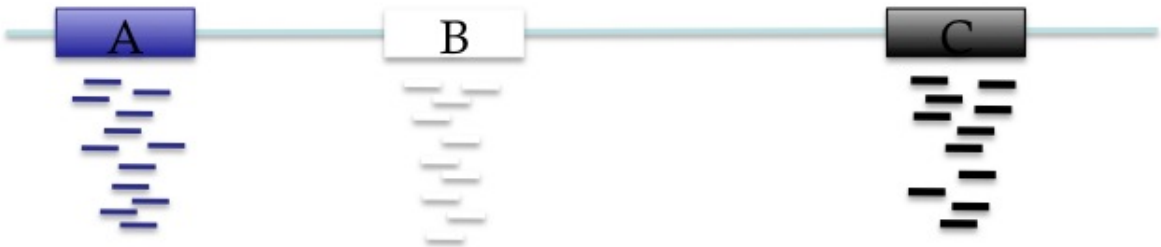
- *blast* [Altschul et al., 1990], *spliced alignments* [Gelfand et al., 1996], *sim4* [Florea et al., 1998], *GeneSeqer* [Usuka et al., 2000], *Spidey* [Wheelan SJ, 2001], *blat* [Kent, 2002], *exalin* [Zhang and Gish, 2006], *Palma* [Schulze et al., 2007]

$\Rightarrow$**Too slow for RNA-seq read alignment**Variety of new tools

specific for spliced NGS read alignment:

- Erange [Mortazavi et al., 2008], GEM [Ribeca], MapNext [Bao et al., 2009], MapSplice [Prins], PALMapper [Rätsch et al., 2010] (=GenomeMapper/QPALMA [Schneeberger et al., 2009b, De Bona et al., 2008]), PASS [Campagna et al., 2009], Star (Dobin), TopHat [Trapnell et al., 2009], ...
- Issues:
  - Assumptions on splice consensus
  - Accuracy of intron predictions
  - Speed (often higher than for unspliced alignments)
  - Memory consumption (similar to unspliced mappers)

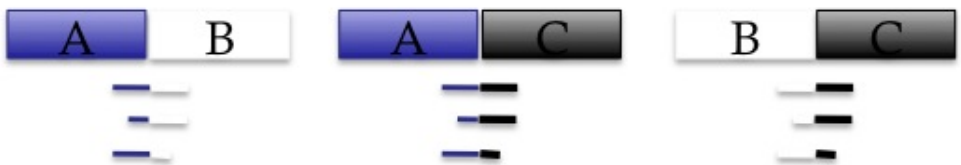
Aligning reads with TopHat



identify candidate exons via genomic mapping

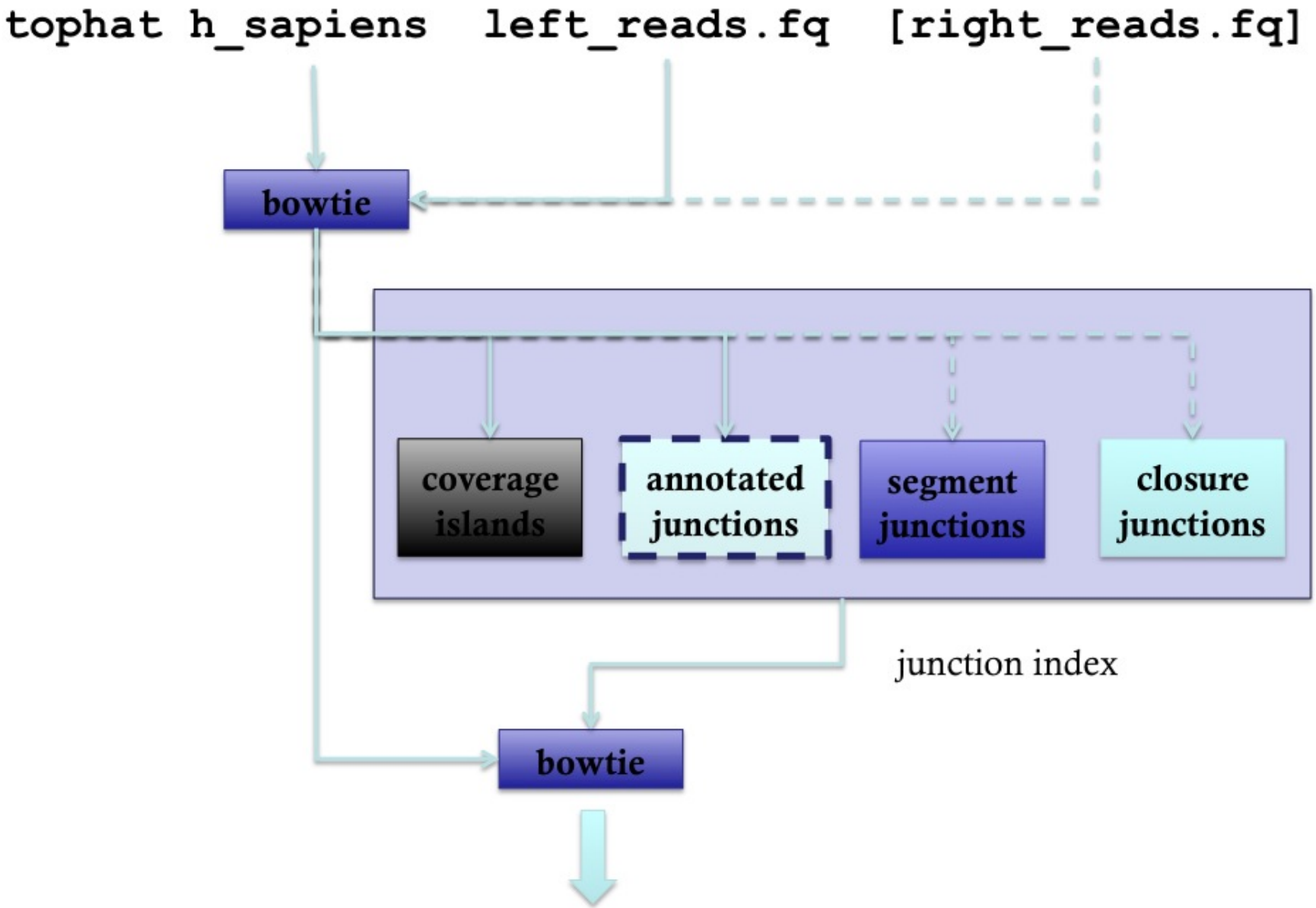


Generate possible pairings of exons



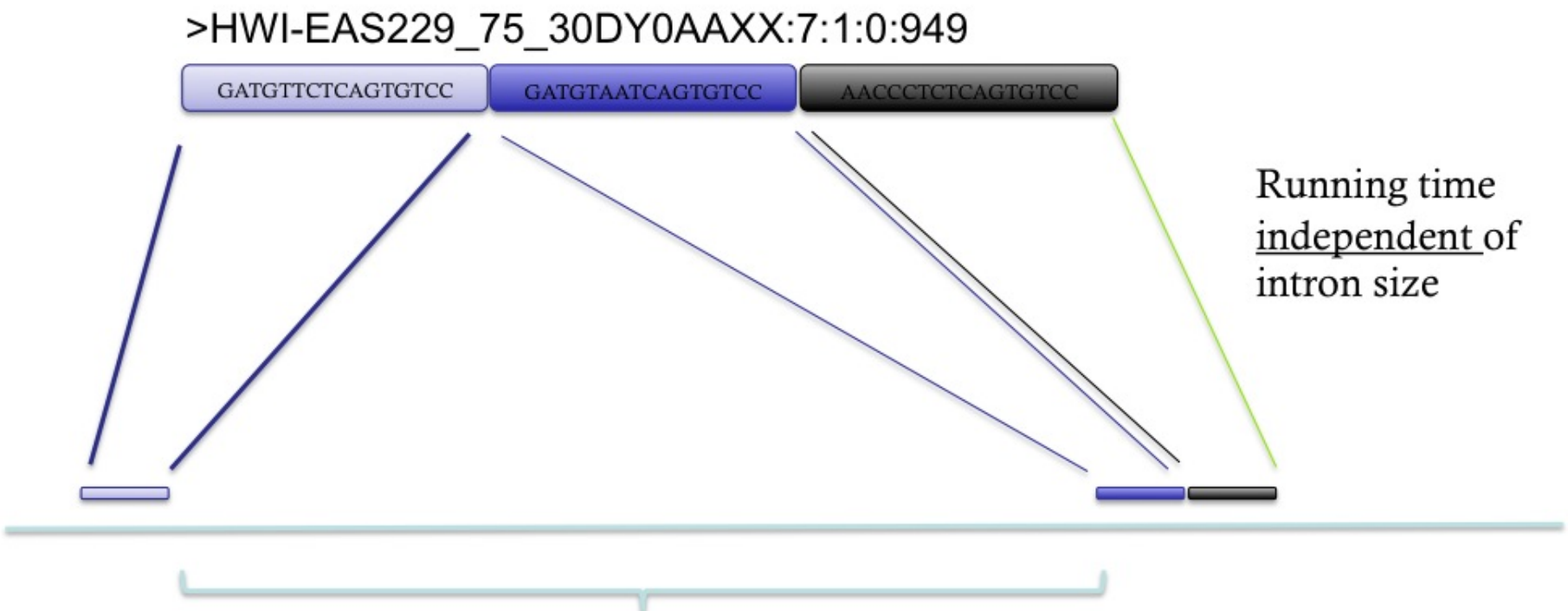
Align reads to possible junctions

(Trapnell, 2010)


tophat h_sapiens left_reads.fq [right_reads.fq]
bowtie
coverage islands
annotated junctions
segment junctions
closure junctions
junction index
bowtie
Spliced alignments
(Trapnell, 2010)

Longer reads

By segmenting the long reads, and mapping the segments independently, we can look harder for junctions we might have missed with shorter reads



(Trapnell, 2010)

Read Alignment with PALMapper



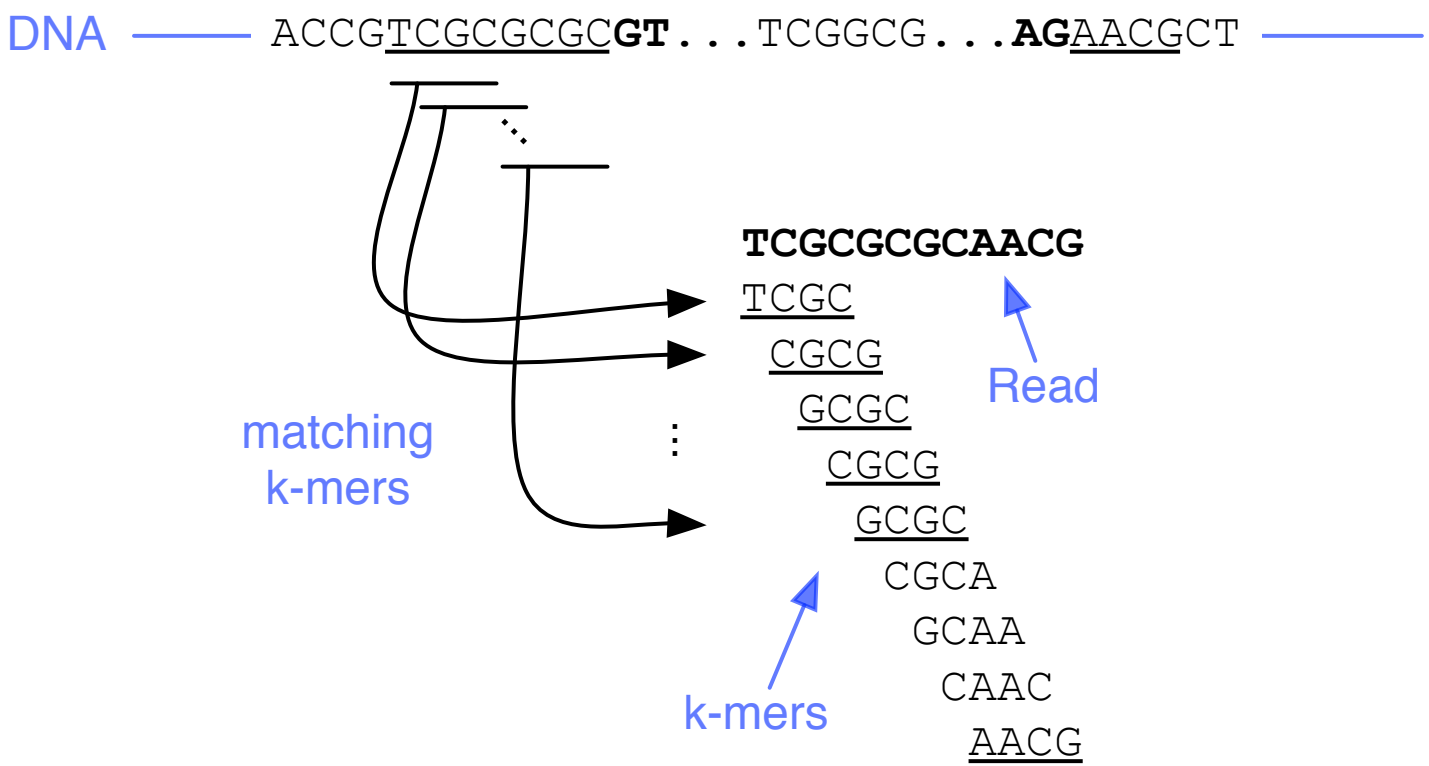
GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009b]
- *k*-mer based index, well suited for smaller genomes

Read Alignment with PALMapper

GenomeMapper for (unspliced) read mapping:

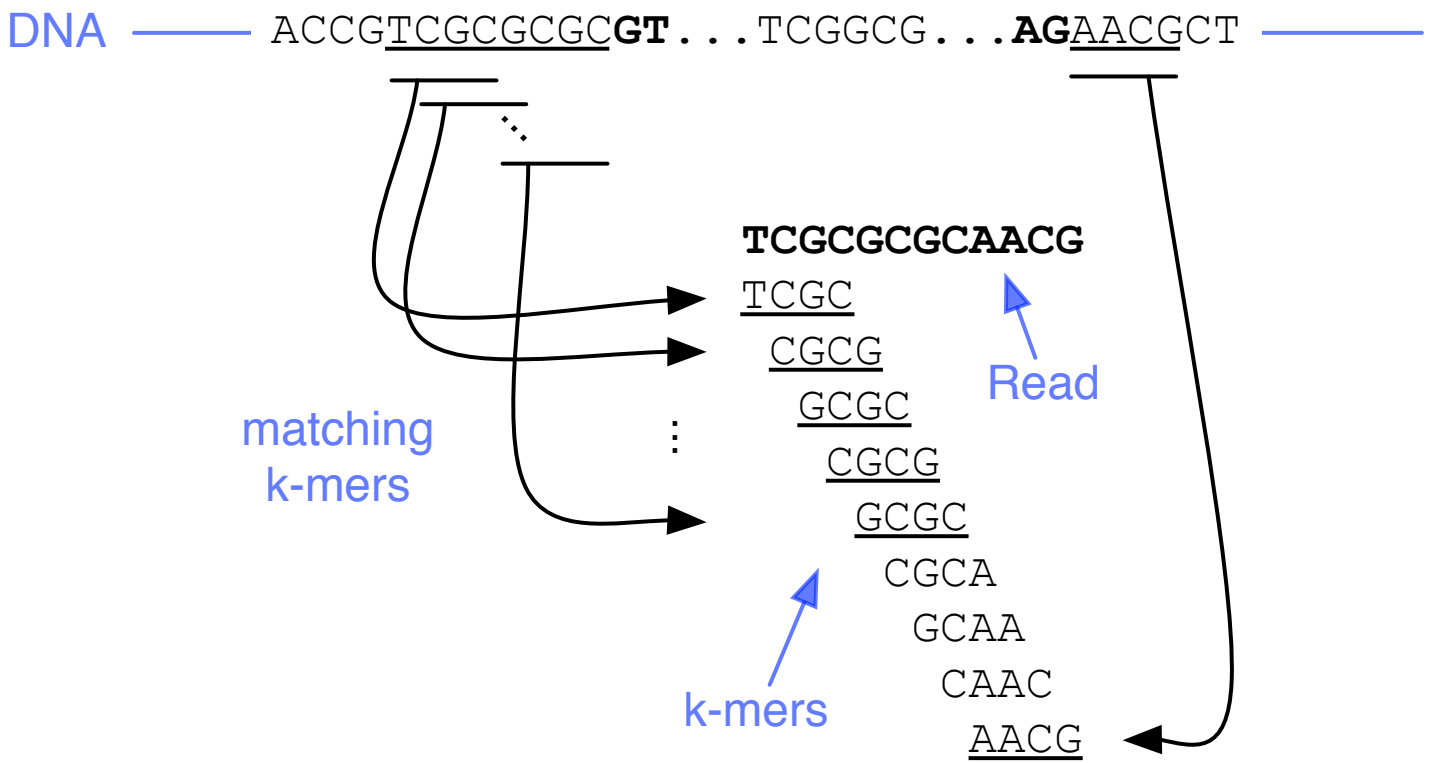
- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009b]
- *k*-mer based index, well suited for smaller genomes



Read Alignment with PALMapper

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009b]
- *k*-mer based index, well suited for smaller genomes



Read Alignment with PALMapper



GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009b]
- *k*-mer based index, well suited for smaller genomes

QPALMA for spliced read alignments:

- GenomeMapper identifies *seed regions*
- *Spliced alignments* by QPALMA [De Bona et al., 2008]

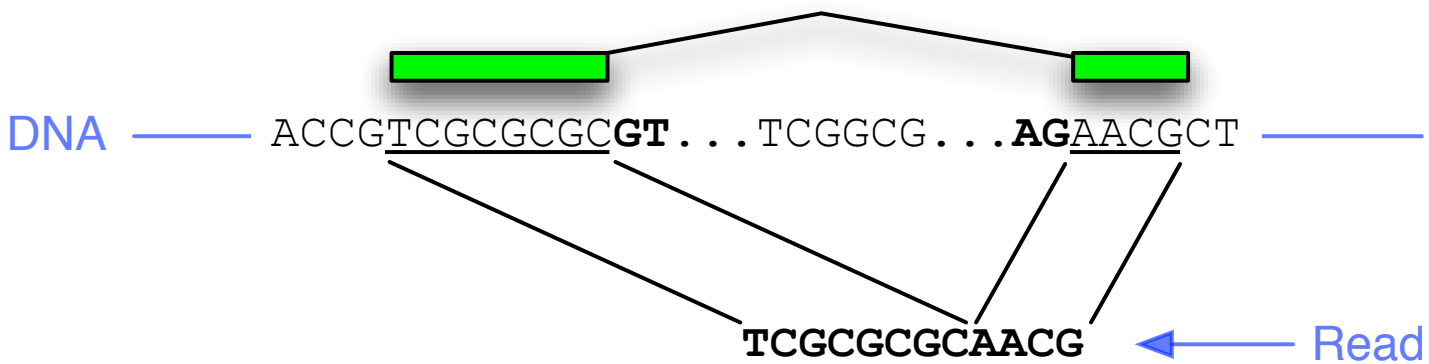
Read Alignment with PALMapper

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009b]
- *k*-mer based index, well suited for smaller genomes

QPALMA for spliced read alignments:

- GenomeMapper identifies *seed regions*
- *Spliced alignments* by QPALMA [De Bona et al., 2008]



QPALMA: Extended Smith-Waterman Alg.



	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

Classical scoring $f : \Sigma \times \Sigma \rightarrow \mathbb{R}$

QPALMA: Extended Smith-Waterman Alg.



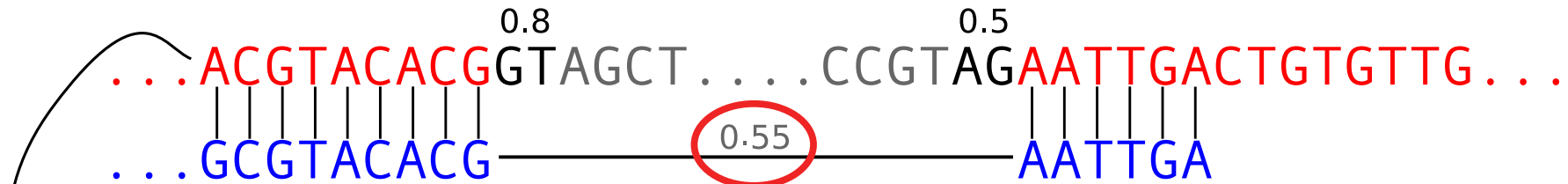
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

Classical scoring $f : \Sigma \times \Sigma \rightarrow \mathbb{R}$

QPALMA: Extended Smith-Waterman Alg.



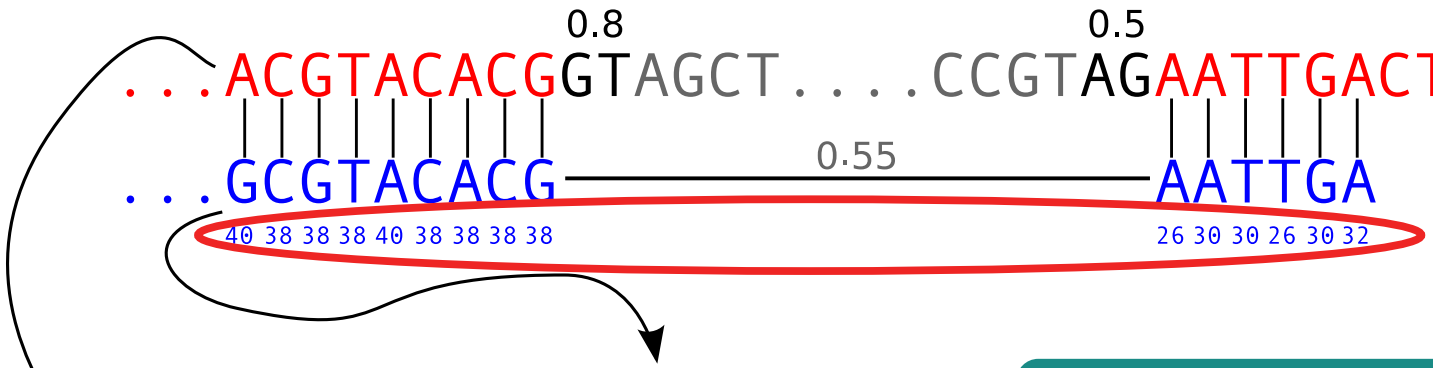
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

Classical scoring $f : \Sigma \times \Sigma \rightarrow \mathbb{R}$

QPALMA: Extended Smith-Waterman Alg.



	gap	A	C	G	T	N
gap	0.33	$f_{-A}(\cdot)$	$f_{-C}(\cdot)$	$f_{-G}(\cdot)$	$f_{-T}(\cdot)$	$f_{-N}(\cdot)$
A	0.31	$f_{AA}(\cdot)$	$f_{AC}(\cdot)$	$f_{AG}(\cdot)$	$f_{AT}(\cdot)$	$f_{AN}(\cdot)$
C	0.44	$f_{CA}(\cdot)$	$f_{CC}(\cdot)$	$f_{CG}(\cdot)$	$f_{CT}(\cdot)$	$f_{CN}(\cdot)$
G	0.13	$f_{GA}(\cdot)$	$f_{GC}(\cdot)$	$f_{GG}(\cdot)$	$f_{GT}(\cdot)$	$f_{GN}(\cdot)$
T	0.55	$f_{TA}(\cdot)$	$f_{TC}(\cdot)$	$f_{TG}(\cdot)$	$f_{TT}(\cdot)$	$f_{TN}(\cdot)$
N	0.12	$f_{NA}(\cdot)$	$f_{NC}(\cdot)$	$f_{NG}(\cdot)$	$f_{NT}(\cdot)$	$f_{NN}(\cdot)$

Source of information

- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

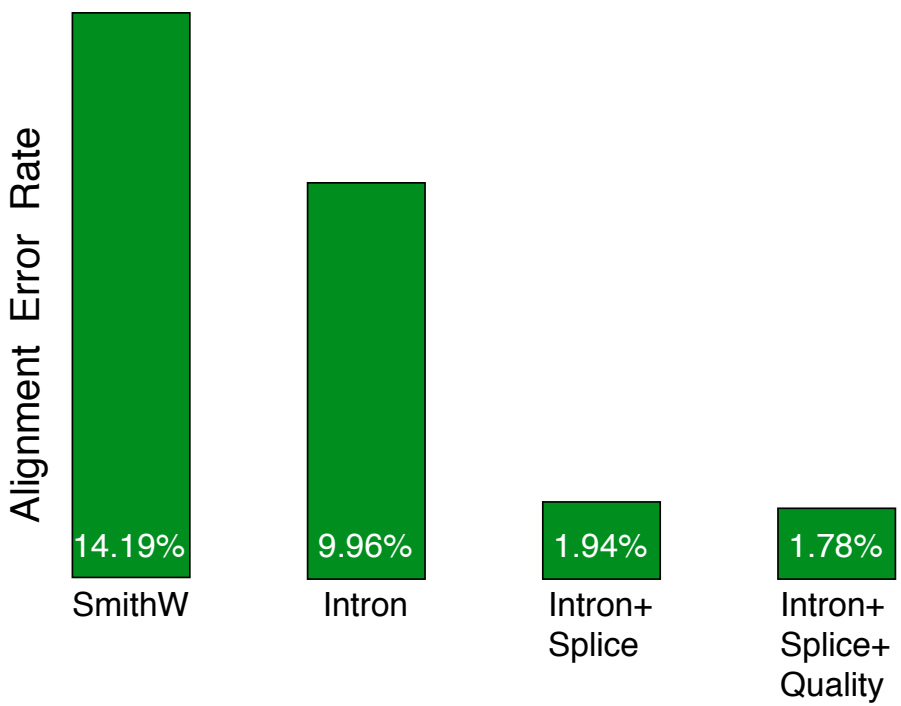
Quality scoring $f : (\Sigma \times \mathbb{R}) \times \Sigma \rightarrow \mathbb{R}$

[De Bona et al., 2008]

RNA-Seq Read Alignment with QPALMA

Generate set of artificially spliced reads

- Genomic reads with quality information
- Genome annotation for artificially splicing the reads
- Use 10,000 reads for training and 30,000 for testing

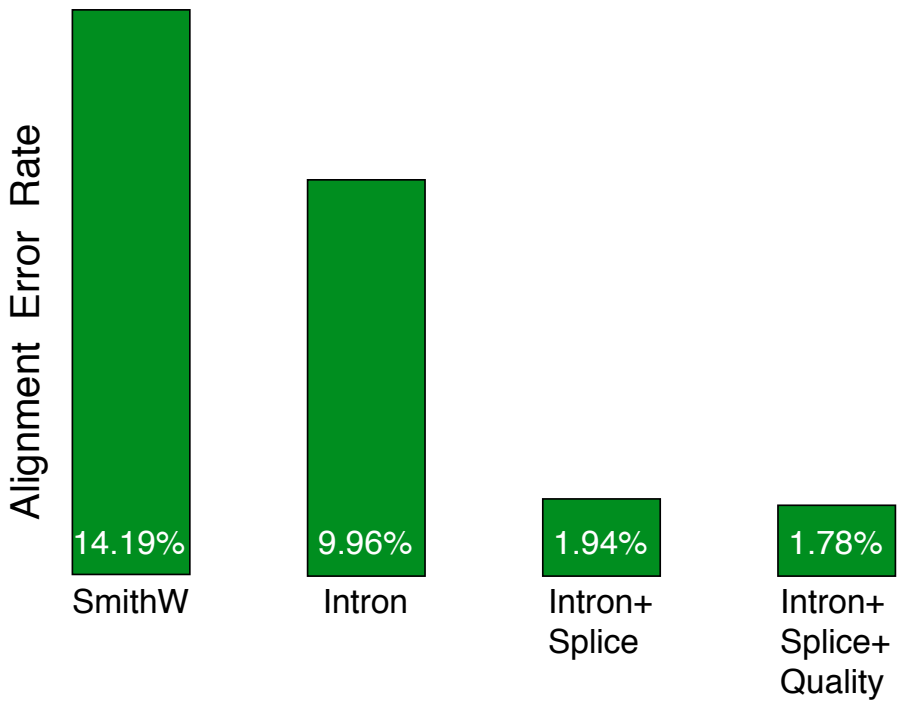


[De Bona et al., 2008]

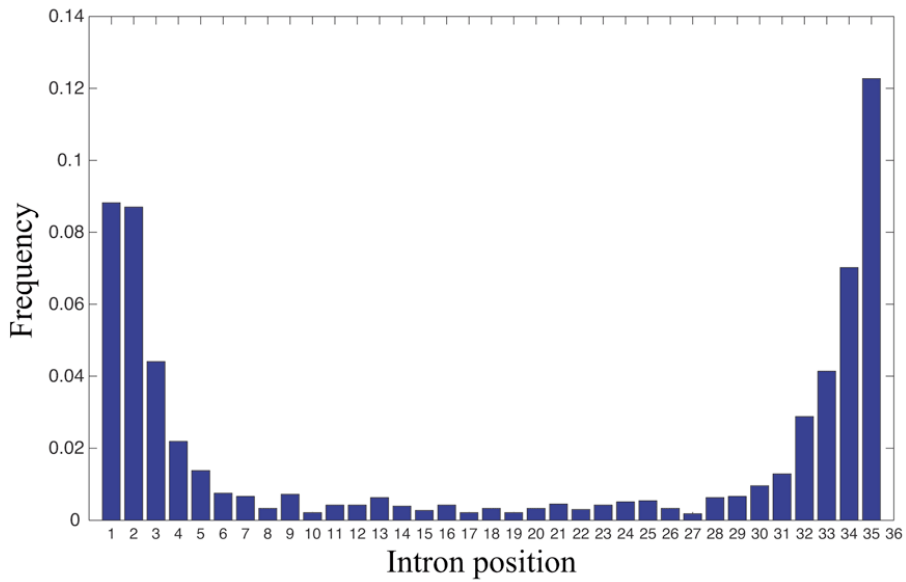
RNA-Seq Read Alignment with QPALMA

Generate set of artificially spliced reads

- Genomic reads with quality information
- Genome annotation for artificially splicing the reads
- Use 10,000 reads for training and 30,000 for testing



Error vs. intron position

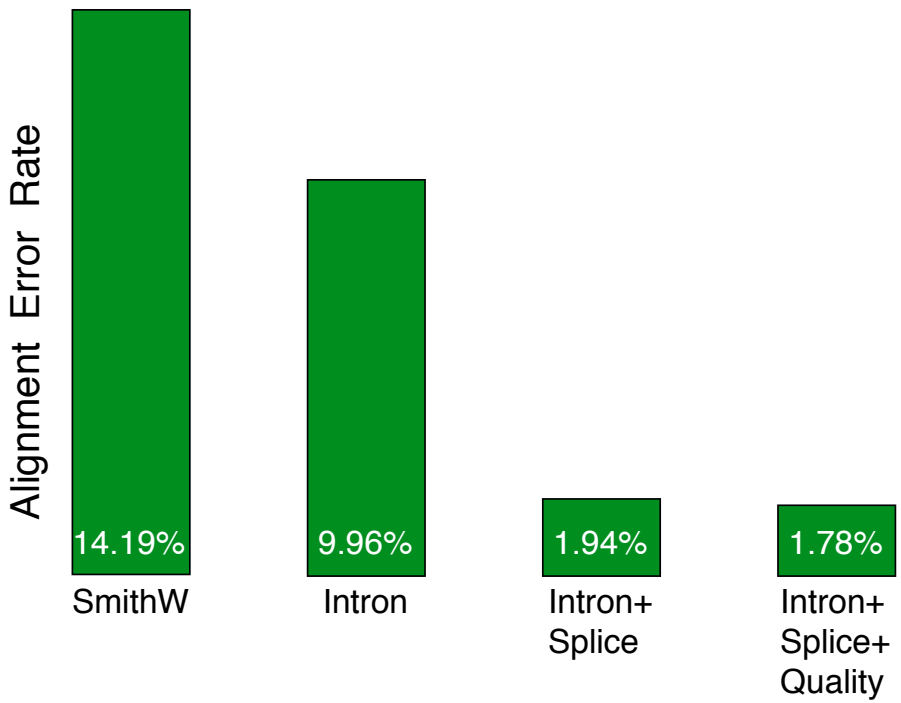


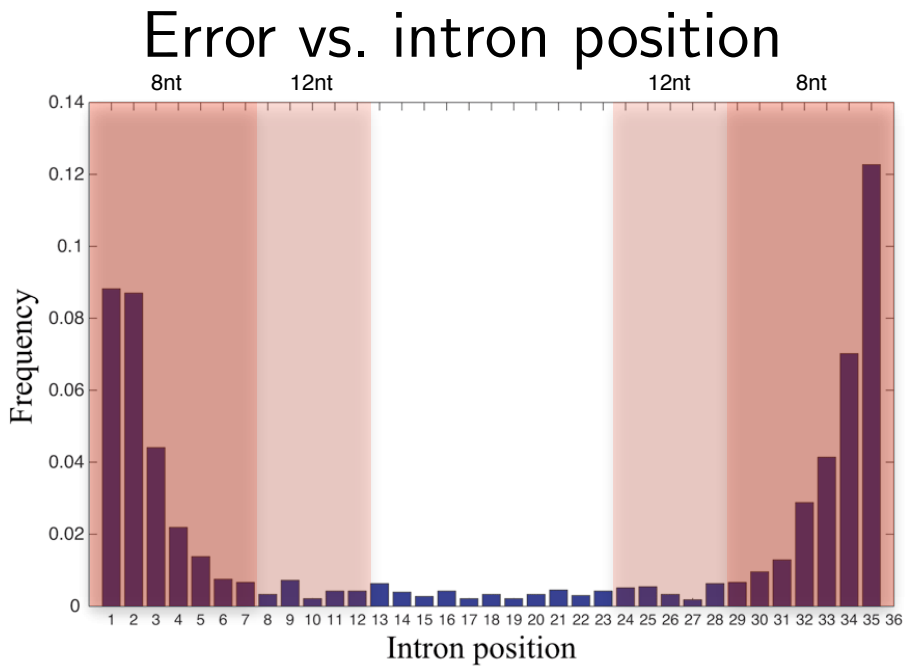
[De Bona et al., 2008]

RNA-Seq Read Alignment with QPALMA

Generate set of artificially spliced reads

- Genomic reads with quality information
- Genome annotation for artificially splicing the reads
- Use 10,000 reads for training and 30,000 for testing





[De Bona et al., 2008]

Quantitation

Estimate Gene Expression



Idea: Use the number of reads mapping to a gene as estimate for the gene expression.

Problem: Read number scales with total number of reads and transcript length

Approach: Normalize read count, by

- Length of the transcript (sum of exonic regions in kilobases)
- Total number of reads (in million)

$\Rightarrow$ **Reads per kilobase per million mapped reads (RPKM)**

Alternative quantity for paired end sequencing (2 reads/fragment):

$\Rightarrow$ **Fragments per kilobase per million mapped reads (FPKM)**

RPKM/FPKM values are strongly dependent on the expression level of the highest expressed genes (largest fraction of reads). Robuster estimates are necessary, e.g. based on the median.

Estimate Gene Expression



Idea: Use the number of reads mapping to a gene as estimate for the gene expression.

Problem: Read number scales with total number of reads and transcript length

Approach: Normalize read count, by

- Length of the transcript (sum of exonic regions in kilobases)
- Total number of reads (in million)

$\Rightarrow$ **Reads per kilobase per million mapped reads (RPKM)**

Alternative quantity for paired end sequencing (2 reads/fragment):

$\Rightarrow$ **Fragments per kilobase per million mapped reads (FPKM)**

RPKM/FPKM values are strongly dependent on the expression level of the highest expressed genes (largest fraction of reads). Robuster estimates are necessary, e.g. based on the median.

Estimate Gene Expression



Idea: Use the number of reads mapping to a gene as estimate for the gene expression.

Problem: Read number scales with total number of reads and transcript length

Approach: Normalize read count, by

- Length of the transcript (sum of exonic regions in kilobases)
- Total number of reads (in million)

$\Rightarrow$ **Reads per kilobase per million mapped reads (RPKM)**

Alternative quantity for paired end sequencing (2 reads/fragment):

$\Rightarrow$ **Fragments per kilobase per million mapped reads (FPKM)**

RPKM/FPKM values are strongly dependent on the expression level of the highest expressed genes (largest fraction of reads). Robuster estimates are necessary, e.g. based on the median.

Estimate Gene Expression



Idea: Use the number of reads mapping to a gene as estimate for the gene expression.

Problem: Read number scales with total number of reads and transcript length

Approach: Normalize read count, by

- Length of the transcript (sum of exonic regions in kilobases)
- Total number of reads (in million)

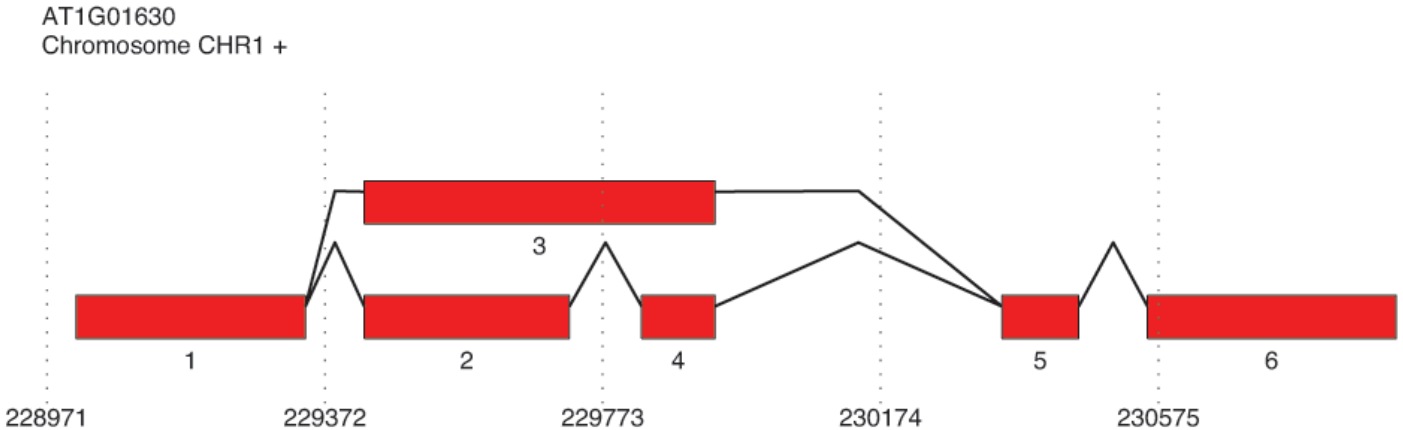
$\Rightarrow$ **Reads per kilobase per million mapped reads (RPKM)**

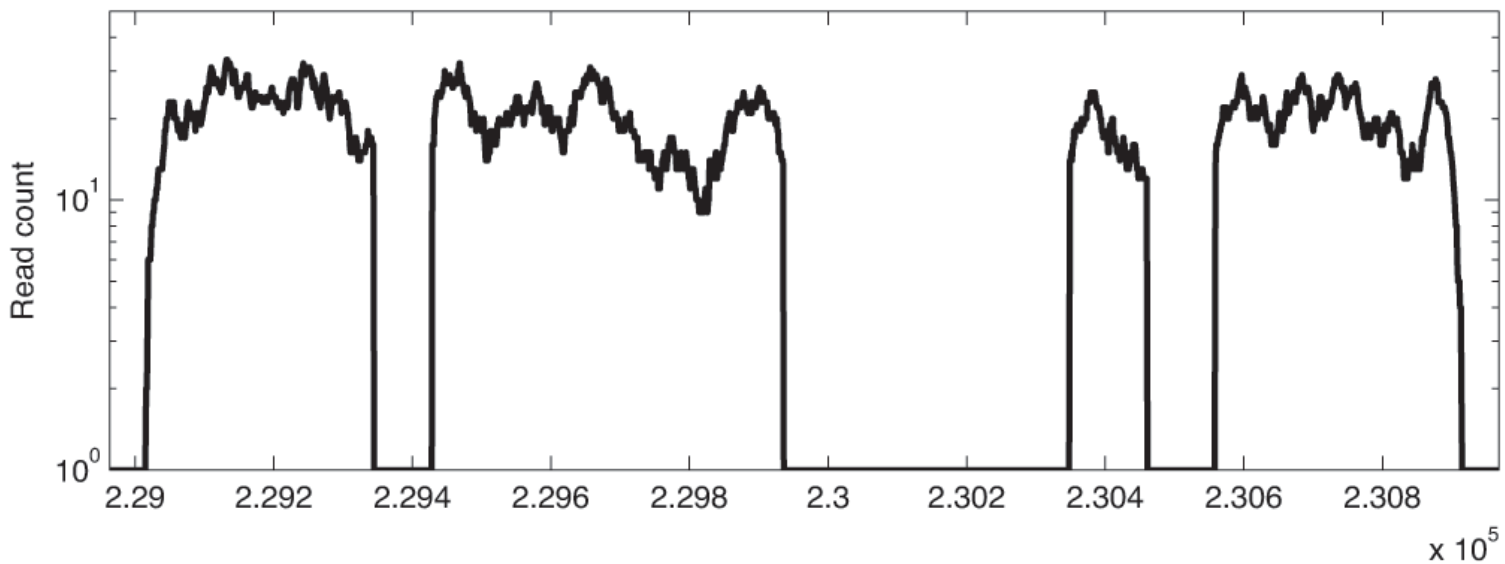
Alternative quantity for paired end sequencing (2 reads/fragment):

$\Rightarrow$ **Fragments per kilobase per million mapped reads (FPKM)**

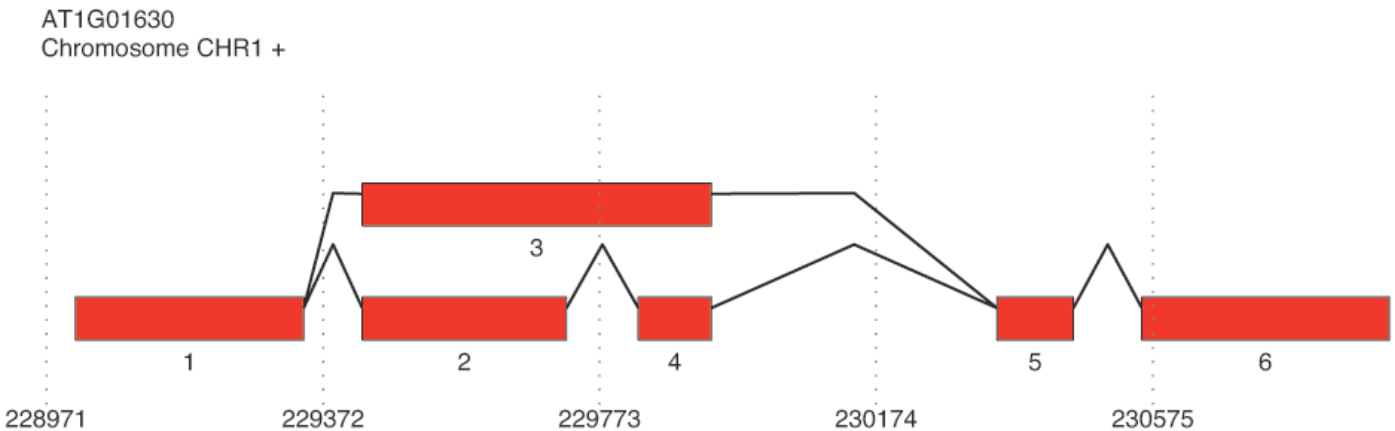
RPKM/FPKM values are strongly dependent on the expression level of the highest expressed genes (largest fraction of reads). Robuster estimates are necessary, e.g. based on the median.

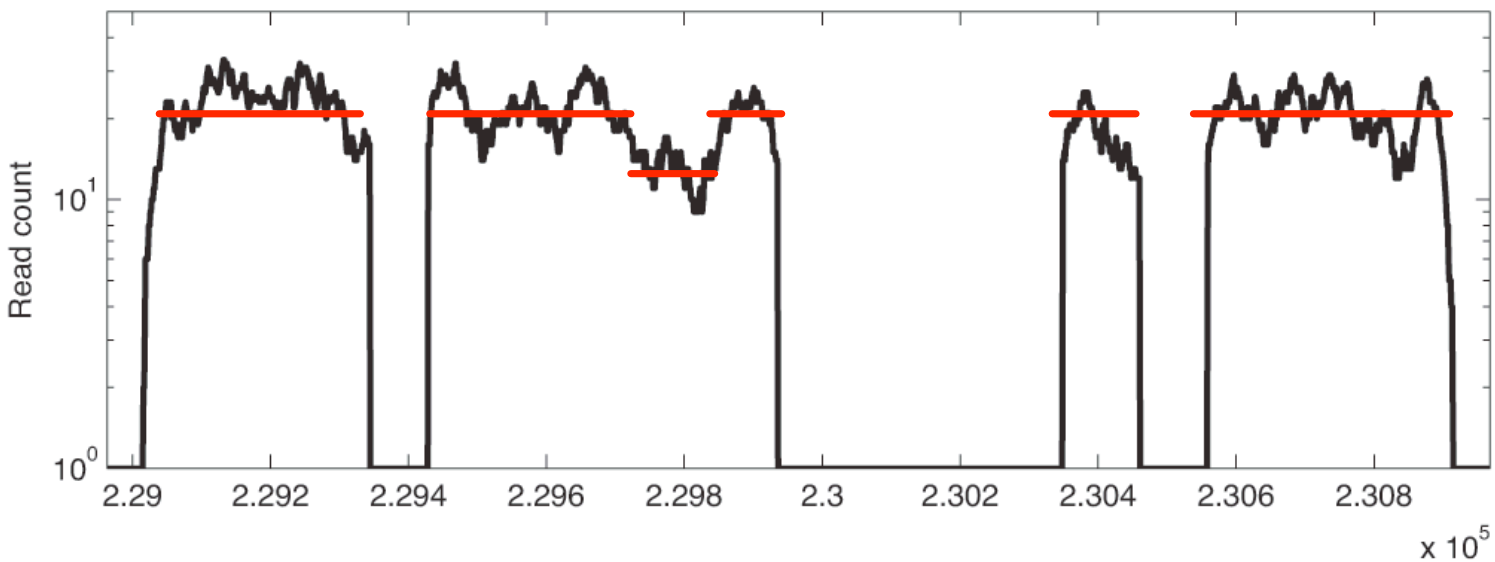
Quantitation of Transcripts





Quantitation of Transcripts





Quantitation of Transcripts



Given short reads alignments and a set of known transcripts, can we disentangle transcript abundances?

Solve an optimization problem:

- Optimizing weights w_t for each transcript $t = 1, \ldots, T$
- Exploiting additive nature of the read coverage
- Minimizing residual error (e.g., squared error)

$$(w_1, \ldots, w_T) = \underset{w_1,\ldots,w_T \geq 0}{\operatorname{argmin}} \sum_{p \in P} \left(R_p - \sum_{t=1}^{T} w_t D_{t,p} \right)^2 ,$$

with

- P: set of considered genomic positions
- R_p: observed read coverage (number of reads covering pos. p)
- $D_{t,p}$: expected read coverage for transcript t at position p

Quantitation of Transcripts



Different approaches rely on similar basic ideas with different models of how to use read count differences and optimization techniques:

- Poisson distributions [Jiang and Wong, 2009]
- Absolute differences using a flow-network [Sammeth, 2009a]
- Squared differences using quadratic programming [Bohnert et al., 2009]

Other methods: [Li et al., 2010a], [Richard et al., 2010], [Trapnell et al., 2010]

Problems:

- Abundances cannot unambiguously be determined with single end reads, better chances with paired ends [Lacroix et al., 2008]
- Solution may not be stable: a few reads more or less may completely change abundance estimates
- Read coverage is not uniform over the transcript

Quantitation of Transcripts



Different approaches rely on similar basic ideas with different models of how to use read count differences and optimization techniques:

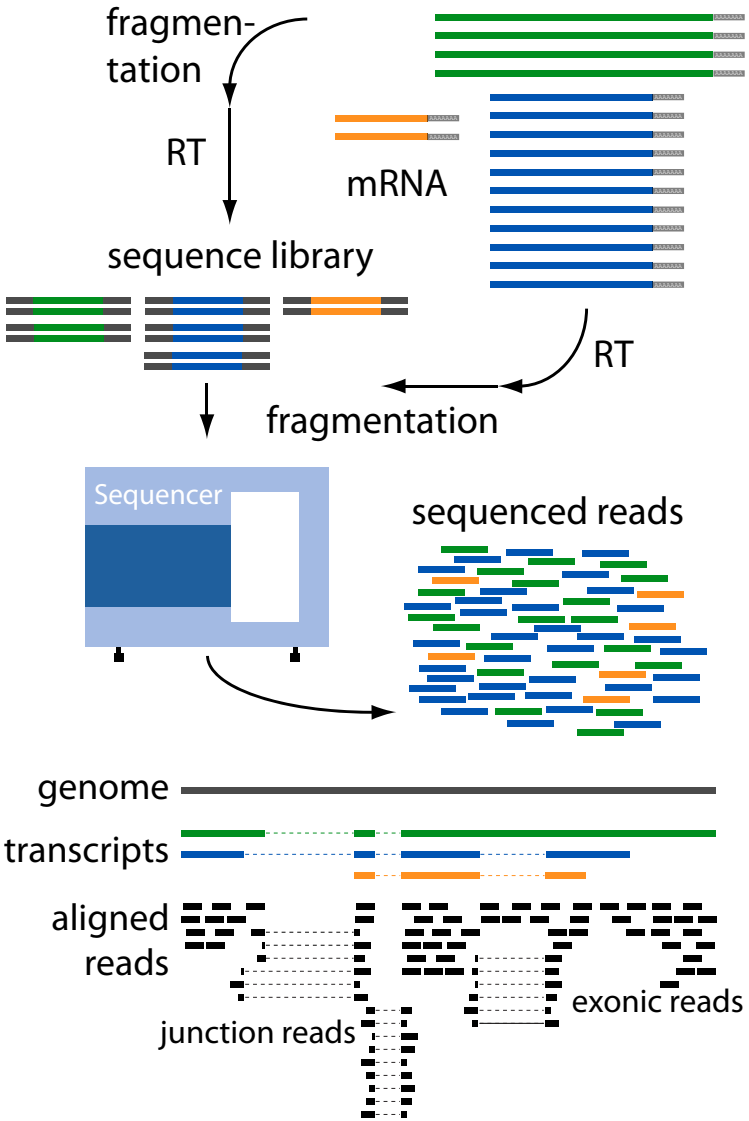
- Poisson distributions [Jiang and Wong, 2009]
- Absolute differences using a flow-network [Sammeth, 2009a]
- Squared differences using quadratic programming [Bohnert et al., 2009]

Other methods: [Li et al., 2010a], [Richard et al., 2010], [Trapnell et al., 2010]

Problems:

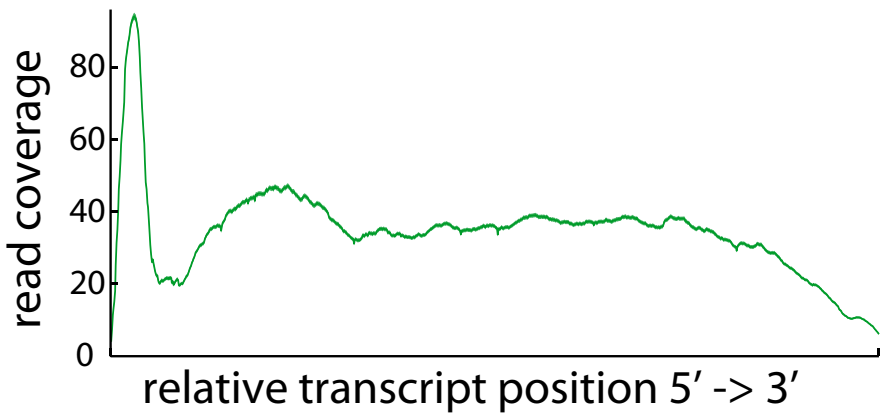
- Abundances cannot unambiguously be determined with single end reads, better chances with paired ends [Lacroix et al., 2008]
- Solution may not be stable: a few reads more or less may completely change abundance estimates
- Read coverage is not uniform over the transcript

RNA-Seq Biases and Quantitation



Biases due to ...

- cDNA library construction
- Sequencing
- Read mapping



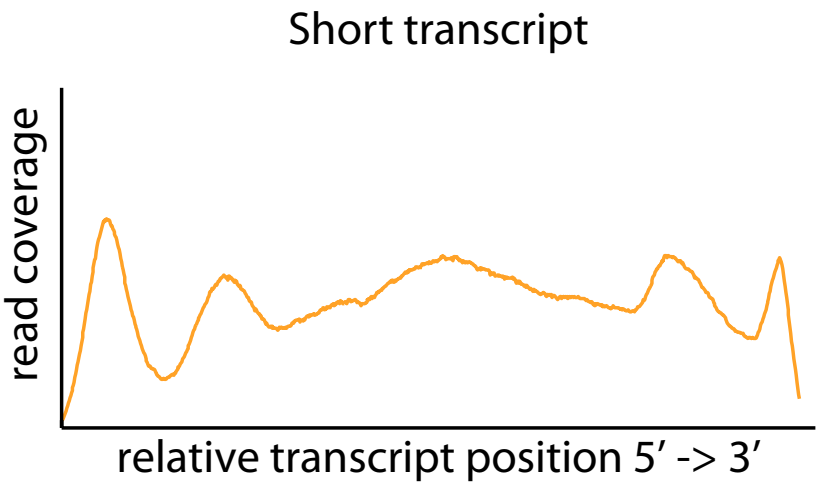
(average over annotated transcripts of length ≈1kb for the *C. elegans* SRX001872 dataset)

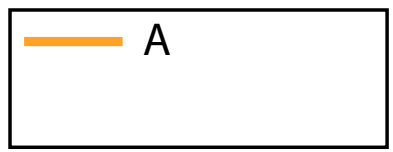
RNA-Seq Biases – Publications



- Transcript length bias [Oshlack and Wakefield, 2009]
- Bias in small RNA sequencing [Linsen et al., 2009]
- Sequence bias in Illumina sequencer [Dohm et al., 2008]
- Biases in nucleotide composition induced by random hexamer priming [Hansen et al., 2010]
- Modeling non-uniformity [Li et al., 2010b, Bohnert et al., 2009]

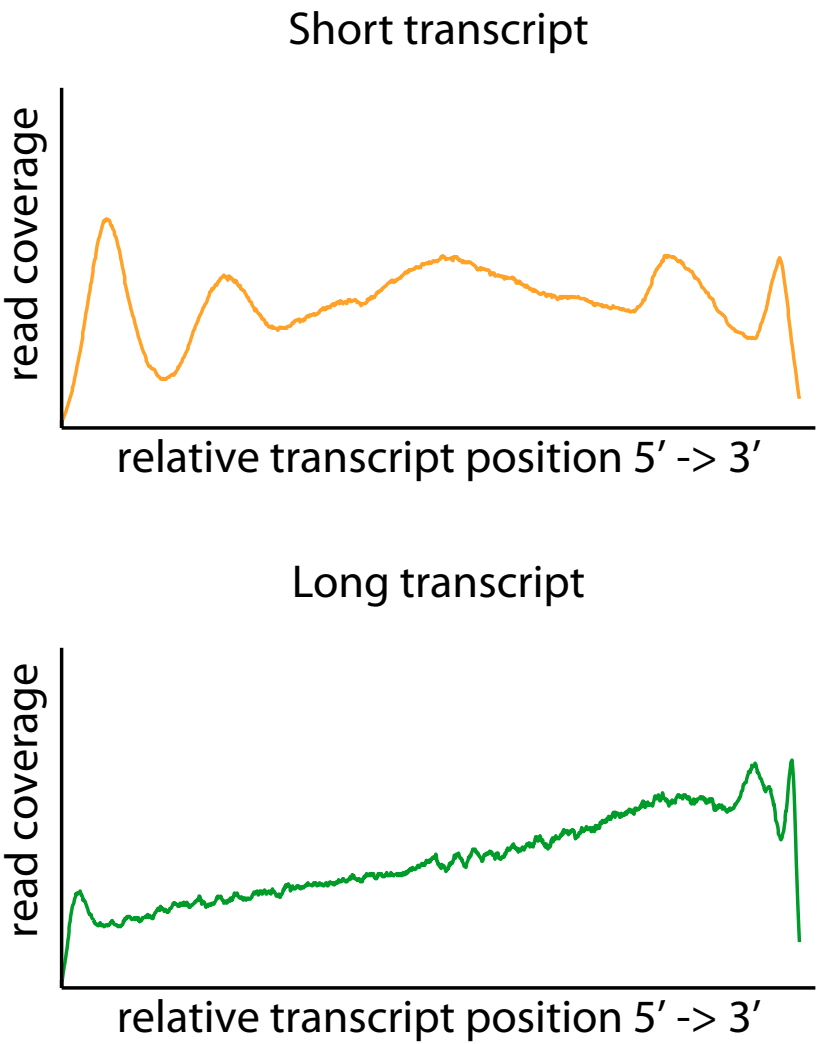
rQuant – Basic Idea





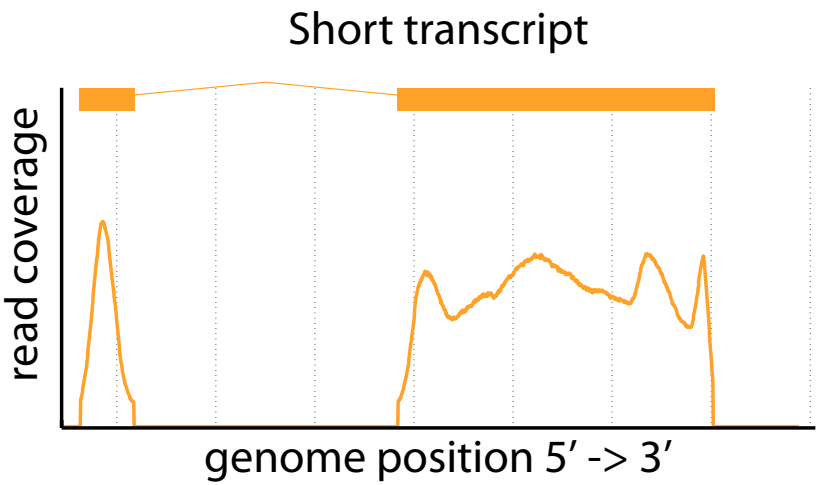
$$M_p = w_A D_{A,p} + w_B D_{B,p} \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell(M_p, R_p)$$

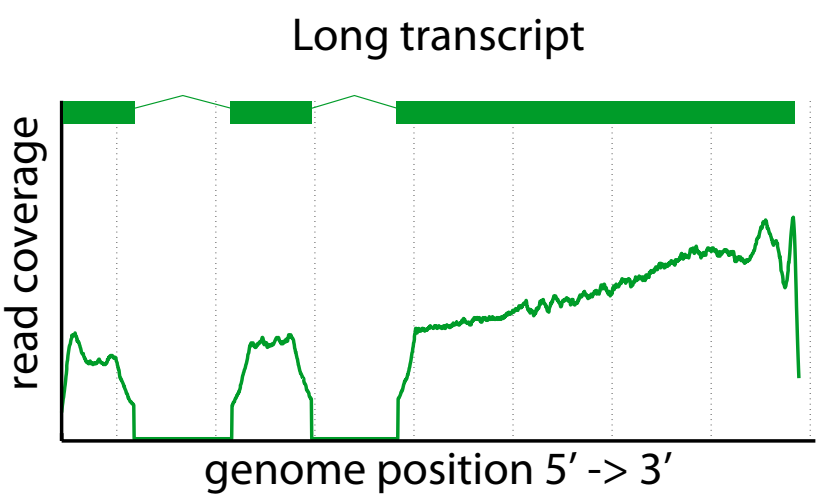
rQuant – Basic Idea

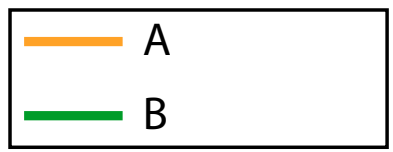


$$M_p = w_A D_{A,p} + w_B D_{B,p} \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell(M_p, R_p)$$

rQuant – Basic Idea

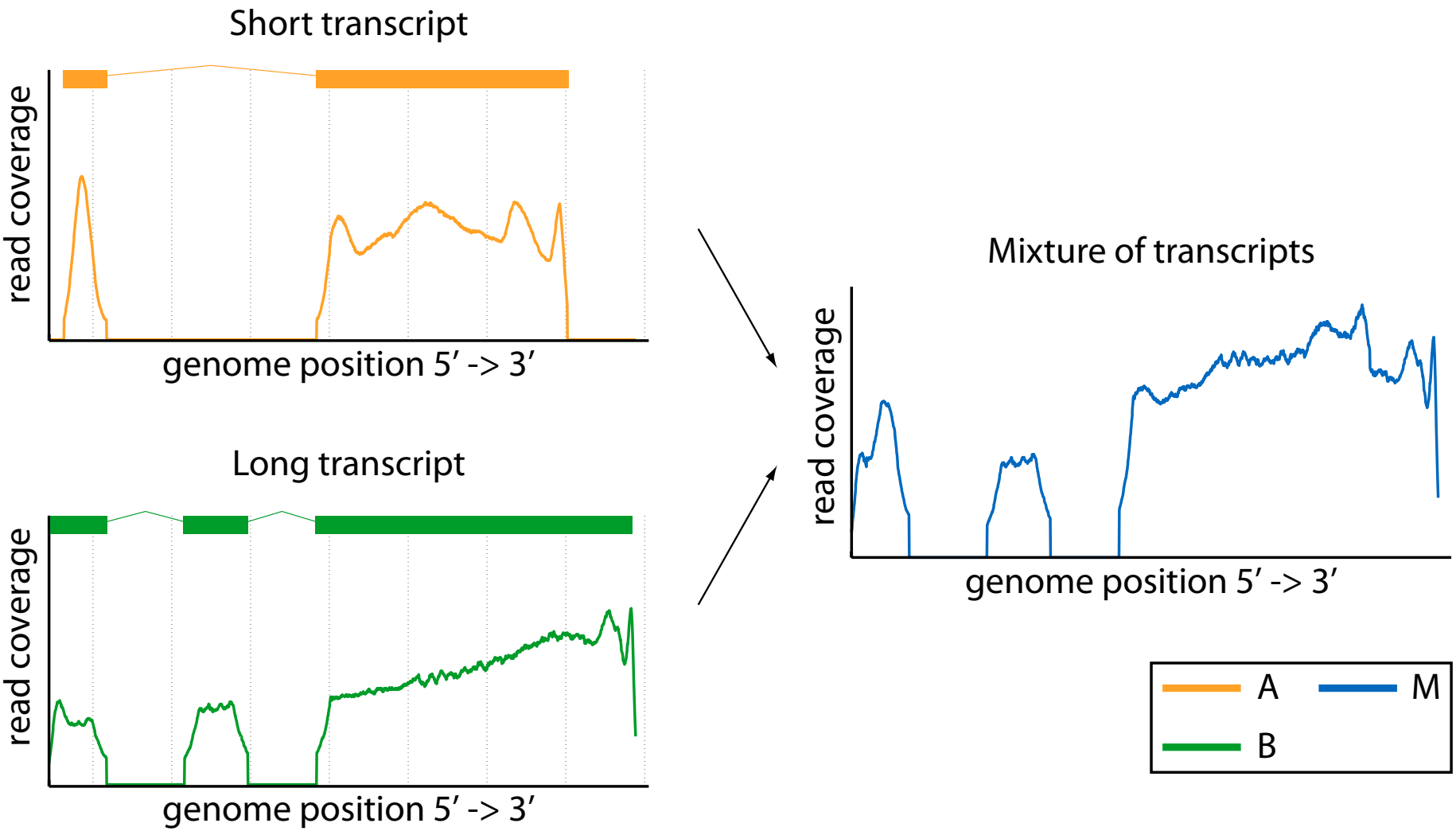






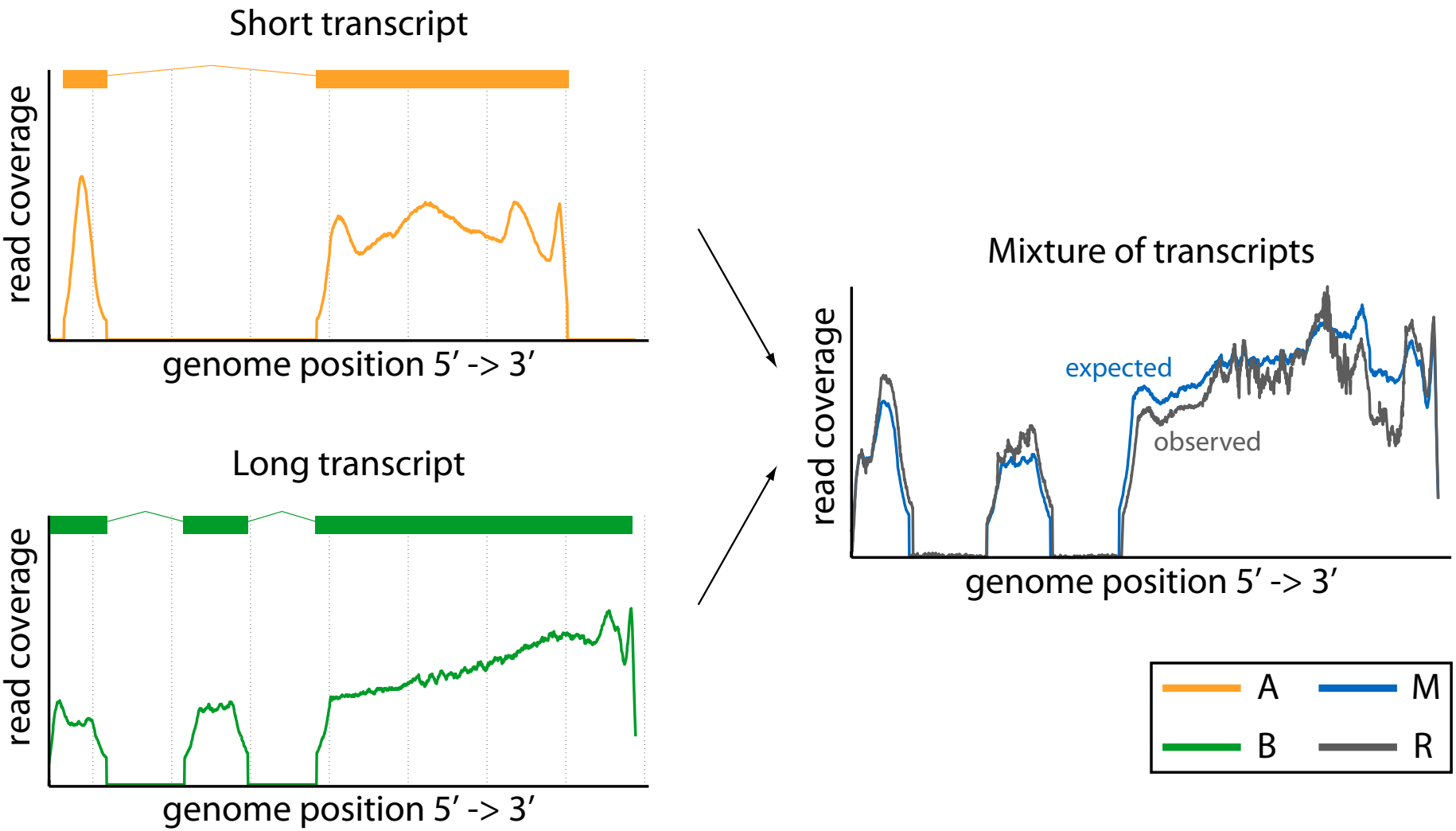
$$M_p = w_A D_{A,p} + w_B D_{B,p} \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell(M_p, R_p)$$

rQuant – Basic Idea



$$M_p = w_A D_{A,p} + w_B D_{B,p} \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell(M_p, R_p)$$

rQuant – Basic Idea

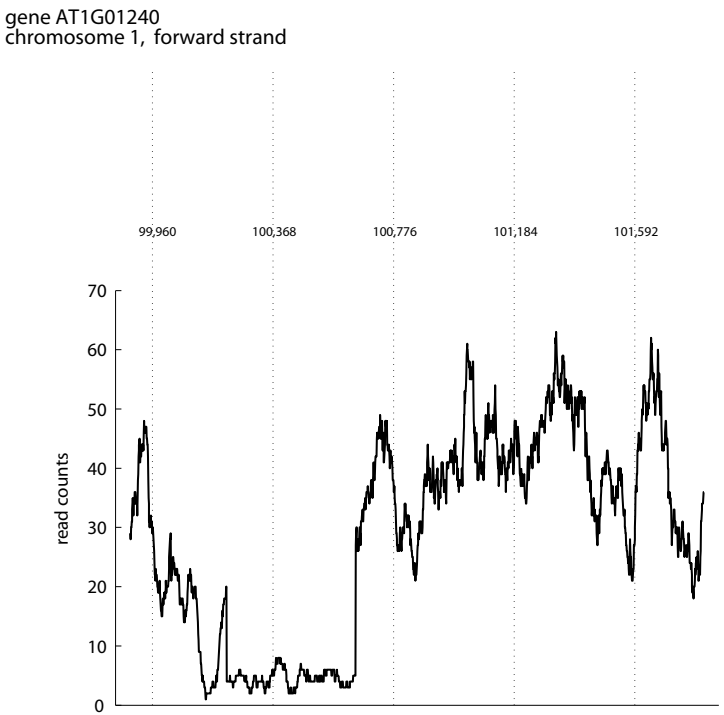


$$M_p = w_A D_{A,p} + w_B D_{B,p} \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell\left(M_p, R_p\right)$$

rQuant – Iterative Algorithm

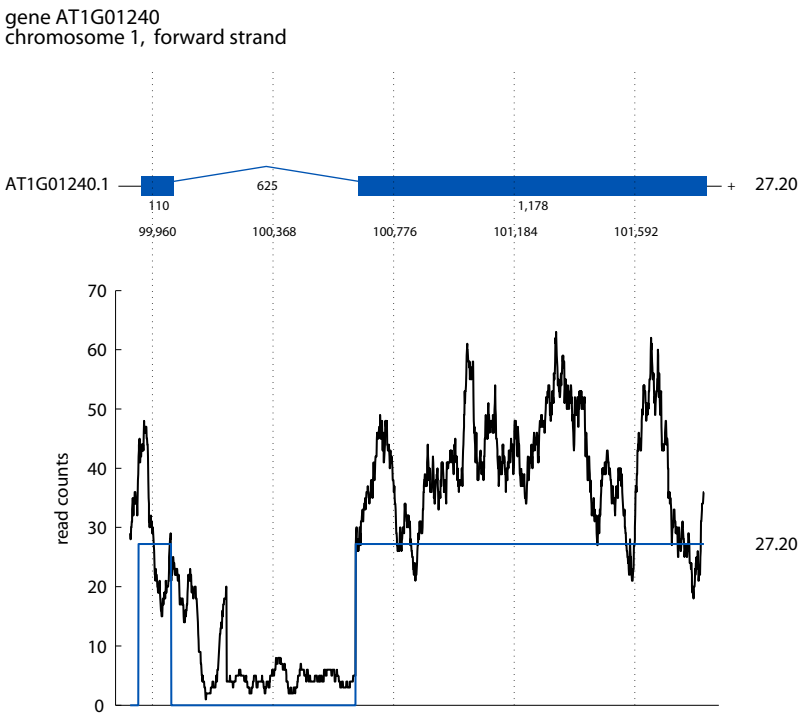


1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence.



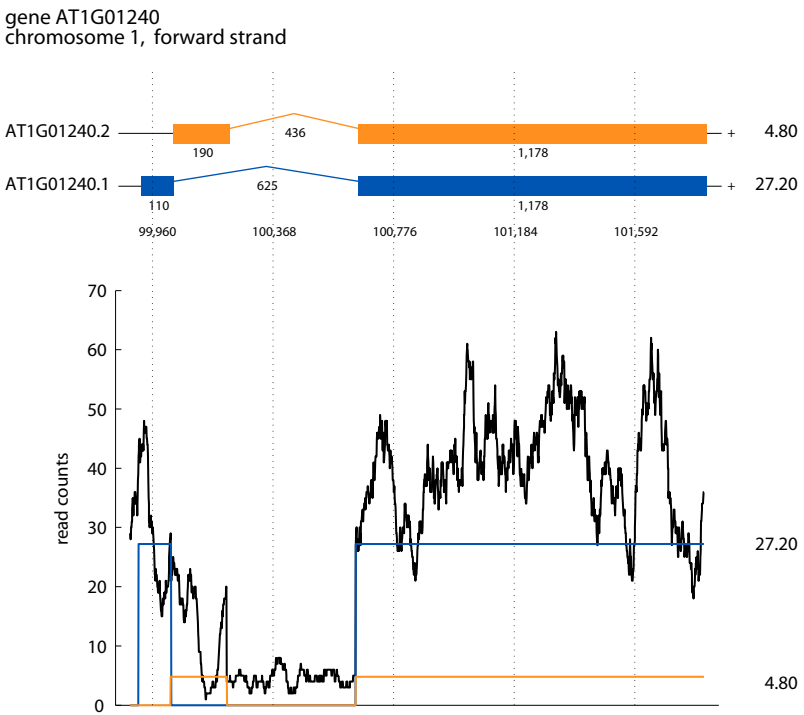
rQuant – Iterative Algorithm

1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence.



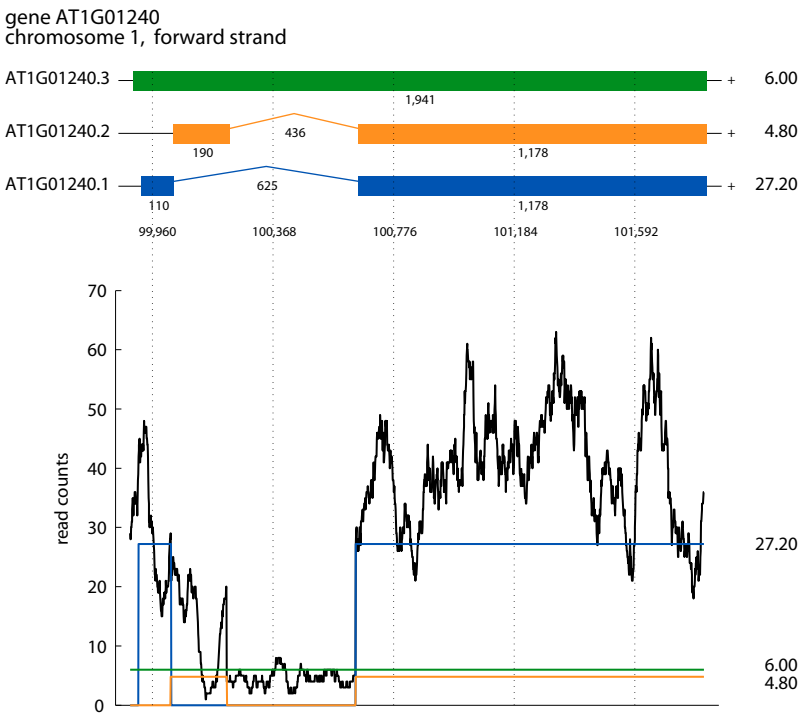
rQuant – Iterative Algorithm

1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence.



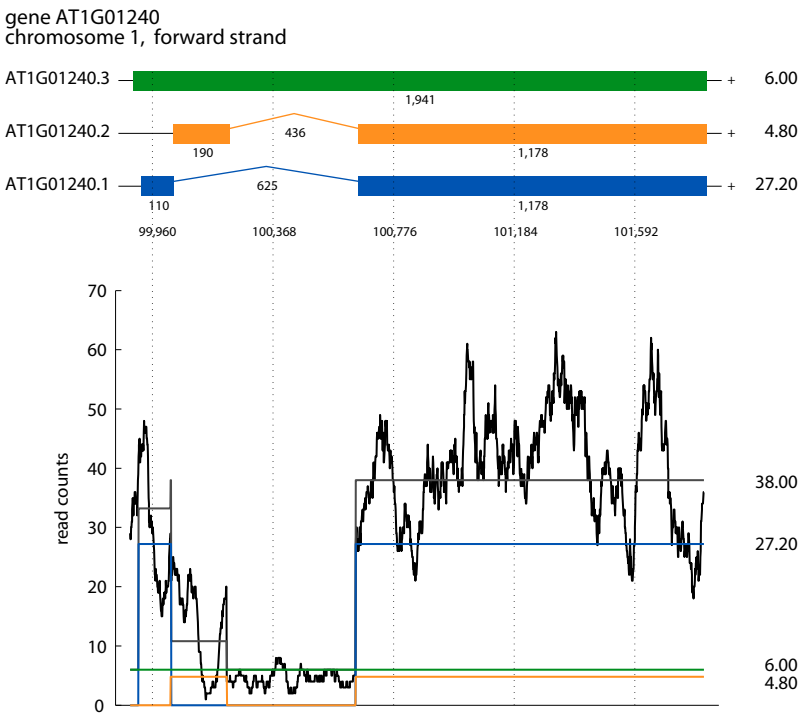
rQuant – Iterative Algorithm

1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence.



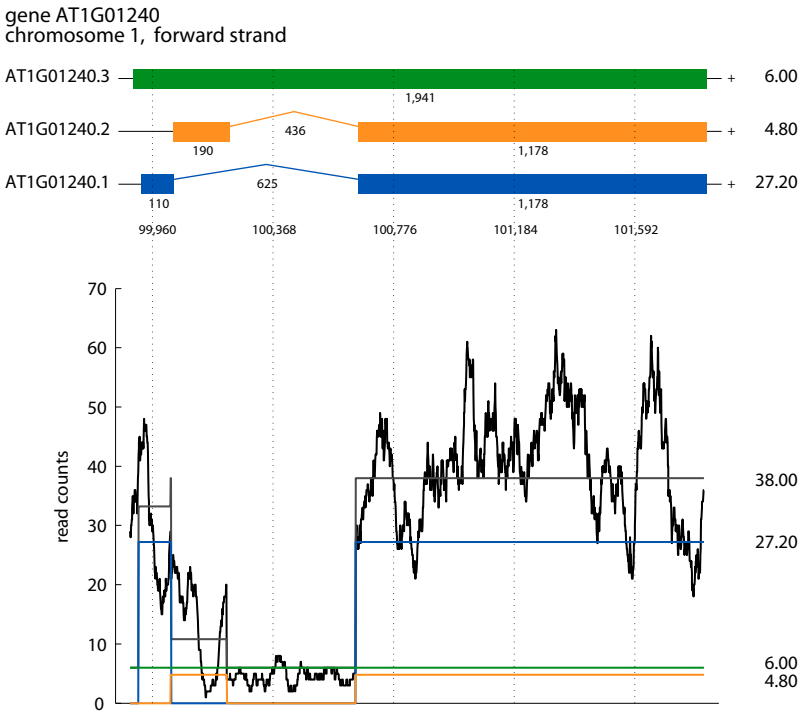
rQuant – Iterative Algorithm

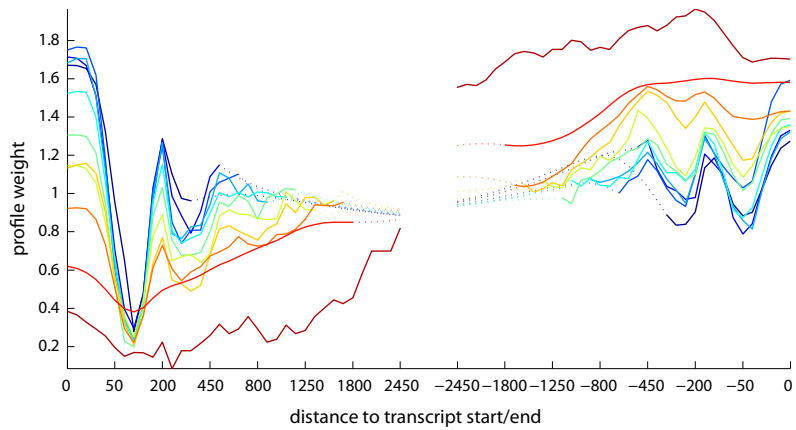
1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence.



rQuant – Iterative Algorithm

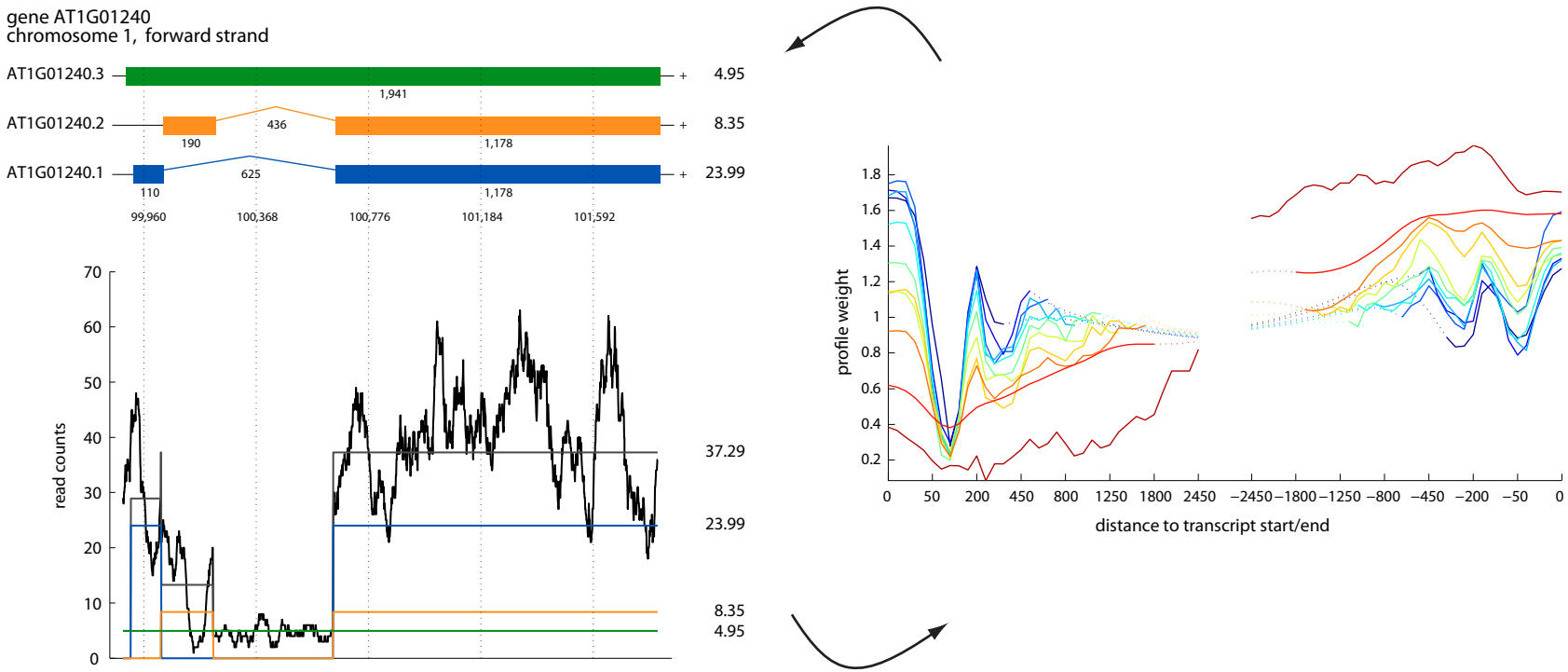
1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence.





rQuant – Iterative Algorithm

1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence.



rQuant Evaluation I



Method rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I



Method rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I

Method rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I

Method rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I

Method rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I



Method rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

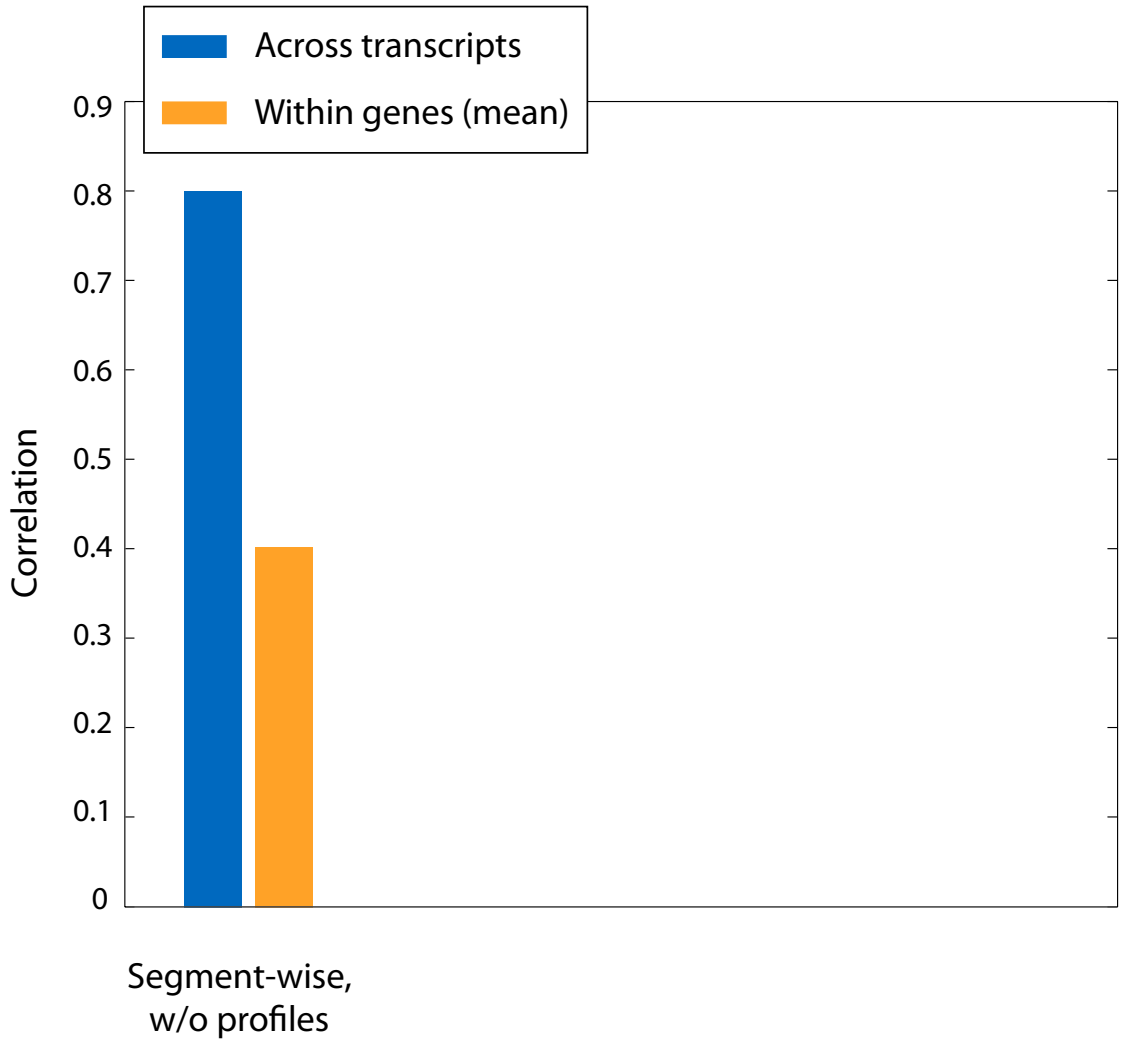
Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

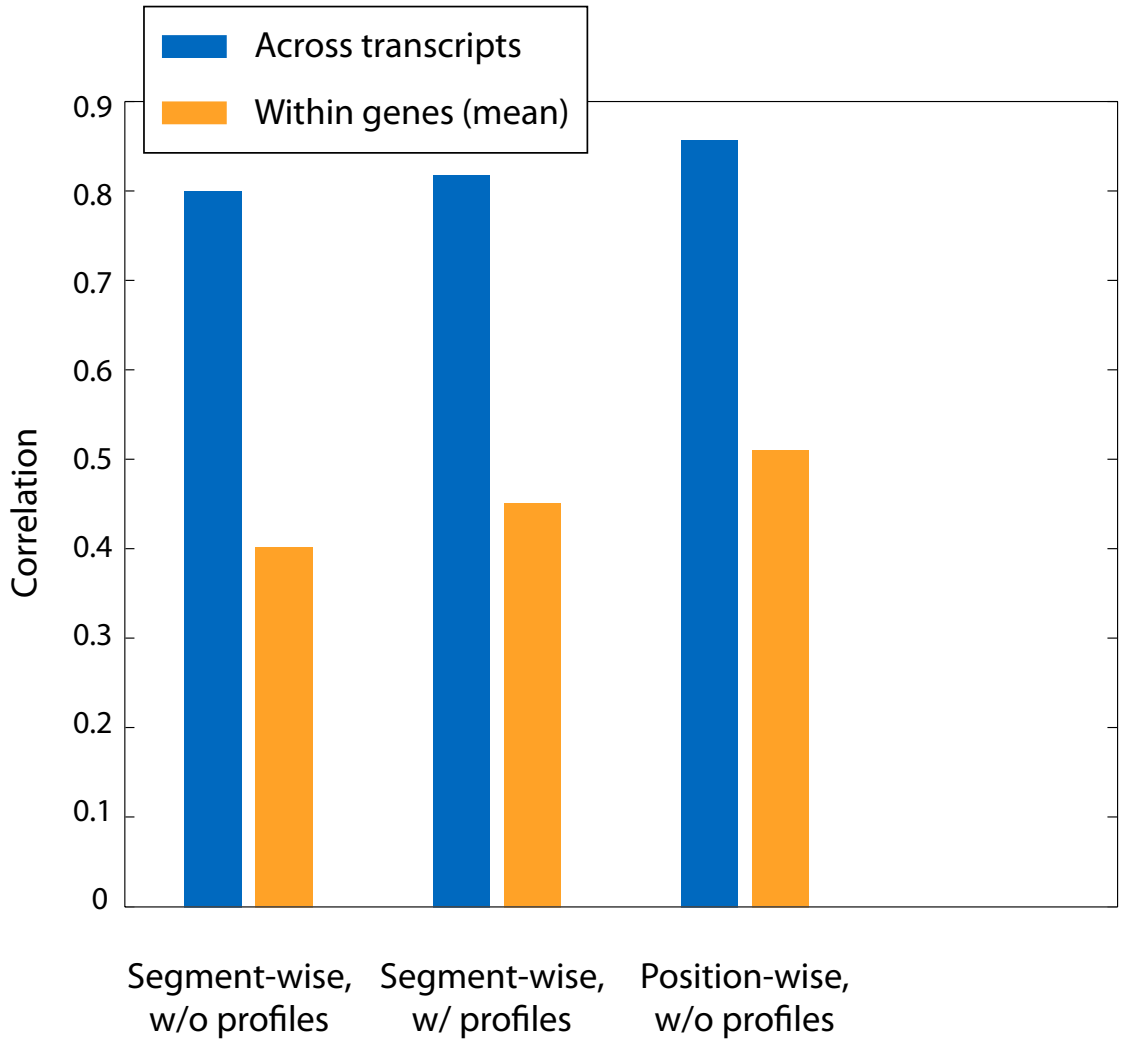
- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation II



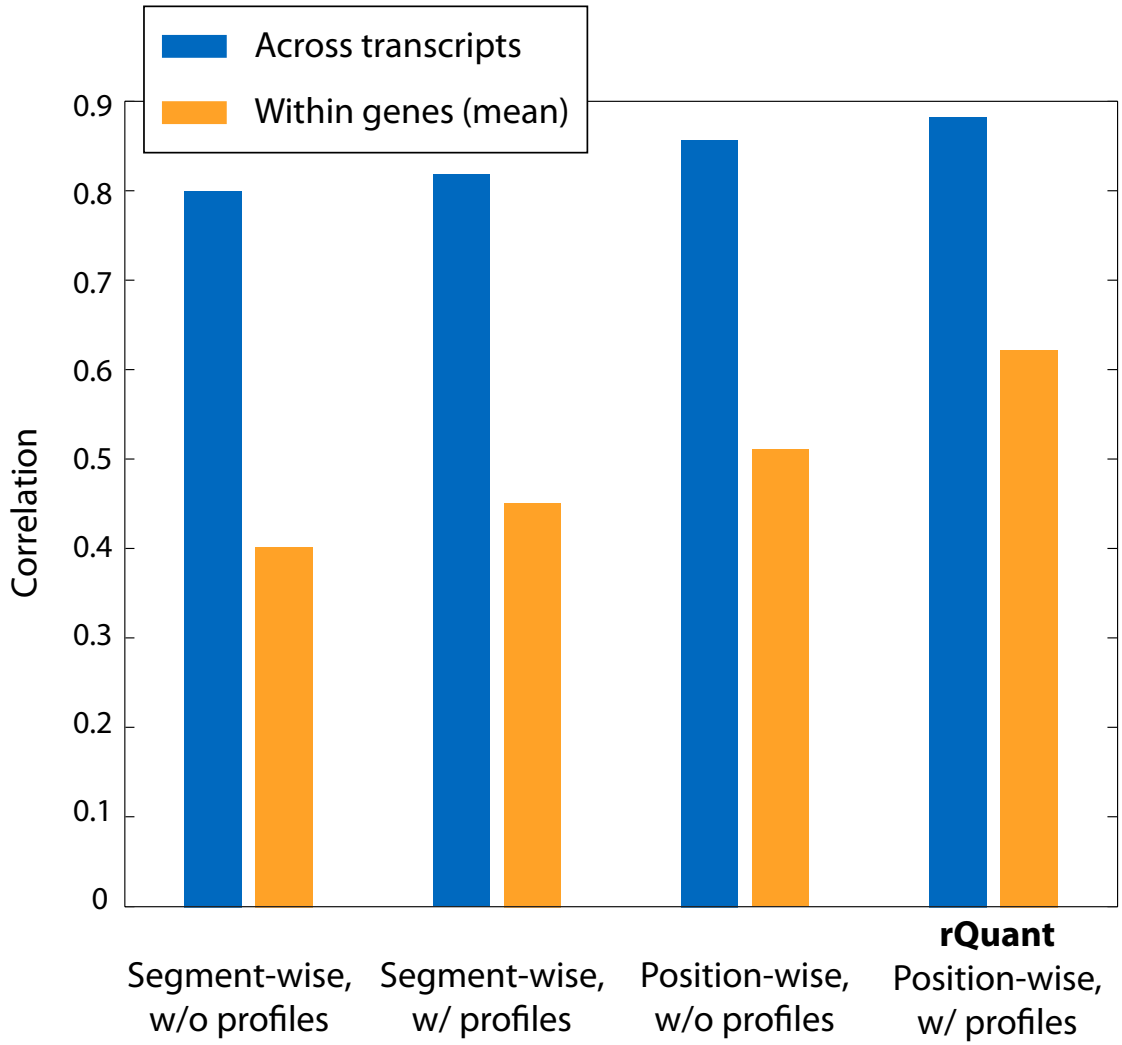
(Bohnert et al., in preparation, 2010)

rQuant Evaluation II



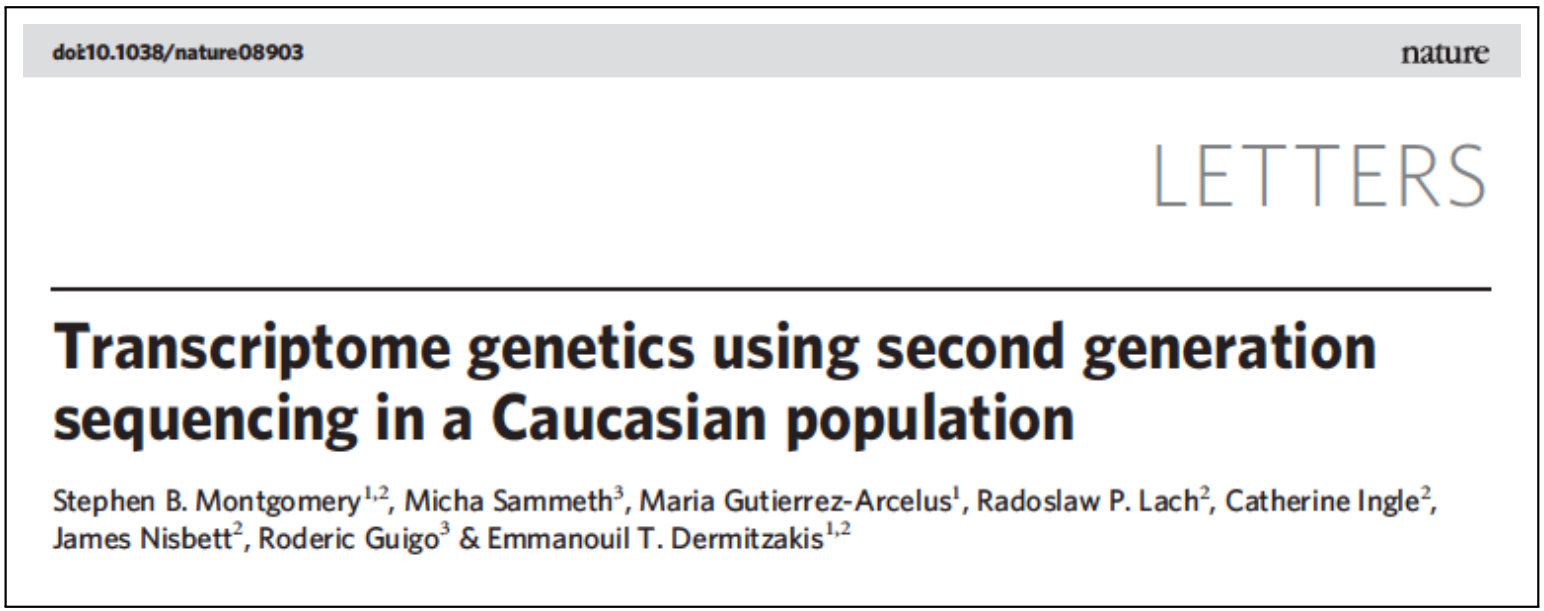
(Bohnert et al., in preparation, 2010)

rQuant Evaluation II



(Bohnert et al., in preparation, 2010)

Transcriptome Variation Studies



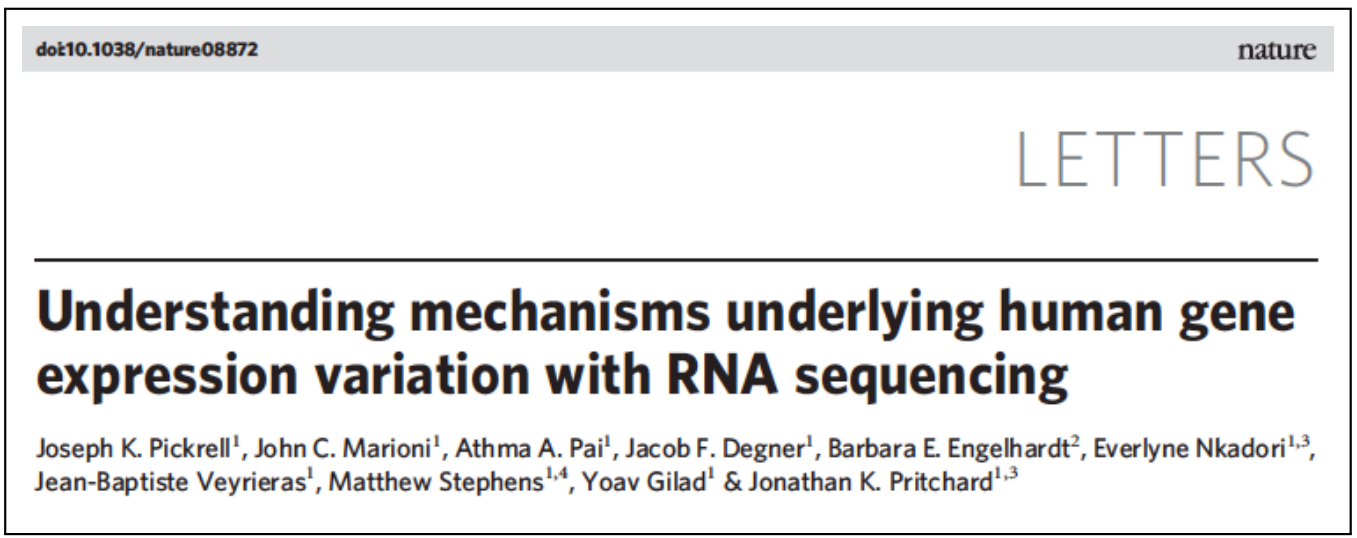
doi:10.1038/nature08903 nature

LETTERS

Transcriptome genetics using second generation sequencing in a Caucasian population

Stephen B. Montgomery[1,2], Micha Sammeth[3], Maria Gutierrez-Arcelus[1], Radoslaw P. Lach[2], Catherine Ingle[2], James Nisbett[2], Roderic Guigo[3] & Emmanouil T. Dermitzakis[1,2]

[Montgomery et al., 2010]



doi:10.1038/nature08872 nature

LETTERS

Understanding mechanisms underlying human gene expression variation with RNA sequencing

Joseph K. Pickrell[1], John C. Marioni[1], Athma A. Pai[1], Jacob F. Degner[1], Barbara E. Engelhardt[2], Everlyne Nkadori[1,3], Jean-Baptiste Veyrieras[1], Matthew Stephens[1,4], Yoav Gilad[1] & Jonathan K. Pritchard[1,3]

[Pickrell et al., 2010]

Testing for differential gene expression

Differential Gene Expression

Idea: Count reads mapping to a gene in two different samples
Problem: Which count differences are significant?

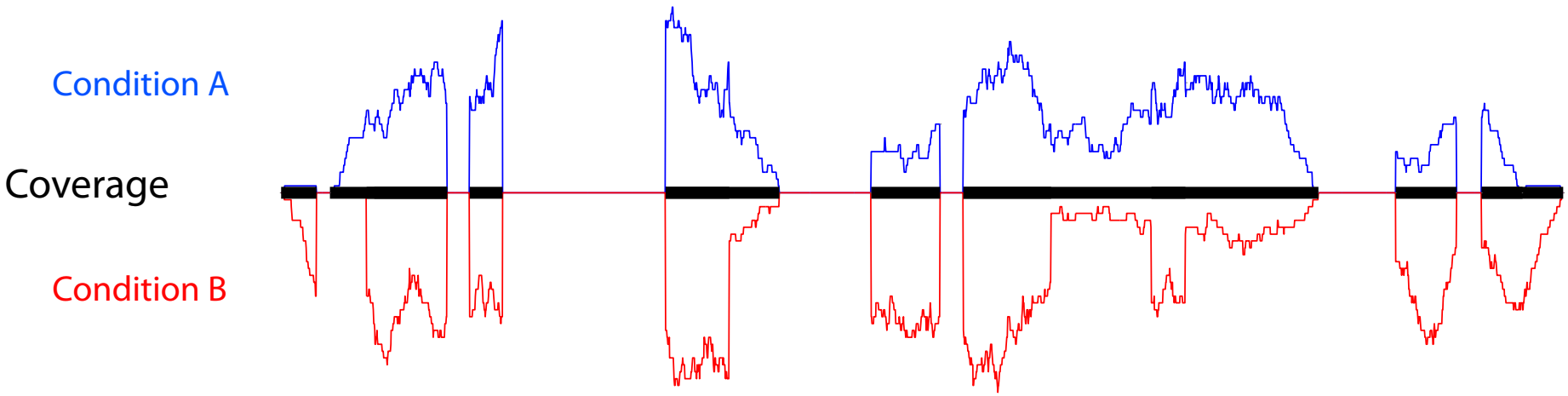
Test for differential gene expression between two conditions by

- Model distribution of reads from one gene by a gamma-mixture of Poisson distributions
- Variance is considered to be a function of the mean which is estimated by a local regression on all genes in all replicates
- *P*-values are obtained by using a standard conditional test

[Anders and Huber, 2010]

Differential Transcript Expression

How can we determine differential expression of alternative transcripts?

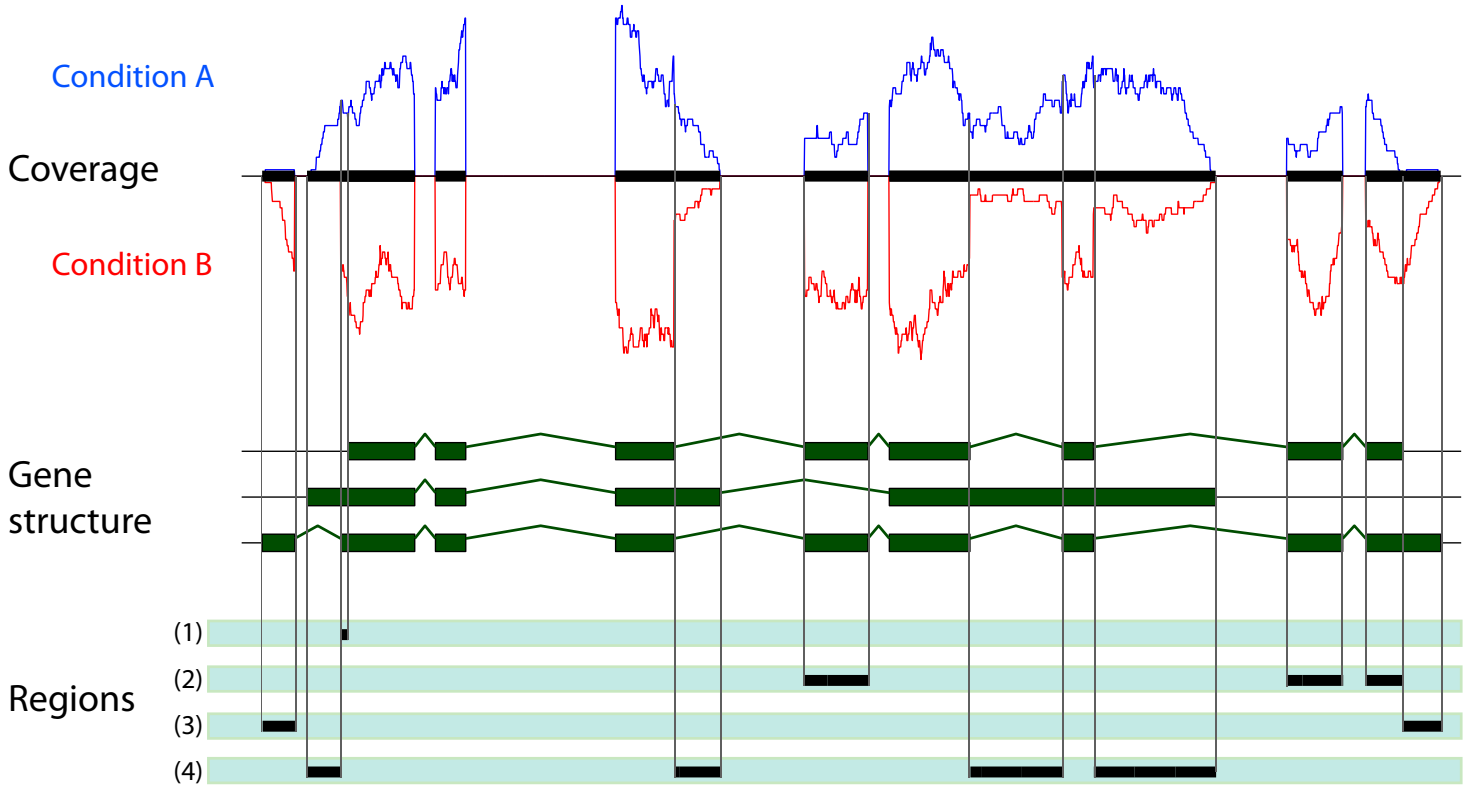


Differential Transcript Expression

Tests for differential transcript expression between two conditions:

- Fully known transcript annotation:
  - *Transcript quantitation* and subsequent testing
  - Statistical test on read counts in *differential regions* [Stegle et al., 2010]
- Incomplete or missing transcript annotation
  - Non-parametric *Maximum mean discrepancy* test on a genomic region [Stegle et al., 2010]

Poisson Test on Differential Regions

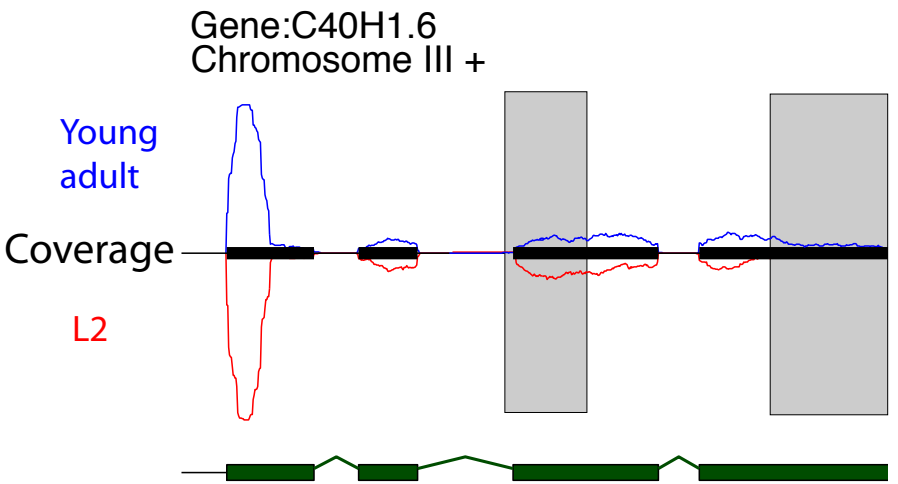


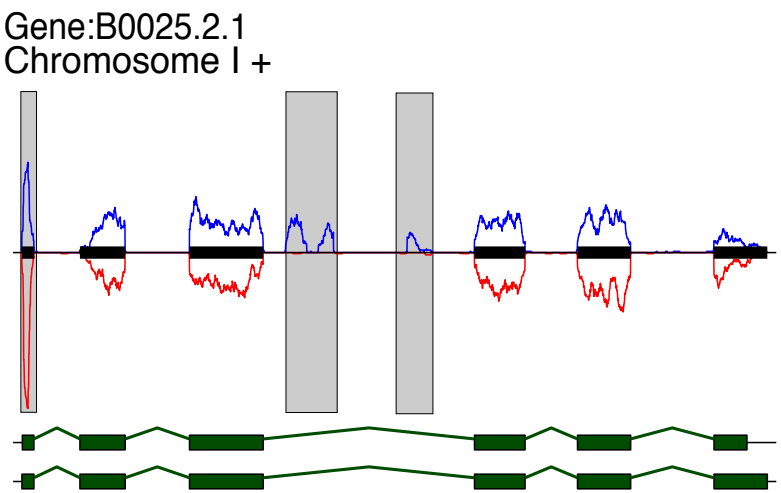
- Considers only regions where differential "expression" can occur
- Models read counts in those regions by Poisson distributions
- *P*-values are obtained by using a standard conditional test and Bonferroni-correction

Maximum Mean Discrepancy Test

When the set of transcripts is incomplete or not available, use the Maximum Mean Discrepancy (MMD) test [Stegle et al., 2010, Gretton et al., 2007]

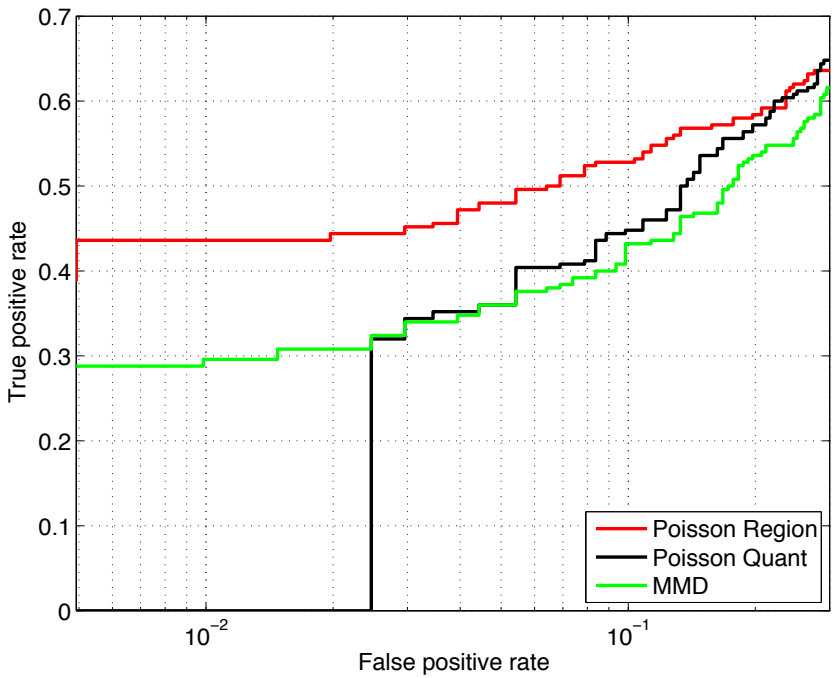
- Test whether the read distributions in a region are the same
- No assumptions on the distribution of reads in a region
- Useful to detect differential expression of yet unknown transcripts (e.g., alternative splicing, transcription starts, etc.)

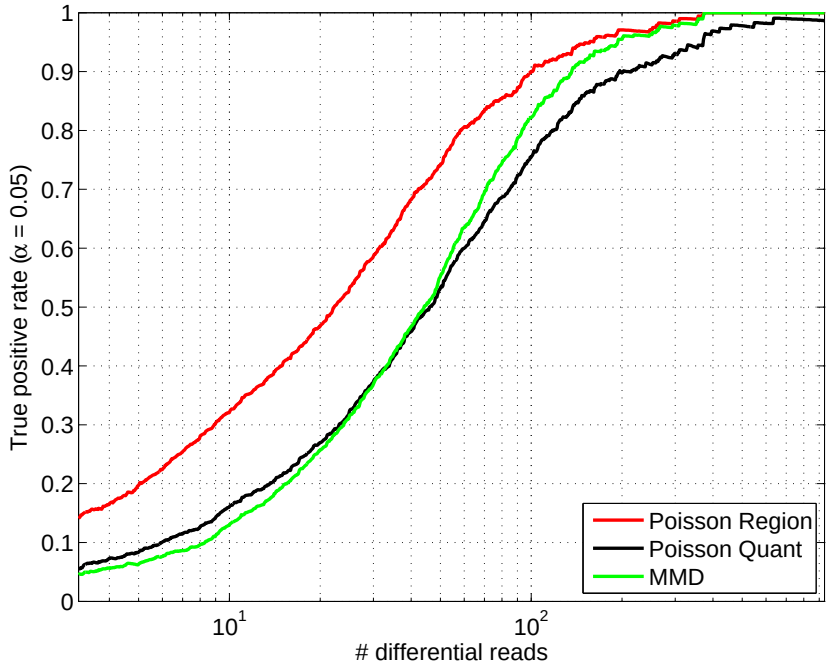




Comparison of Differential Expression Detection

Results on simulated expression and read data [Stegle et al., 2010]

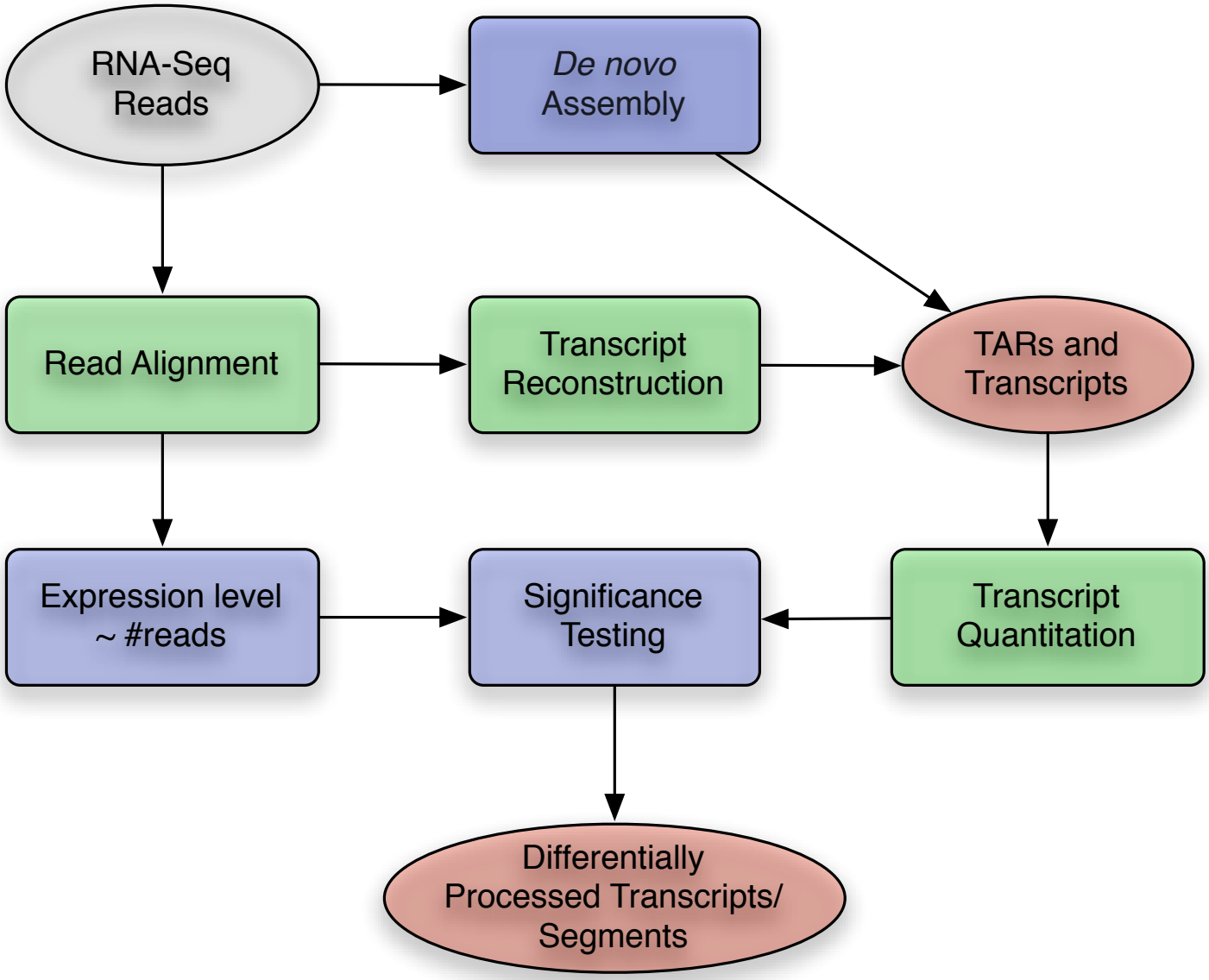




- *Poisson Region*: Poisson test on differential regions
- *Poisson Quant*: Quantitation with subsequent Poisson test
- *MMD*: Maximum Mean Discrepancy test

Transcript Reconstruction

Common RNA-Seq Analysis Steps



Transcript Identification Strategies

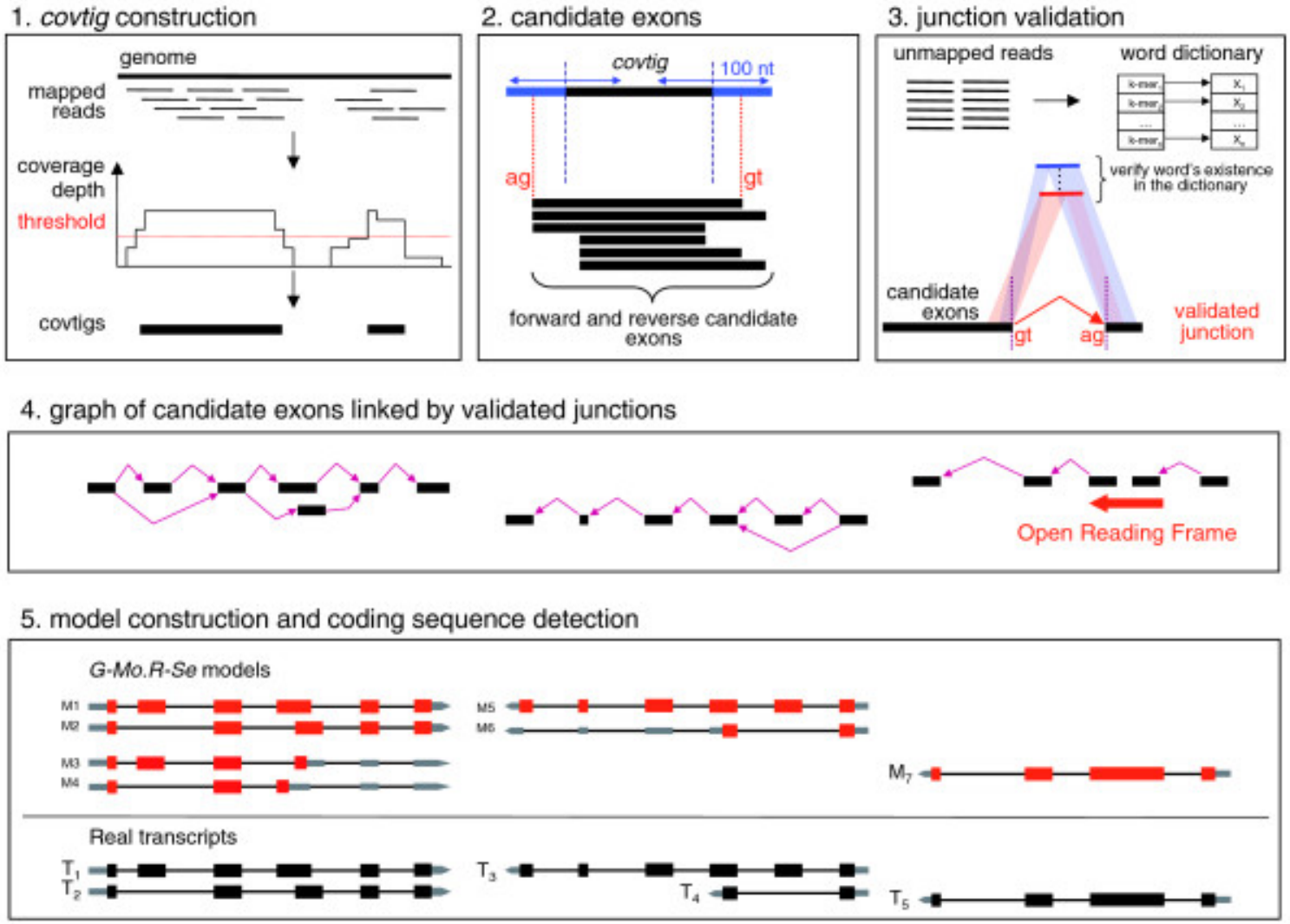


- Segmentation strategies: Based on read coverage only
  → Little assumptions
  → Find non-coding transcripts
  → Differential expression
- Trade-off read information with genomic sequence features
  → Find comprehensive set of genes
  → More accurate due to additional information

Methods:

- G.Mo.R-Se (threshold based exon identification)
- mTIM ("advanced thresholding")
- mGene.NGS (gene finding with RNA-seq)
- Cufflinks, Scripture (genome-guided assembly)

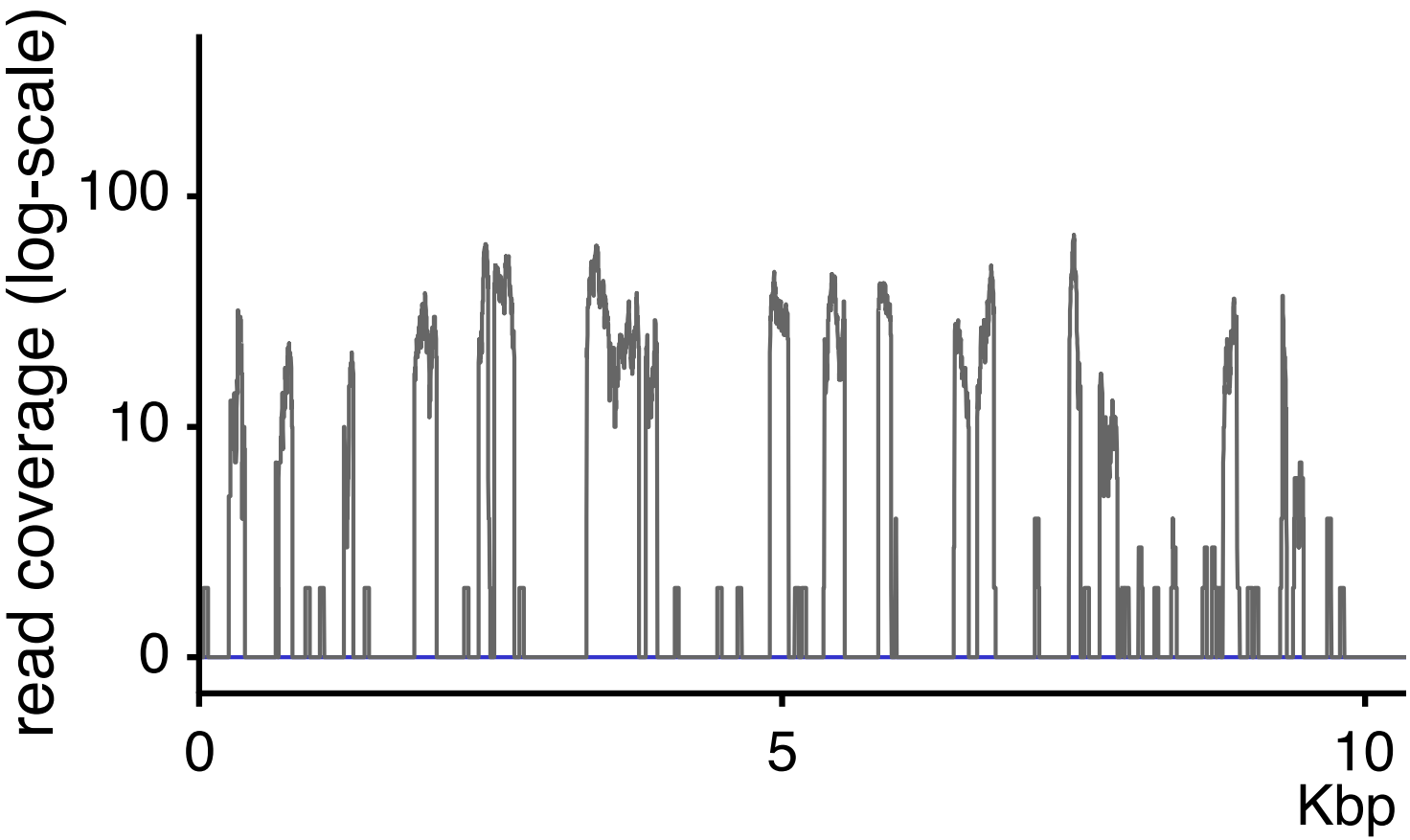
G.Mo.R-Se Splice Graph Reconstruction



Denoeud F. et al. Annotating genomes with massive-scale RNA sequencing. Genome Biology(2008) 9:R175

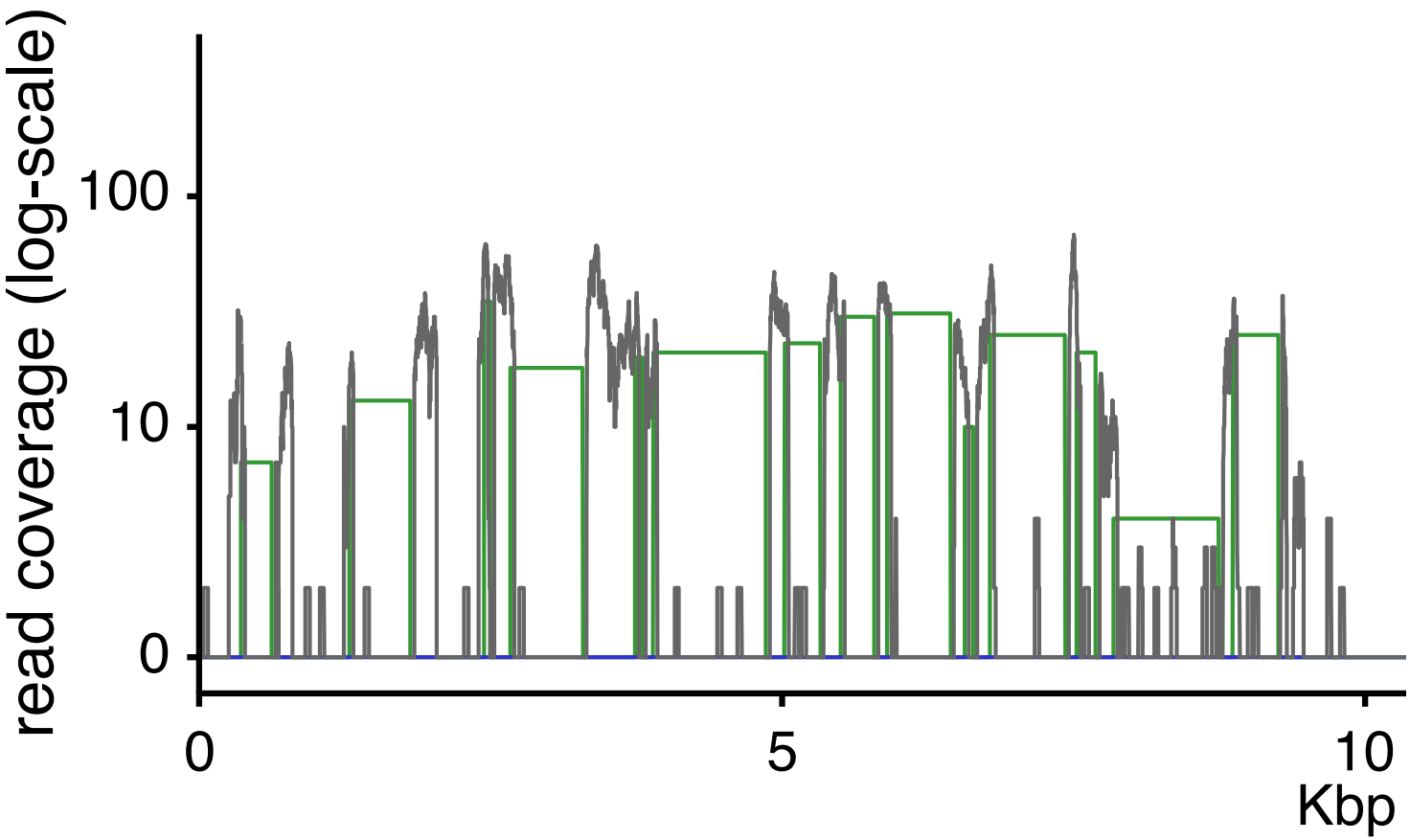
mTiM: Read Coverage Segmentation

Goal: Characterize each base as *intergenic*, *exonic*, or *intronic*



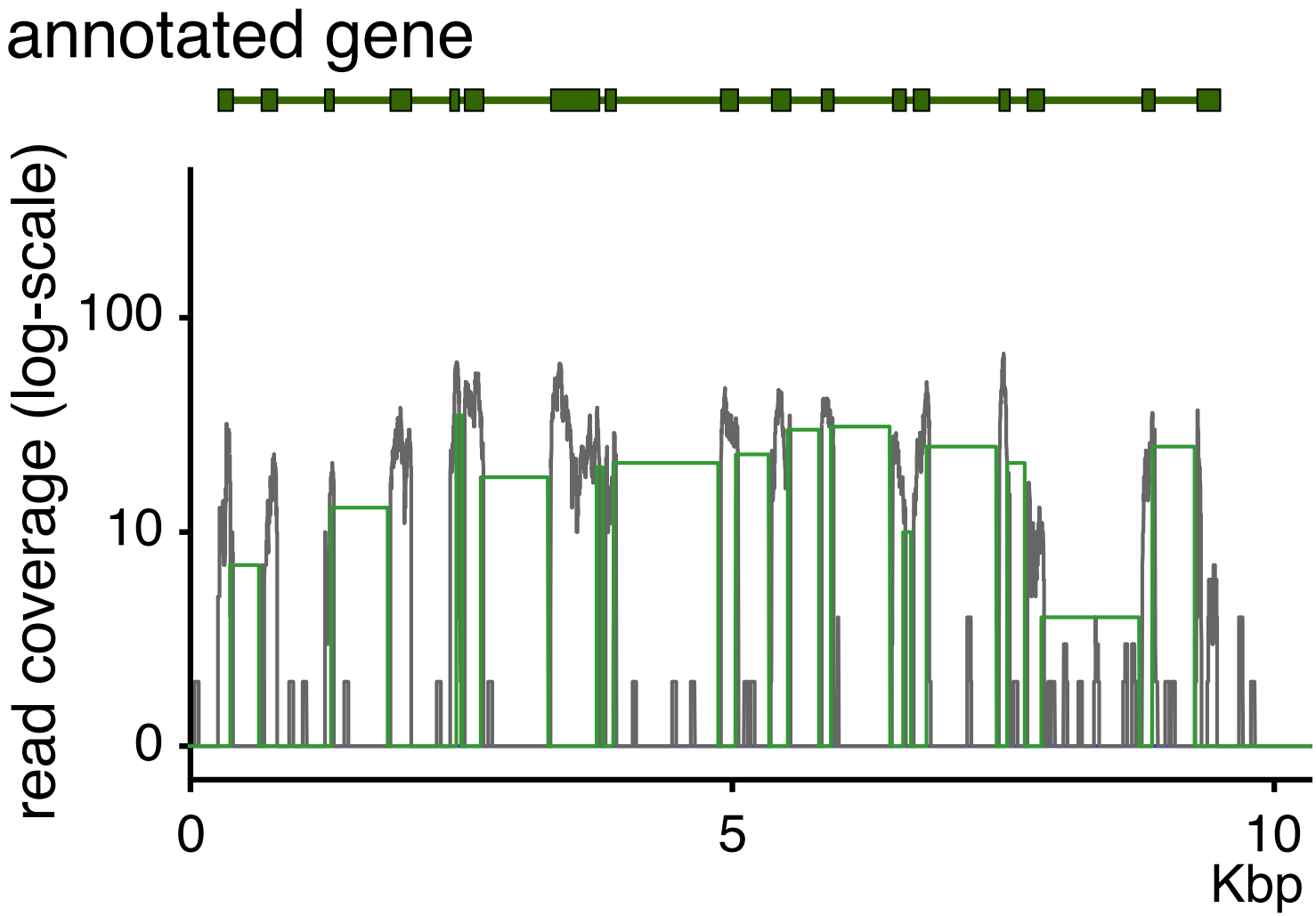
mTiM: Read Coverage Segmentation

Goal: Characterize each base as *intergenic*, *exonic*, or *intronic*

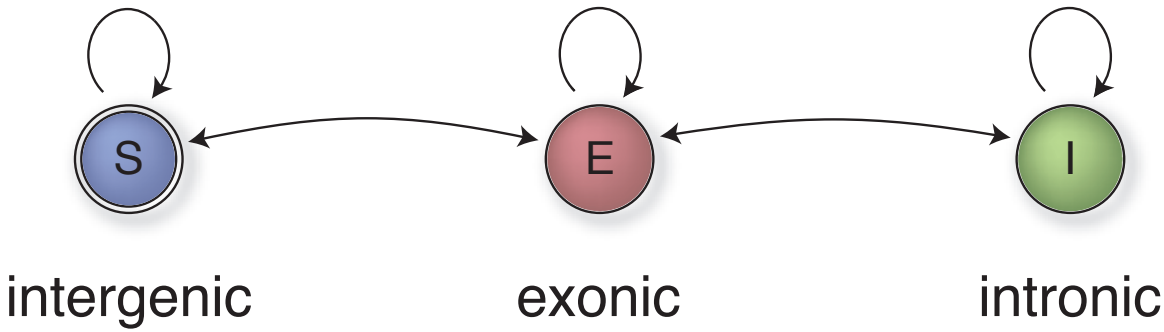


mTiM: Read Coverage Segmentation

Goal: Characterize each base as *intergenic*, *exonic*, or *intronic*



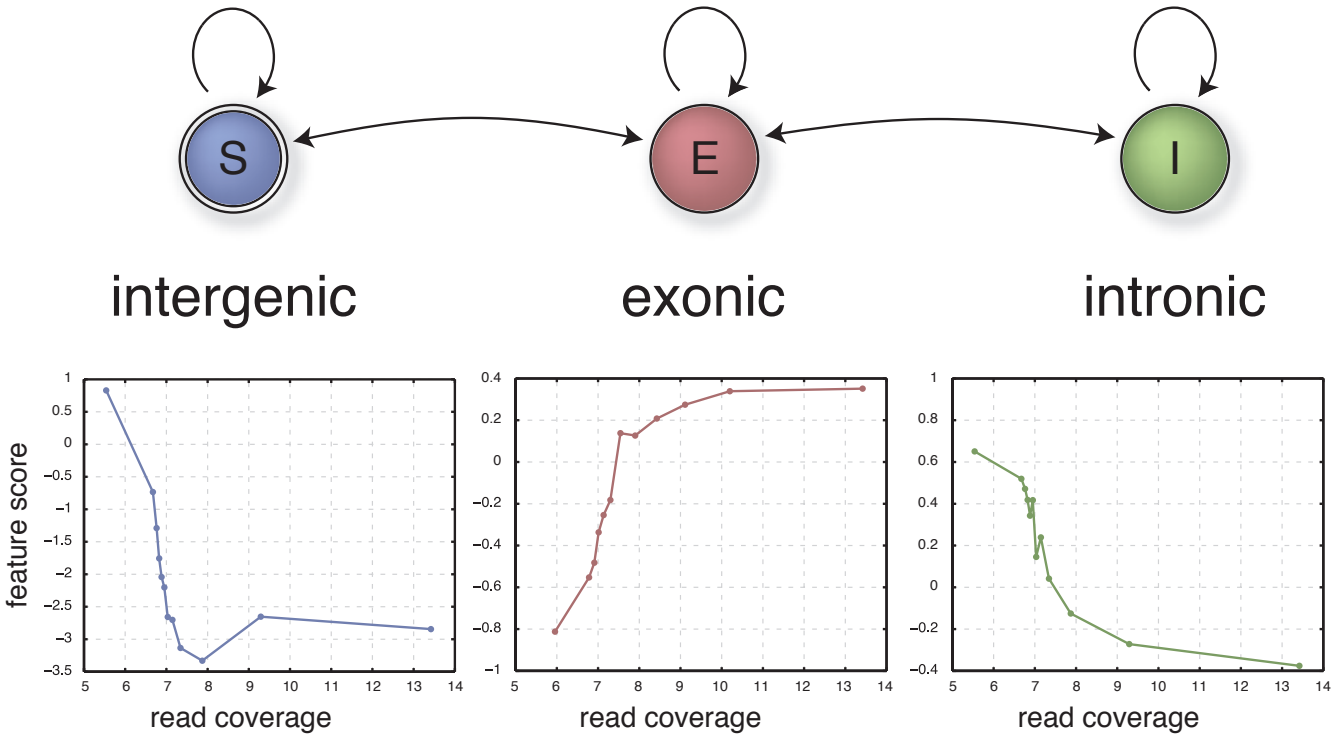
The mTiM Segmentation Approach



- Learn to associate a state with each position given its read coverage and local context
- HM-SVM training: Optimize transformations: signal → score
- Extension: Score spliced reads and splice sites

(G. Zeller et al., Pac. Symp. Biocomp., 2008; G. Zeller et al., unpublished, 2010)

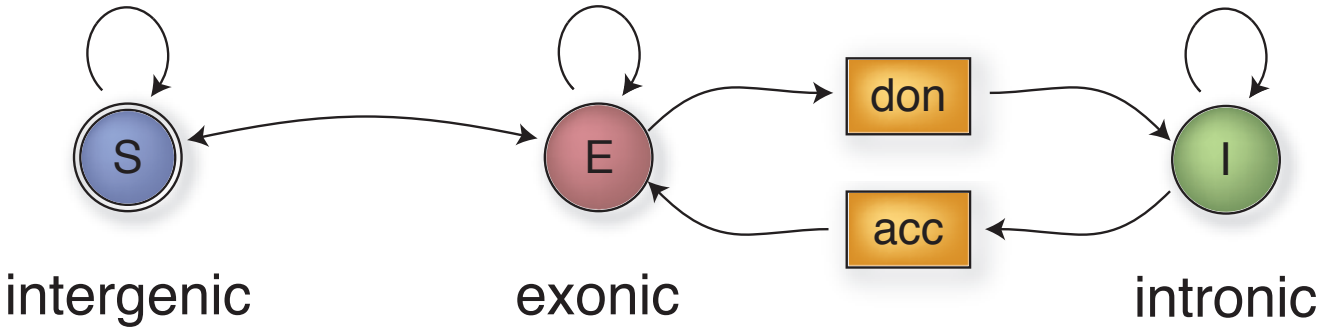
The mTiM Segmentation Approach

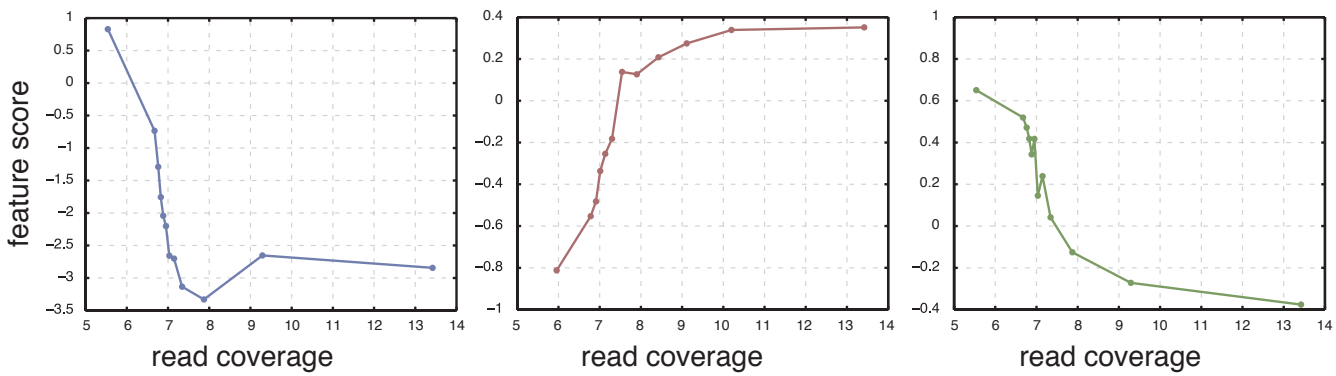


- Learn to associate a state with each position given its read coverage and local context
- HM-SVM training: Optimize transformations: signal $\longrightarrow$ score
- Extension: Score spliced reads and splice sites

(G. Zeller et al., Pac. Symp. Biocomp., 2008; G. Zeller et al., unpublished, 2010)

The mTiM Segmentation Approach



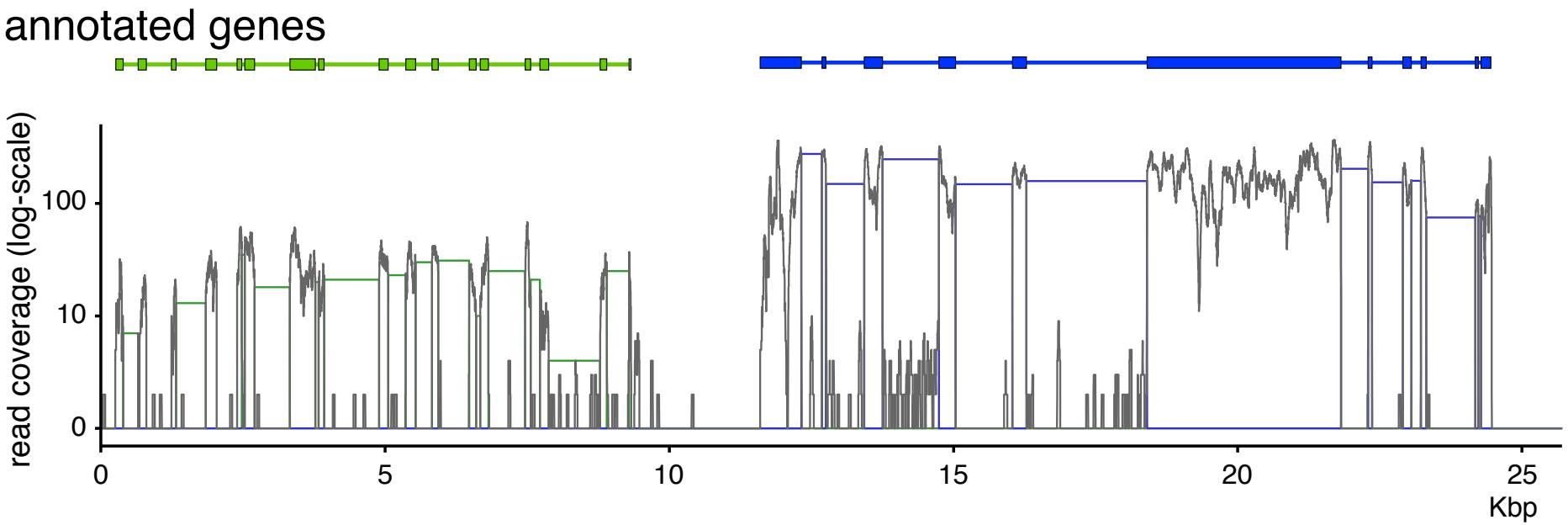


- Learn to associate a state with each position given its read coverage and local context
- HM-SVM training: Optimize transformations: signal $\longrightarrow$ score
- Extension: Score spliced reads and splice sites

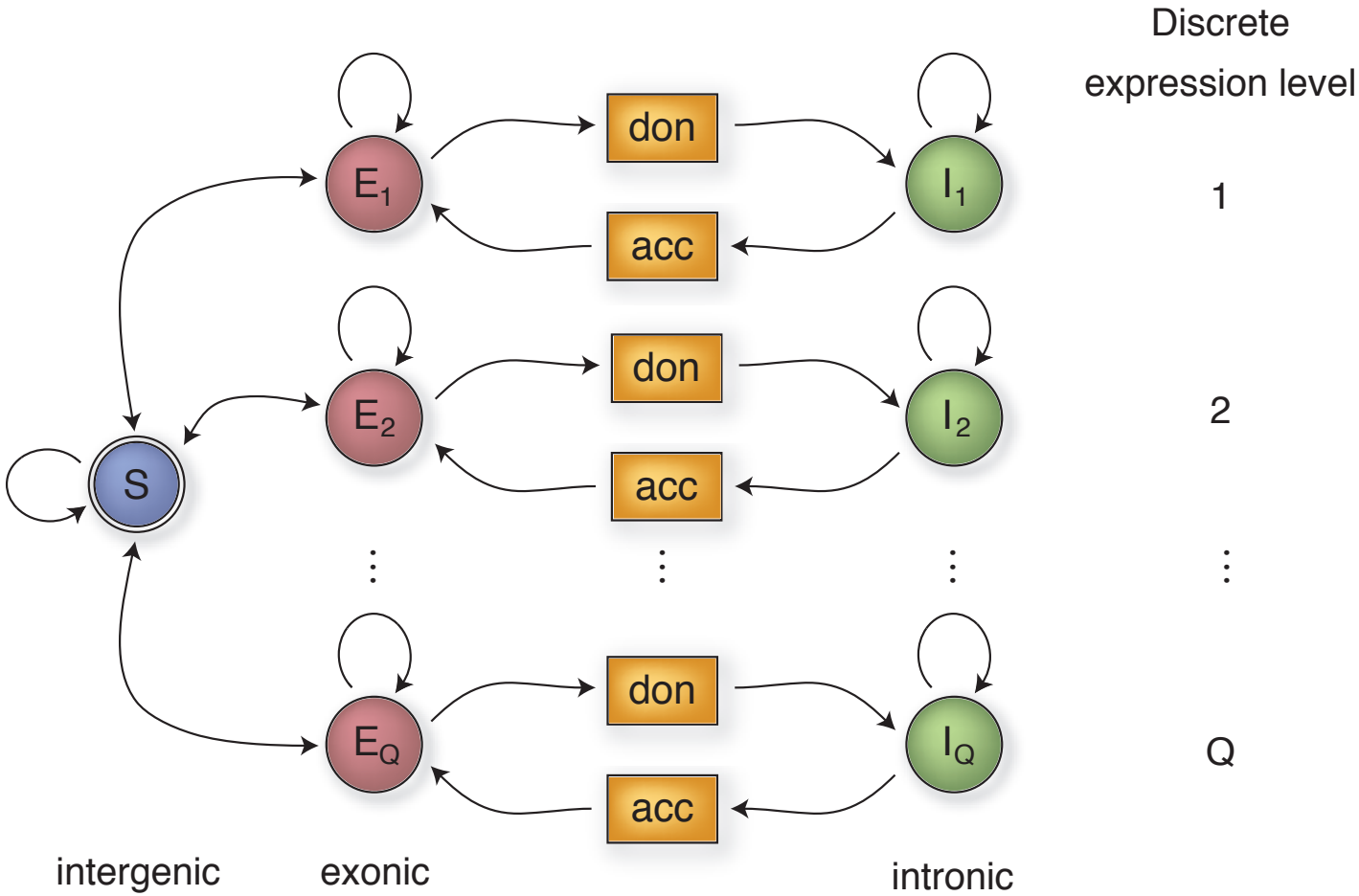
(G. Zeller et al., Pac. Symp. Biocomp., 2008; G. Zeller et al., unpublished, 2010)

The mTiM Segmentation Approach

Idea: Assume uniform read coverage within exons of same transcript



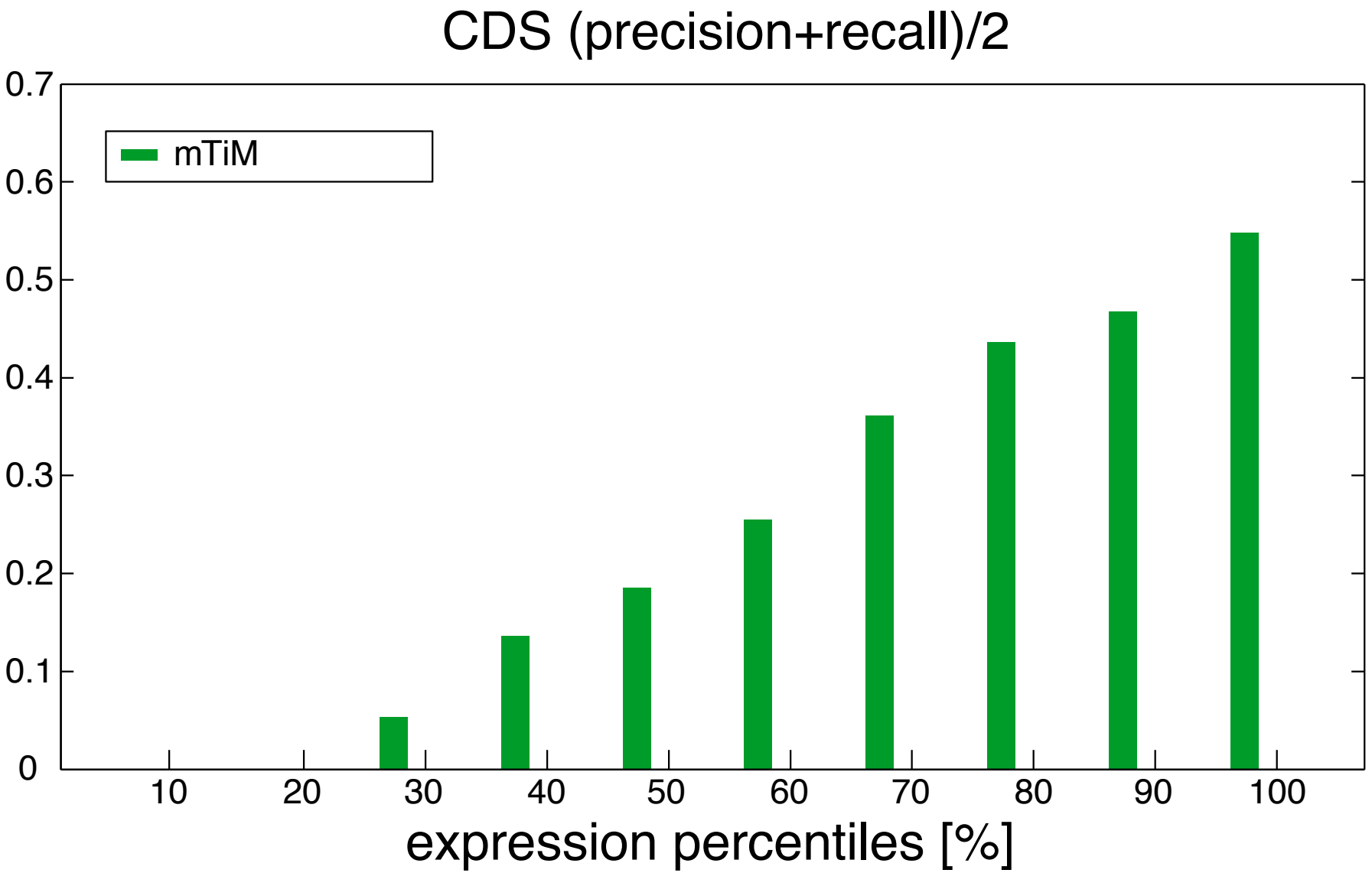
The mTiM Segmentation Approach



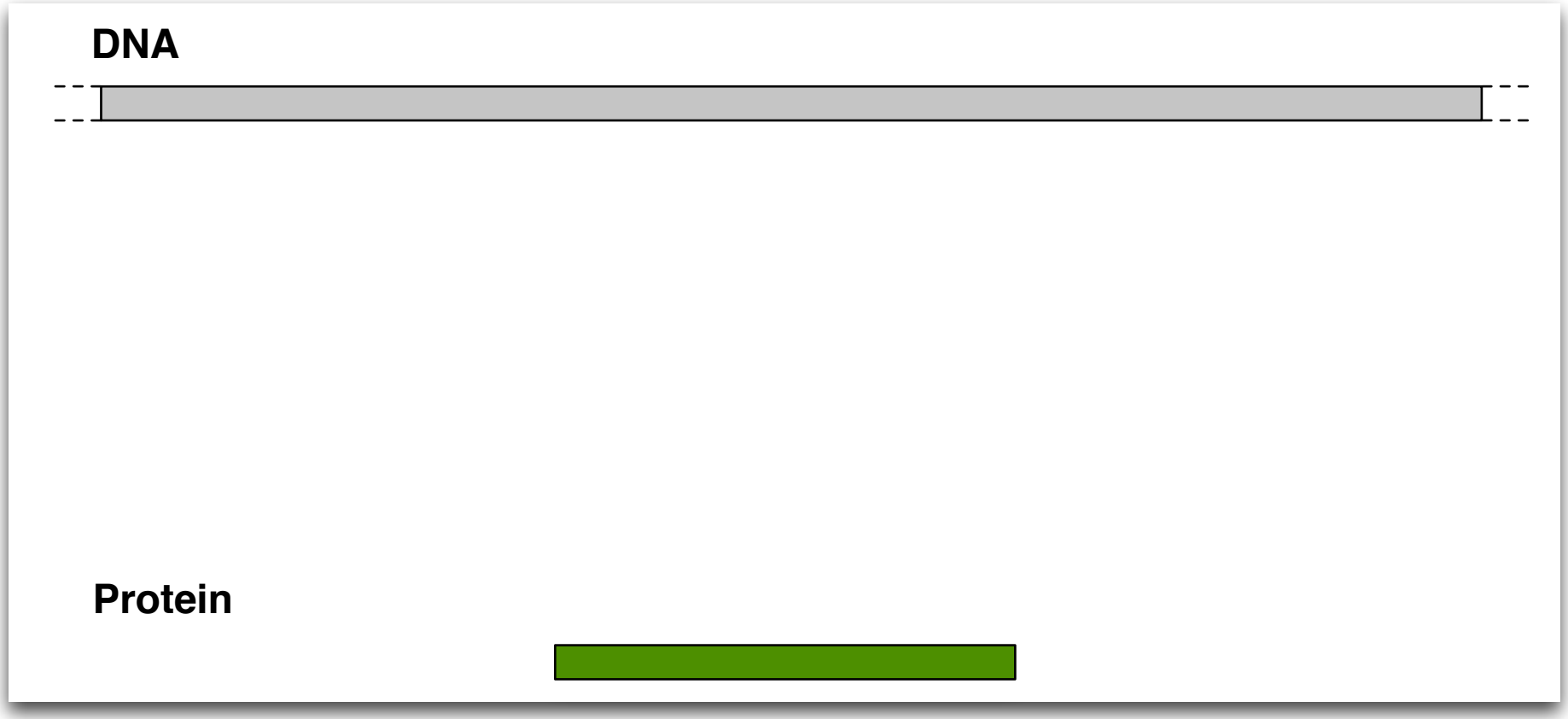
Carry "expression level" information between exons of same transcript

(G. Zeller et al., Pac. Symp. Biocomp., 2008; G. Zeller et al., unpublished, 2010)

Preliminary Evaluation (*C. elegans*)

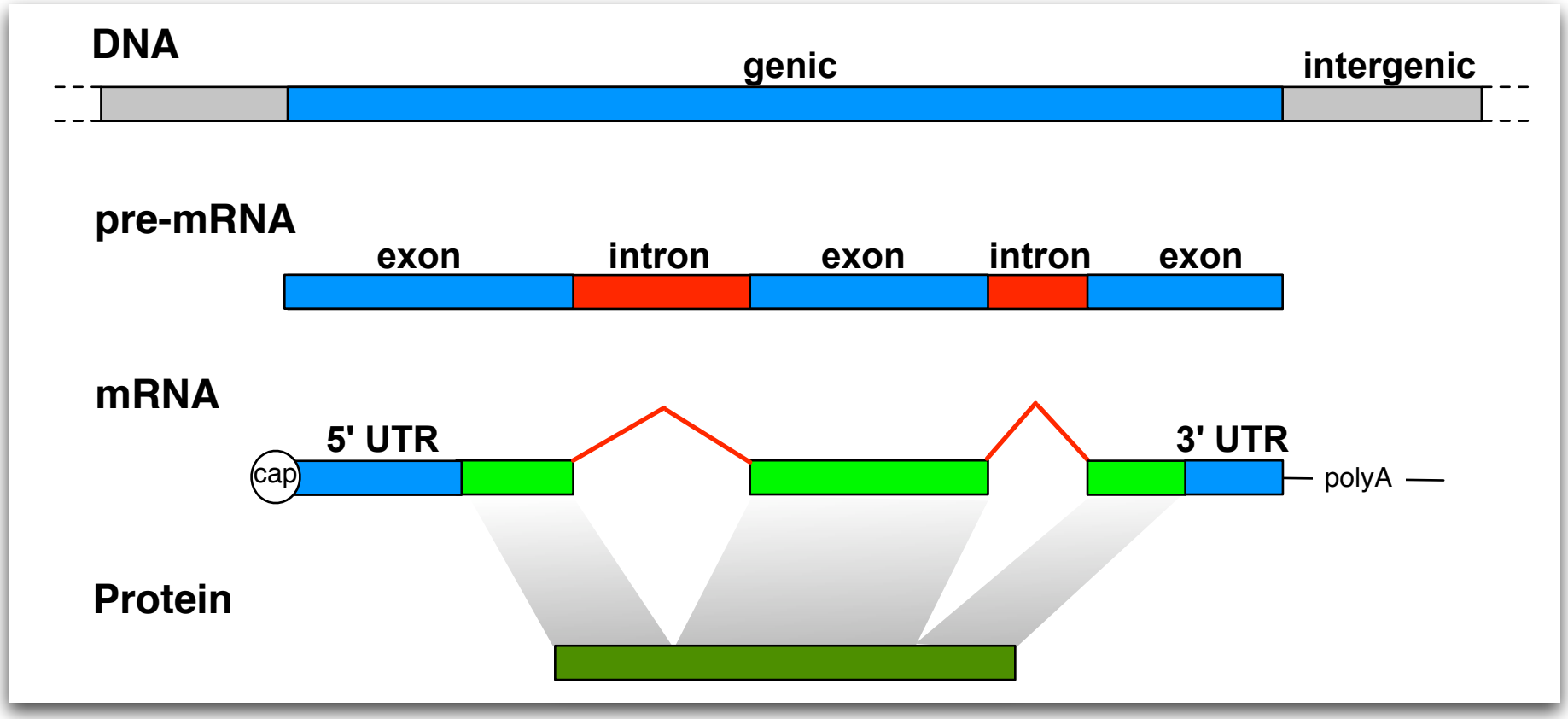


Gene Finding



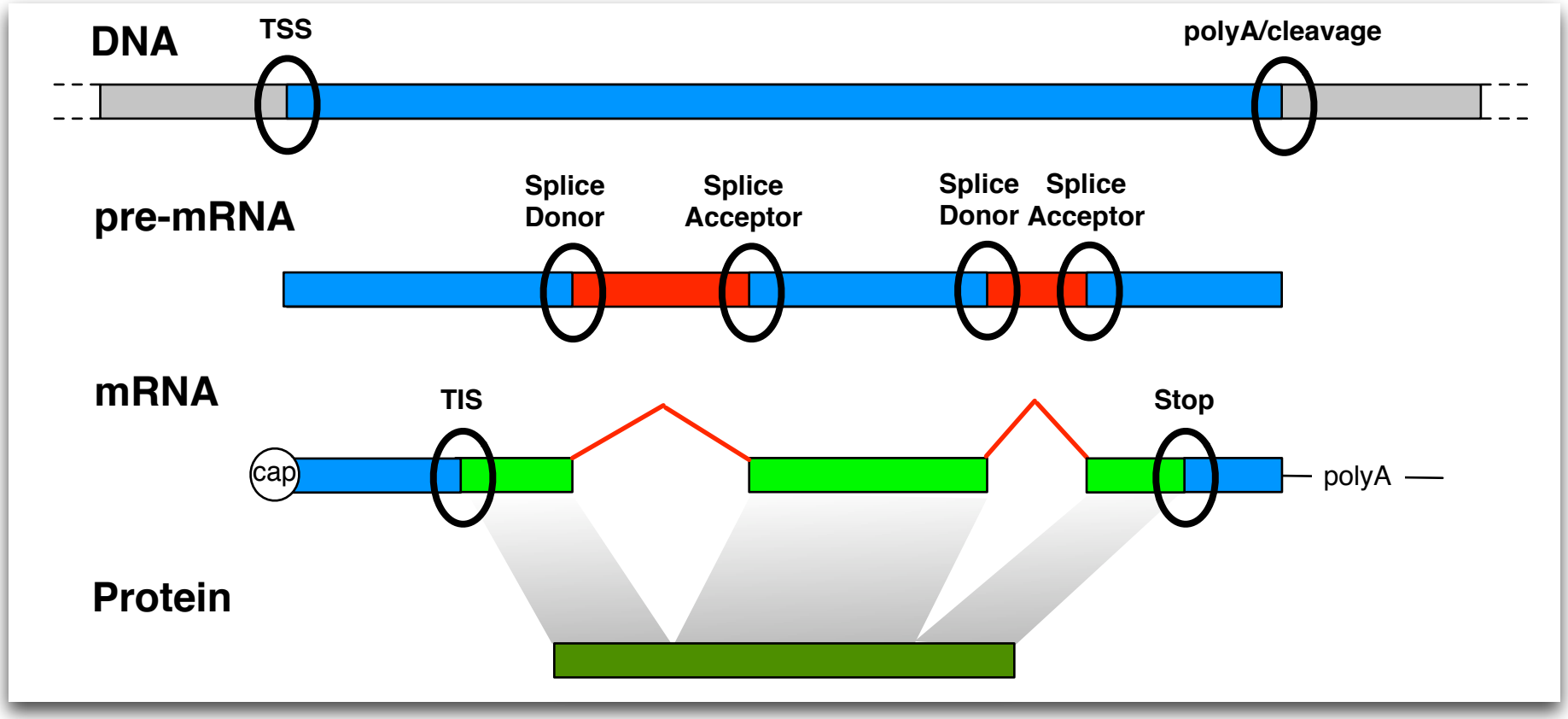
- Given a piece of DNA sequence
- Predict proteins (or non-coding RNAs)

Gene Finding

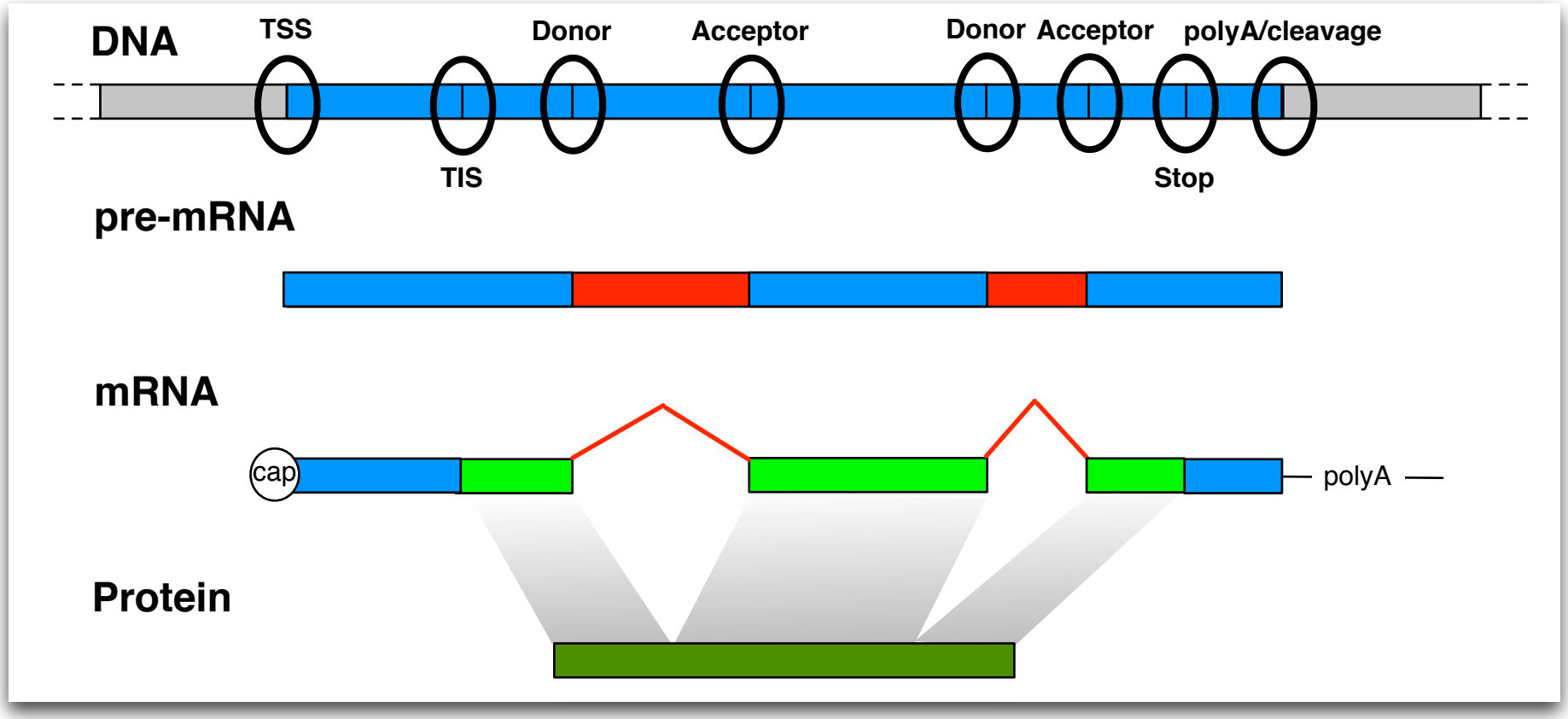


- Given a piece of DNA sequence
- Predict the correct corresponding **label sequence** with labels "intergenic", "exon", "intron", "5' UTR", etc.

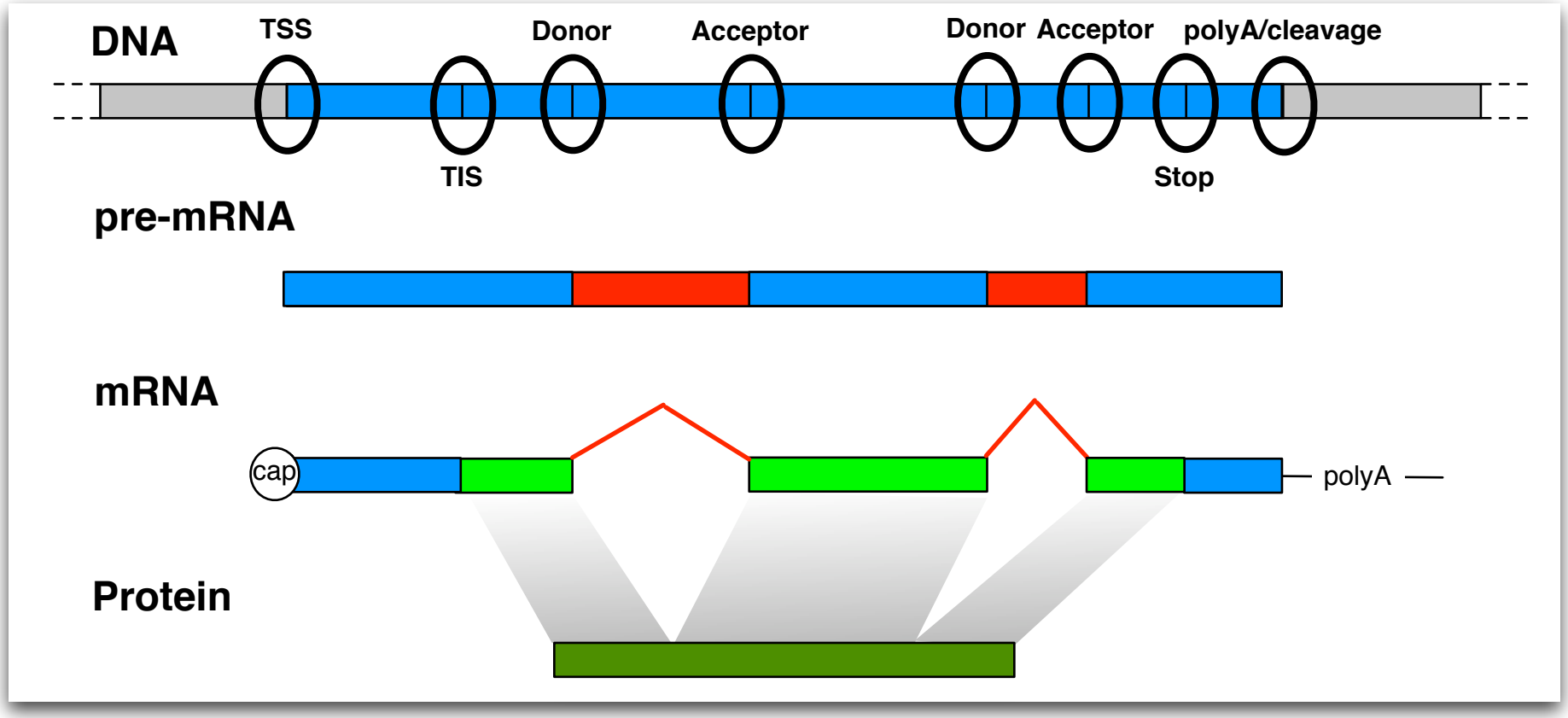
Computational Gene Finding

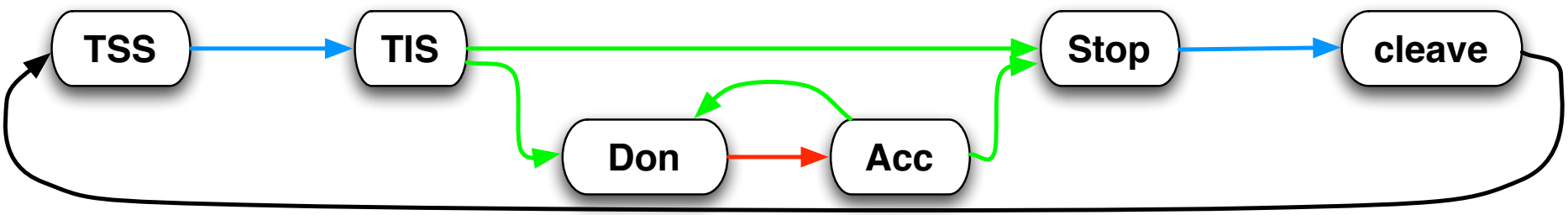


Computational Gene Finding

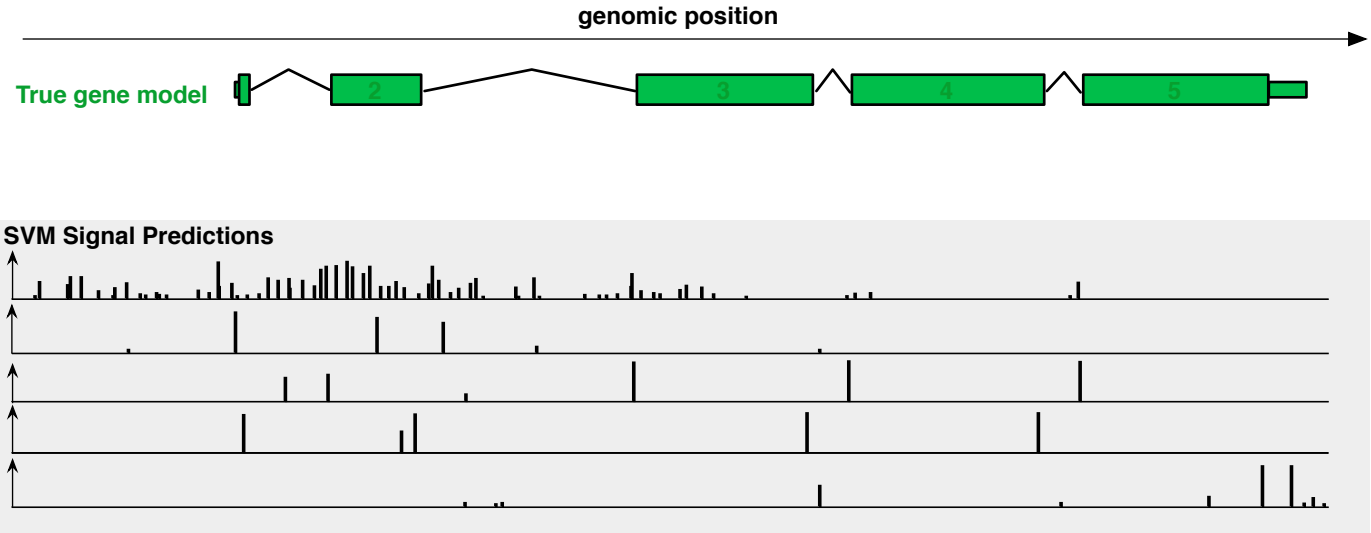


Computational Gene Finding

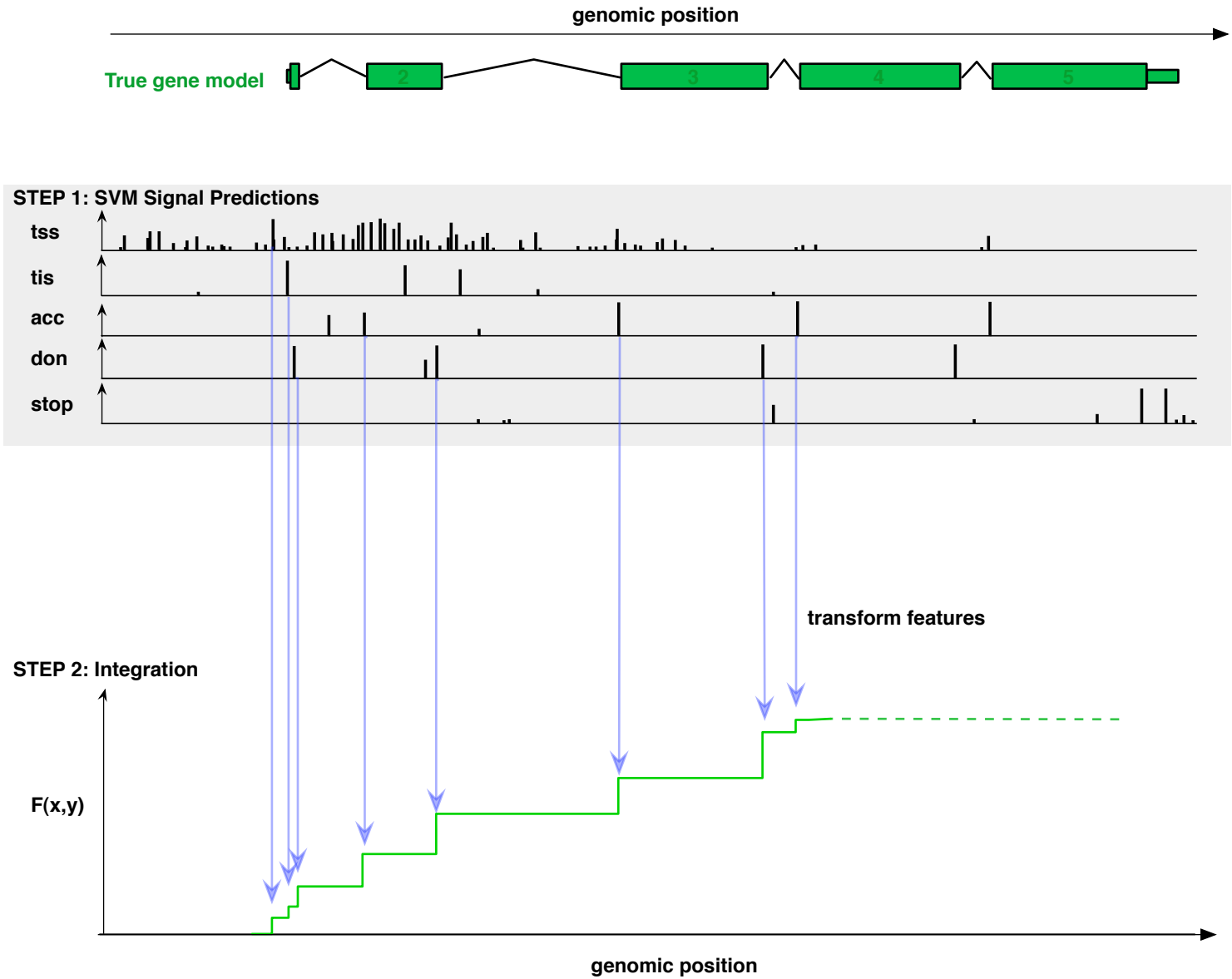




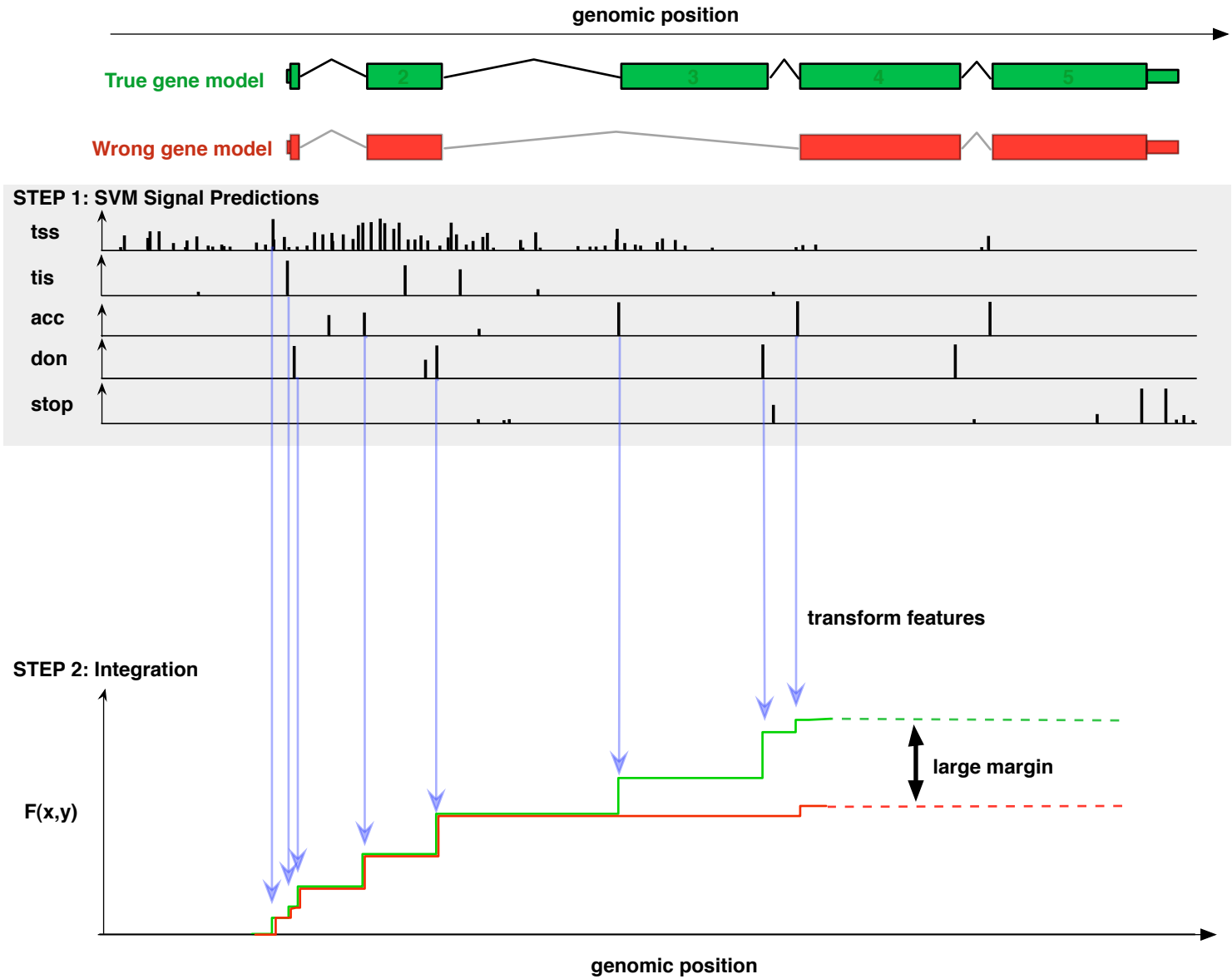
mGene-based Transcript Prediction I



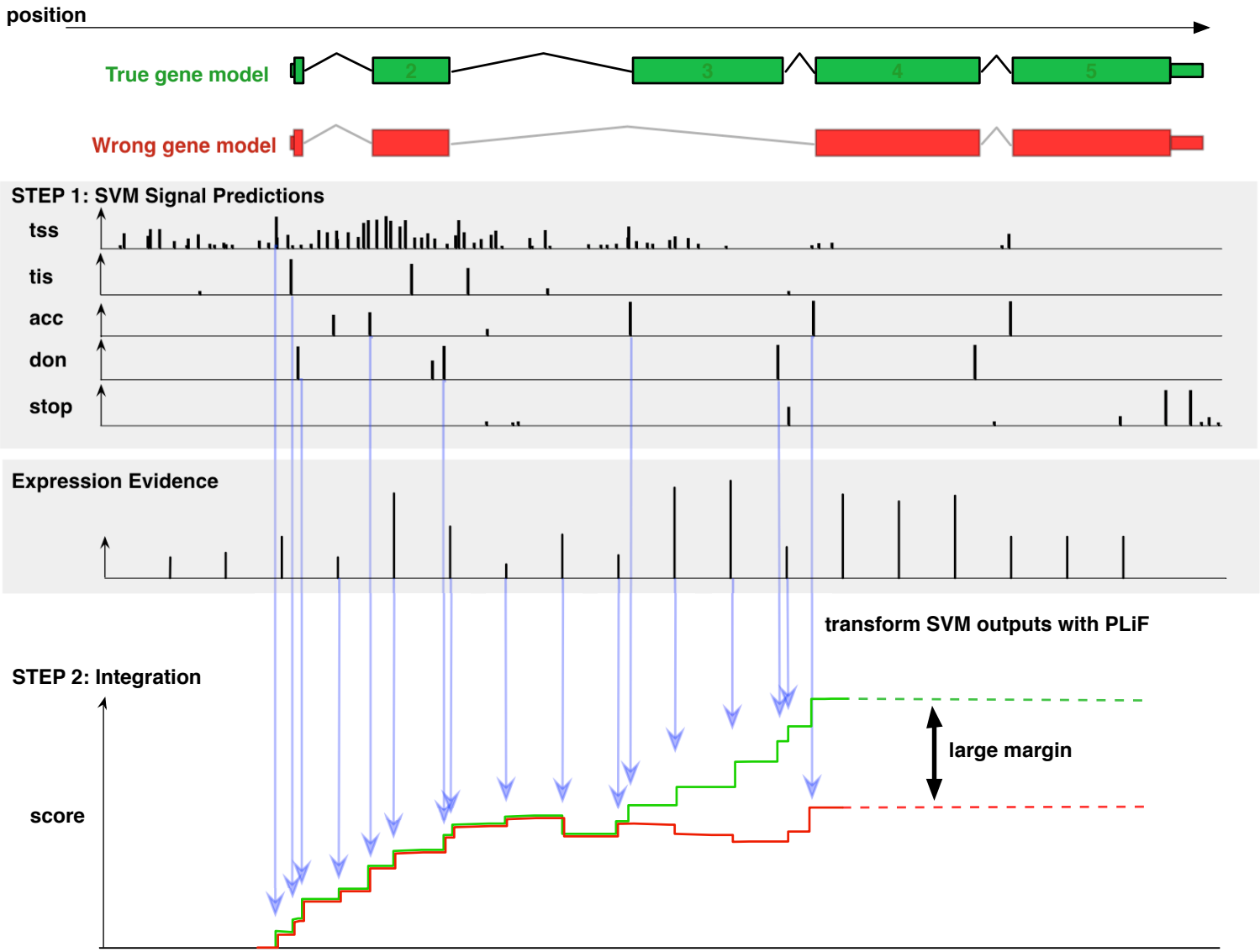
mGene-based Transcript Prediction I



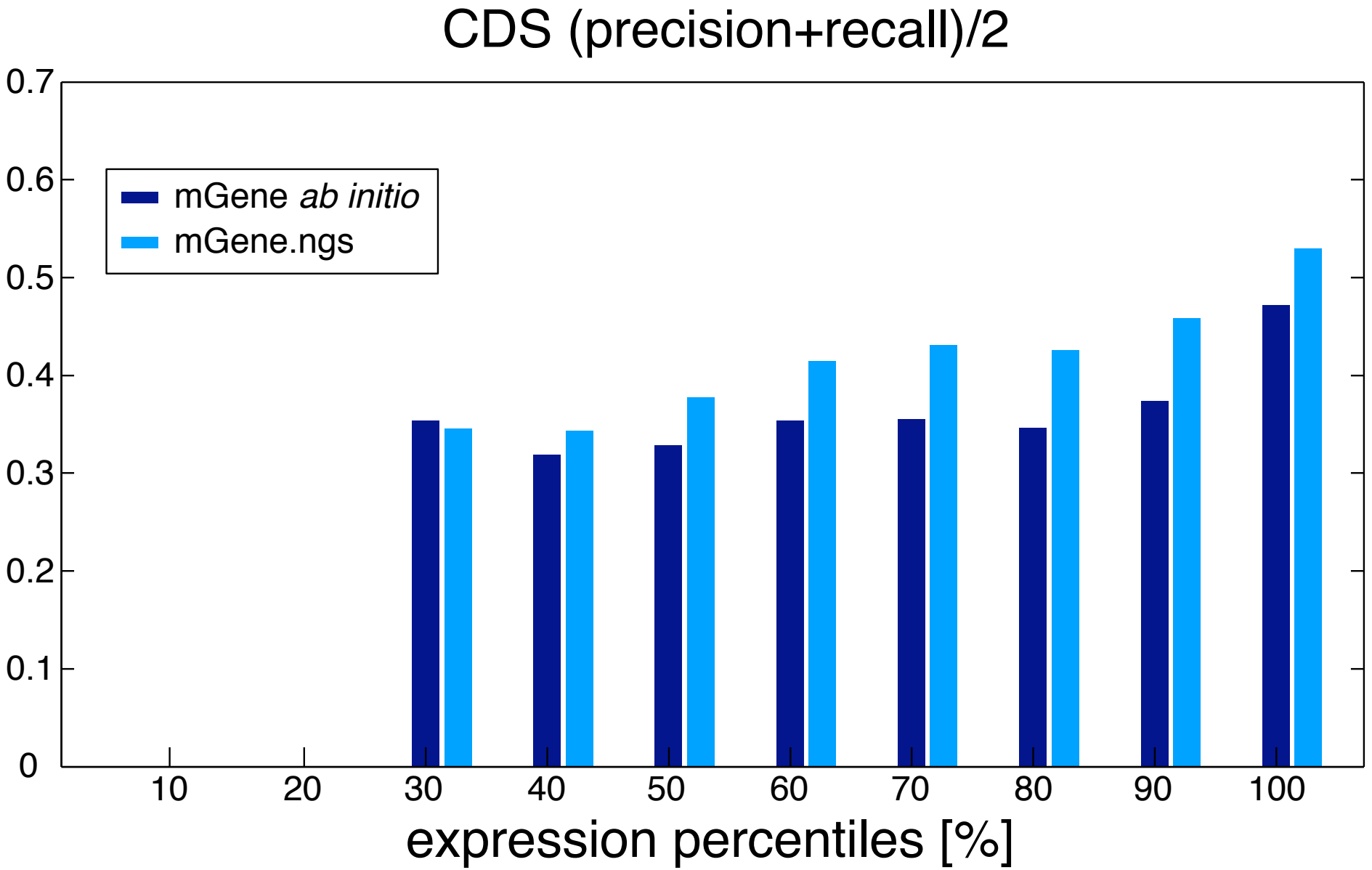
mGene-based Transcript Prediction I



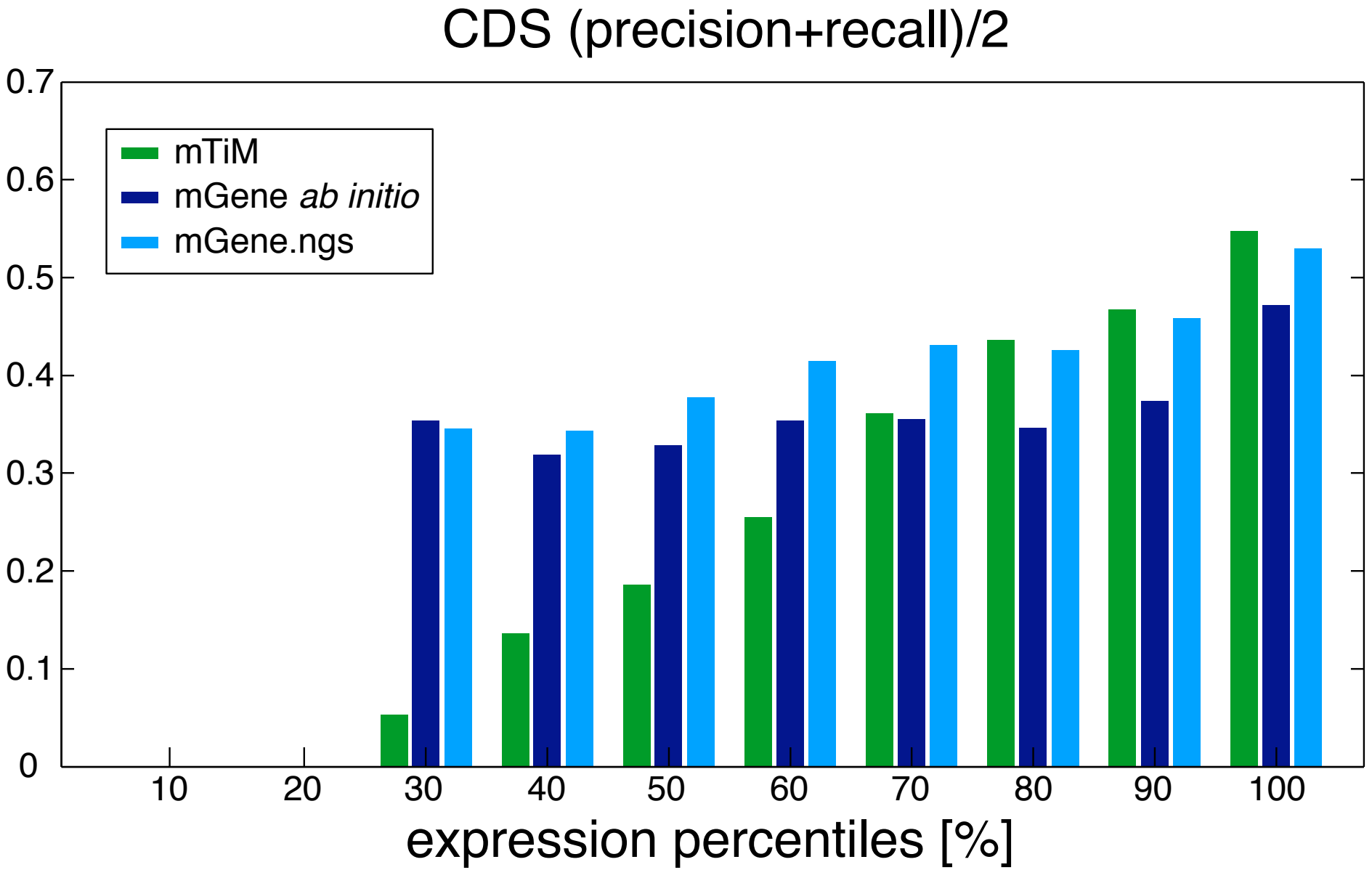
mGene-based Transcript Prediction I



Preliminary Evaluation (*C. elegans*)



Preliminary Evaluation (*C. elegans*)



Digestion



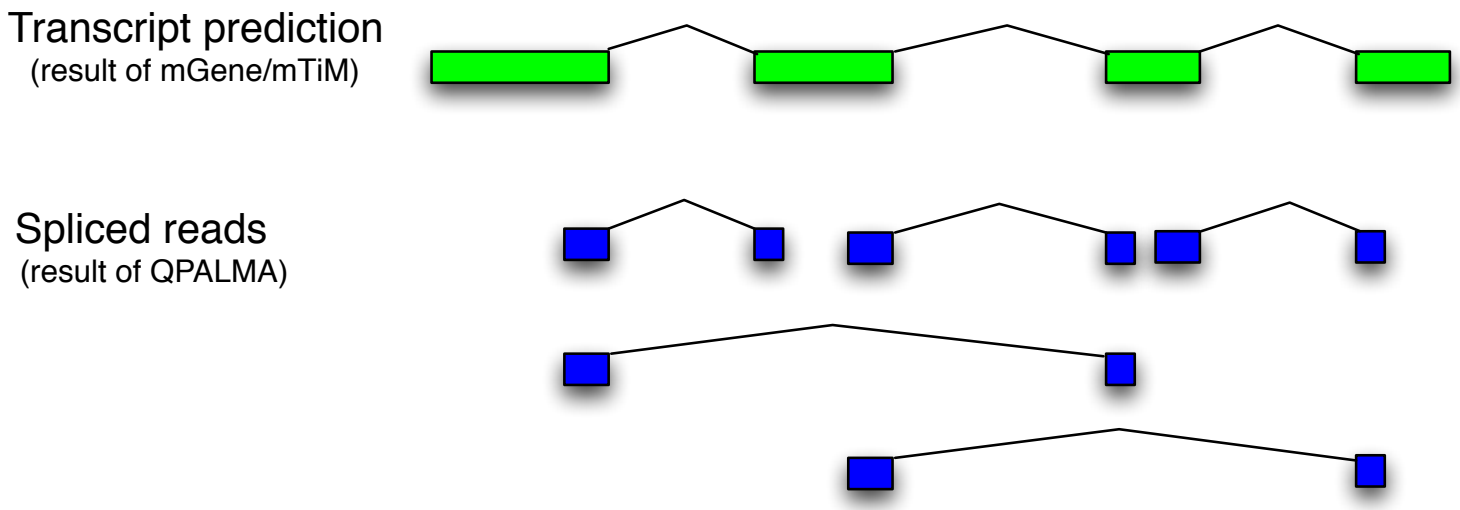
- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:
- Paths through splicing graph define *alternative transcripts*

Digestion

- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:
- Paths through splicing graph define *alternative transcripts*

Digestion

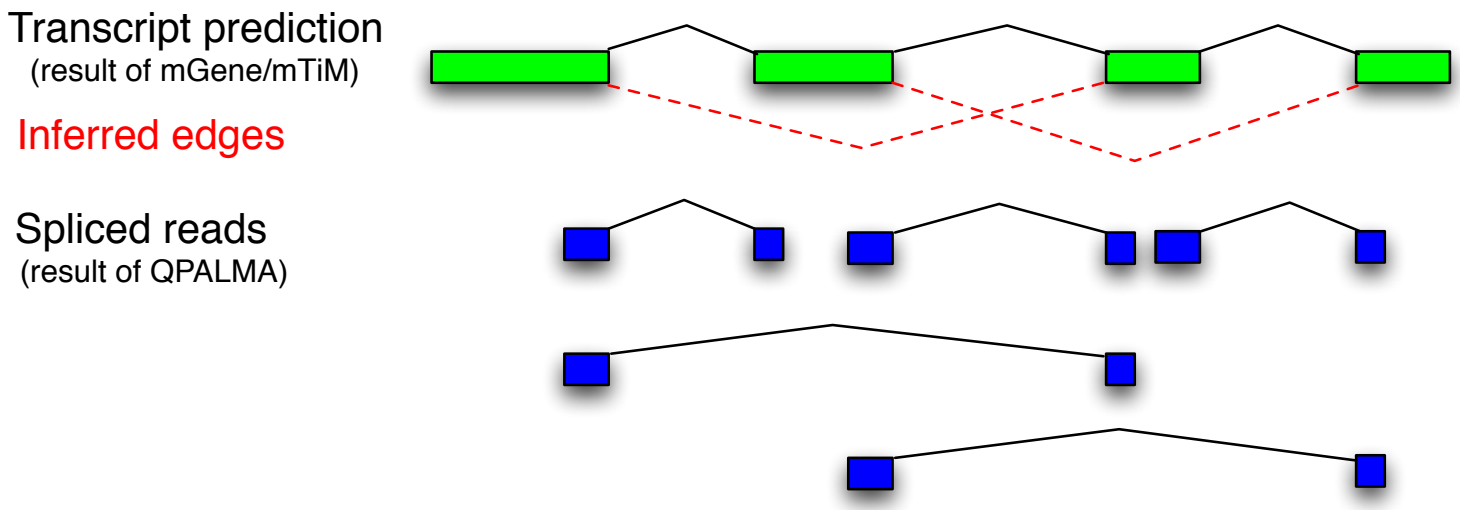
- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:



- Paths through splicing graph define *alternative transcripts*

Digestion

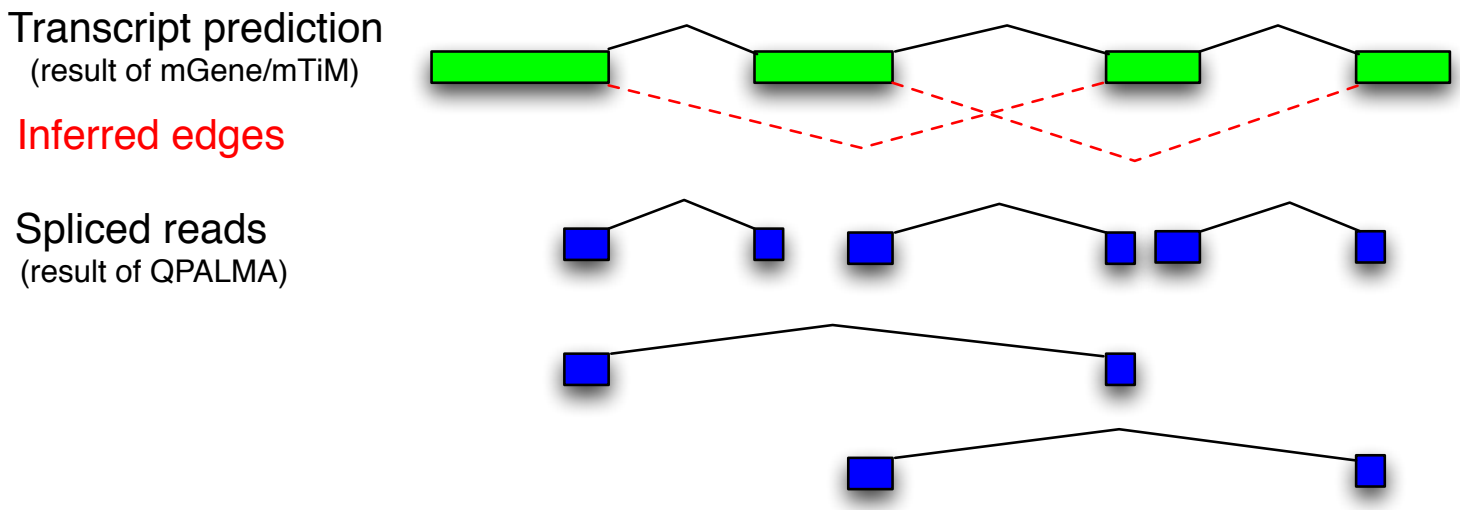
- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:



- Paths through splicing graph define *alternative transcripts*

Digestion

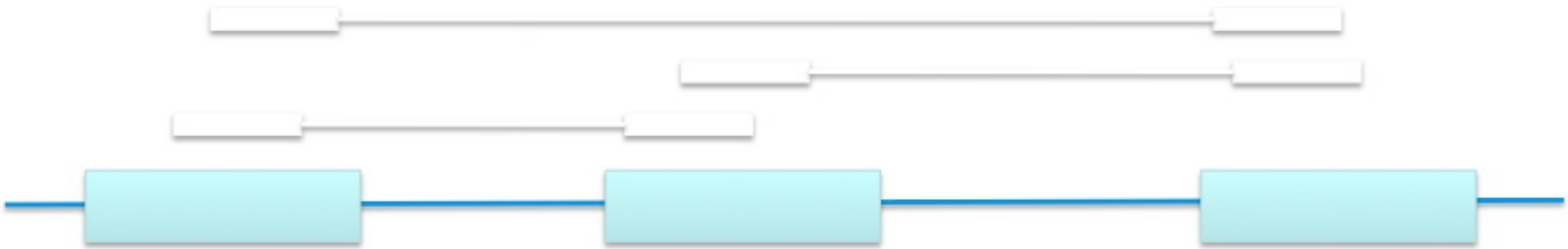
- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:



- Paths through splicing graph define *alternative transcripts*

Cufflinks

Assemble as many reads into as few (highly probable) transcripts as possible.



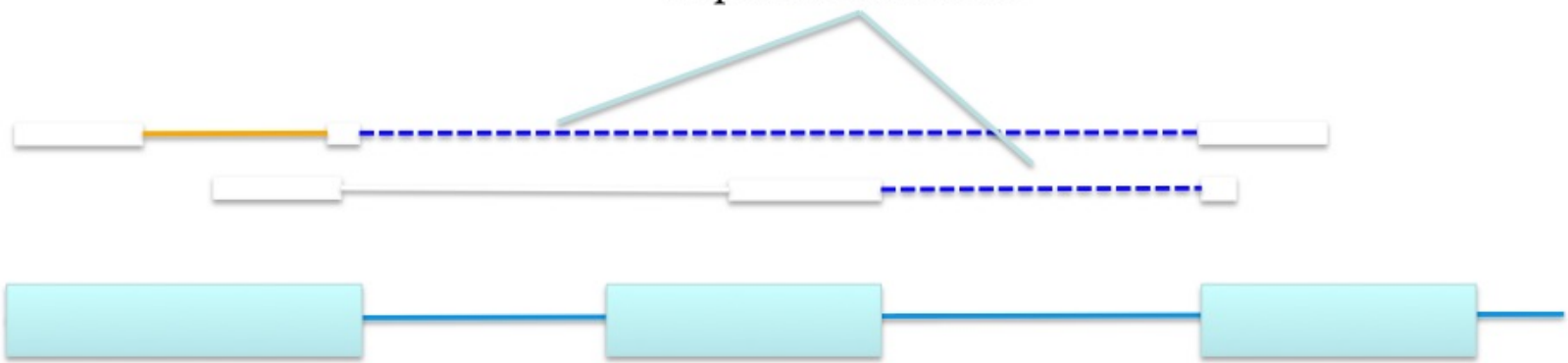
How many transcripts?

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Compatible reads

Don't know that two reads came from the same transcript, but sometimes we know that they came from **different** transcripts.

Gaps across introns

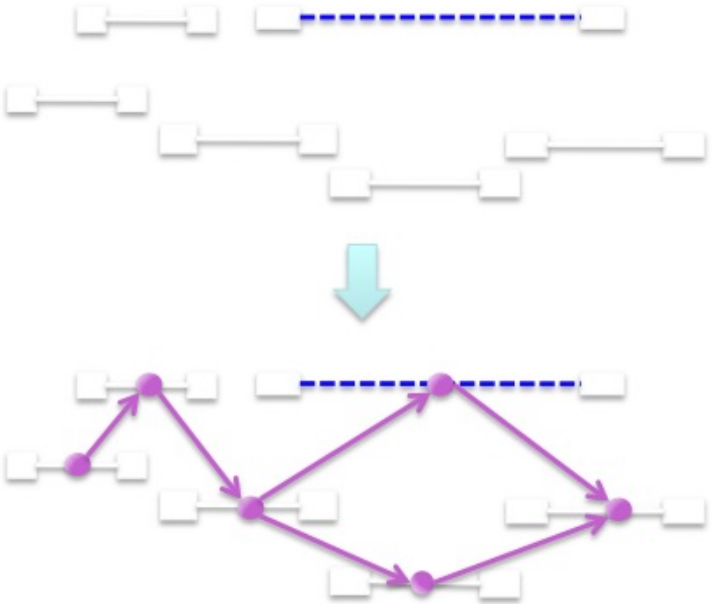


How many transcripts?

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Partially ordering mates

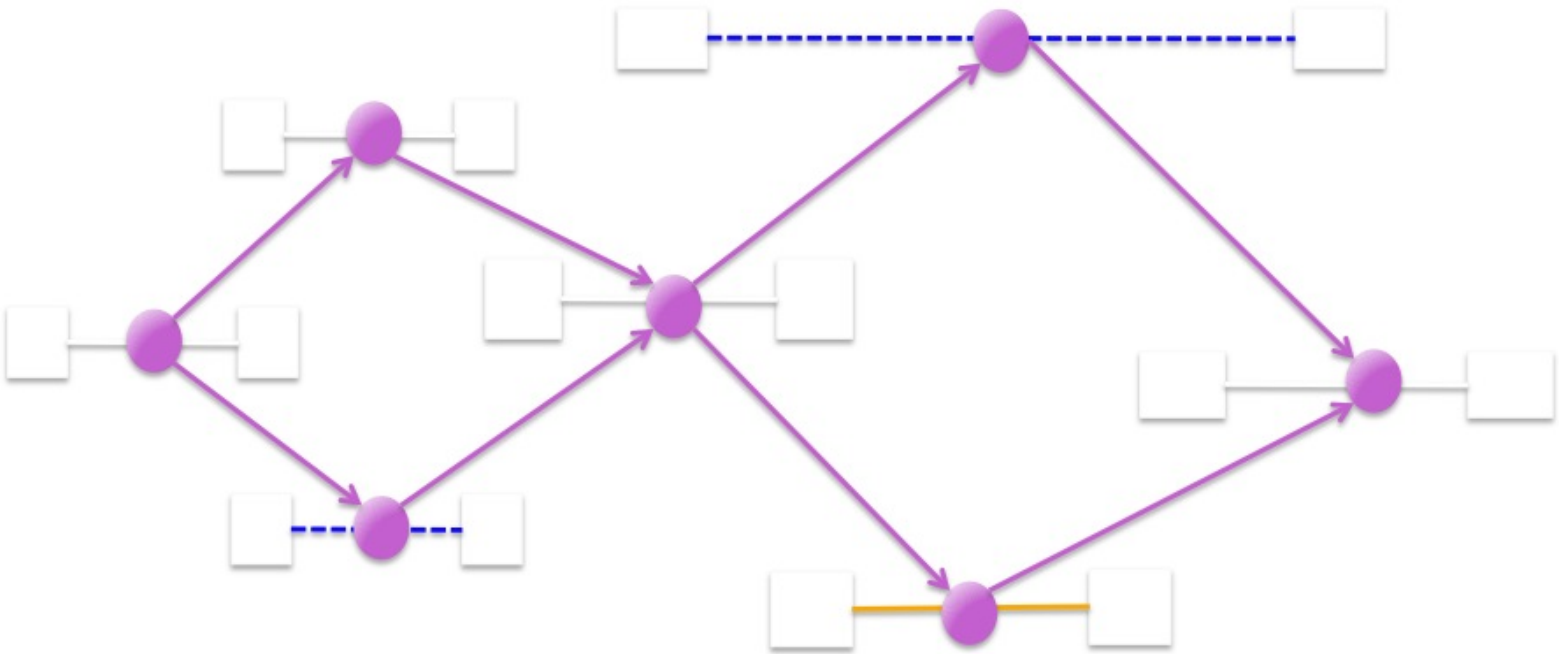
- alignment $x \leq y$ when
  - x starts to the left of y in the reference
  - x and y overlap consistently
  - y is not contained in x
- In other words, $x \leq y$ when they could have come from the same transcript



Building a DAG for the partial order

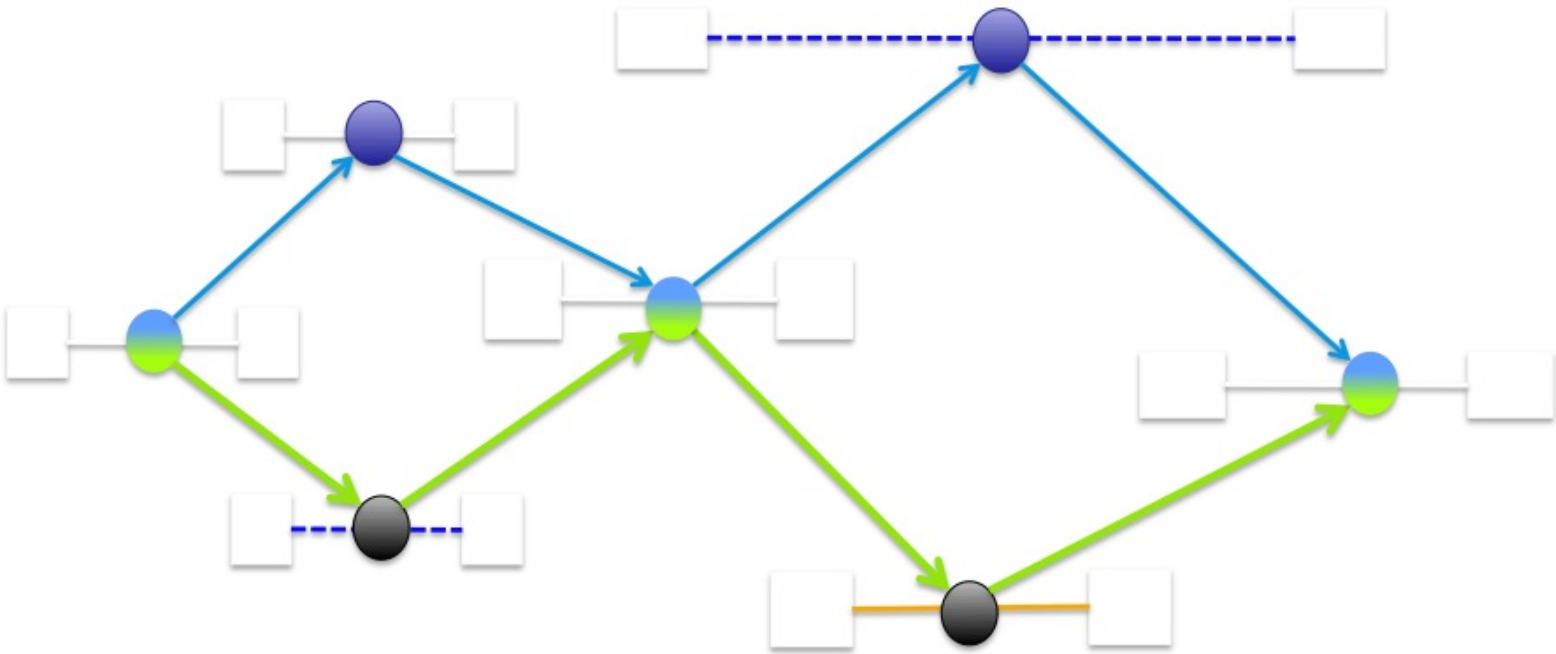
(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Path cover of the DAG



(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Path cover of the DAG



(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

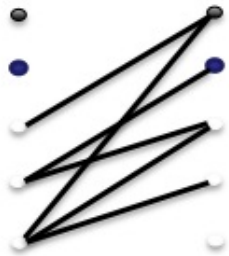
Transcript assembly via Dilworth's theorem

1



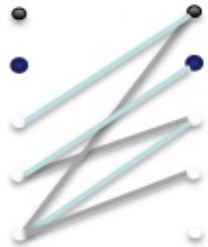
Construct DAG on alignments

2



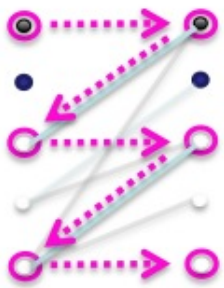
Express DAG transitive closure as bipartite graph

3



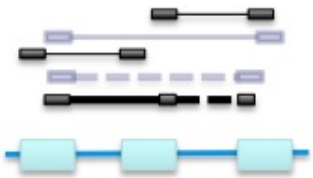
Maximum matching on BP graph

4



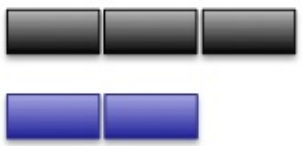
Chain decomposition from matching

5



Cover of alignments from chains

6



Transcripts from cover

Dilworth's Theorem

- Definition: In a partially ordered set of alignments, an *antichain* of alignments is a subset where no two alignments could have come from the same transcript
- Theorem: The size of the *largest* antichain is the minimum number of transcripts we need to explain the alignments

R.P. Dilworth, "A Decomposition Theorem for Partially Ordered Sets", *Annals of Mathematics*, 1950, **51**: 161–166.

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Antichains

Don't know that two reads came from the same transcript, but sometimes we know that they came from **different** transcripts.

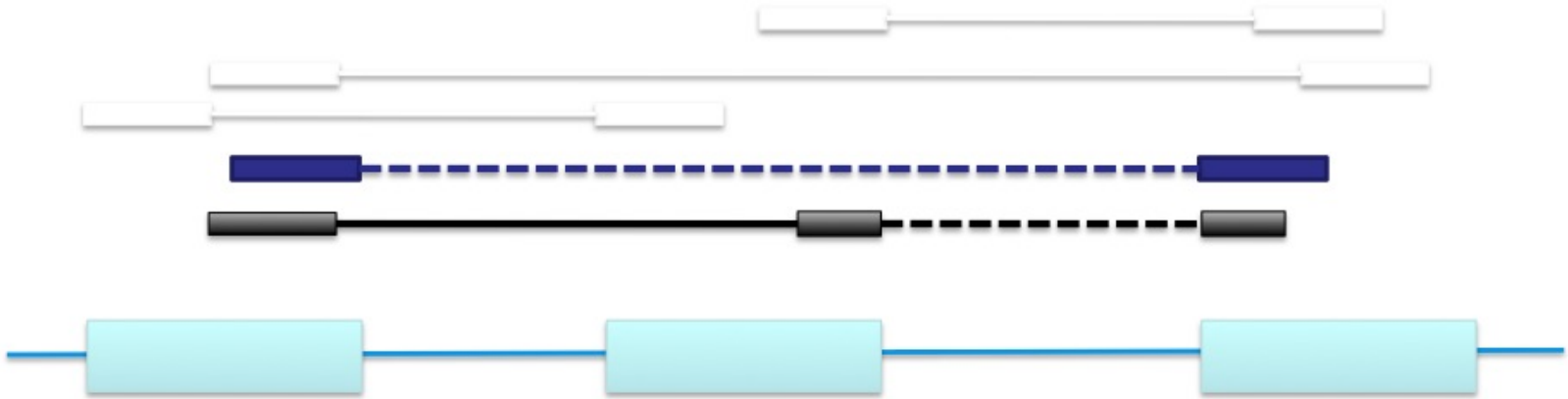


How many transcripts?

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Antichains

Don't know that two reads came from the same transcript, but sometimes we know that they came from **different** transcripts.



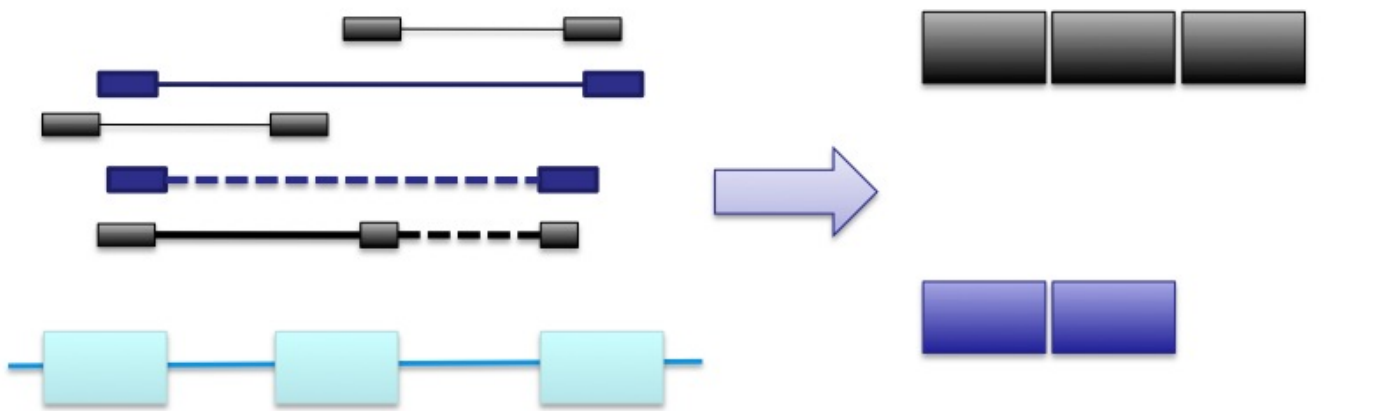
How many transcripts?
Two!

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Chains

From chain decomposition of alignments…

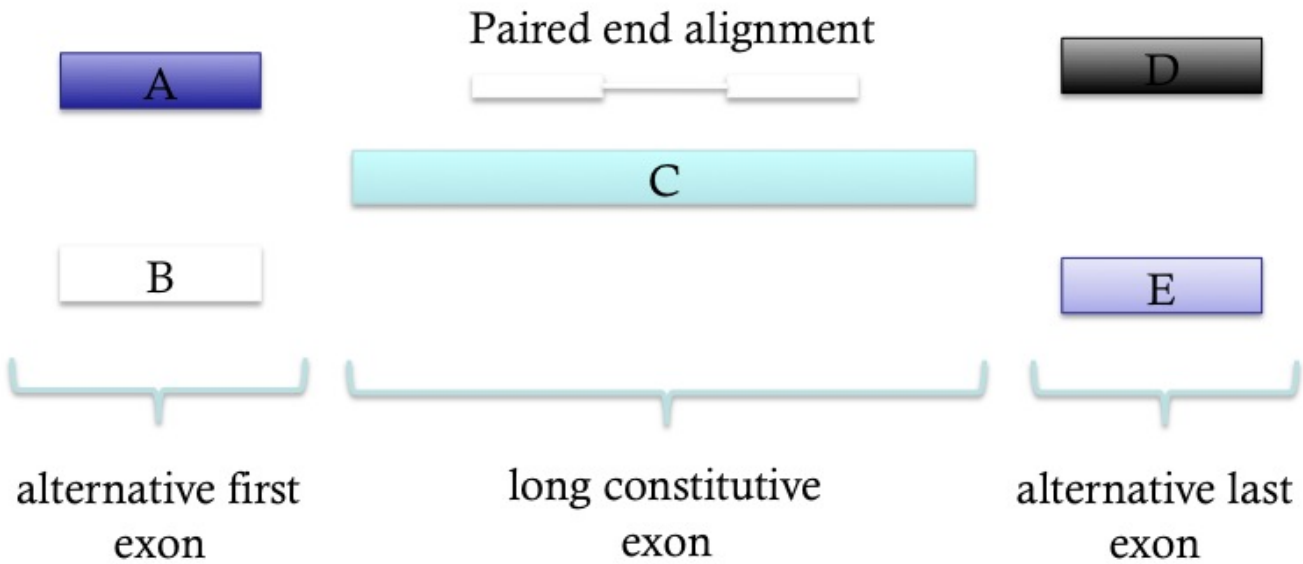
… build transcript assemblies



The chains are sets of mutually compatible alignments, so we can smash them together into transcripts at will.

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

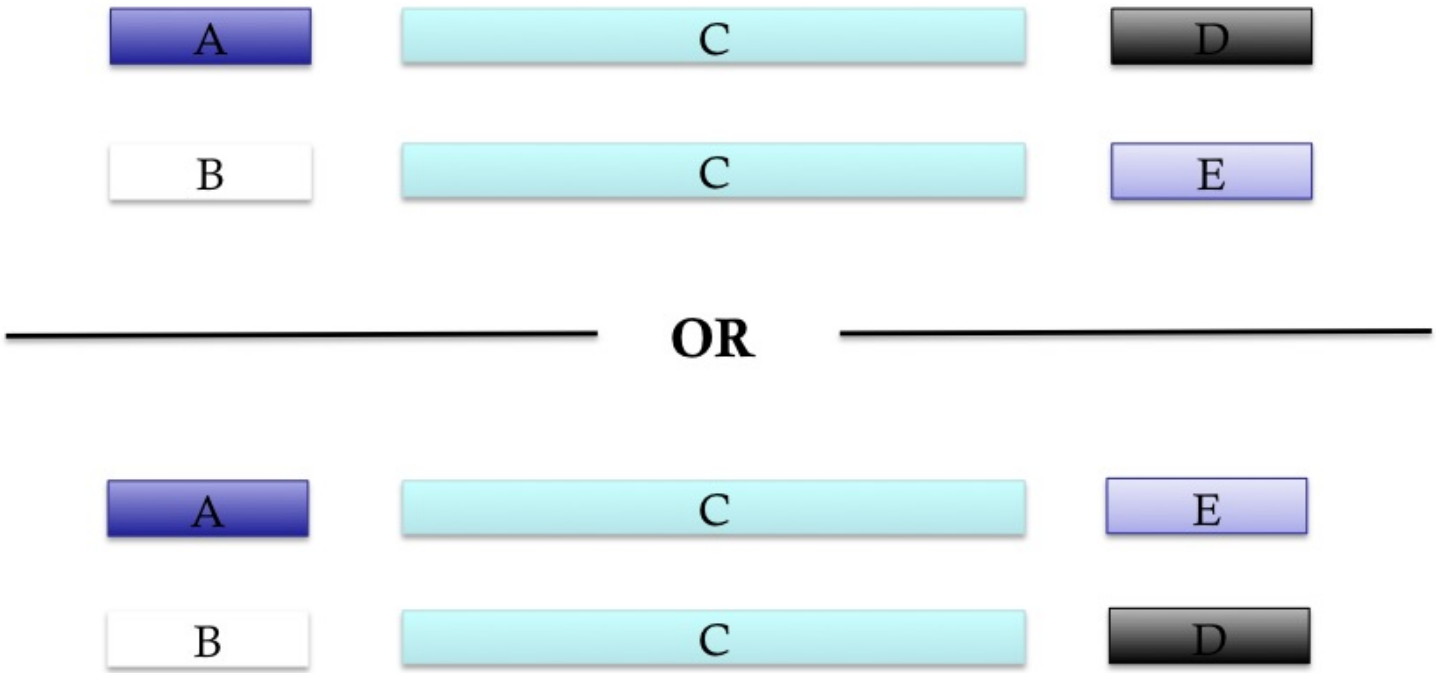
Phasing splicing events



Two parsimonious assemblies of these splicing event outcomes. The maximum cardinality matching *may not be unique*.

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

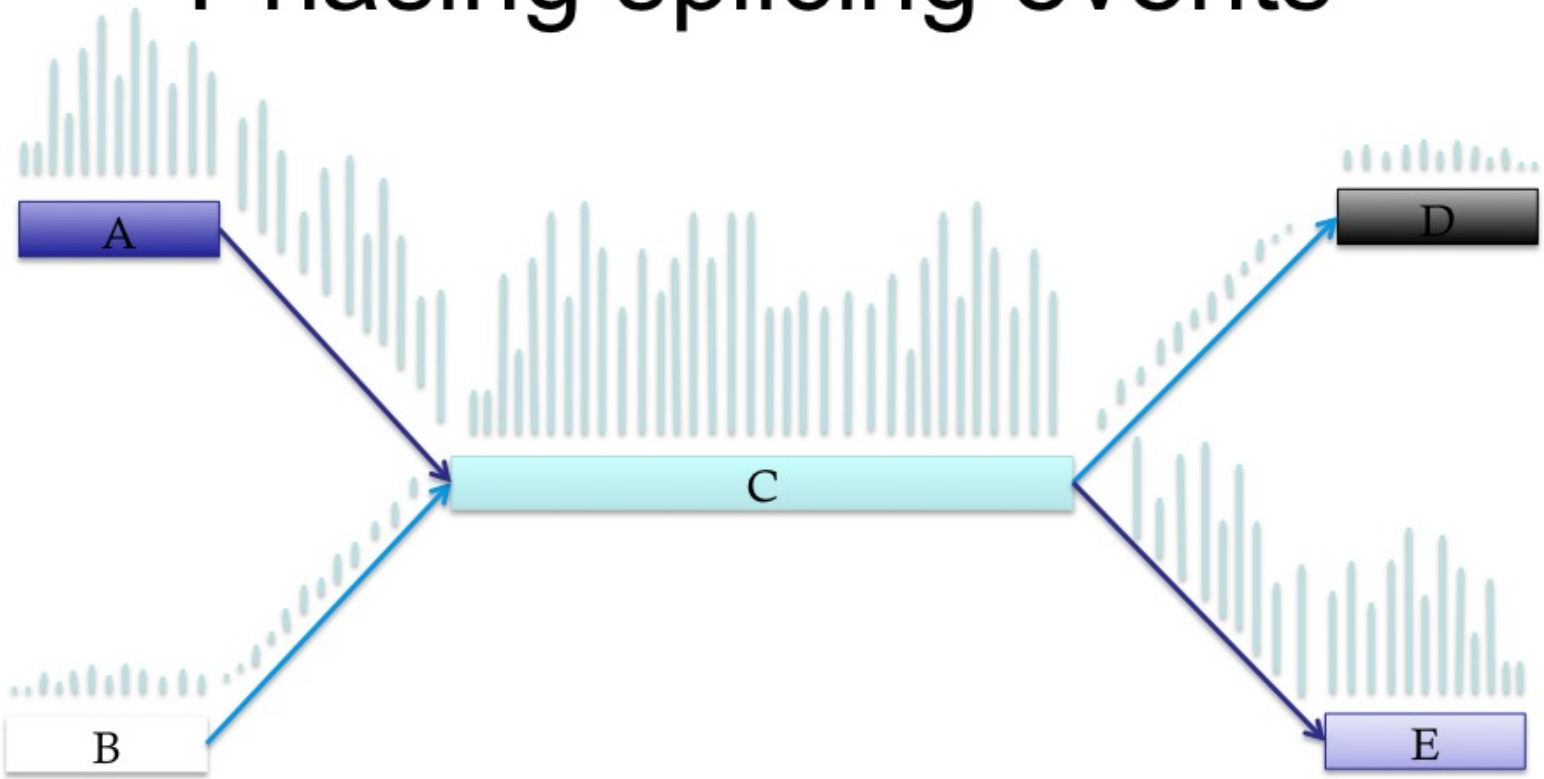
Phasing splicing events



Two parsimonious assemblies of these splicing event outcomes. The maximum cardinality matching *may not be unique*.

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Phasing splicing events

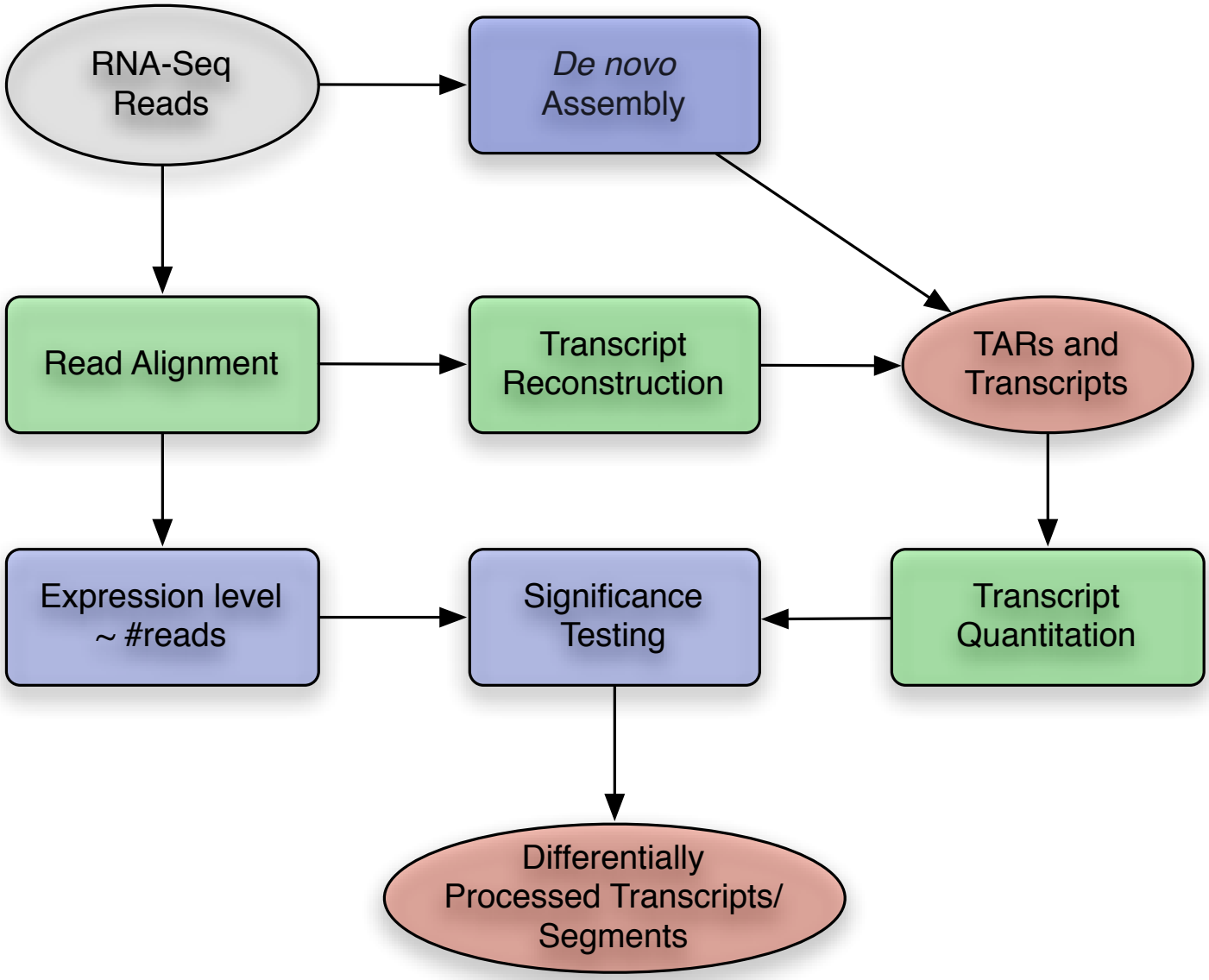


Can use coverage on exons and splice junctions to phase splicing events, via **min cost** bipartite matching, instead of max cardinality.

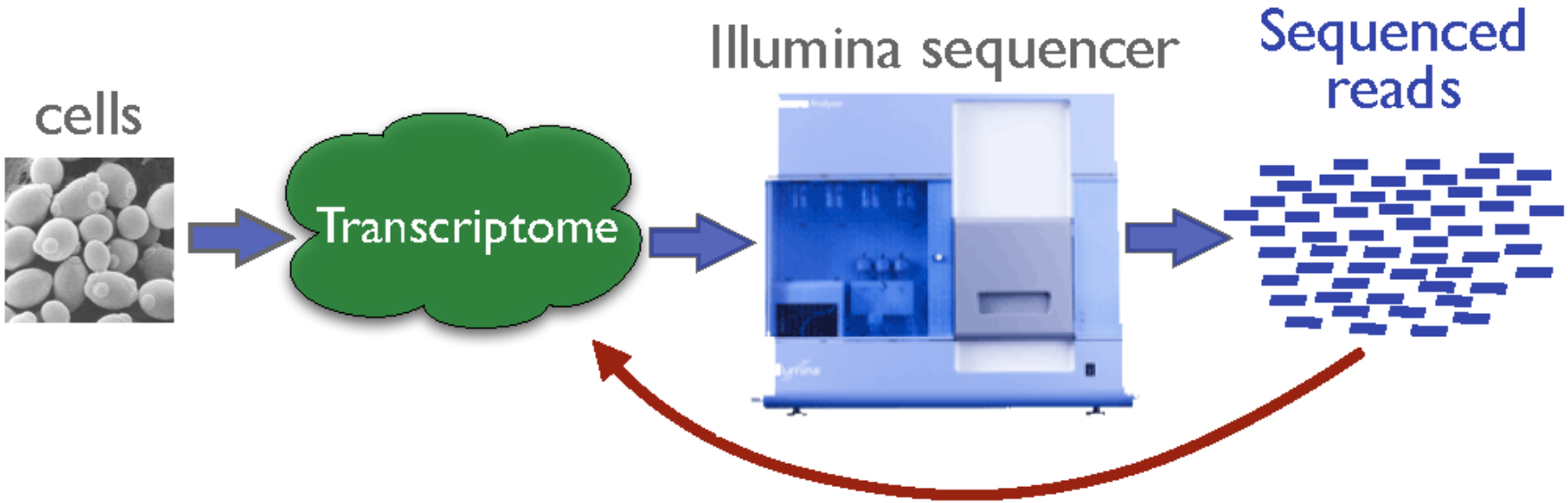
(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Transcript Assembly

Common RNA-Seq Analysis Steps



RNA-seq



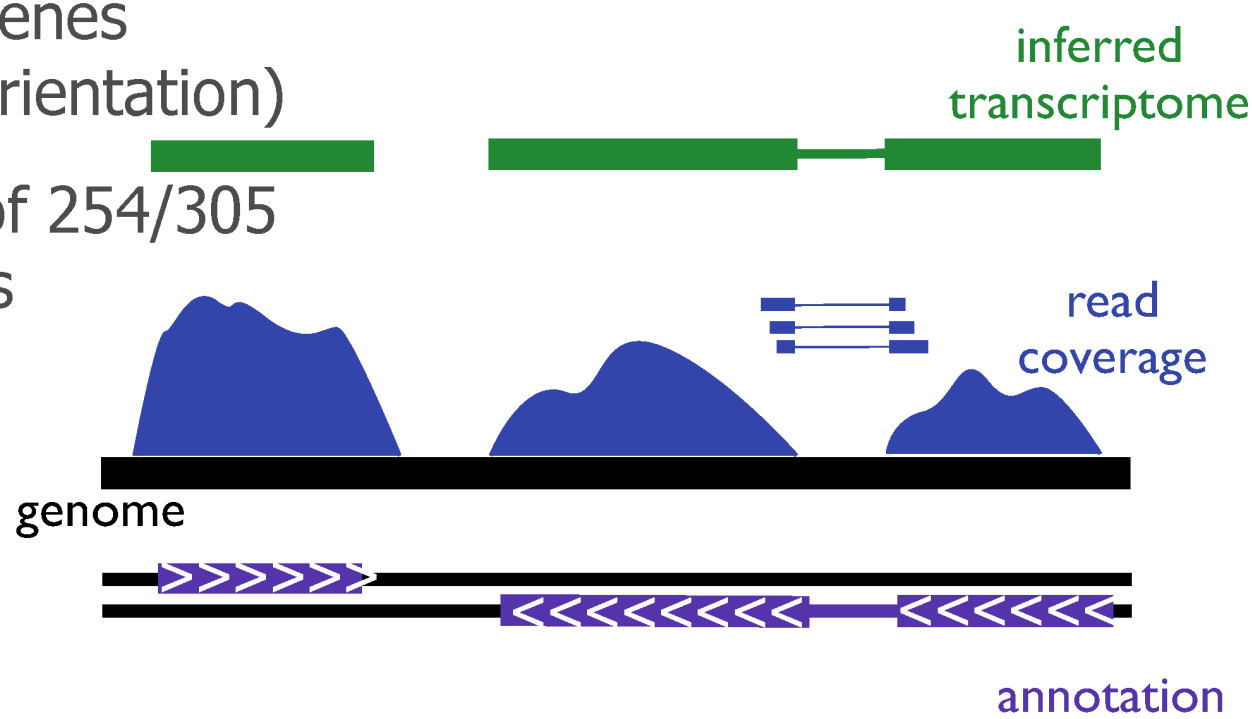
RNA-seq assembly

(Slides provided by Moran Yassour, Hebrew University and Broad)

Mapping first approach (S. cerevisiae case)

S. cerevisiae results:

- Reconstruction of 86% of expressed genes (with correct orientation)
- Identification of 254/305 splice junctions



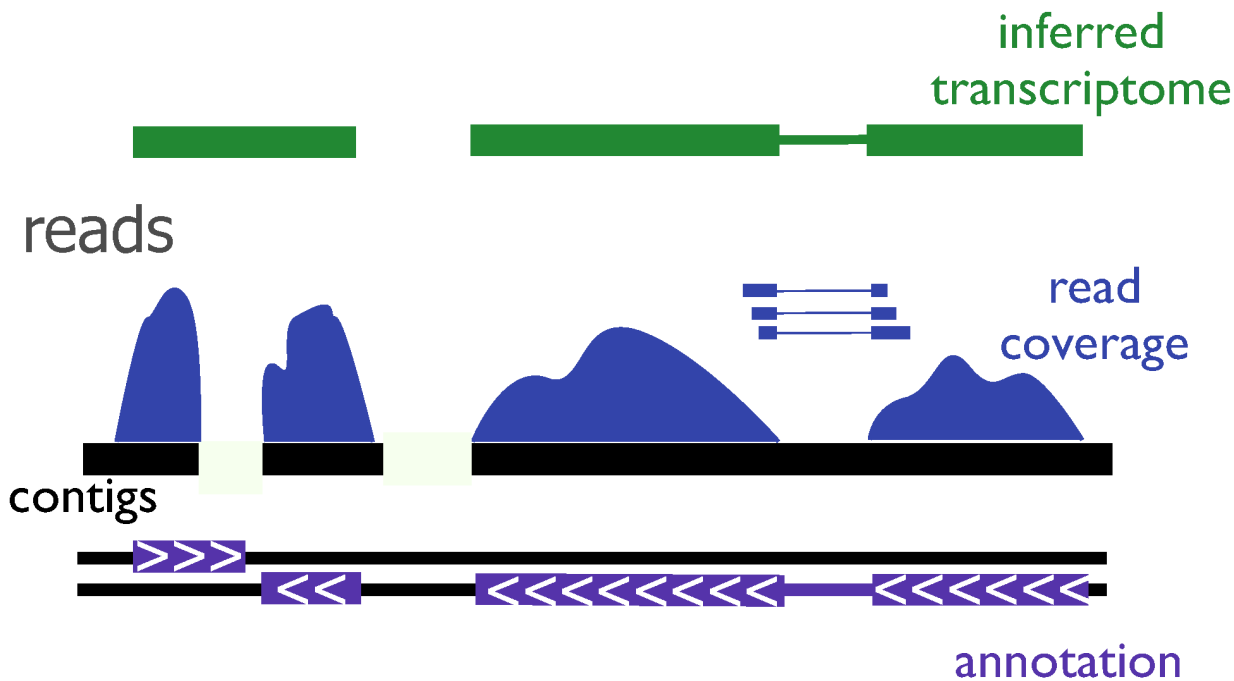
Yassour *et al.*, PNAS, Jan. 09

(Slides provided by Moran Yassour, Hebrew University and Broad)

Mapping first approach

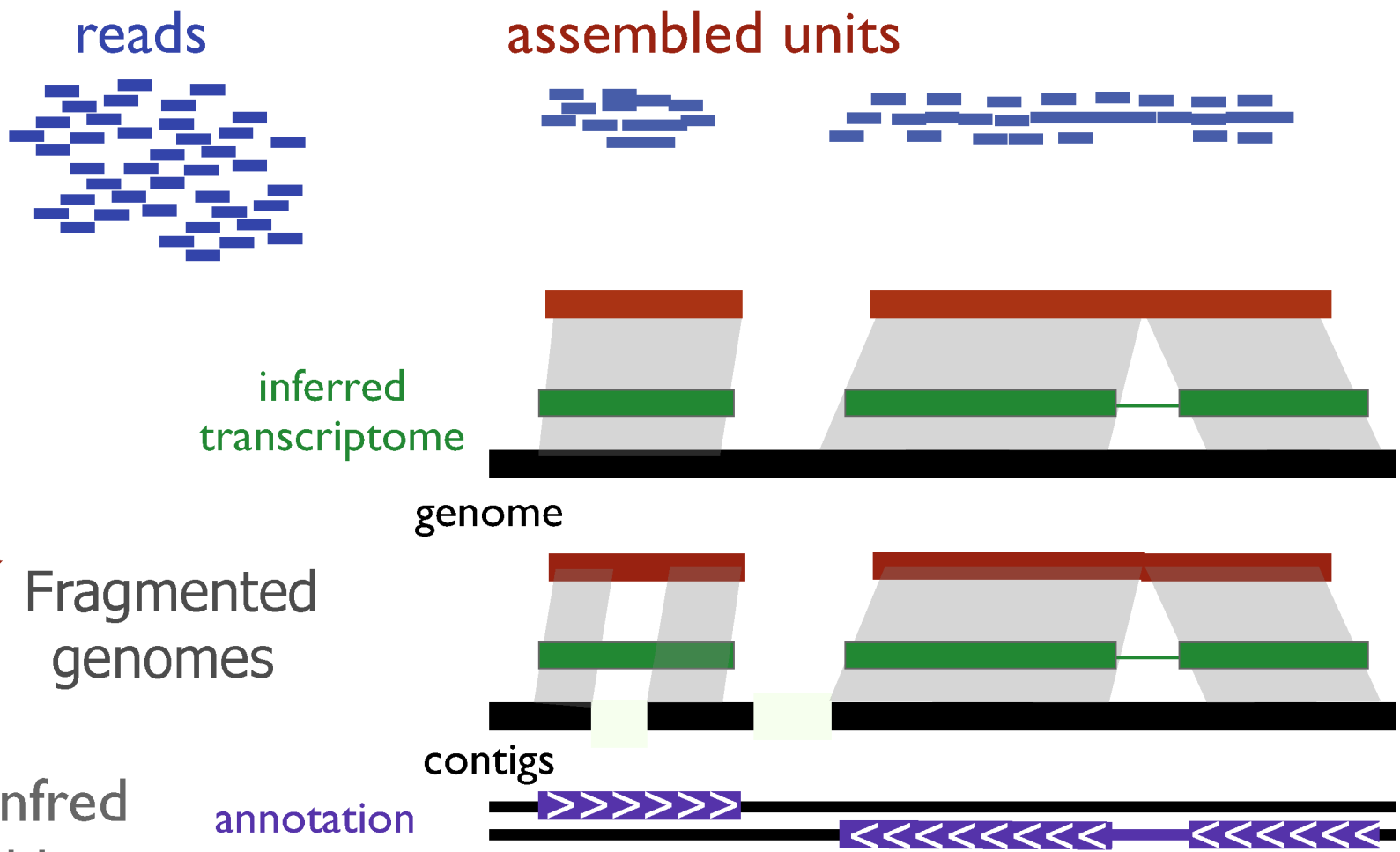
Problems:

- Clear boundaries
  - Strand specific libraries
- Mapping spliced reads
- Fragmented genomes



(Slides provided by Moran Yassour, Hebrew University and Broad)

Assembling first approach



✓ Fragmented genomes

Manfred Grabherr

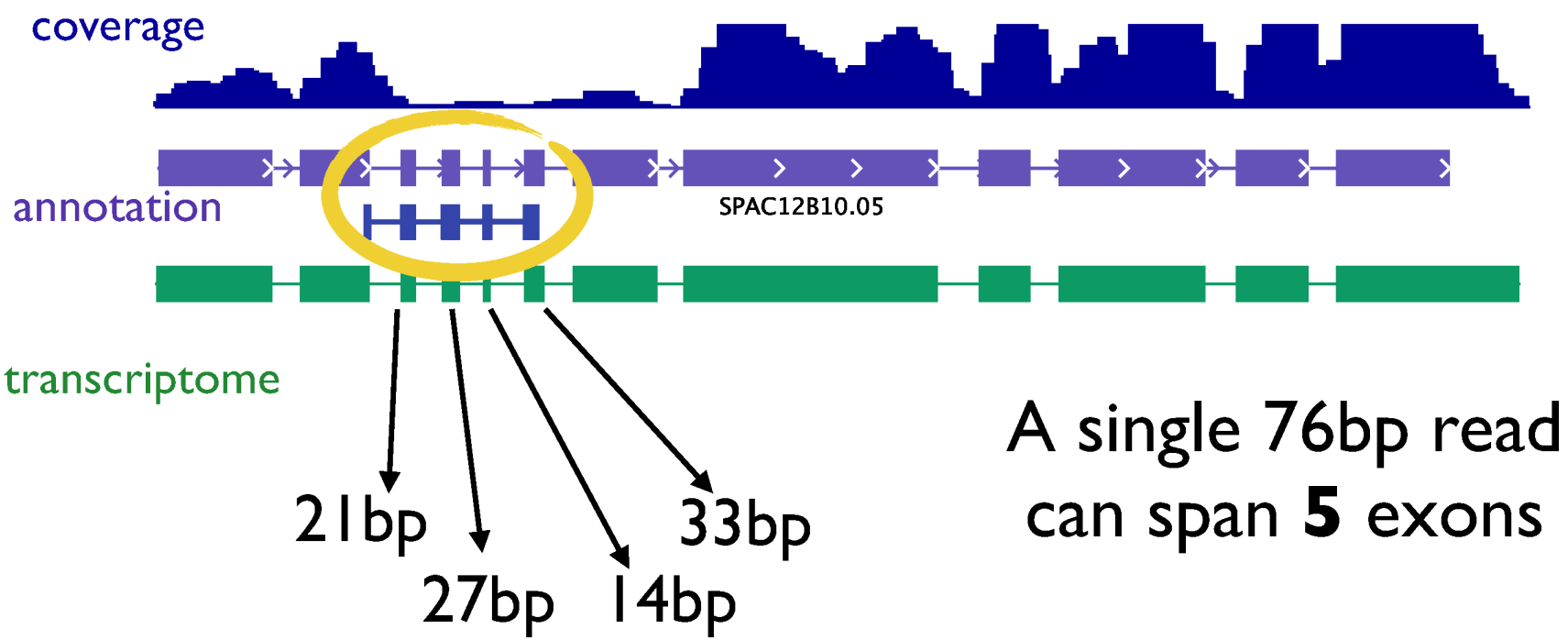
(Slides provided by Moran Yassour, Hebrew University and Broad)

When should we use the "assembly first" approach?

- Highly spliced transcriptome
- Fragmented genome
- Cancer genome
- No sequenced genome
- Meta genomes

(Slides provided by Moran Yassour, Hebrew University and Broad)

Ab-initio reconstruction of S. pombe gene with 12 exons



(Slides provided by Moran Yassour, Hebrew University and Broad)

Read Assembly



Problem

Given a set of reads (from genome or transcriptome), find maximally overlapping parts in order to reconstruct the original sequences

Classical assembly ⇒ **Too inefficient for NGS reads**

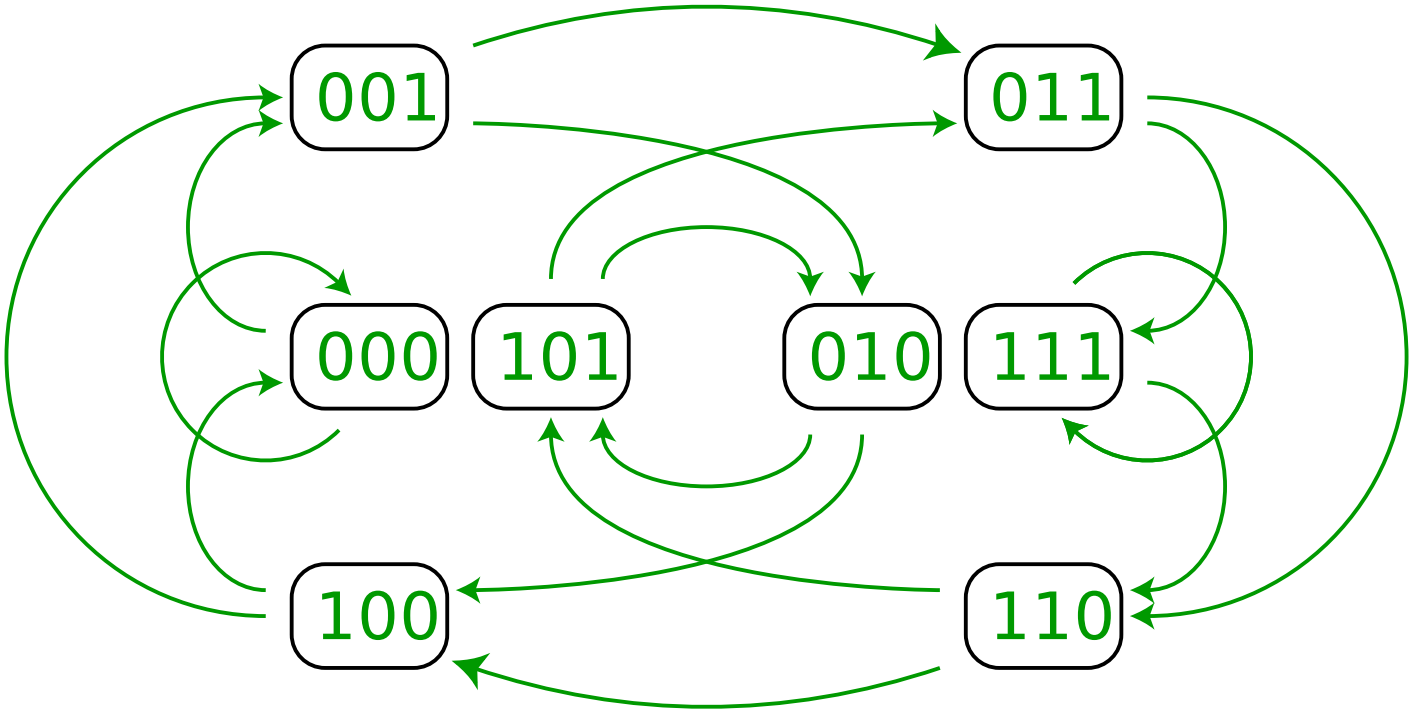
1. *Overlap phase*: Every read is compared with every other read and the overlap graph is computed
2. *Layout phase*: Pairs are determined that position every read in the assembly
3. *Consensus phase*: Multi-alignment of all the placed reads is produced to obtain the final sequence

New techniques based on **de Bruijn Graphs**

- Genome assembly: EULER [Pevzner et al., 2001], Velvet [Zerbino and Birney, 2008], Allpaths [Maccallum et al., 2009], SHARCGS [Dohm et al., 2007], SSAKE/VCAKE [Warren et al., 2007], ...
- Transcriptome assembly: Oases, Ananas

Read Assembly - de Bruijn Graphs

Example:

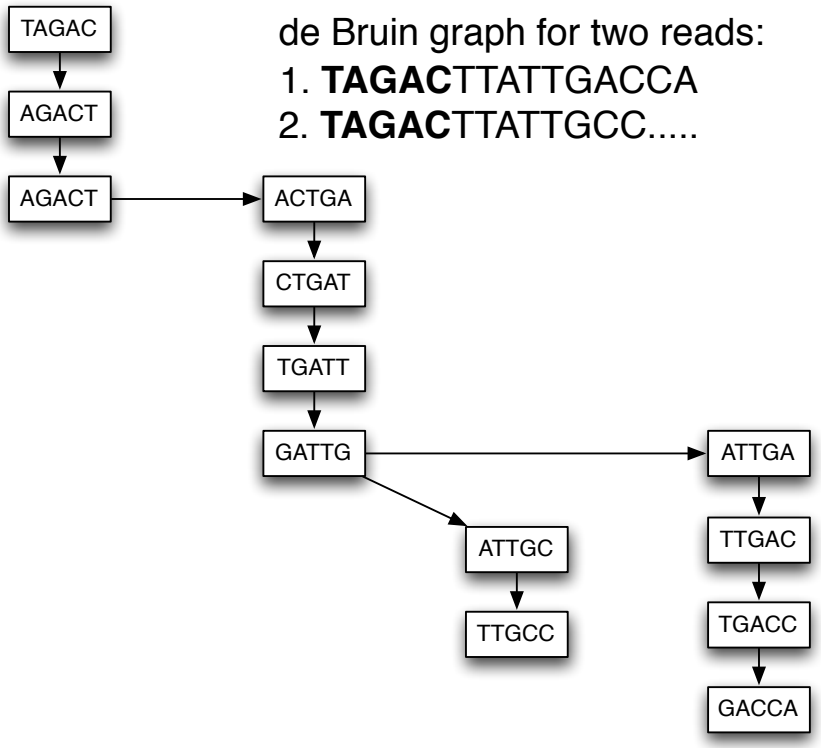


- Reads are mapped as a path in the graph
- Number of reads does not influence number of nodes
  $\Rightarrow$ Use de-Bruijn graphs to solve the problem

[Wikipedia]

de Bruijn Graphs

Example:

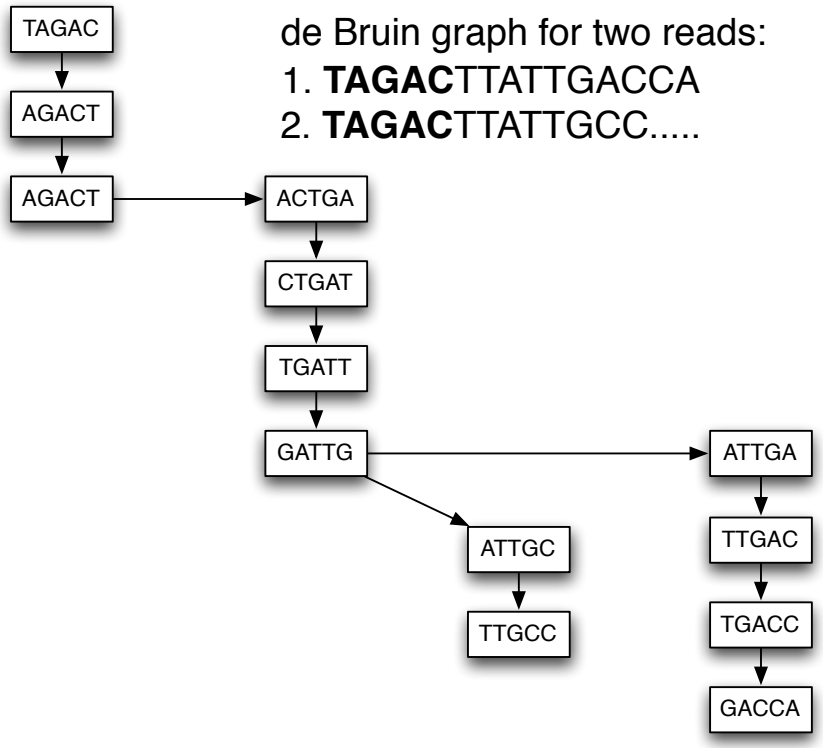


[Zerbino and Birney, 2008]

- Nodes represent *k*-mers smaller than read length
- A k-mer can refer to thousands of reads containing it
- Read errors or ambiguities lead to branching of paths
- Each node also stores the reverse complement

de Bruijn Graphs

Example:

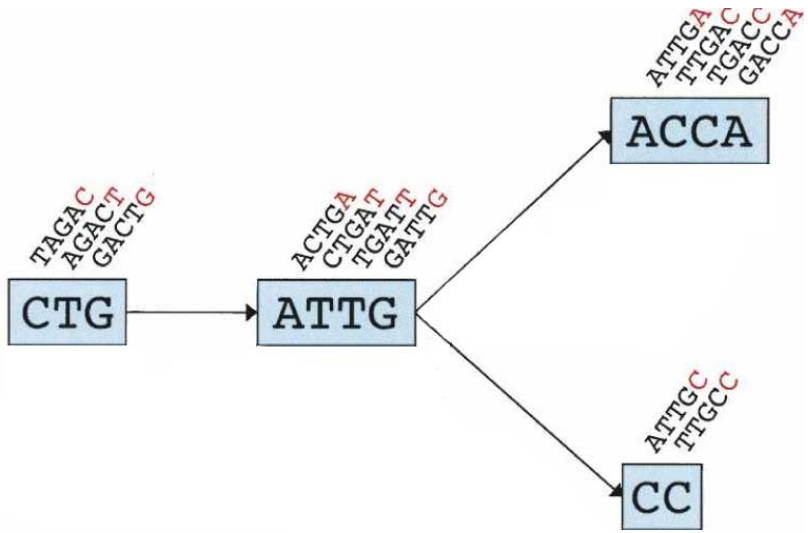


[Zerbino and Birney, 2008]

- Nodes represent *k*-mers smaller than read length
- A k-mer can refer to thousands of reads containing it
- Read errors or ambiguities lead to branching of paths
- Each node also stores the reverse complement

de Bruijn Graphs

Example:

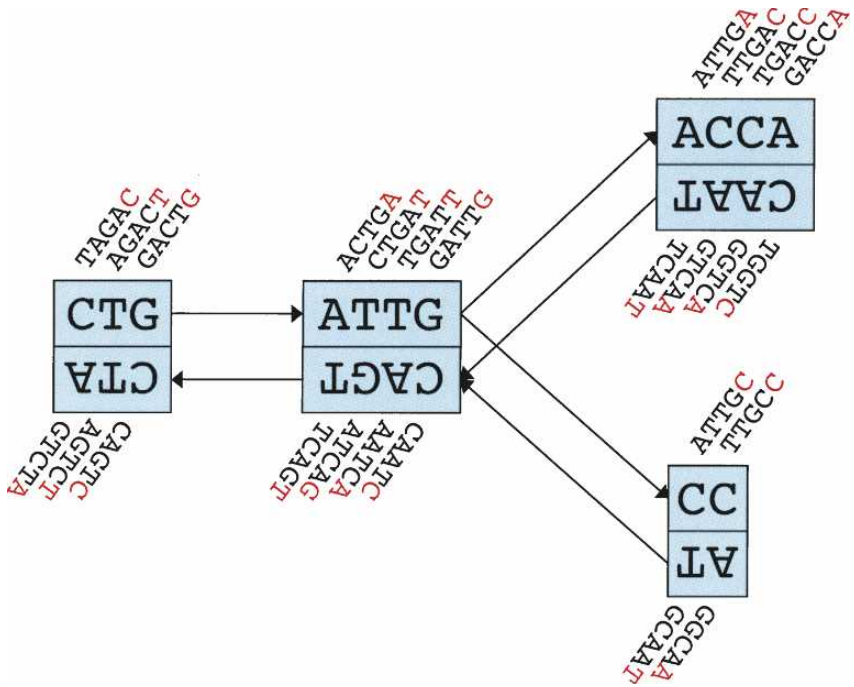


[Zerbino and Birney, 2008]

- Nodes represent *k*-mers smaller than read length
- A k-mer can refer to thousands of reads containing it
- Read errors or ambiguities lead to branching of paths
- Each node also stores the reverse complement

de Bruijn Graphs

Example:



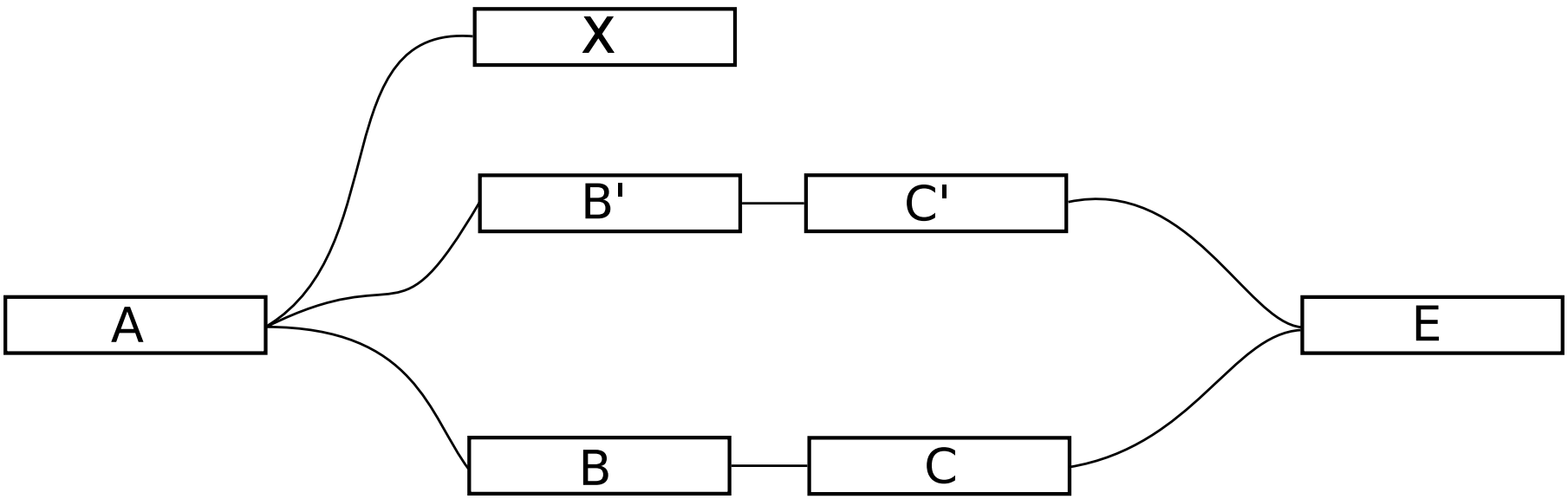
[Zerbino and Birney, 2008]

- Nodes represent *k*-mers smaller than read length
- A k-mer can refer to thousands of reads containing it
- Read errors or ambiguities lead to branching of paths
- Each node also stores the reverse complement

Read Assembly - Example

Consider a constructed de Bruijn graph

- Unconnected nodes
- Ambiguous paths
- Erroneous edges



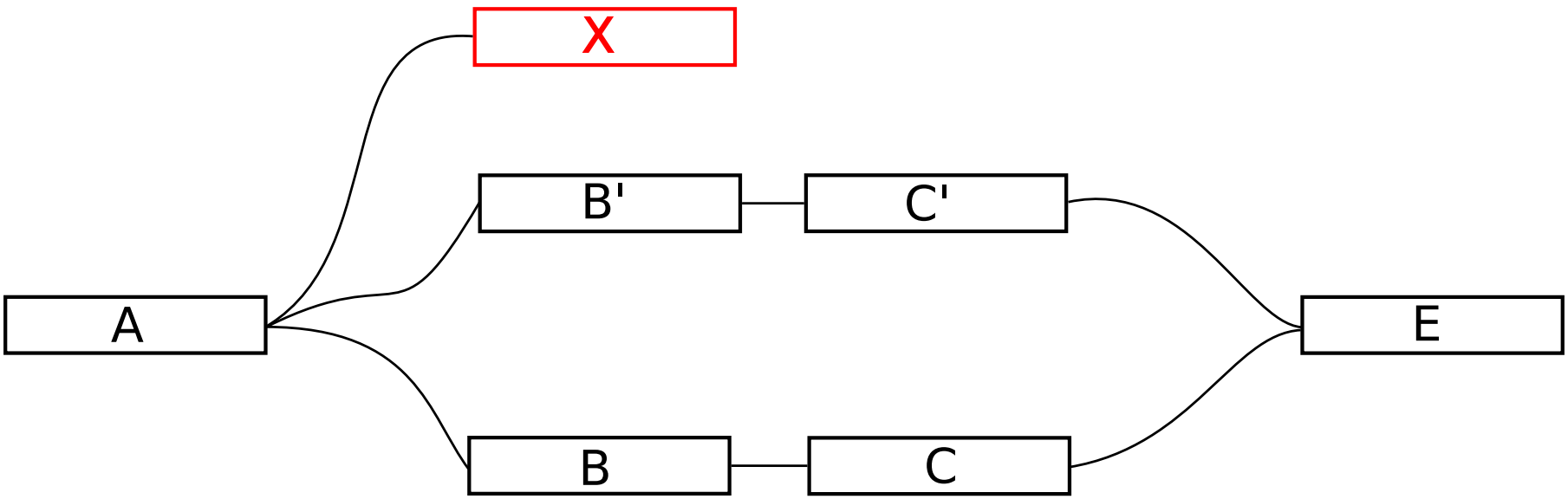
Read off genome sequence (if everything goes well ;-)

[Zerbino and Birney, 2008]

Read Assembly - Example

Consider a constructed de Bruijn graph

- Unconnected nodes
- Ambiguous paths
- Erroneous edges



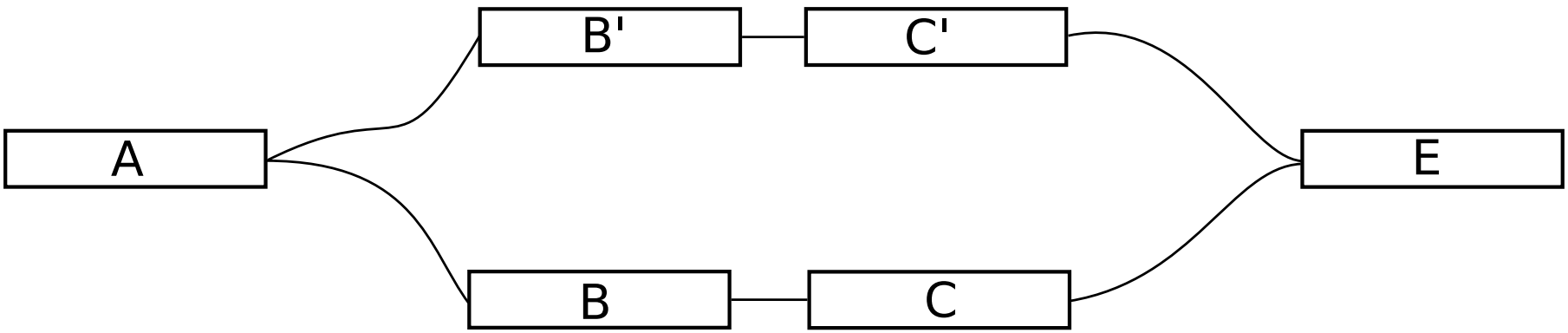
Read off genome sequence (if everything goes well ;-)

[Zerbino and Birney, 2008]

Read Assembly - Example

Consider a constructed de Bruijn graph

- Unconnected nodes
- Ambiguous paths
- Erroneous edges



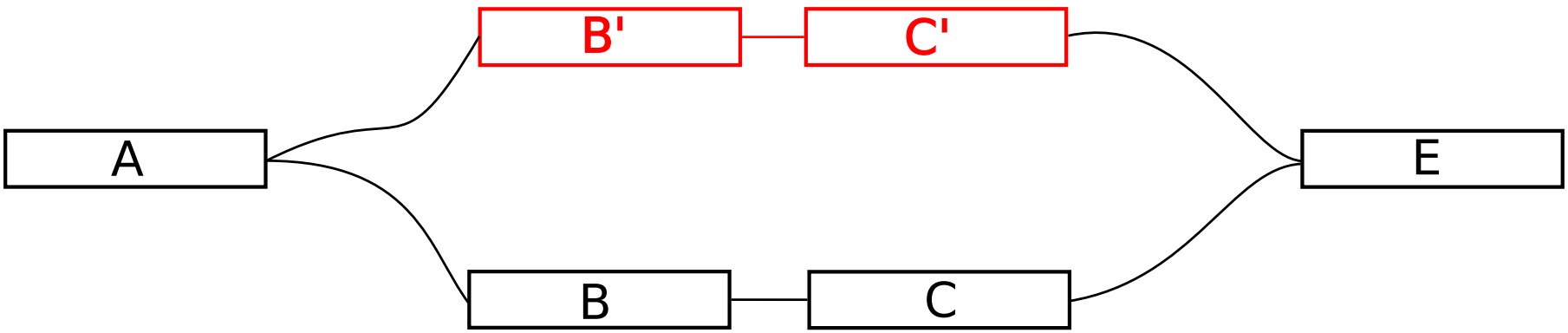
Read off genome sequence (if everything goes well ;-)

[Zerbino and Birney, 2008]

Read Assembly - Example

Consider a constructed de Bruijn graph

- Unconnected nodes
- Ambiguous paths
- Erroneous edges



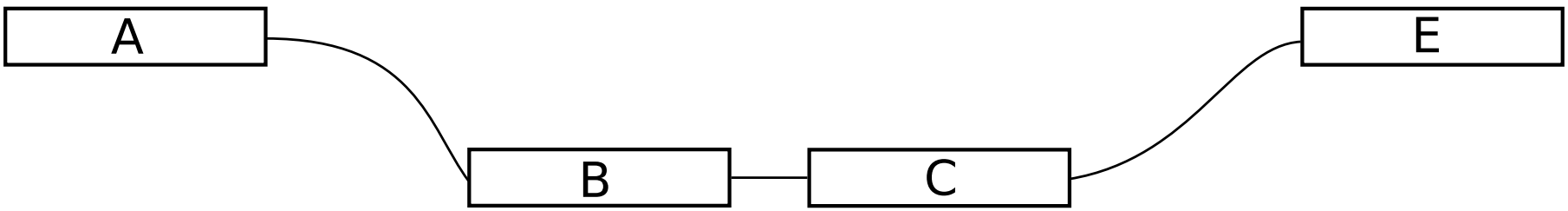
Read off genome sequence (if everything goes well ;-)

[Zerbino and Birney, 2008]

Read Assembly - Example

Consider a constructed de Bruijn graph

- Unconnected nodes
- Ambiguous paths
- Erroneous edges



Read off genome sequence (if everything goes well ;-)

[Zerbino and Birney, 2008]

Results for Genome Assembly



Table 3. Comparison of short read assemblers on experimental *Streptococcus suis* Solexa reads

Assembler	No. of contigs	N50	Average error rate	Memory	Time	Seq. Cov.
Velvet 0.3	470	8661 bp	0.02%	2.0G	2 min 57 sec	97%
SSAKE 2.0	265	1727 bp	0.20%	1.7G	1 h 47 min	16%
VCAKE 1.0	7675	1137 bp	0.64%	1.8G	4 h 25 min	134%

Considerably shorter fragments for larger genomes

[Zerbino and Birney, 2008]

Results for Genome Assembly



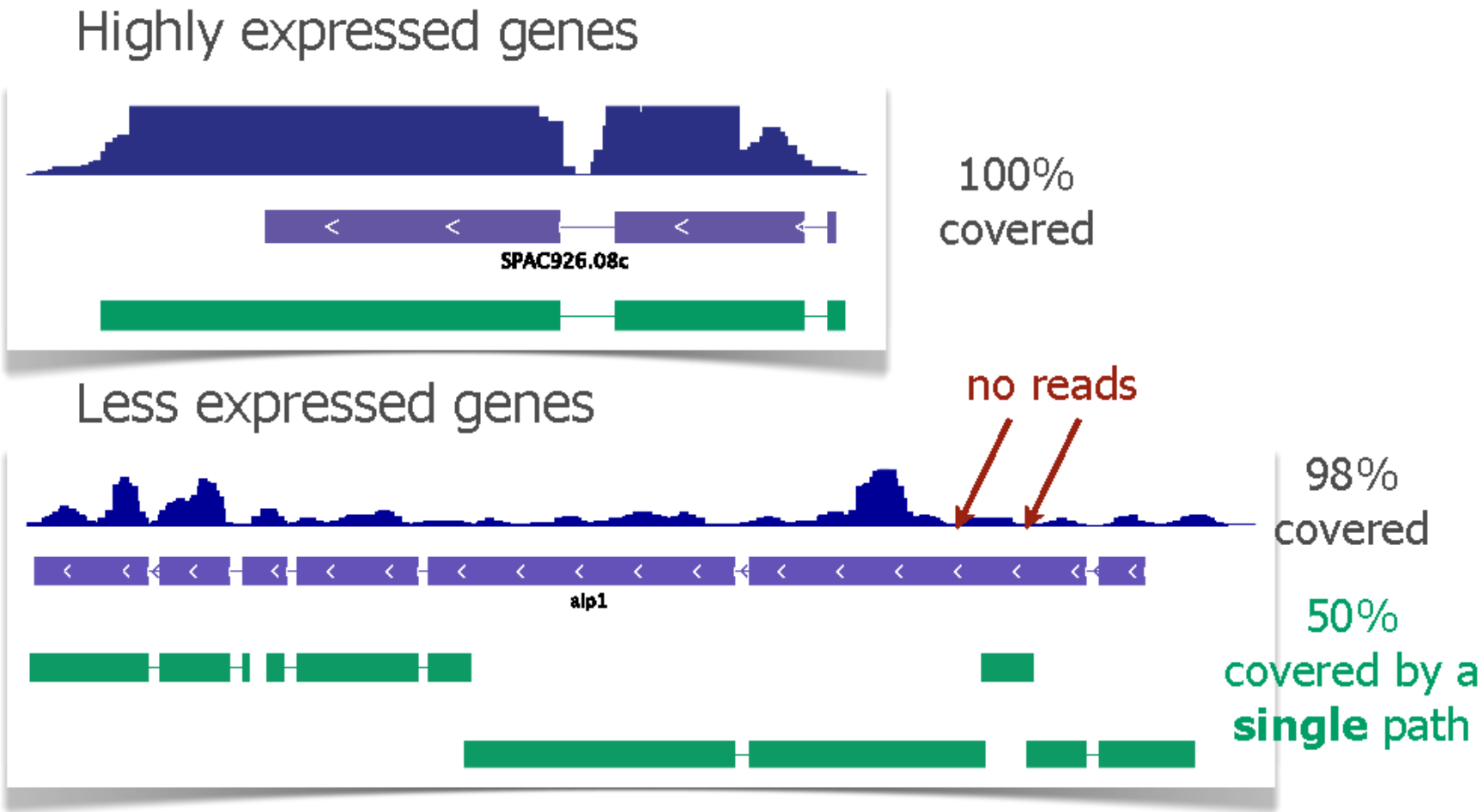
Table 3. Comparison of short read assemblers on experimental *Streptococcus suis* Solexa reads

Assembler	No. of contigs	N50	Average error rate	Memory	Time	Seq. Cov.
Velvet 0.3	470	8661 bp	0.02%	2.0G	2 min 57 sec	97%
SSAKE 2.0	265	1727 bp	0.20%	1.7G	1 h 47 min	16%
VCAKE 1.0	7675	1137 bp	0.64%	1.8G	4 h 25 min	134%

Considerably shorter fragments for larger genomes

[Zerbino and Birney, 2008]

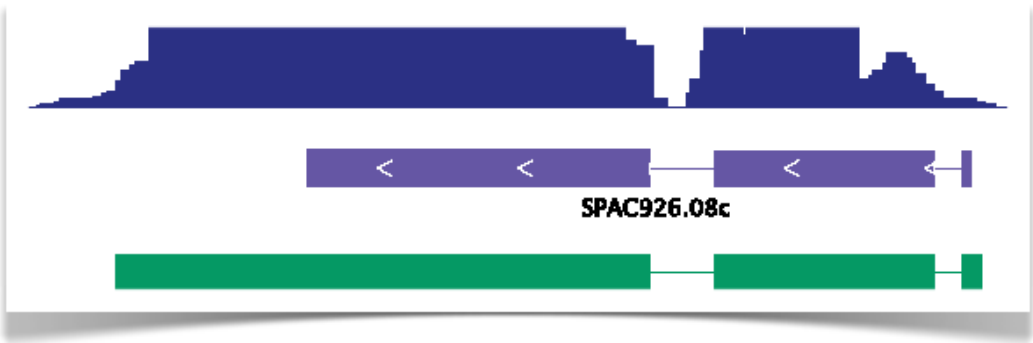
Easier to reconstruct highly expressed genes



(Slides provided by Moran Yassour, Hebrew University and Broad)

Easier to reconstruct highly expressed genes

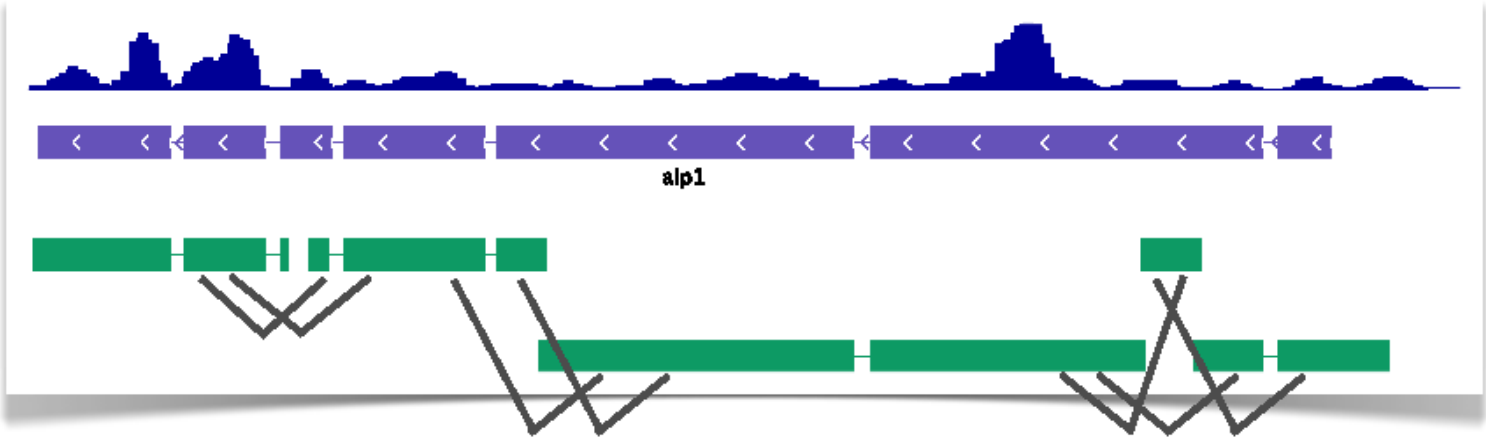
Highly expressed genes



100% covered

Use the paired-end reads

Less expressed genes



(Slides provided by Moran Yassour, Hebrew University and Broad)

Roadmap

Introduction
- Quantitation in biology
- The Rise of the Sequencers
- Sequence-counting Assays
- Important Considerations

Transcriptome Analysis
- Motivation
- Common Steps and Challenges
- RNA-seq Read Mapping
- Quantitation & Diff. Expression
- Transcript Reconstruction
- Transcript Assembly

ChIP-seq
- Introduction
- Mapping the reads
- Taking Advantage of Directionality
- Shifting Reads
- Motif Finding

Integration
- The Combinatorial Problem
- Analyzing Chip-seq Datasets Jointly
- Segmenting the Genome
- Learning Regulatory States
- Joint Analysis of Chip- and RNA-Seq

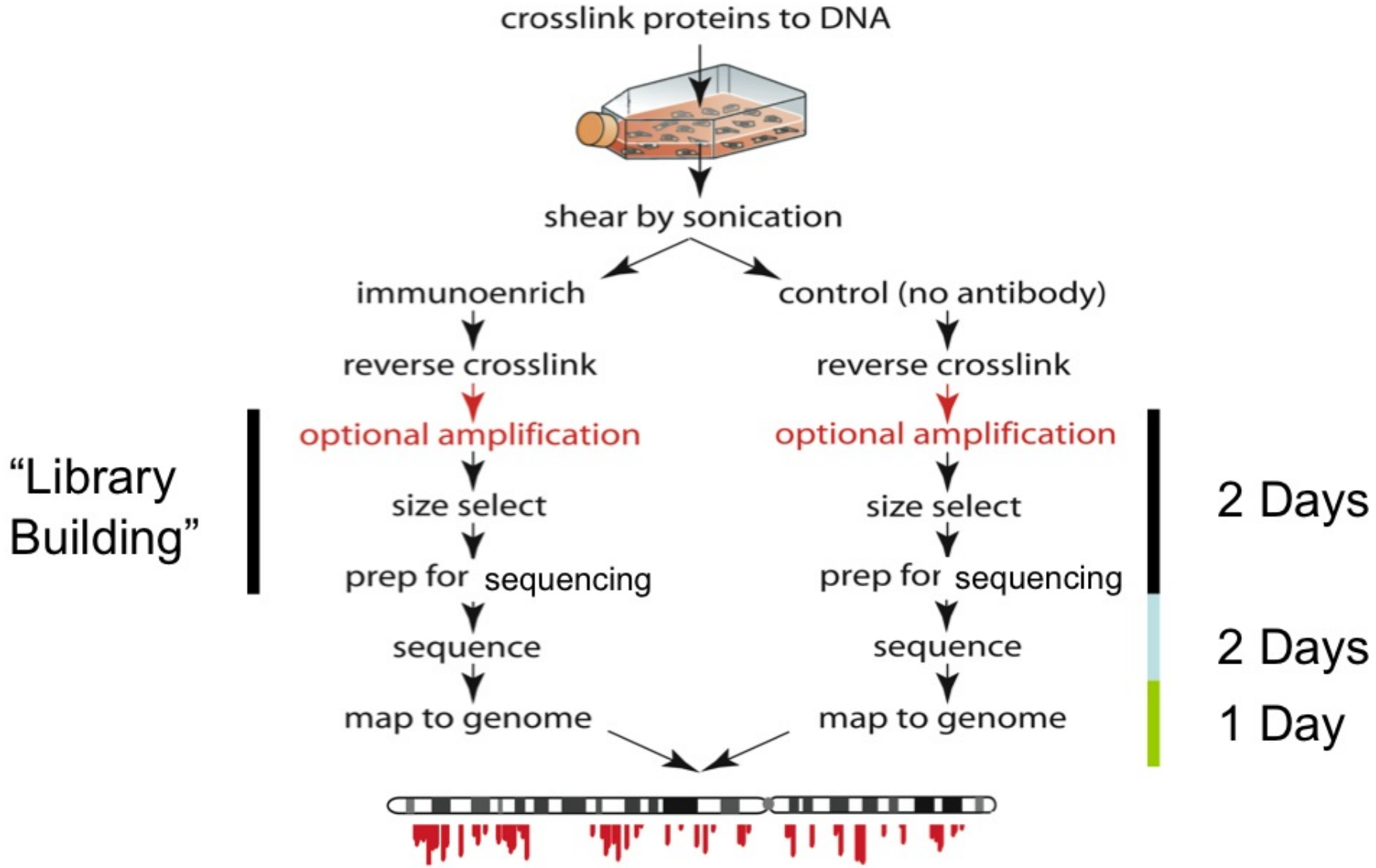
Introduction



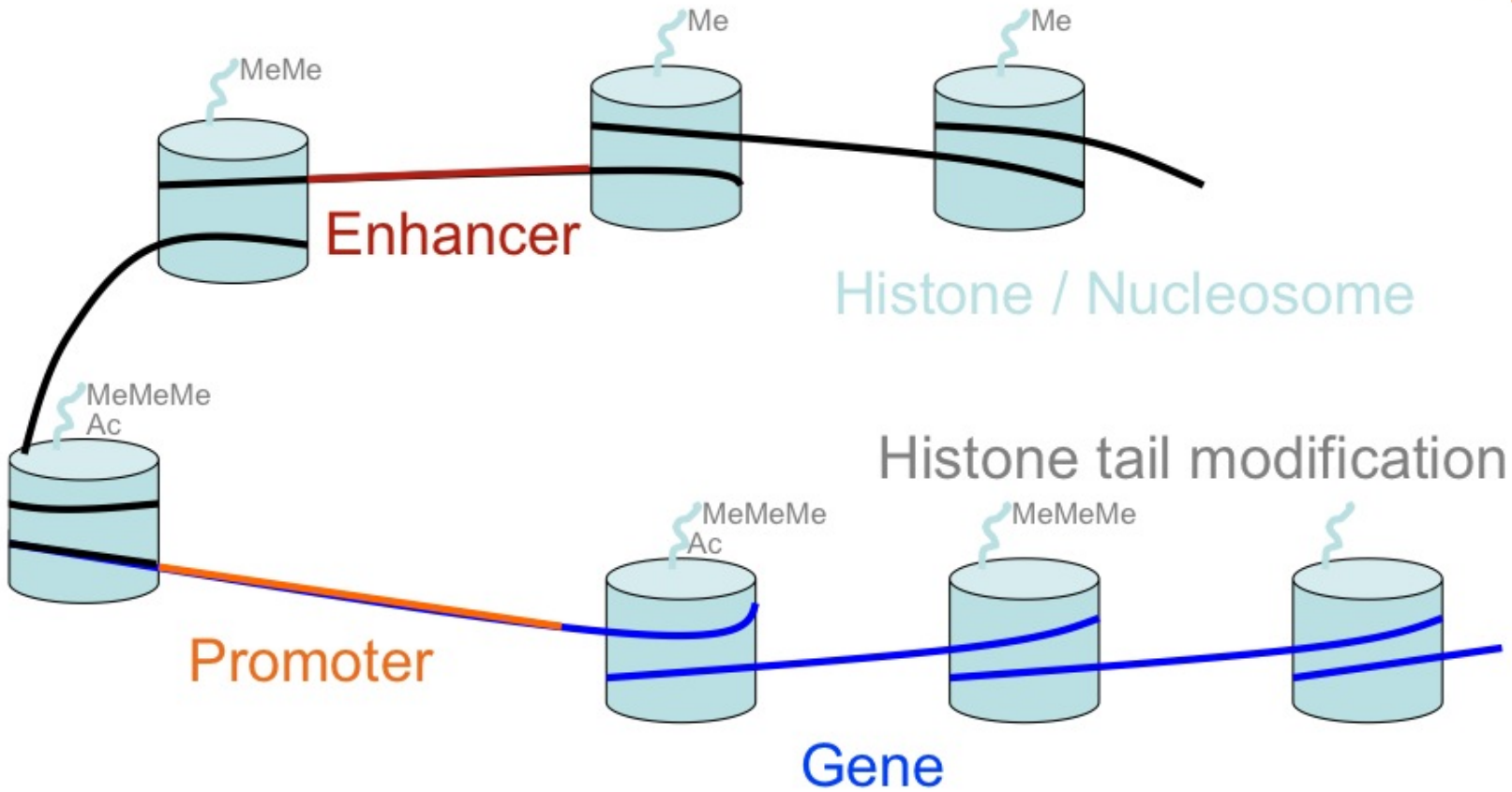
- All the ChIP you need to know
- Mapping reads
- Taking Advantage of read directionality
- Shifting reads
- Motif calling & Gene association

ChIP-seq

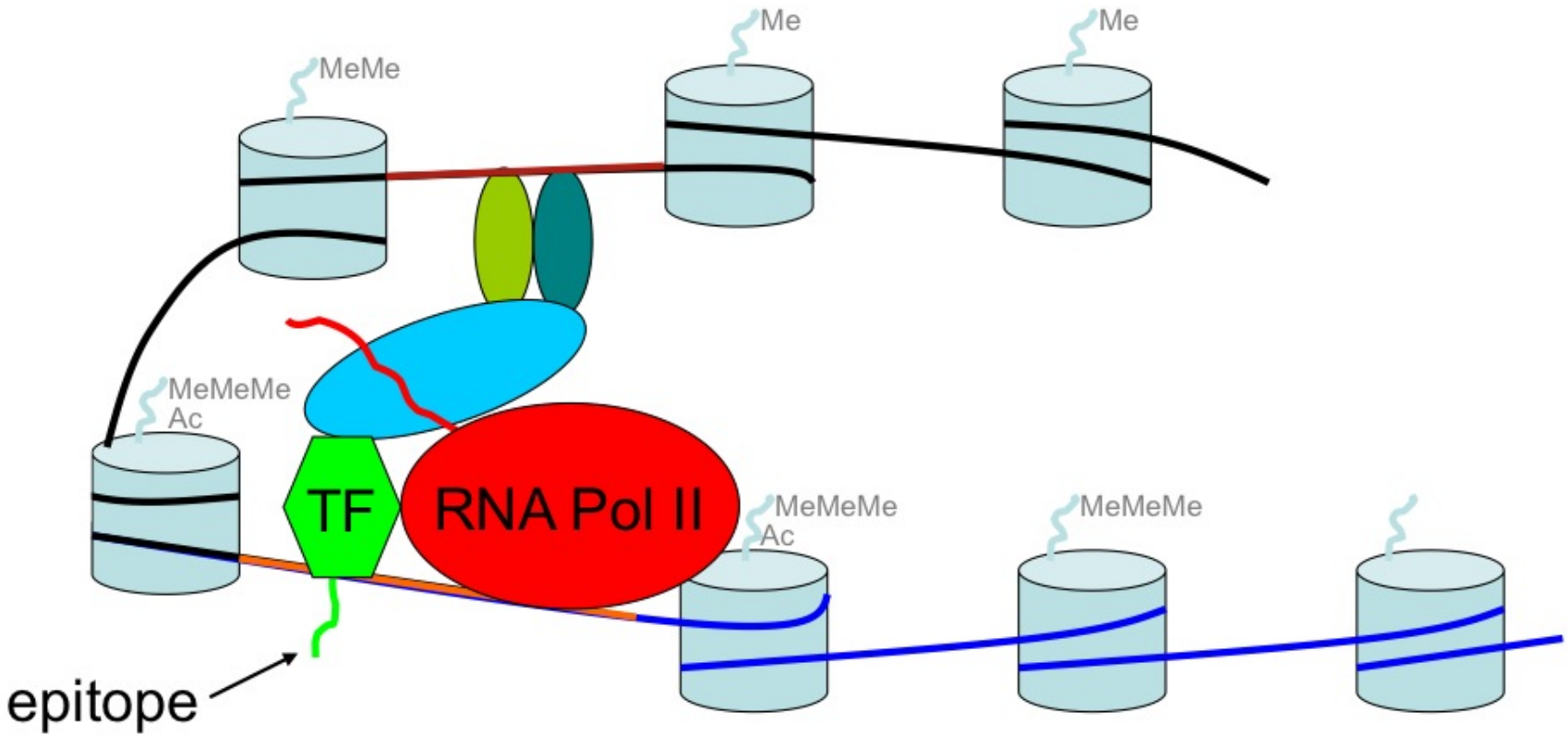
A digital counting method to score site occupancy by DNA binding proteins



Chromatin Immunoprecipitation (ChIP

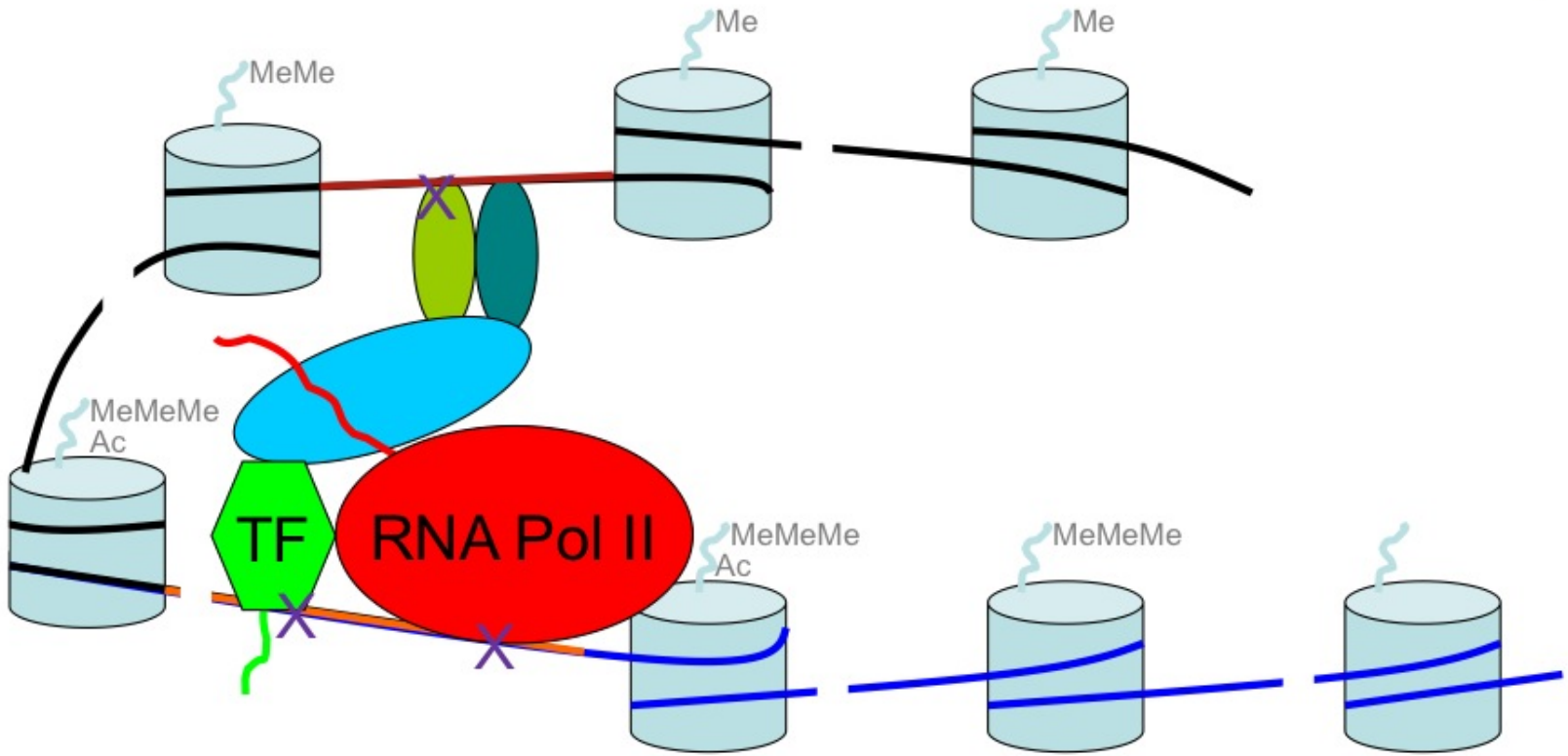


Chromatin Immunoprecipitation (ChIP



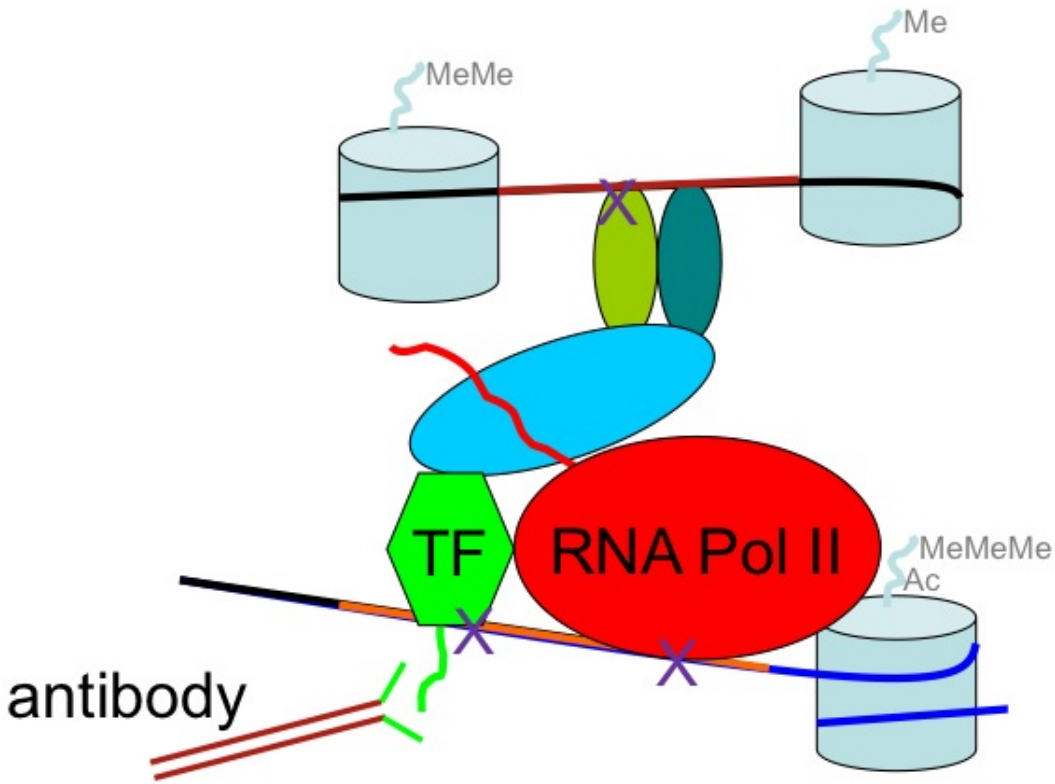
Transcription Factors + histones + DNA = chromatin

Chromatin Immunoprecipitation (ChIP

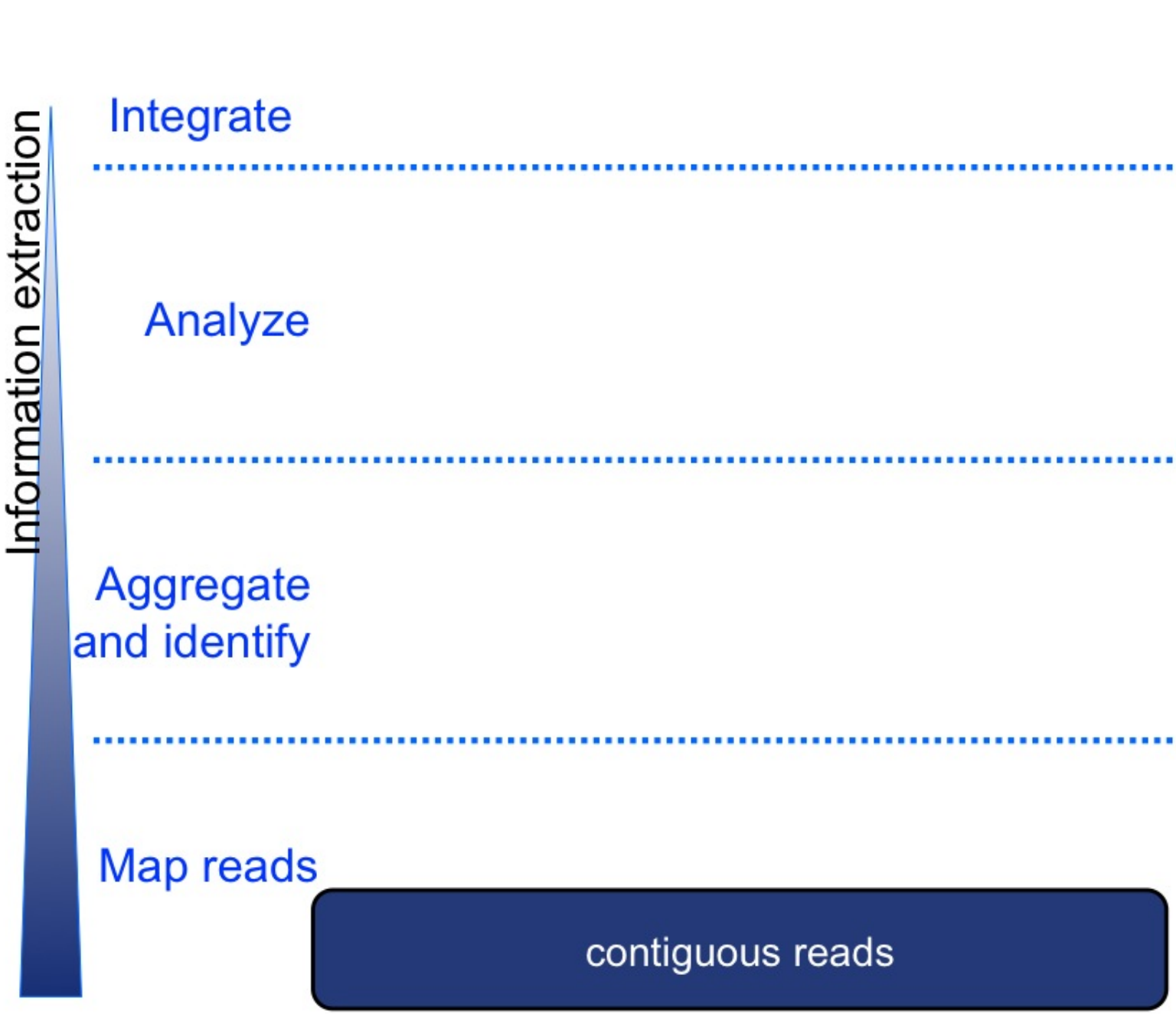


1. Crosslink with formaldehyde
2. Fragment the DNA using sonication or digestion to an average fragment size of 200 (~ 1 nucleosome)

Chromatin Immunoprecipitation (ChIP



3. Use antibody specific to a factor to retrieve DNA fragments that are (not necessarily directly) bound.
4. Reverse crosslinks and sequence ends of fragments.

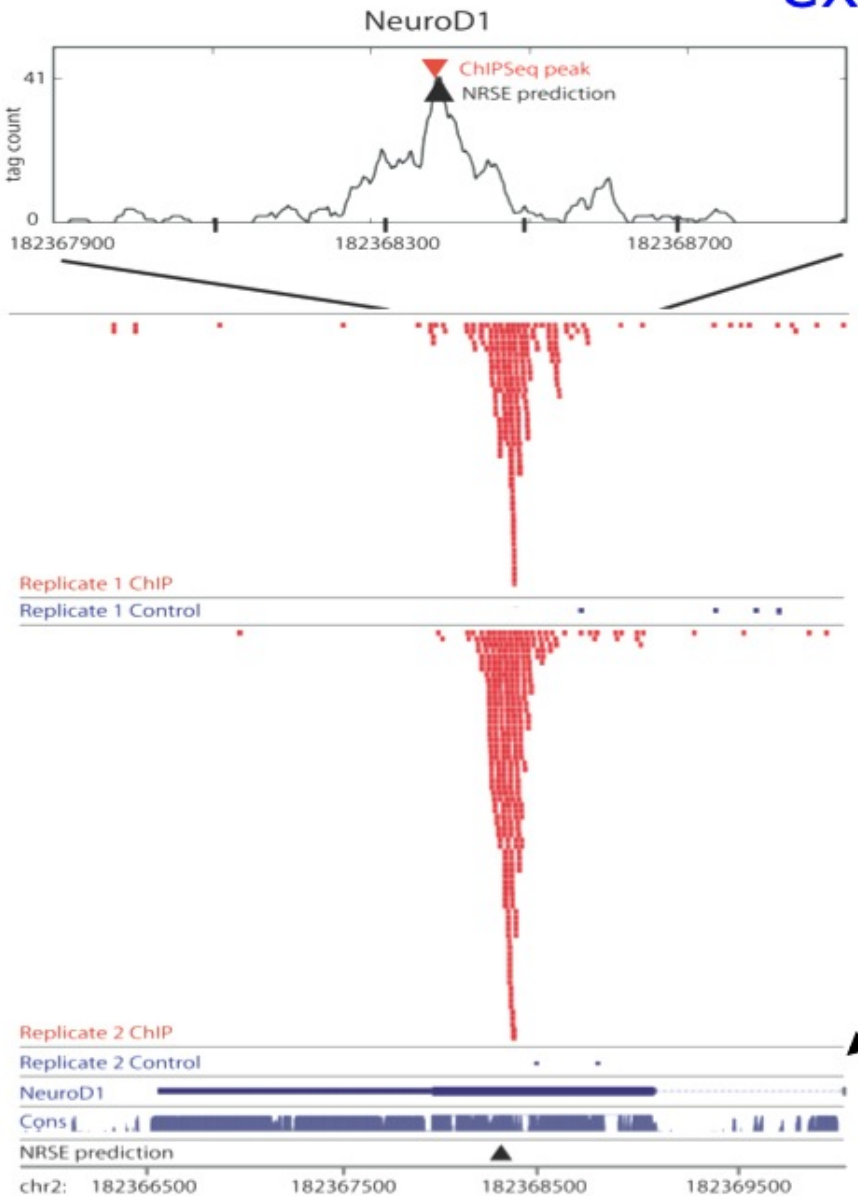

Integrate
Analyze
Aggregate
and identify
Map reads
Information extraction
contiguous reads

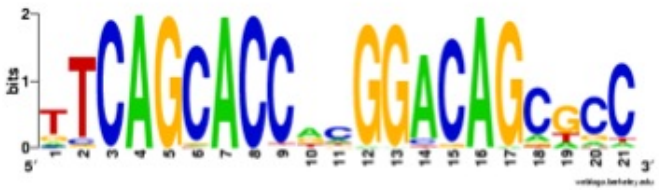
Mapping the reads



- Once the sequencer is run, we will get reads that have possibly been mapped against a genome assembly. Should you remap the reads ?
- Each read mapper makes specific assumptions about:
  - Read quality & number of mismatches allowed
  - Handling of reads that map to more than one place in the genome
- Genome build changes can affect read coordinates drastically
- These mapping settings affect which signals are detectable, the so-called “read mappability”.
- For maximum control, consider remapping from raw reads

ChIP-Seq identifies NRSF occupancy in NeuroD1 exon

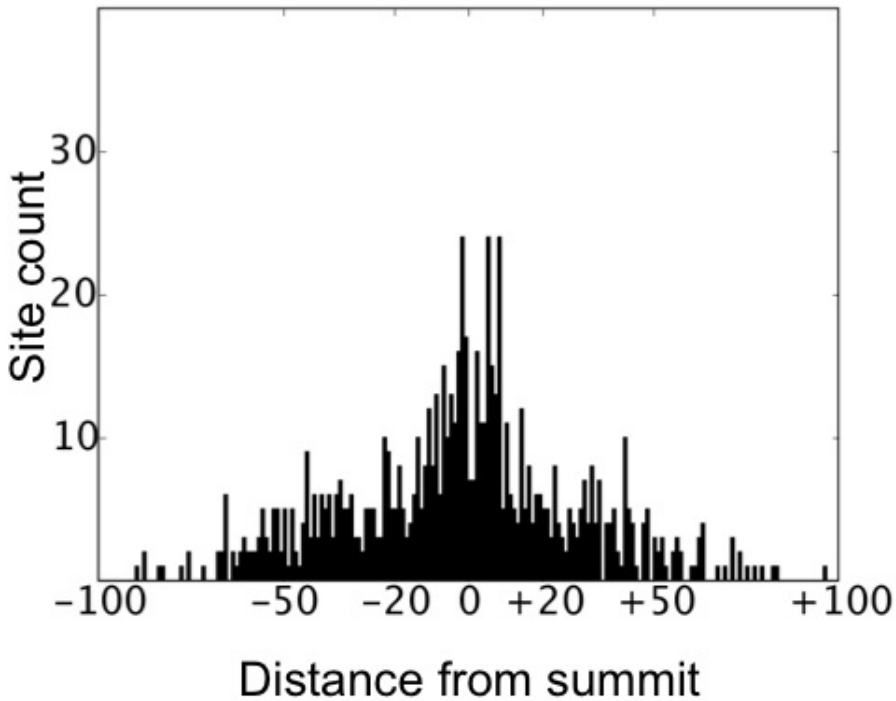




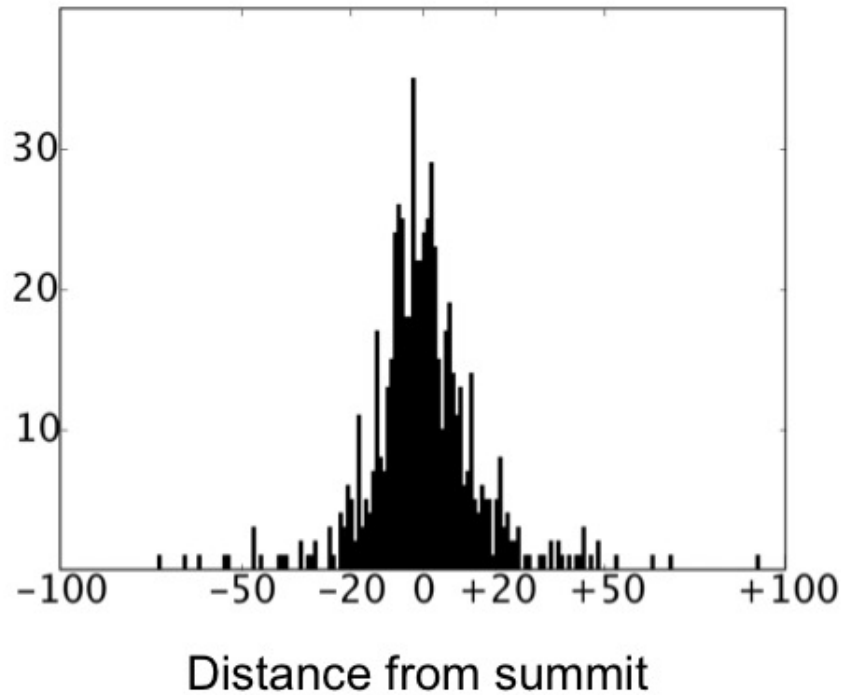
Controls

(Johnson et al, 2007)

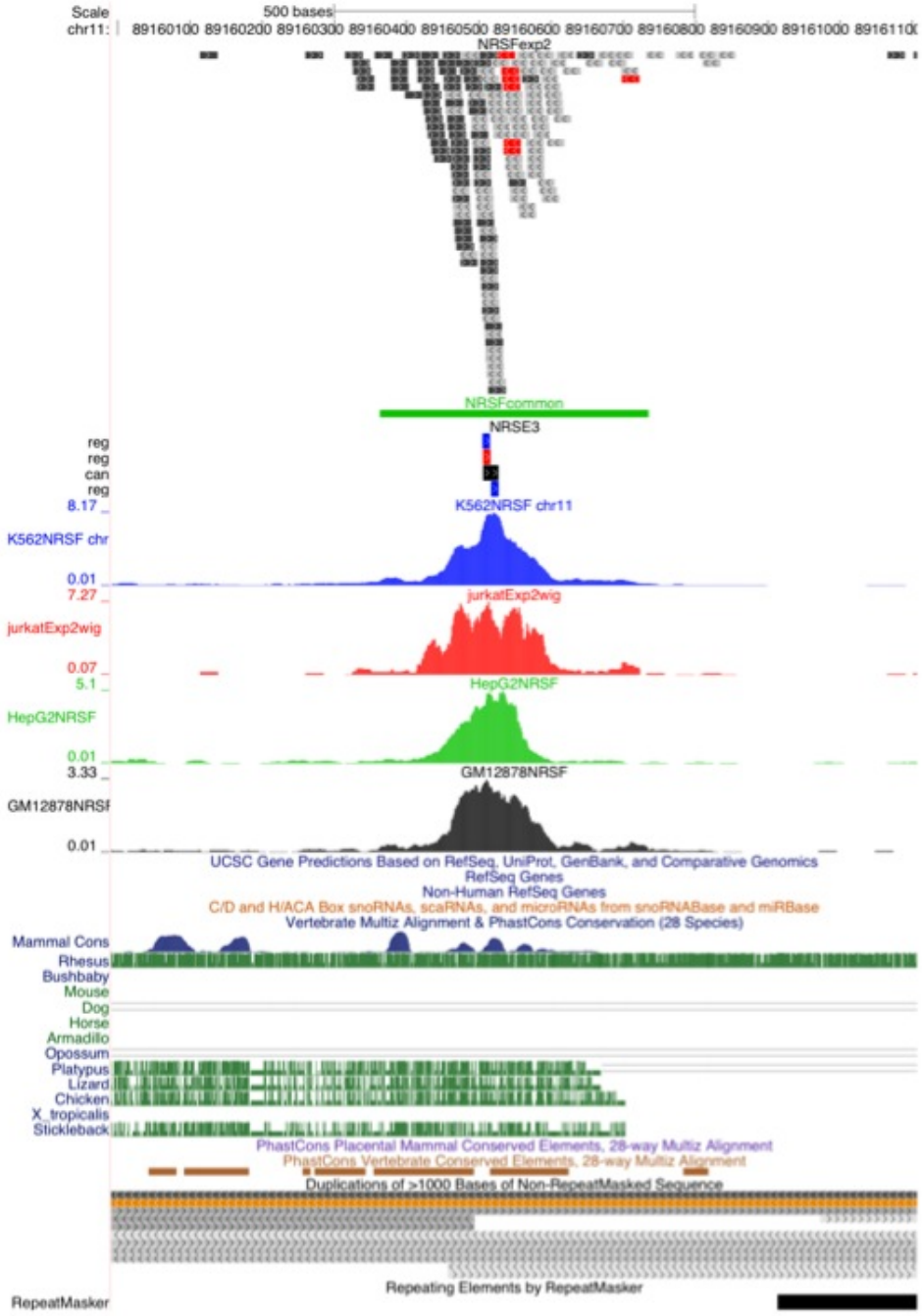
Increasing # of reads improves resolution



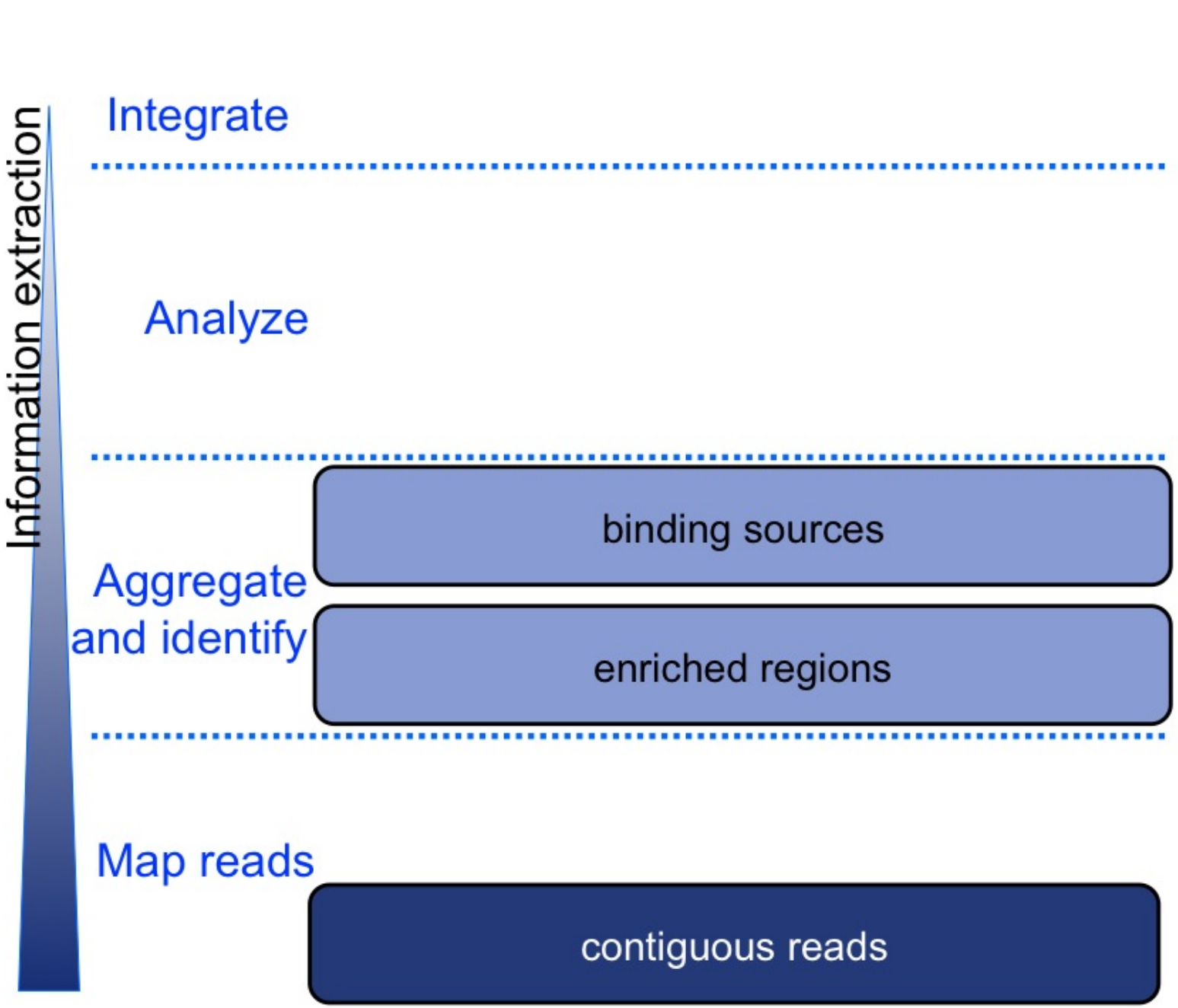
1.7 Million reads



6.4 Million reads



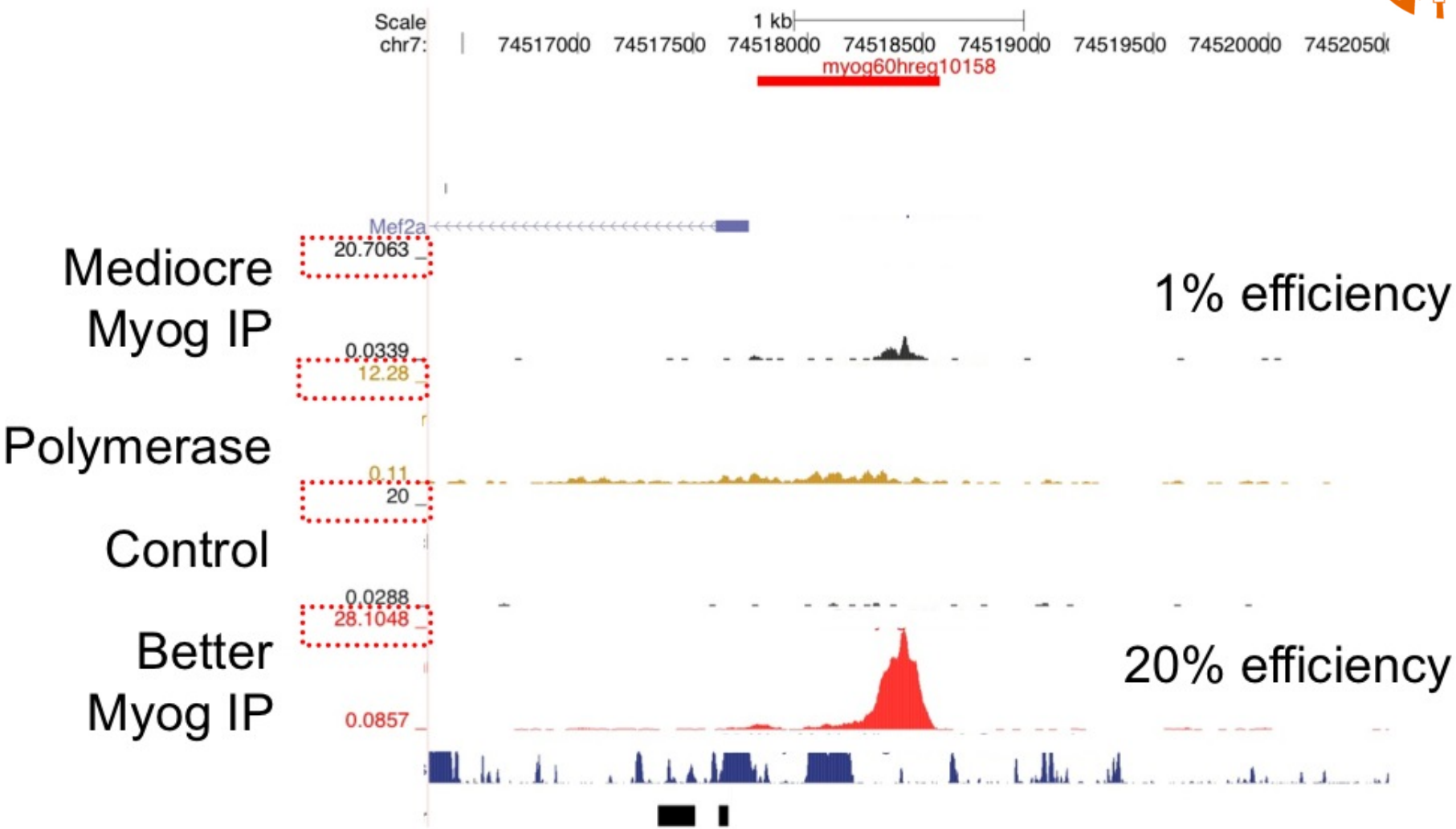
Multireads matt
for ChIP-seq too

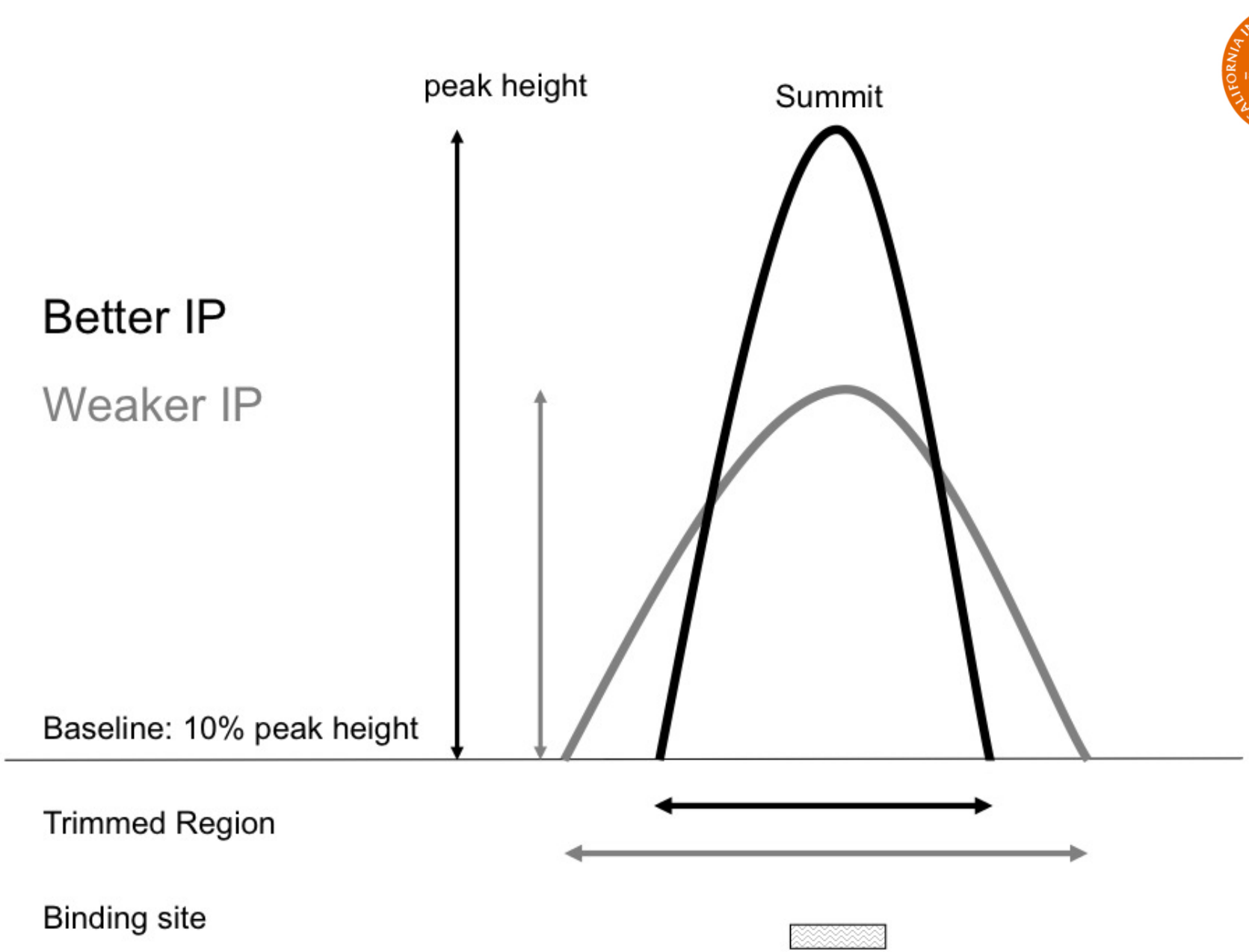

Information extraction
Integrate
Analyze
Aggregate and identify
binding sources
enriched regions
Map reads
contiguous reads

How do we measure the success of a ChIP-seq experiment ?

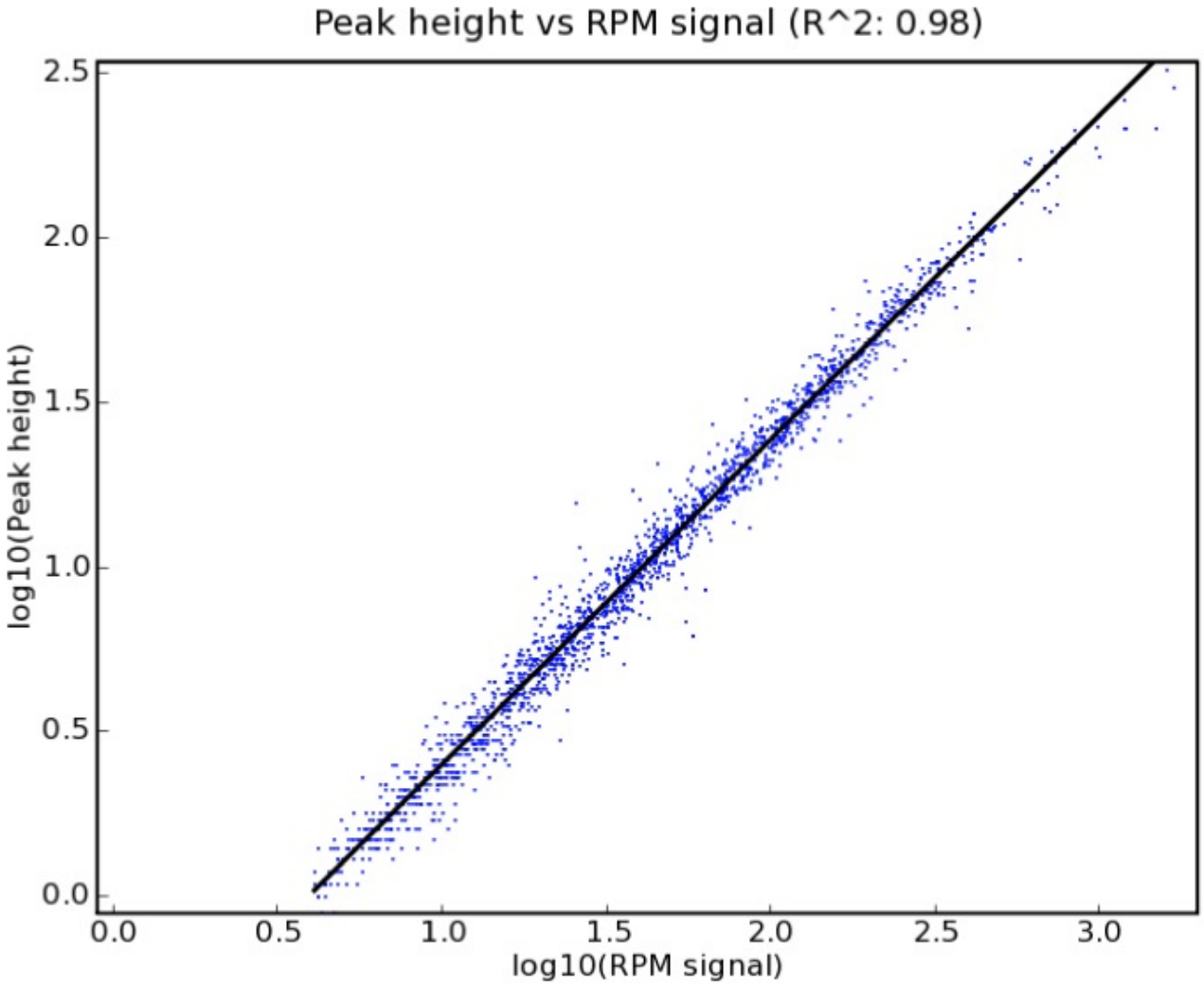


- The number of peaks we recover
  - Most factors have thousands of binding sites
- What fraction of the reads come from these peaks
  - For a good antibody, 10% would be good
  - Hence, most of the reads are from the background!
- Whether we find an explanatory motif, when applicable
- Whether the called peaks and discovered motifs match the existing literature
  - Hopefully, the antibody doesn't cross-react


Scale
chr7:
1 kb
74517000
74517500
74518000
74518500
74519000
74519500
74520000
myog60hreg10158
Mef2a
Mediocre Myog IP
20.7063
0.0339
12.28
1% efficiency
Polymerase
0.11
20
Control
0.0288
28.1048
Better Myog IP
0.0857
20% efficiency


peak height
Summit
Better IP
Weaker IP
Baseline: 10% peak height
Trimmed Region
Binding site

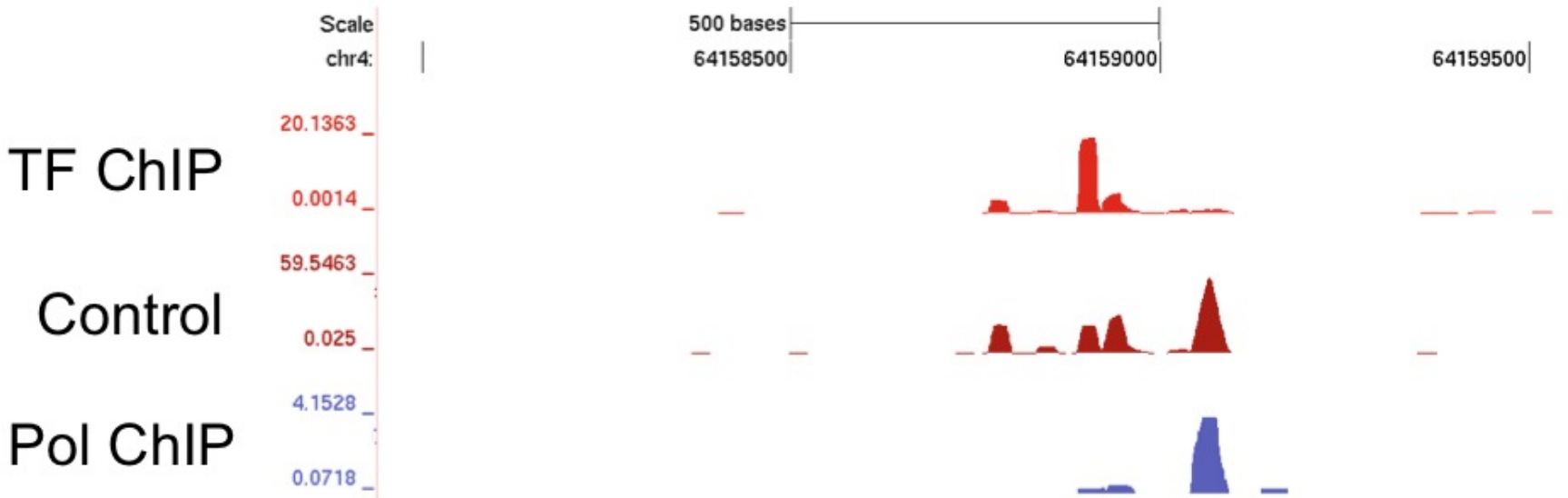
Peak height and signal density correlat

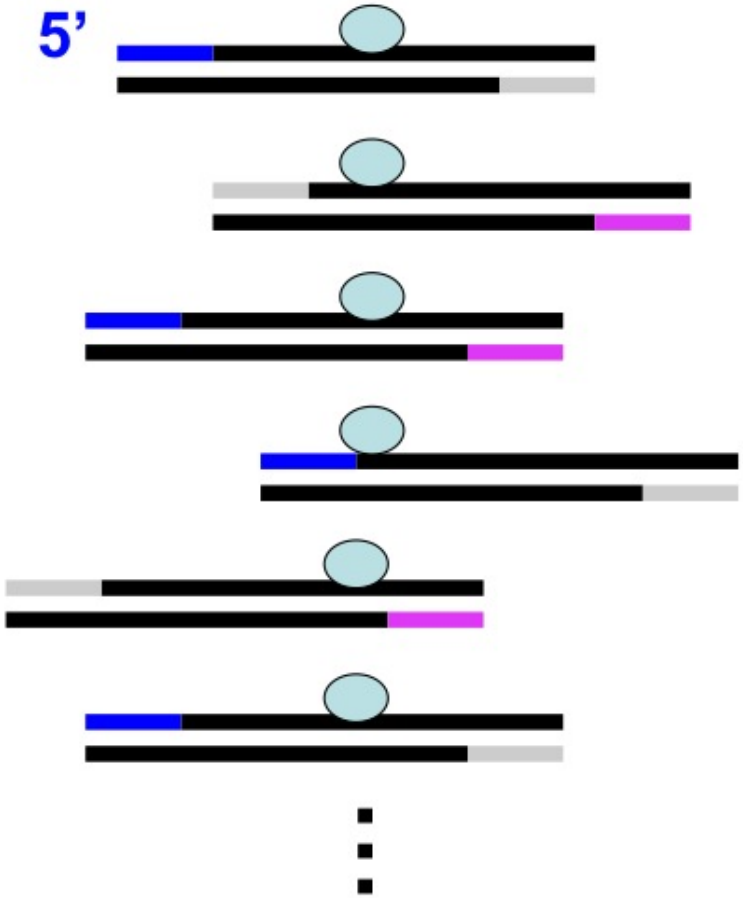


Why do we need a control ?



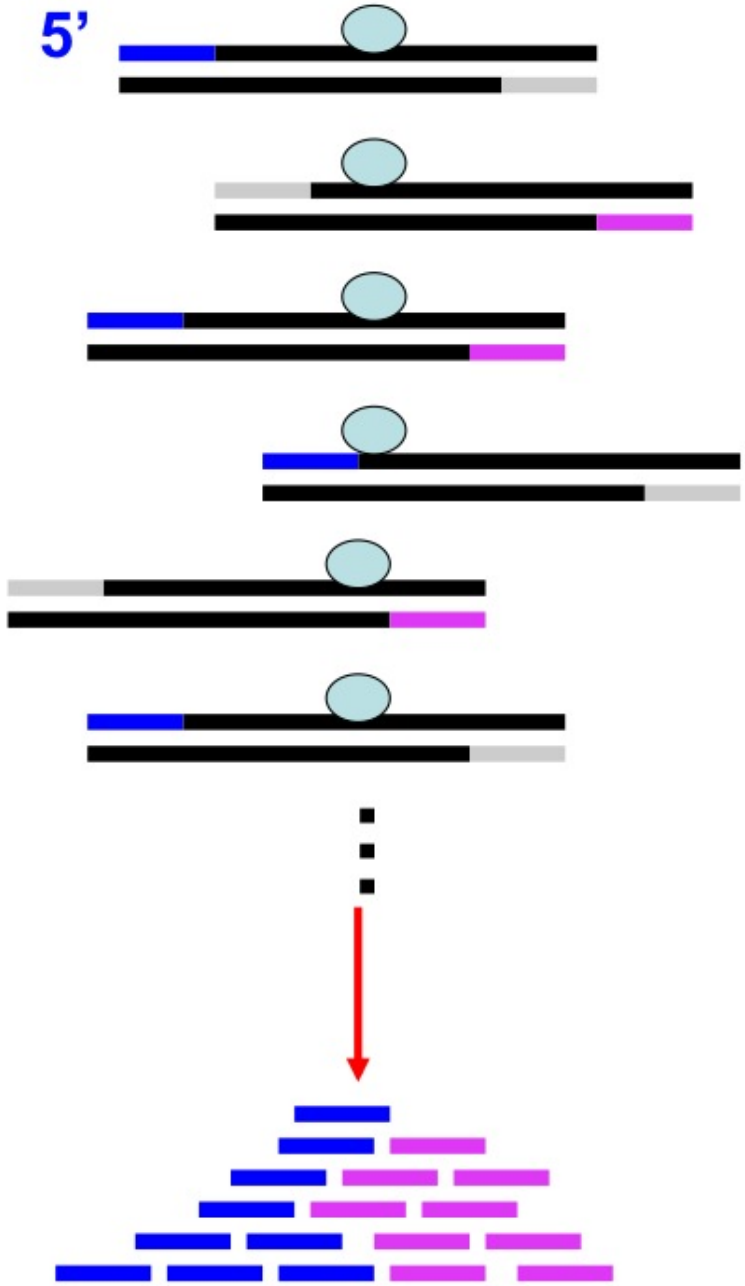
- A significant fraction of the signal is coming from the background.
- Sources of artifacts:
  - Mismapping
  - Repeats





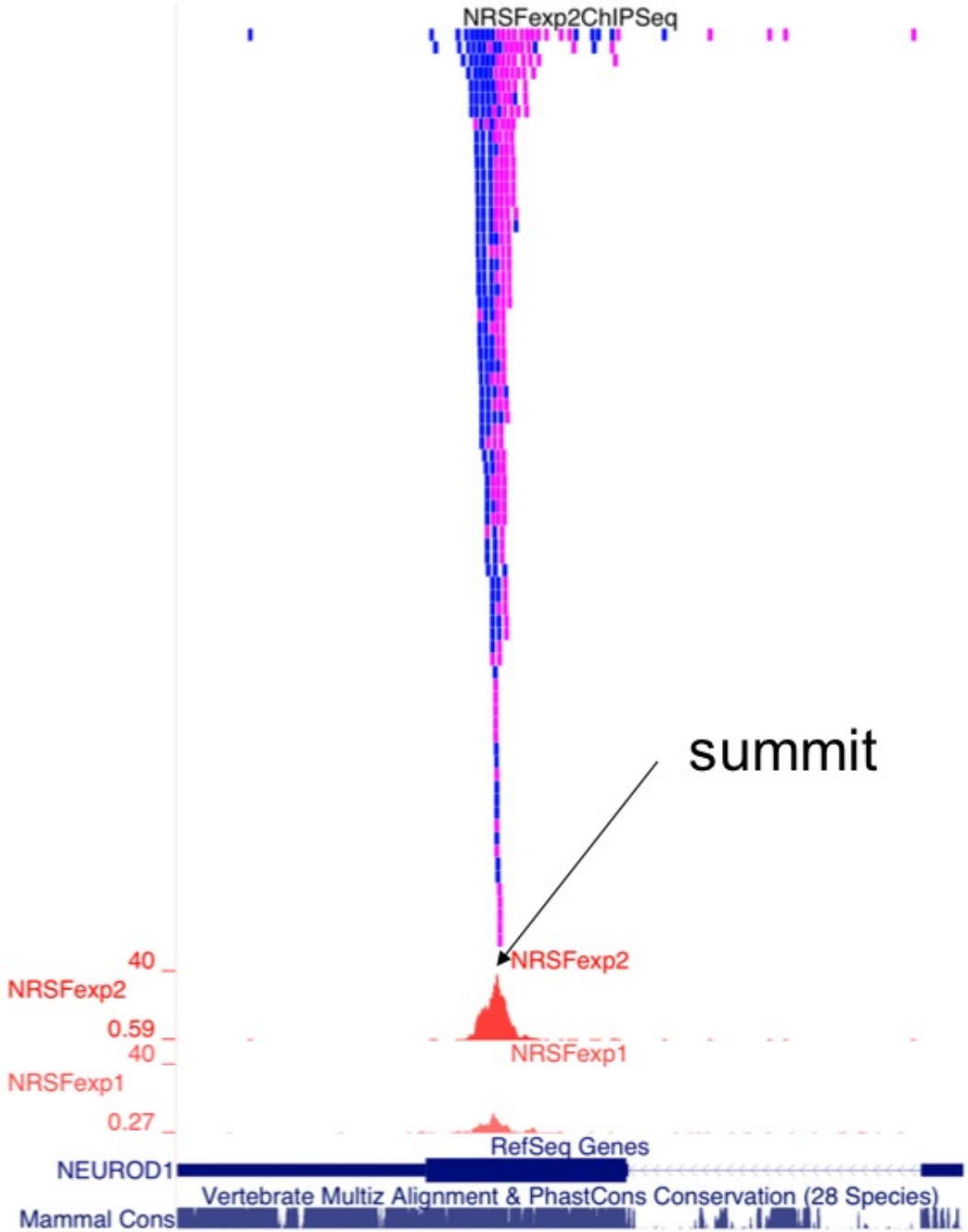


Since Illumina uses polymerase, we are always sequencing from 5’ end of fragment (blue or purple), typically observing one end of the fragment.



Since Illumina uses polymerase, we are always sequencing from 5’ end of fragment (blue or purple), typically observing one end of the fragment.

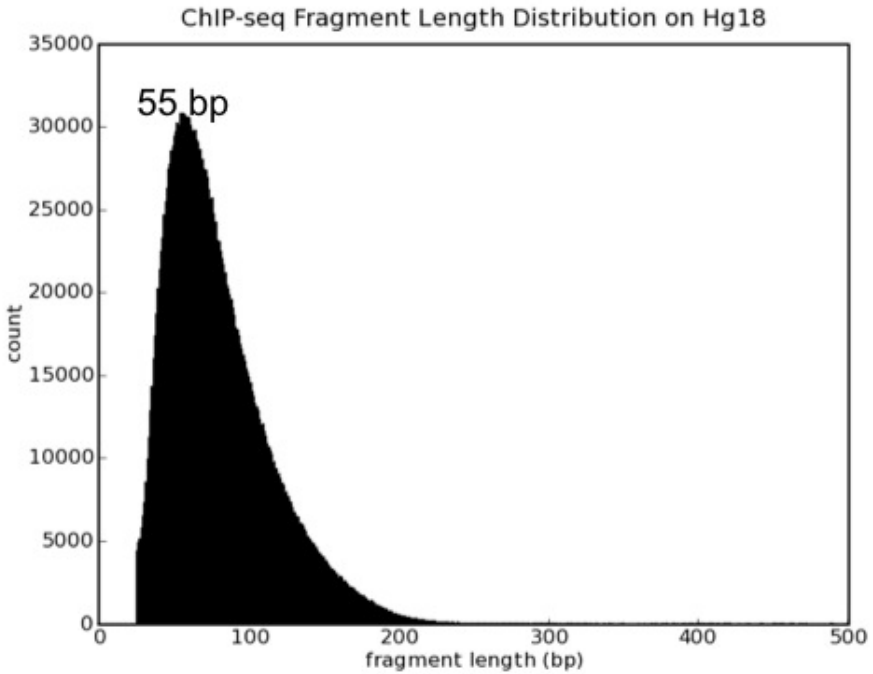
Static sites should be visible as blue to yellow transitions


NRSFexp2ChIPSeq
summit
40
NRSFexp2
0.59
40
NRSFexp1
0.27
NRSFexp2
NRSFexp1
RefSeq Genes
NEUROD1
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons

What is the impact of fragment length



- The larger the fragments sequenced, the further apart
- Three general methods:
  - Use just raw reads
  - Extend reads to be the size of the expected fragment length
  - Estimate half of the fragment length size and "shift" the reads by that amount.
- Best strategy depends on ChIP-seq peak type.
- Don't take fragment size for granted!

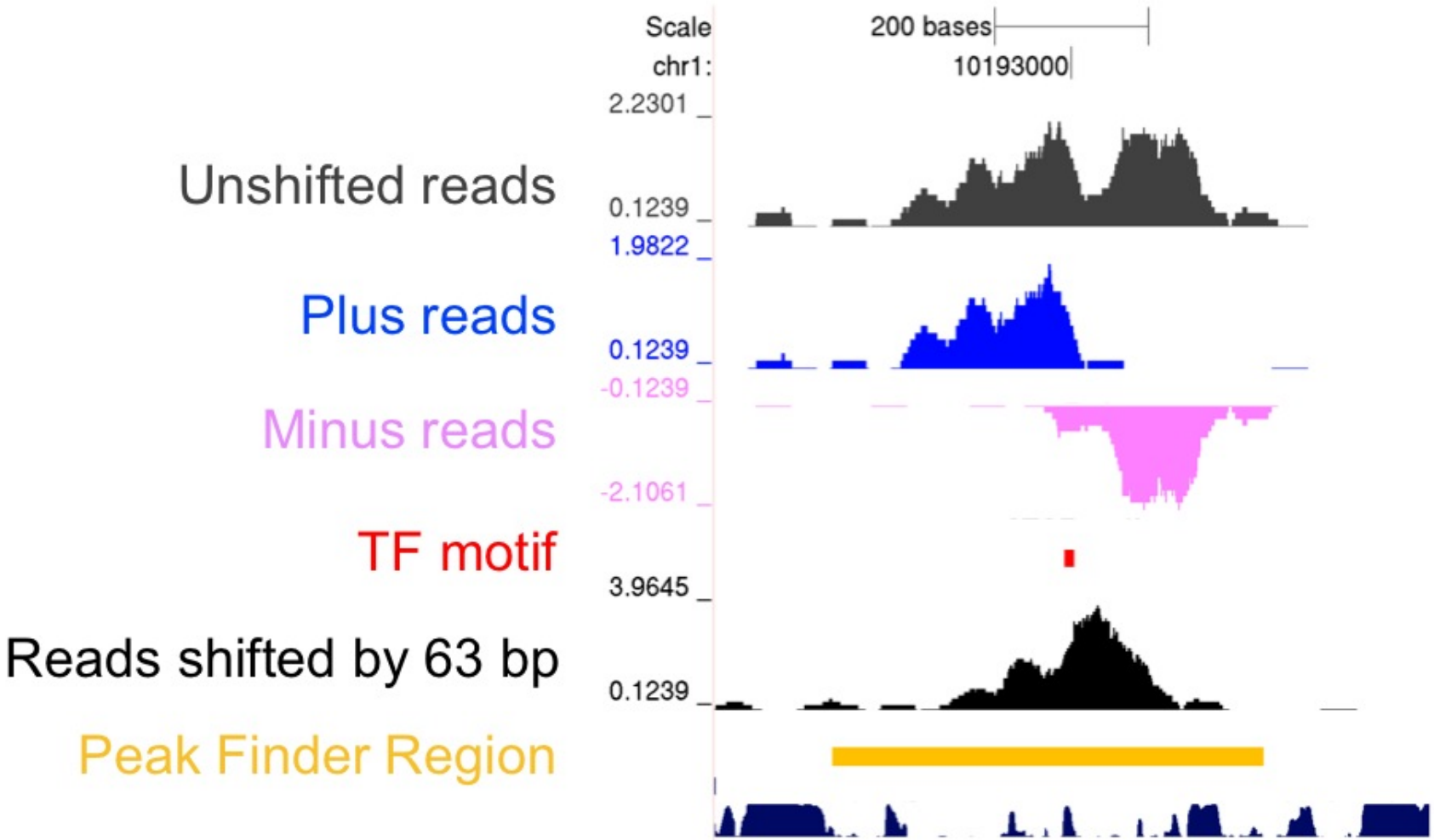


Original Gel Cut: 150-300 bp

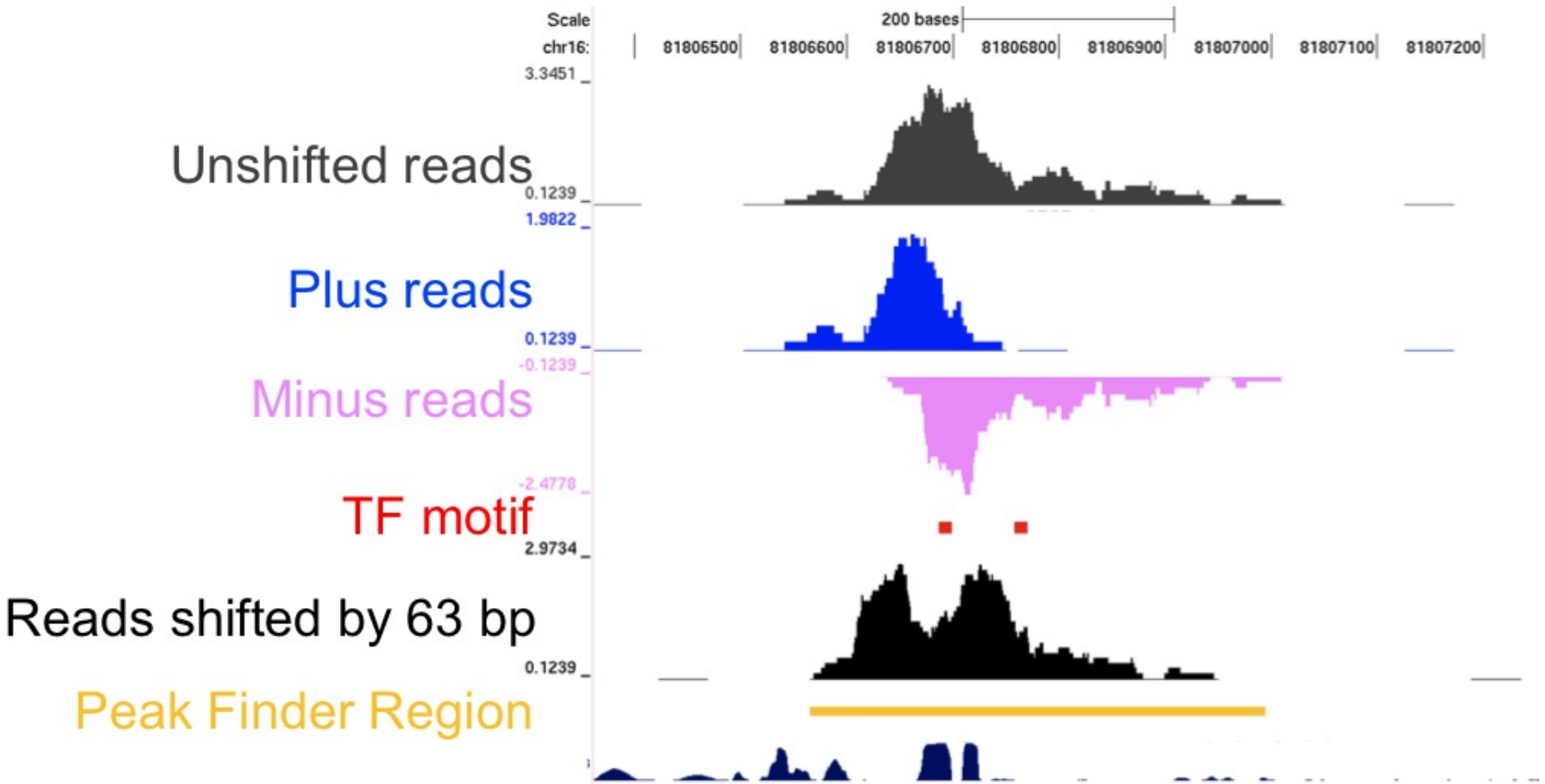
Narrow ChIP-seq peaks



Single source such as a Transcription Factor binding site



Multiple sources closer than the average fragment length are more difficult to resolve



Medium ChIP-seq peaks



Some active promoter and transcription associated factor

Plus reads

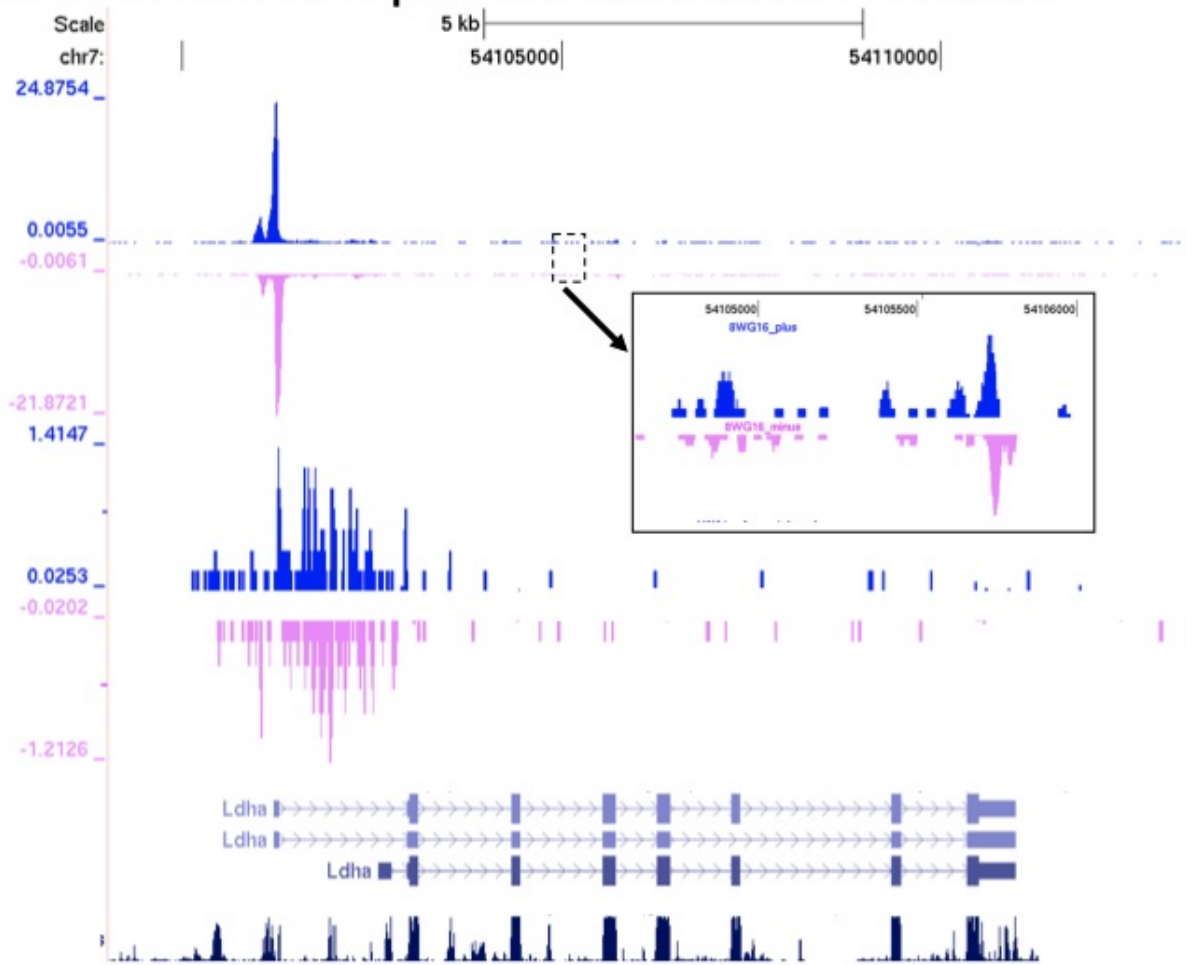
RNA Pol II 8WG16

Minus reads

Plus reads

Histone H3K4me3

Minus reads

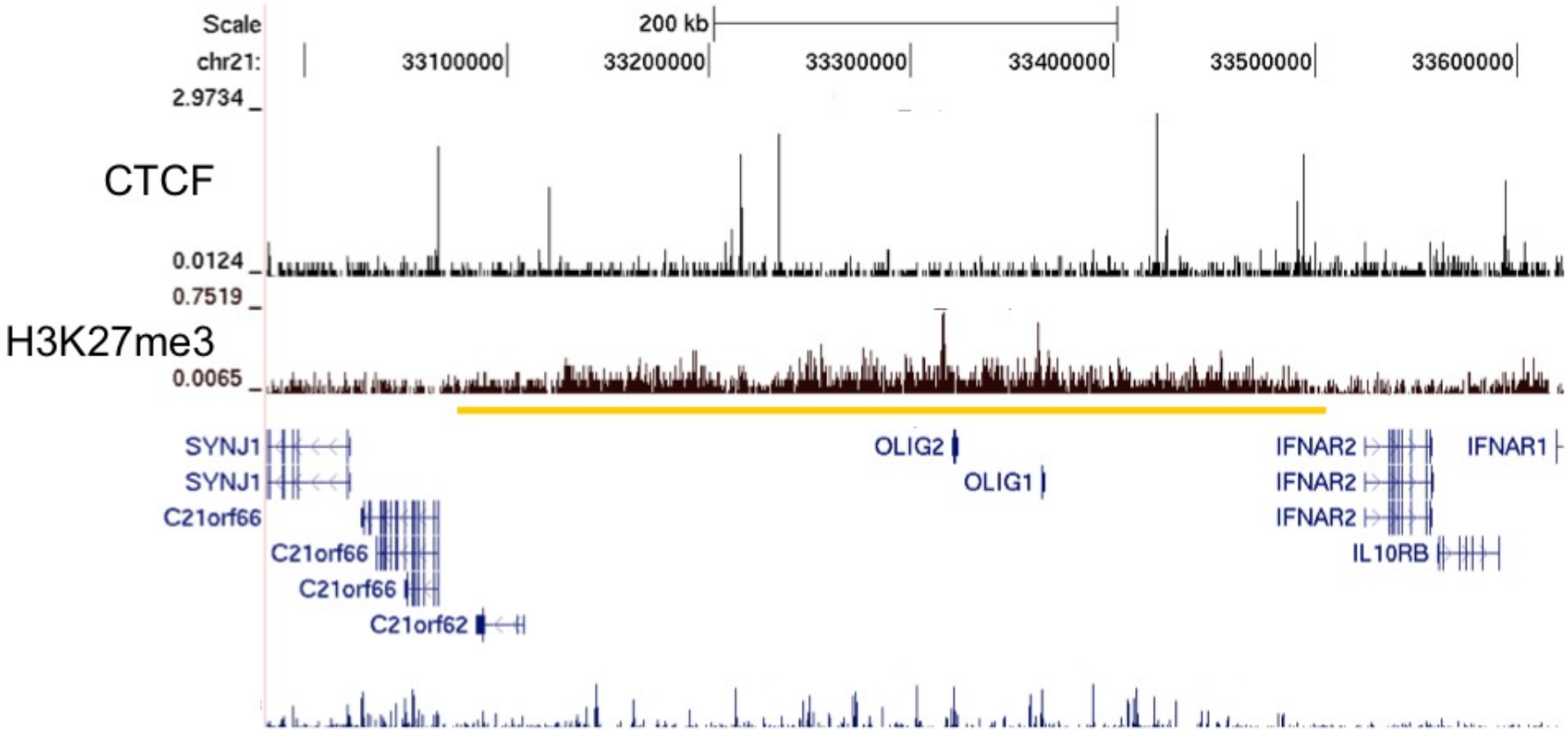


Can either shift or extend

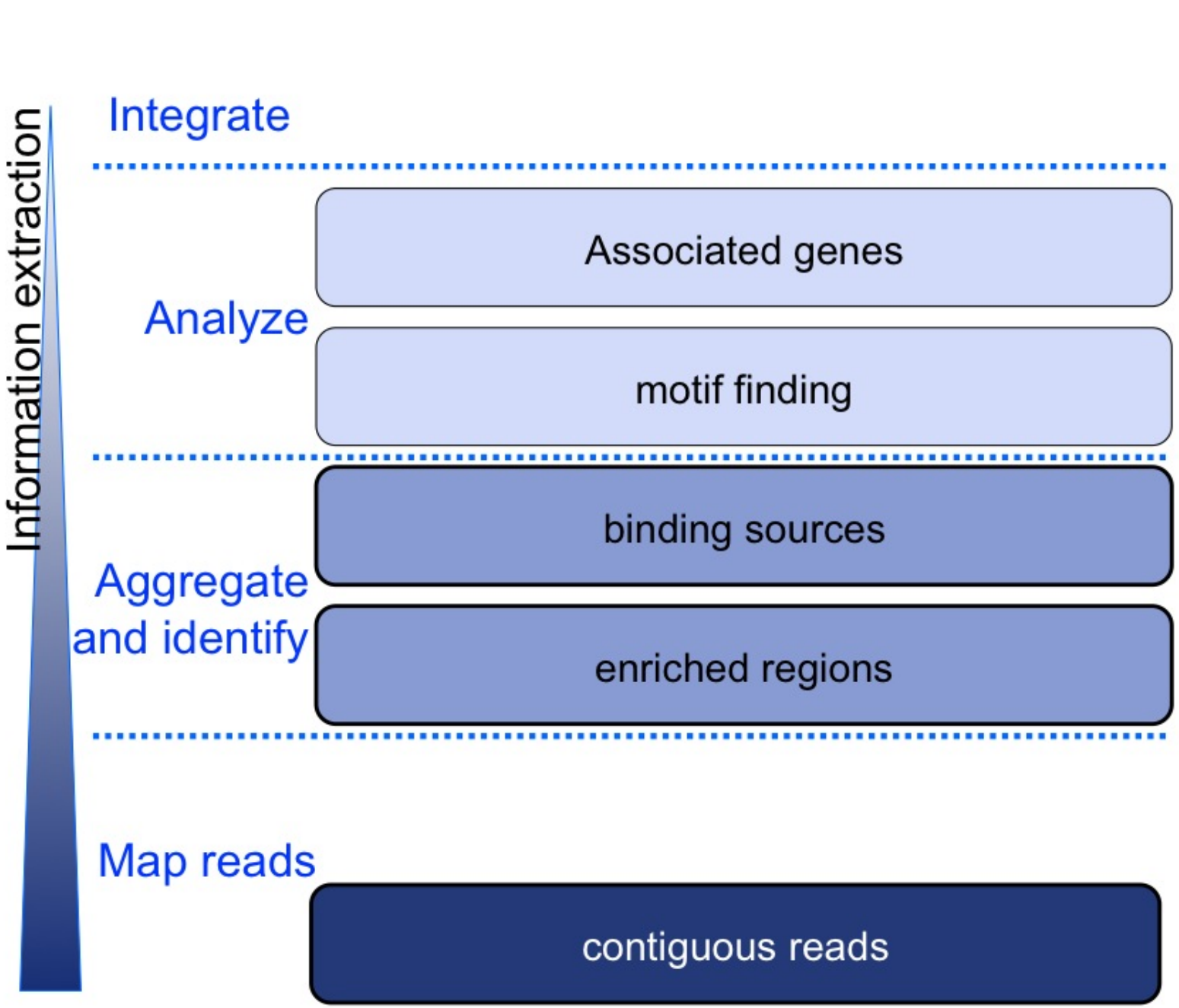
Broad ChIP-seq peaks



Some repressive chromatin histone modification marks



No benefit to shifting or extending - use a sliding window


Information extraction
Integrate
Analyze
Associated genes
motif finding
Aggregate and identify
binding sources
enriched regions
Map reads
contiguous reads

Finding motifs

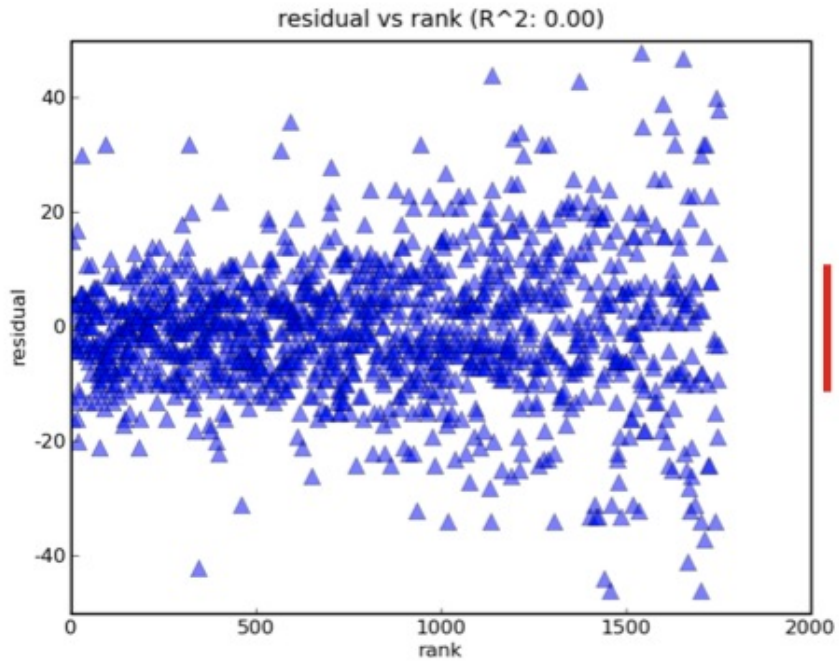


- Once we have regions and summits, we can retrieve the associated DNA and run them through a motif finder such as Meme to discover one or more motifs.
- Can limit ourselves to +/- 50 bp from summit
- If there are large numbers of site, consider stratifying, using peak height or peak total signal for ranking, e.g:
  - 1000 regions with high signal
  - 1000 regions with medium signal
  - 1000 regions with low signal
- Rescan all regions with discovered motifs

How close is the motif to the summit

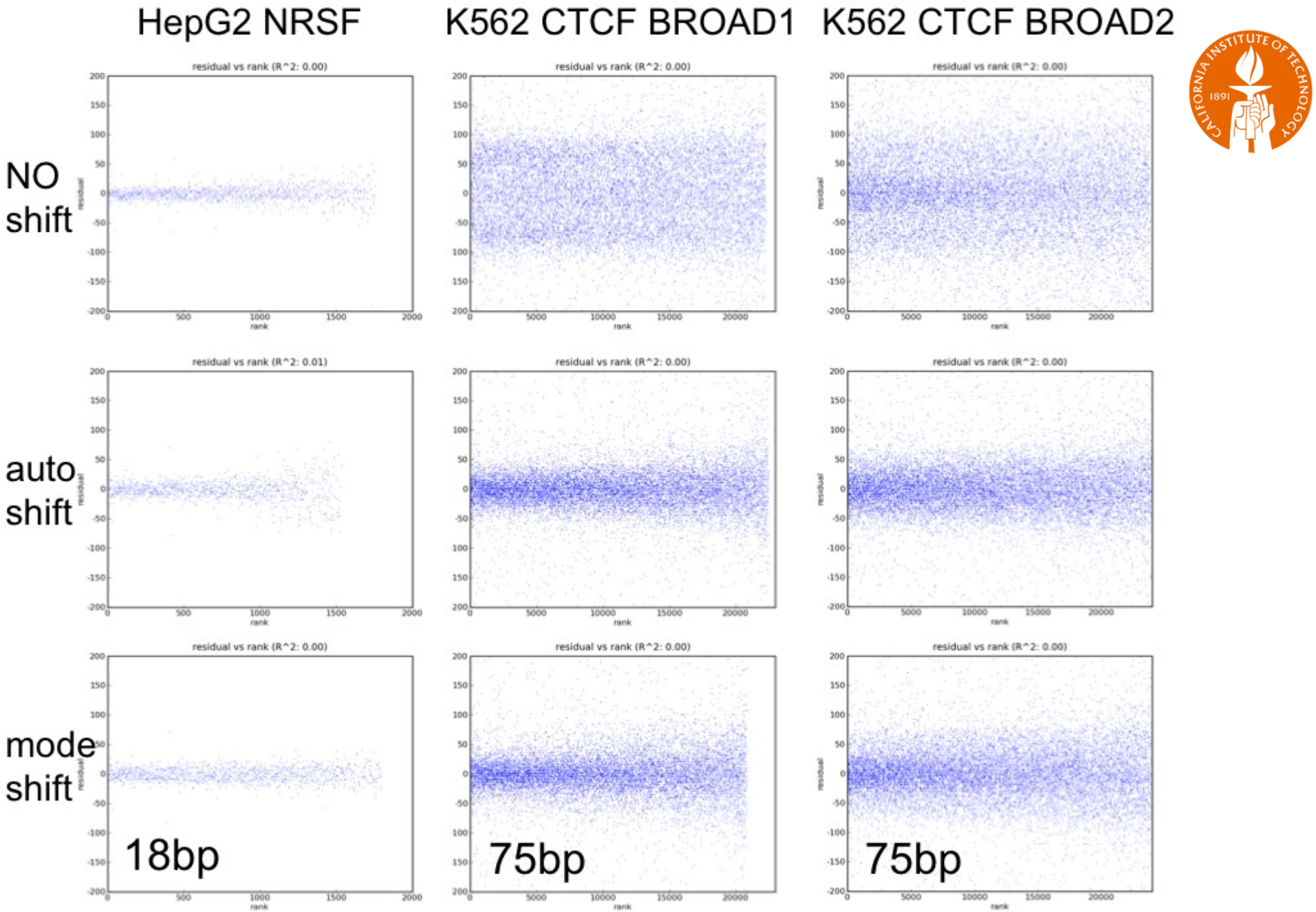


Residual = midpoint of motif - summit



NRSF motif size

The summit is typically less than 30 bp
from the explanatory motif


HepG2 NRSF
K562 CTCF BROAD1
K562 CTCF BROAD2
NO shift
auto shift
mode shift
residual vs rank (R^2: 0.00)
residual vs rank (R^2: 0.01)
residual vs rank (R^2: 0.00)
residual vs rank (R^2: 0.00)
residual vs rank (R^2: 0.00)
residual vs rank (R^2: 0.00)
residual vs rank (R^2: 0.00)
residual vs rank (R^2: 0.00)
residual vs rank (R^2: 0.00)
rank
residual
18bp
75bp
75bp

Associating genes with peaks



- Two simplest strategy are:
  - Nearest gene within a radius
  - All genes within a radius
- Different factors are more or less bidirectional
- Biggest problem is to assign binding at candidate distal enhancers
  - Enhancer for one gene could be in the intron of another gene
  - Binding site that is far from a known gene could really be in the promoter of an unknown transcript.

References



Original Papers:

Johnson, DS, Mortazavi, A, Myers, RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science.

Valouev, A, Johnson, DS, Sundquist, A, Medina, C, Anton, E, Batzoglou, S, Myers, RM, Sidow A (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-seq data. Nature Methods.

Rozowsky, J, Euskirchen, G, Auerbach, RK, Zhang, ZD, Gibson, T, Bjornson, R., Carriero, N, Snyder, M, Gerstein, MB (2009). PeakSeq enables systematic scoring of experiments relative to controls. Nature Biotech.

Reviews:

Park, PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nature Reviews Genetics.

Pepke, S, Wold, B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nature Methods.

Roadmap

Introduction
- Quantitation in biology
- The Rise of the Sequencers
- Sequence-counting Assays
- Important Considerations

Transcriptome Analysis
- Motivation
- Common Steps and Challenges
- RNA-seq Read Mapping
- Quantitation & Diff. Expression
- Transcript Reconstruction
- Transcript Assembly

ChIP-seq
- Introduction
- Mapping the reads
- Taking Advantage of Directionality
- Shifting Reads
- Motif Finding

Integration
- The Combinatorial Problem
- Analyzing Chip-seq Datasets Jointly
- Segmenting the Genome
- Learning Regulatory States
- Joint Analysis of Chip- and RNA-Seq

Introduction

- The combinatorial problem
- The gene assignment problem
- Analyzing multiple ChIP-seq datasets jointly
- Segmenting the genome
- Learning regulatory states
- Joint Analysis of ChIP-seq and Analysis
- An example: Analysis of ChIP-seq and RNA-seq data using Self-Organizing Maps



The combinatorial problem

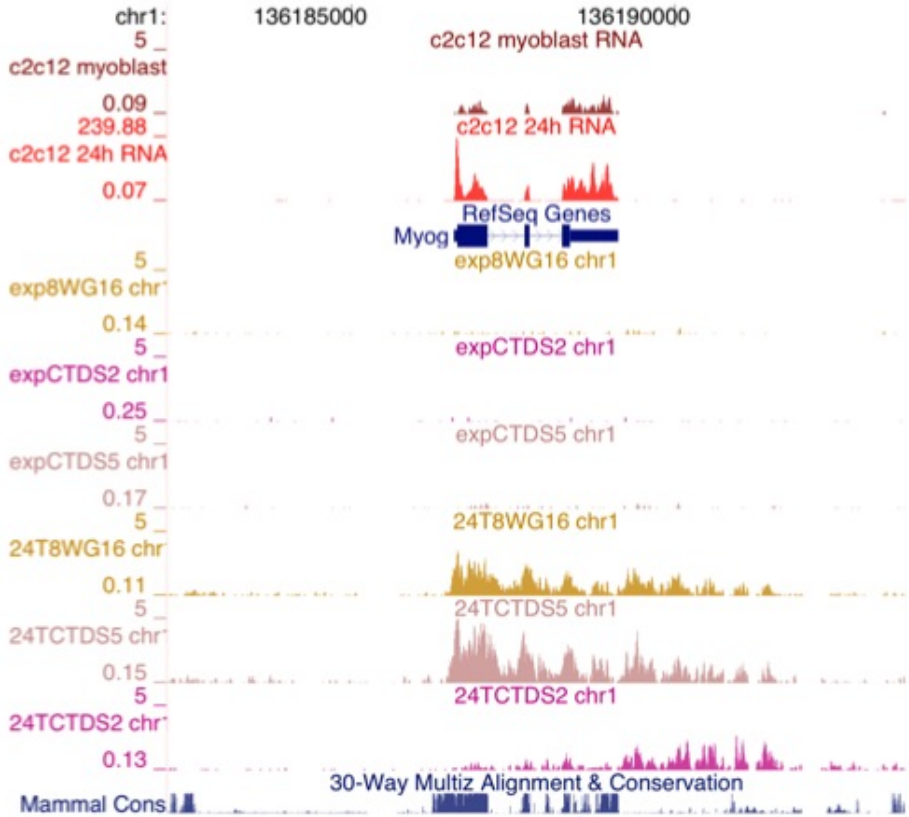
- It is now relatively easy to generate dozens of ChIP-seq and/or RNA-seq for a biological sample of interest and to analyze them singly.

- The problem is exponentially more difficult as we analyze multiple datasets across multiple timepoints and/or cell types

- many custom methods, few tools

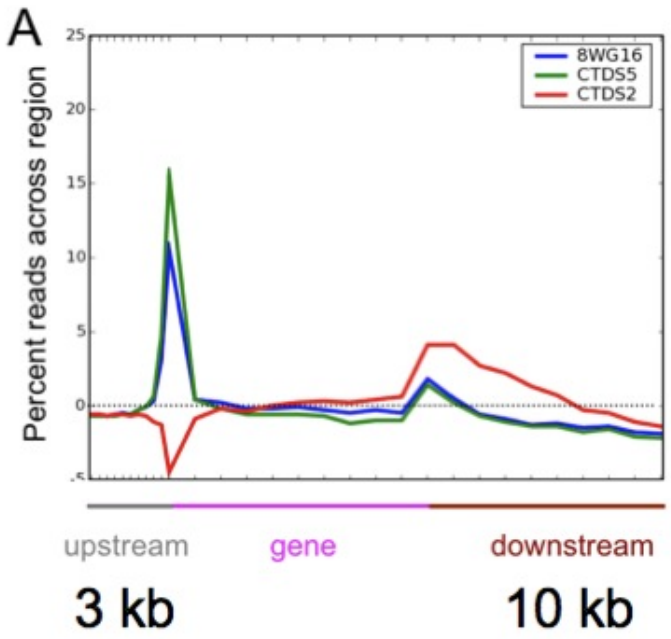
Given N factors, each of which could have M states, then each region of the genome could be in any of M_k^N states.

  - Which are the interesting ones ?
  - What are the region boundaries ?



The gene assignment problem

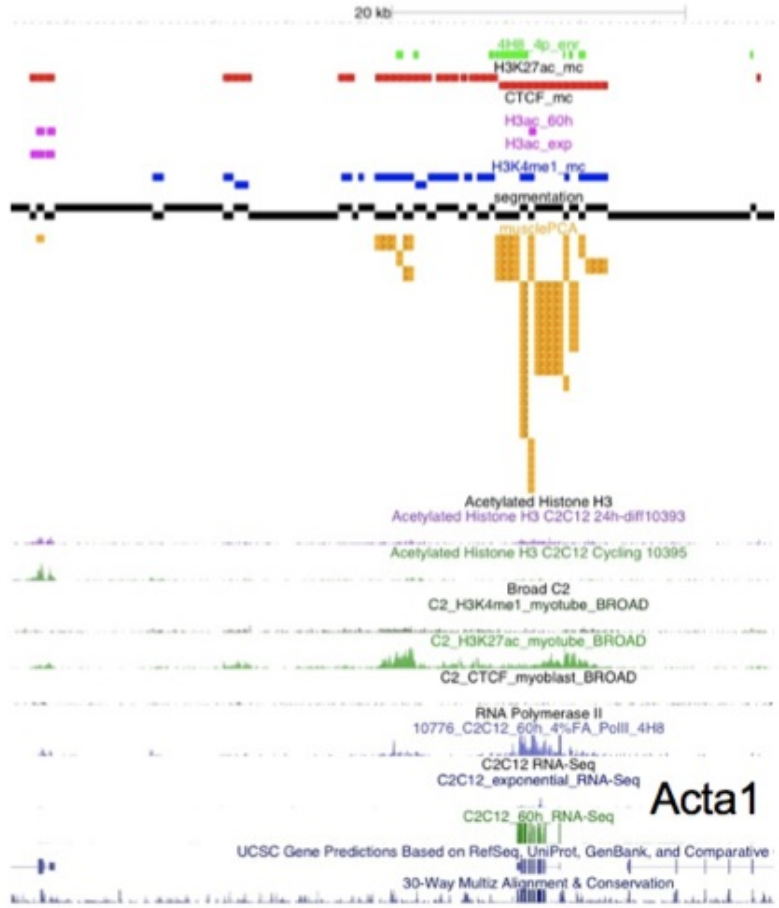
- Gene boundaries are the natural boundaries for annotation.
  - Distance from TSS is a popular representation
- But gene loci come in all sizes (1 kb - 1 Mb), so such display require gene-part binning.
- Unfortunately, cis-regulatory elements do not necessarily regulate the nearest gene



Analyzing multiple ChIP-seq datasets jointly

Most integrative analyses boil down to:

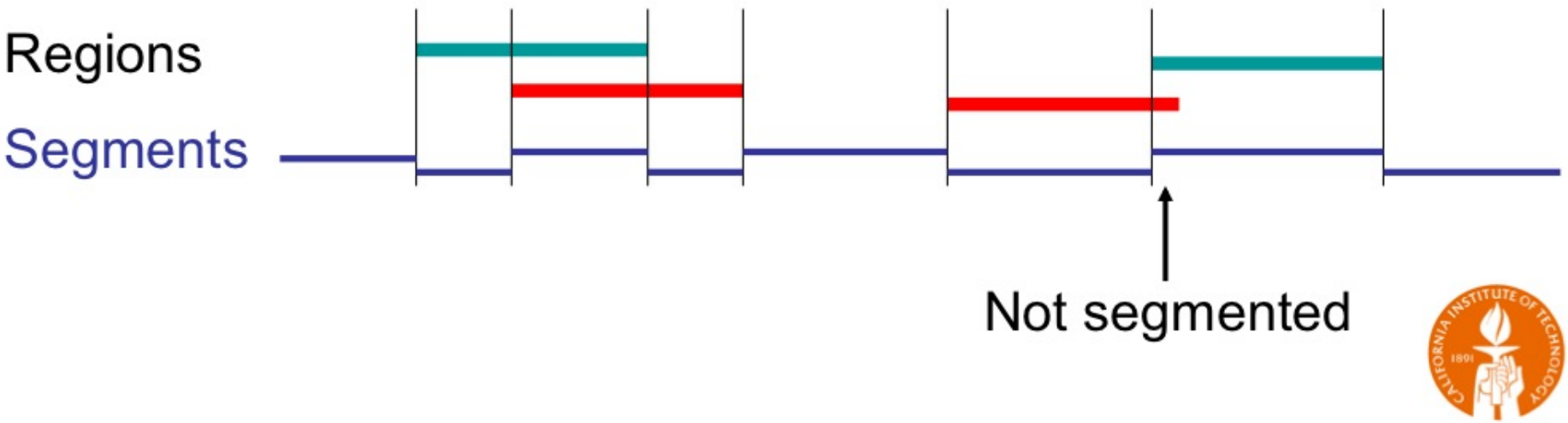
1. Determining the boundaries of regions
2. Scoring the datasets over these regions
3. Using statistical or machine learning techniques to discover combinations of patterns
   1. Supervised (e.g. on TSS)
   2. Unsupervised
4. Analyzing those combinations for functional significance





Segmenting the genome

- Segmentation can be straightforward fixed-length segments:
  - Fixed distance to TSS
  - Every 1kb
- Alternatively, the algorithms are designed to learn variable length segmentation, often with a minimum size constraint:
  - Sliding window with threshold
  - Hidden Markov Models
  - Segmentation based on ChIP-seq peaks and a normalized density measurement (e.g. RPKM)



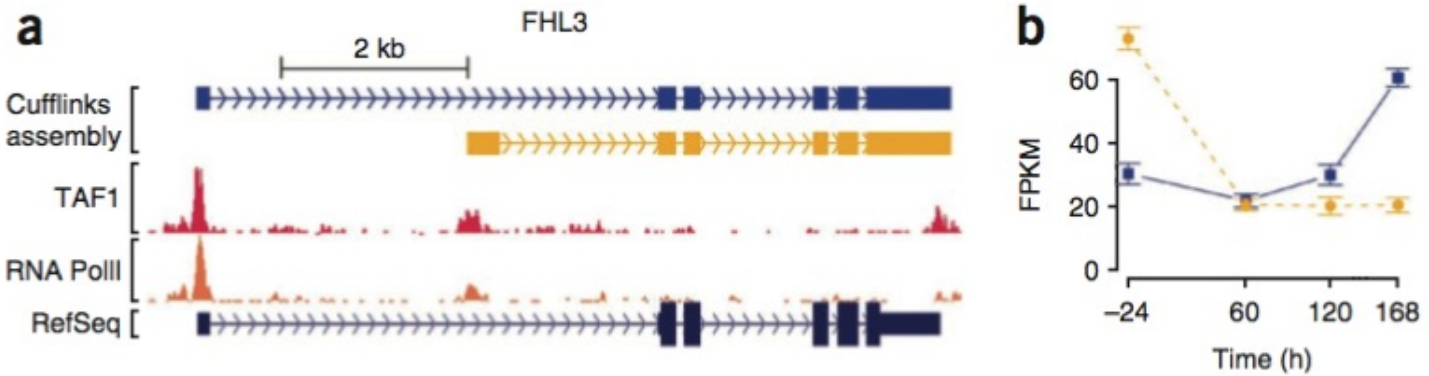
Learning regulatory states

- Once we have scored segments, we would like to identify sets of recurring signal combinations:
  - For supervised approaches, we provide a training set with state labels and derive a classification rule that we apply to the rest of the data.
  - For unsupervised HMM-based approaches, we simply need to look at the states
  - For other approaches, we need to apply a clustering method. The clusters will be our states.
- Once we have identified states of interest, we can analyze them for interesting patterns, such as:
  - Enrichment or lack-thereof in relation to gene body
  - Relation to gene expression / gene function / gene ontology
  - Relation to other states (transition probabilities in HMM)
  - Motif enrichment



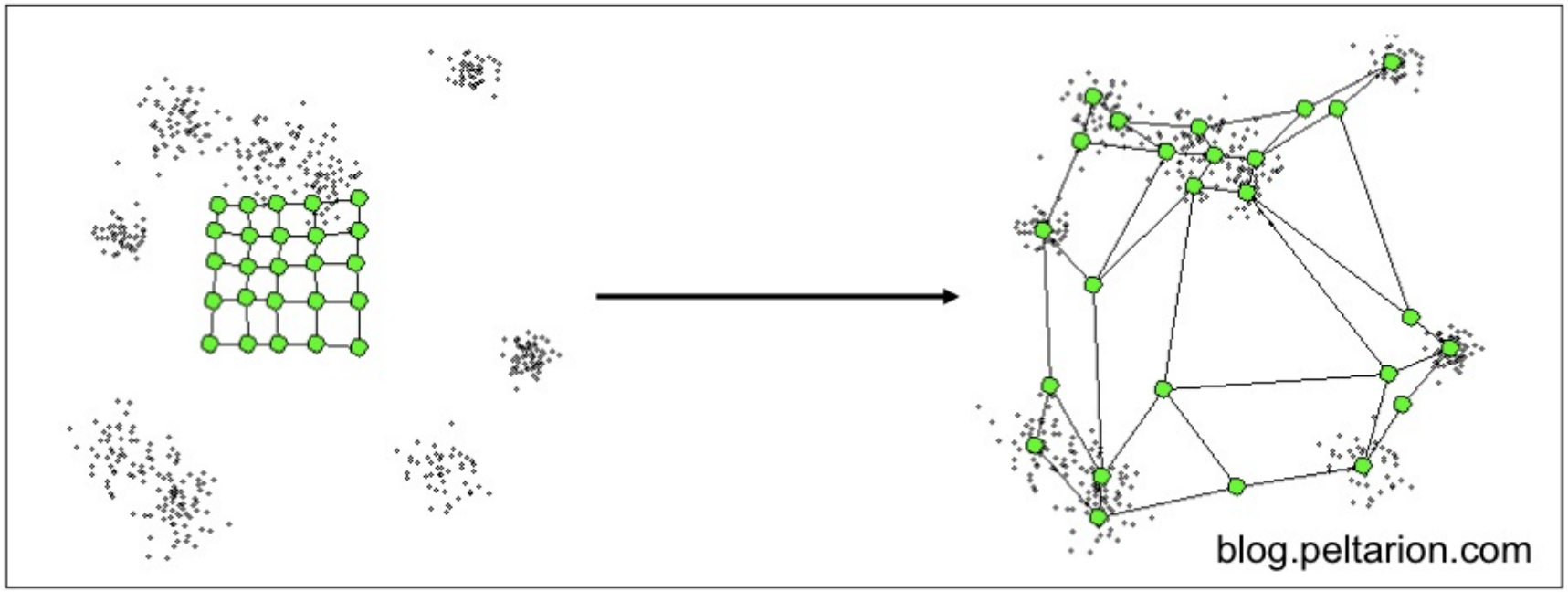
Joint Analysis of ChIP-seq and RNA-seq

- ChIP-seq measures the input into transcription
- RNA-seq measures the (steady-state) output of transcription
- Can we analyze them jointly to learn the rules of transcriptional regulation ?



Self organizing maps (SOM)

An unsupervised learning technique for clustering high dimensional data that preserves and represents topographical relationships

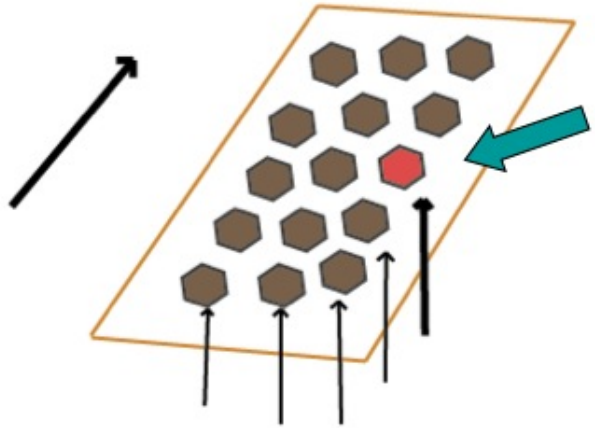


The big idea: data that is close in the input space is close in the output space

Training the units in the SOM

Consider each segment vector one at a time



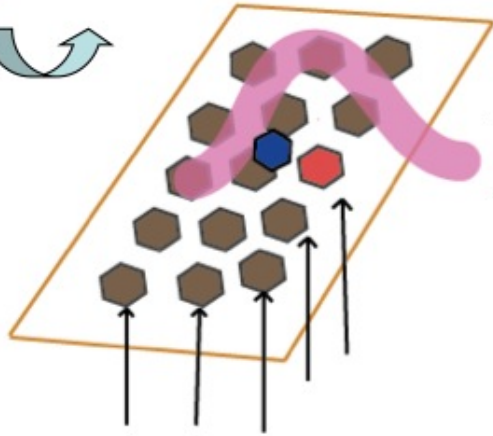


"winner"

Choose a "winner" map unit that is closest to the current input

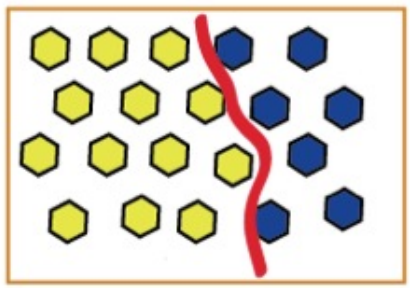
Iterate until converged





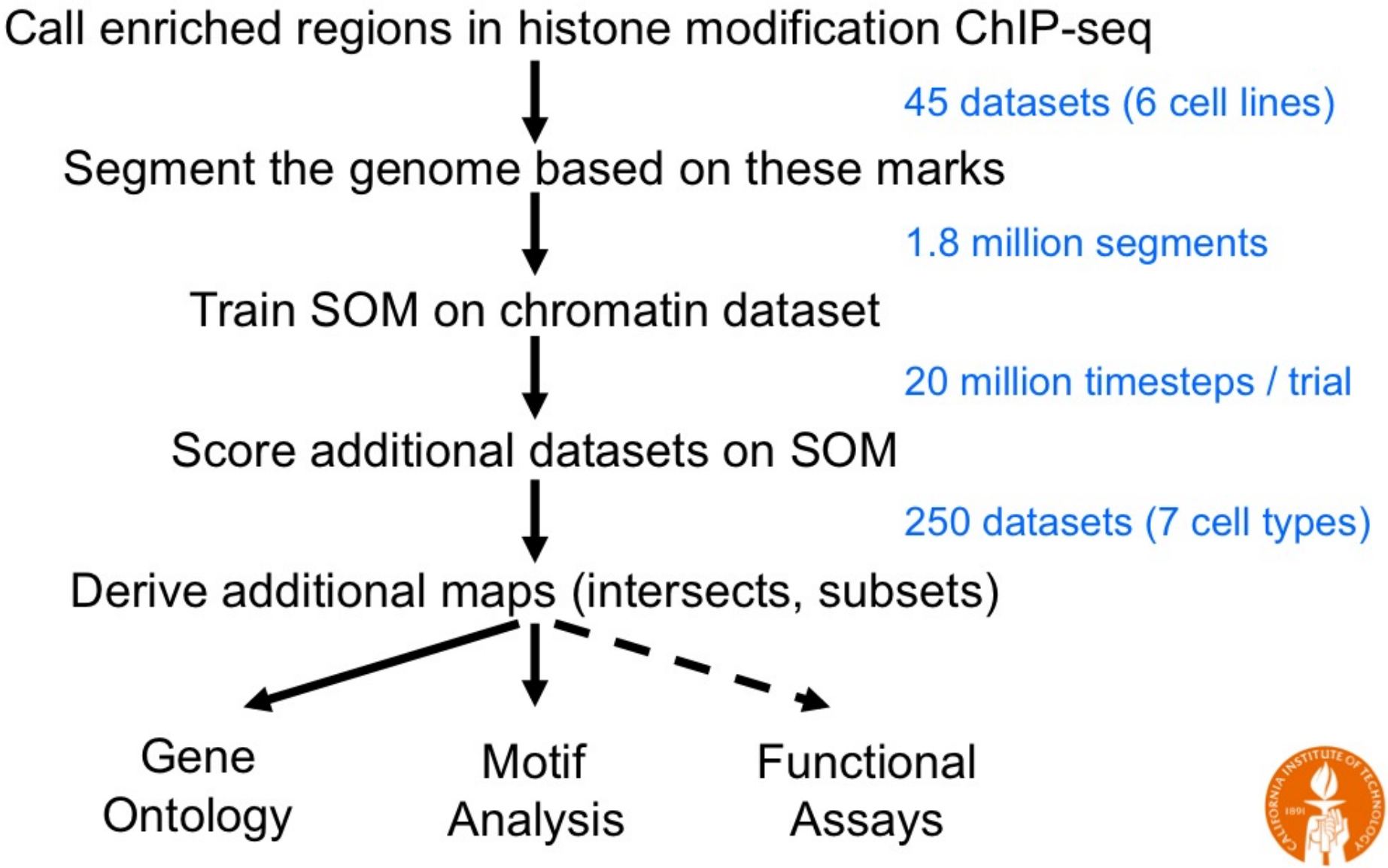
Adjust the winning map unit weights and those as well as those of vectors in the topographic neighborhood

After the map has converged to an organized state, map additional datasets onto the map.

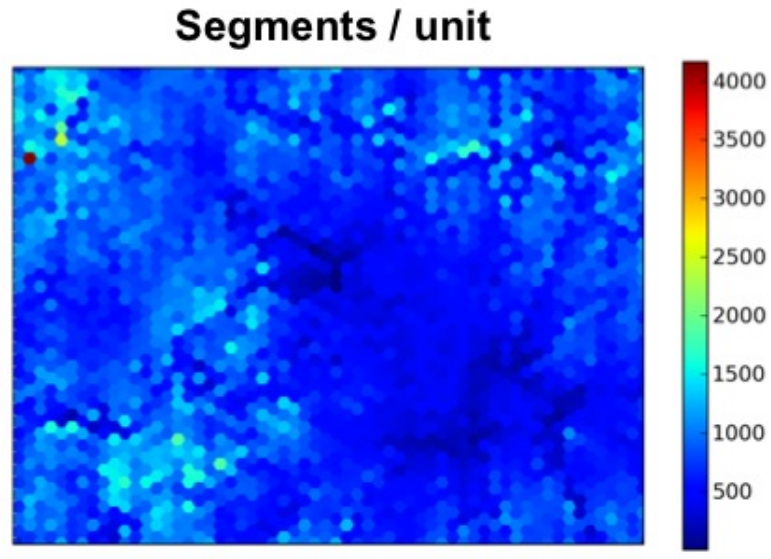


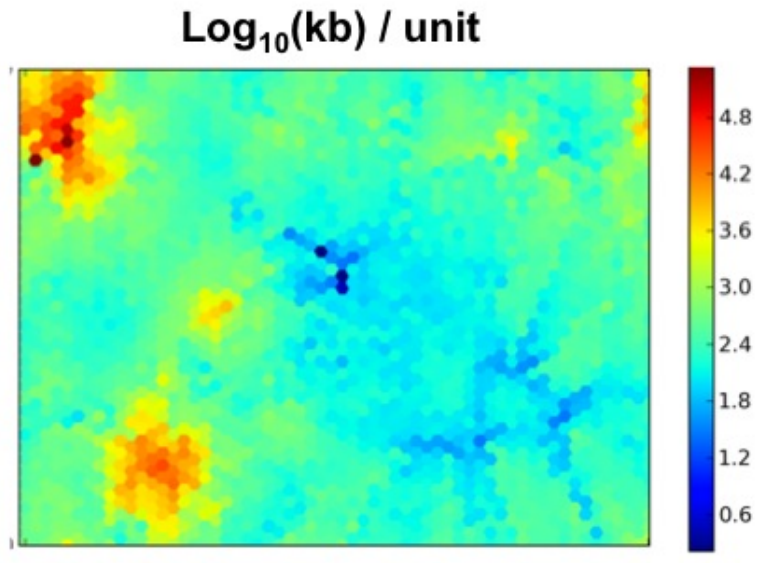


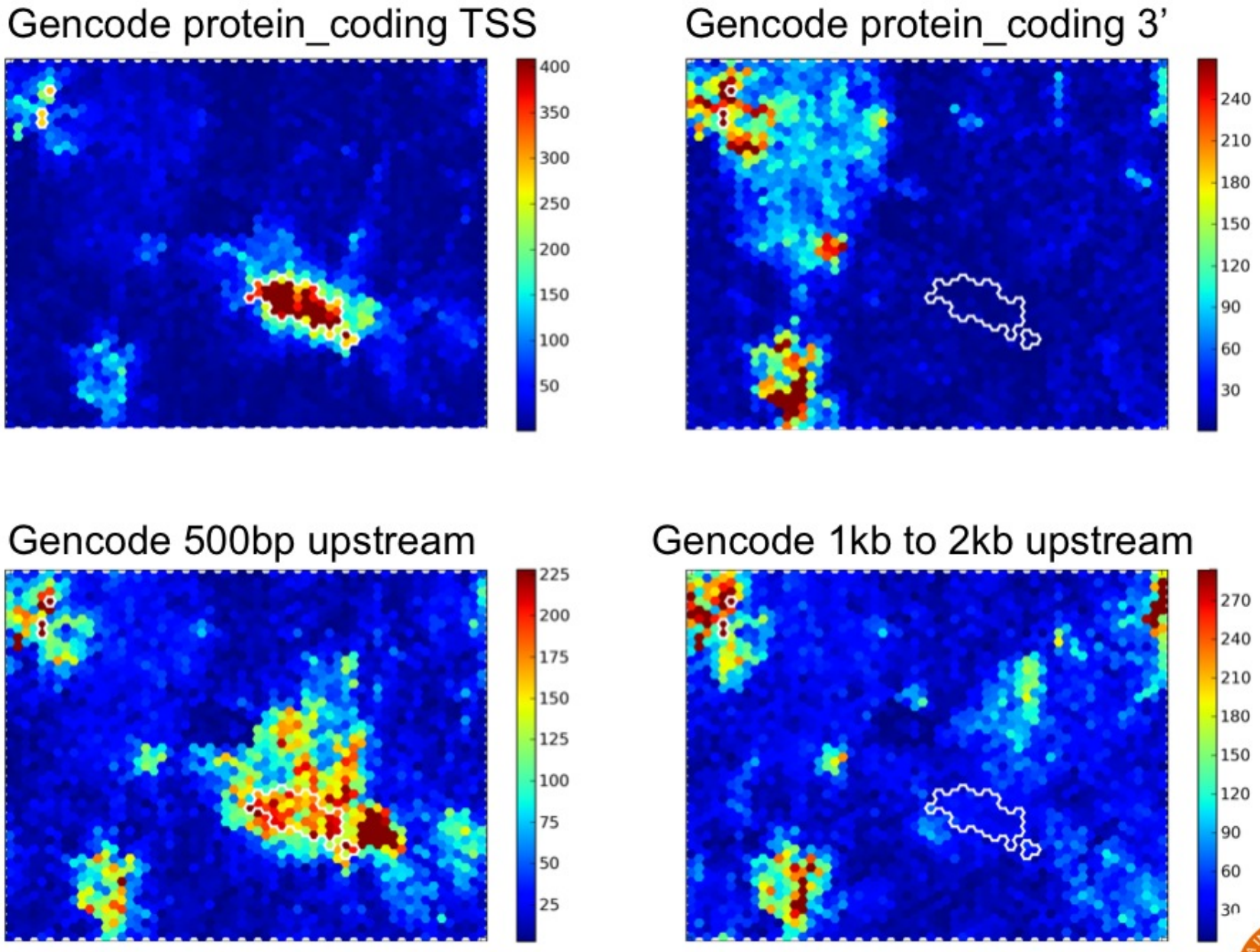
Genomewide SOM Analysis pipeline



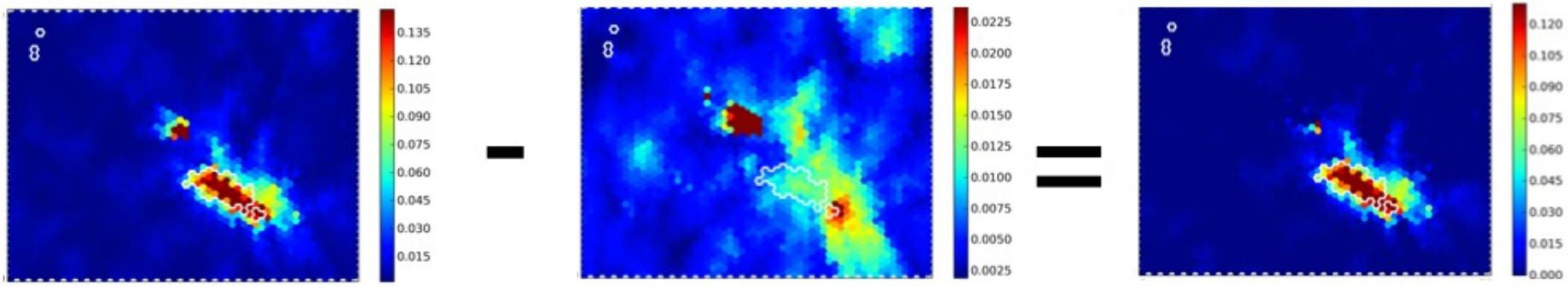
- 40x60 toroid map
  - "no boundaries"
- The distribution of segments per units and nucleotides is uneven
- Segments with little signal are aggregated in few units
- The bulk of the units are finely parsing signal changes





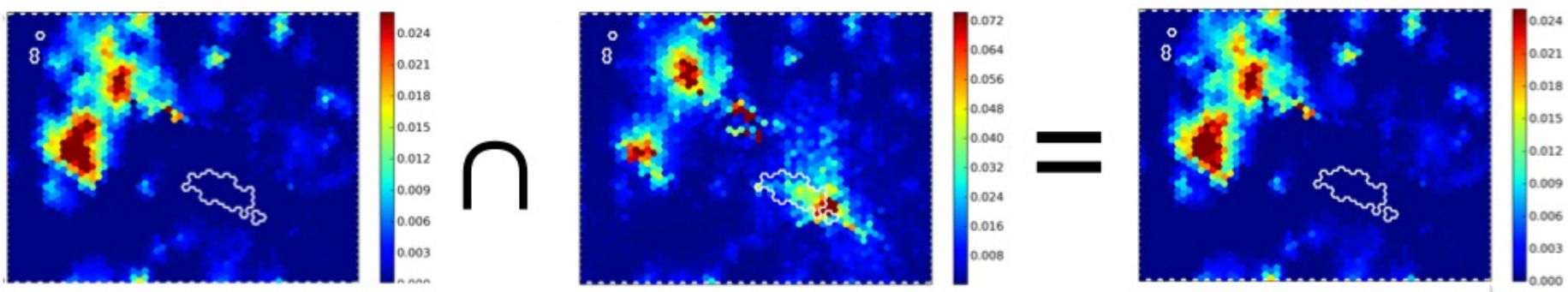

Gencode protein_coding TSS
400
350
300
250
200
150
100
50
Gencode protein_coding 3'
240
210
180
150
120
90
60
30
Gencode 500bp upstream
225
200
175
150
125
100
75
50
25
Gencode 1kb to 2kb upstream
270
240
210
180
150
120
90
60
30
Scale: counts/units

Subtracting maps



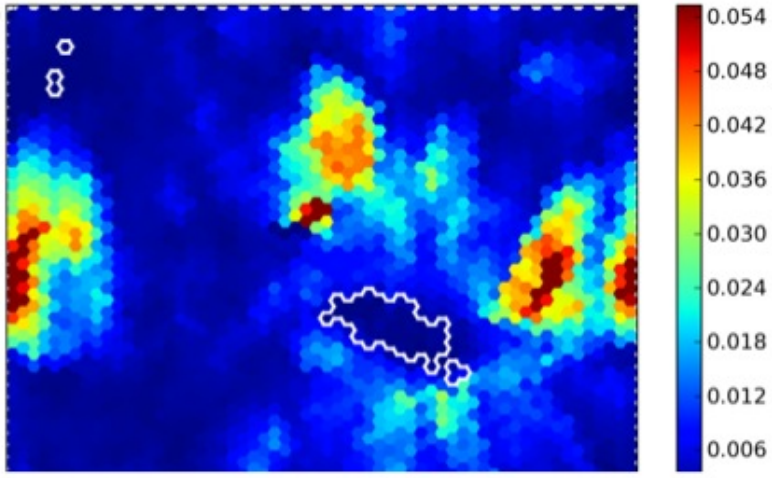
HepG2 TAF1 HepG2 control HepG2 TAF1 clean

Intersecting maps

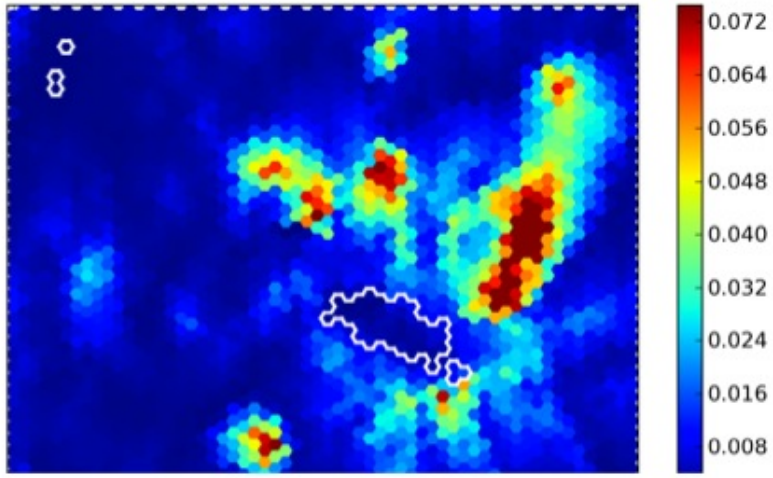


HepG2 H3K36me3 clean HepG2 2x75 RNA-seq R^2: 0.314

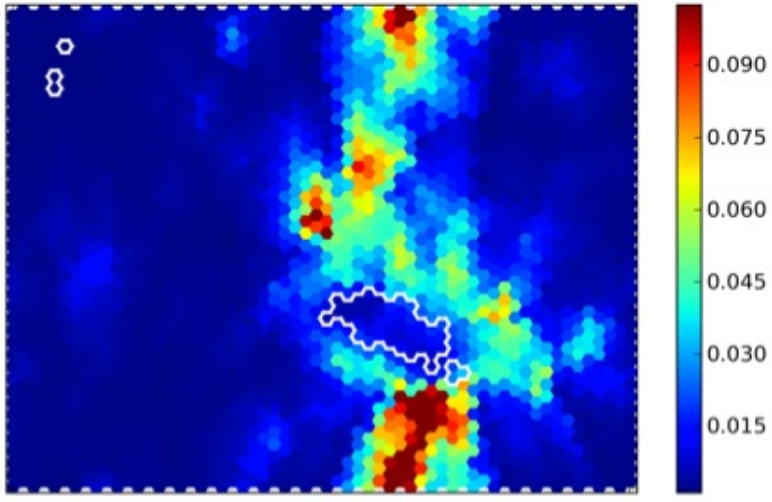
Scale: RPKM/region/units



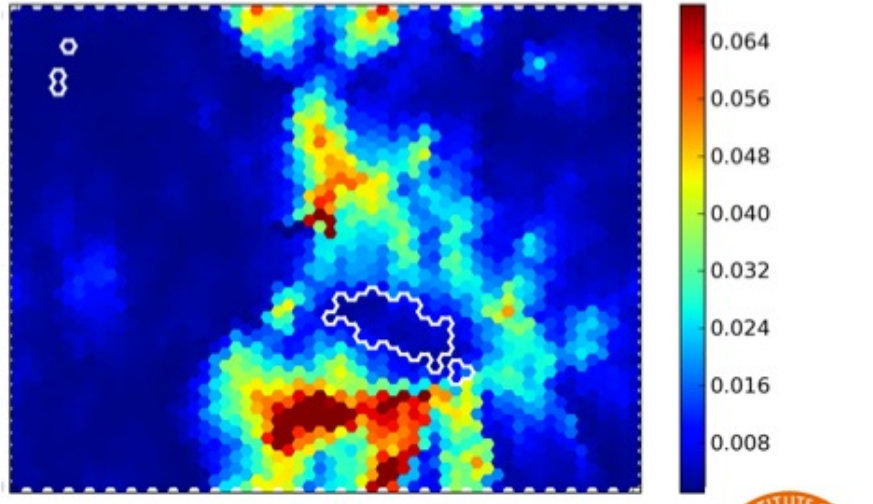
GM12878 H3K4me1



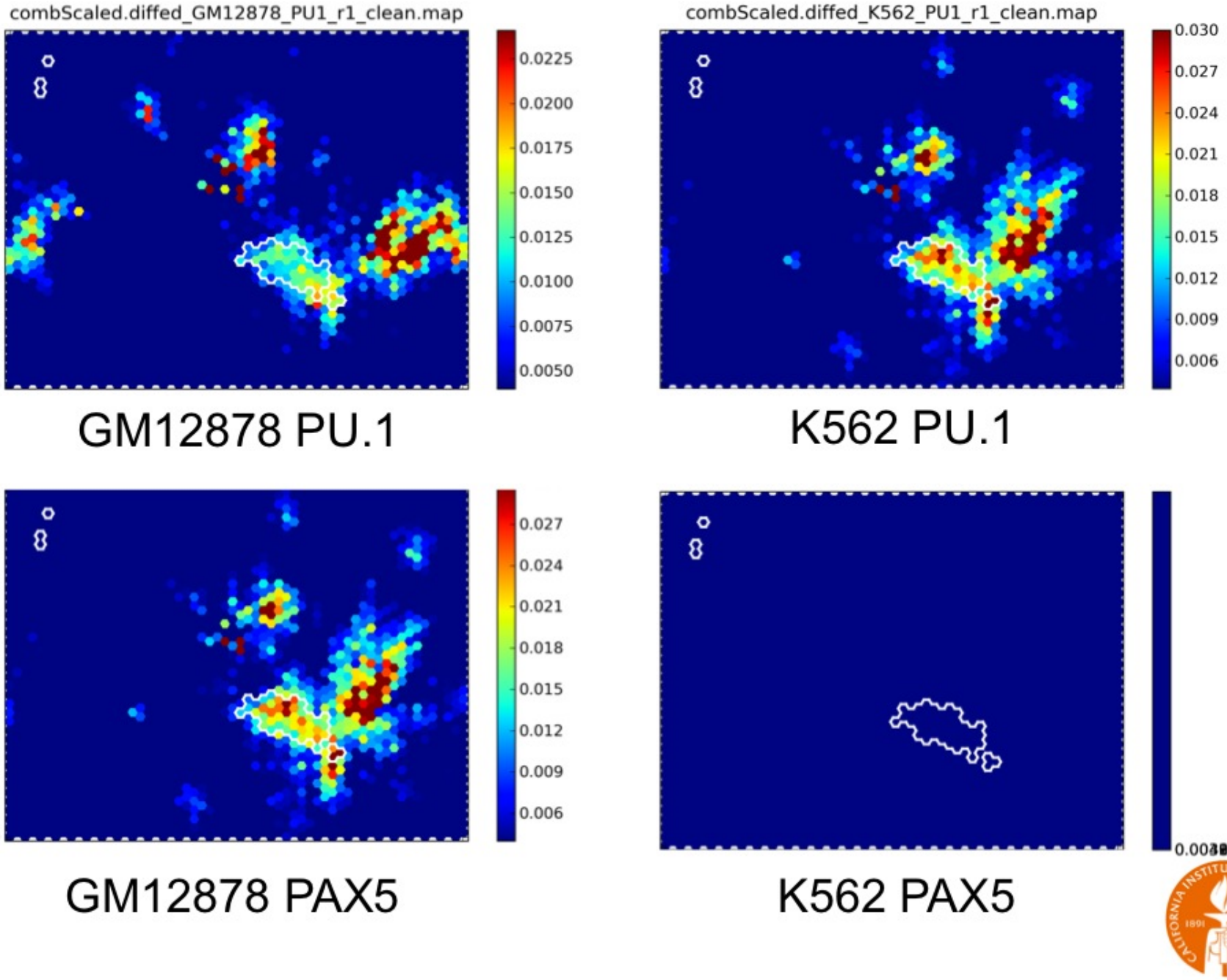
K562 H3K4me1



HUVEC H3K4me1

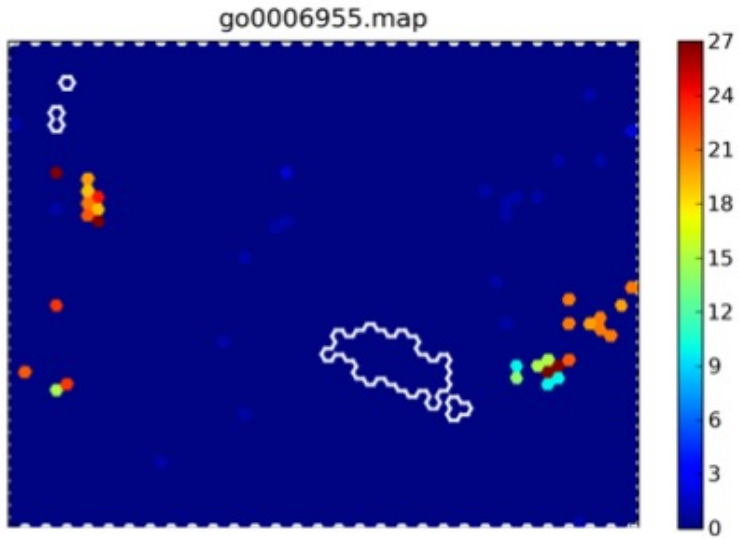


NHEK H3K4me1


combScaled.diffed_GM12878_PU1_r1_clean.map
0.0225
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
GM12878 PU.1
combScaled.diffed_K562_PU1_r1_clean.map
0.030
0.027
0.024
0.021
0.018
0.015
0.012
0.009
0.006
K562 PU.1
0.027
0.024
0.021
0.018
0.015
0.012
0.009
0.006
GM12878 PAX5
K562 PAX5

Zooming into the units

- Gene Ontology
  - Run go analysis on each unit
  - 520 significant GO terms
- We can also mine each unit for motifs



GO Immune Response Enriched

Summary

Introduction
- Quantitation in biology
- The Rise of the Sequencers
- Sequence-counting Assays
- Important Considerations

Transcriptome Analysis
- Motivation
- Common Steps and Challenges
- RNA-seq Read Mapping
- Quantitation & Diff. Expression
- Transcript Reconstruction
- Transcript Assembly

ChIP-seq
- Introduction
- Mapping the reads
- Taking Advantage of Directionality
- Shifting Reads
- Motif Finding

Integration
- The Combinatorial Problem
- Analyzing Chip-seq Datasets Jointly
- Segmenting the Genome
- Learning Regulatory States
- Joint Analysis of Chip- and RNA-Seq

Acknowledgments

We gratefully acknowledge help with the material:

- Anna Abelin, Caltech
- Jonas Behr, Friedrich Miescher Laboratory
- Regina Bohnert, Friedrich Miescher Laboratory
- Philipp Drewe, Friedrich Miescher Laboratory
- Fabio De Bona, Friedrich Miescher Laboratory & Google Research
- André Kahles, Friedrich Miescher Laboratory
- Stefan Henz, MPI for Developmental Biology
- Georgi Marinov, Caltech
- Shirley Pepke, Caltech
- Cole Trapnell, U Maryland & UC Berkeley
- Moran Yassour, Hebrew University & Broad Institute
- Georg Zeller, EMBL Heidelberg

The slides and additional material will be available online at

`http://www.fml.mpg.de/raetsch/lectures/ismb10tutorial`

References I

S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal Molecular Biology*, 215(3):403–10, 1990.

S. Anders and W. Huber. Differential expression analysis for sequence count data. Pre-published on Nature Precedings, March 2010. URL `http://precedings.nature.com/documents/4282/version/1/files/npre20104282-1.pdf`.

H. Bao, Y. Xiong, H. Guo, R. Zhou, X. Lu, Z. Yang, Y. Zhong, and S. Shi. MapNext: a software tool for spliced and unspliced alignments and SNP detection of short sequence reads. *BMC genomics*, 10(Suppl 3):S13, 2009.

J. Behr, G. Schweikert, J. Cao, F. De Bona, G. Zeller, S. Laubinger, S. Ossowski, K. Schneeberger, D. Weigel, and G. Rätsch. Rna-seq and tiling arrays for improved gene finding. Oral presentation at the CSHL Genome Informatics Meeting, September 2008. URL `http://www.fml.tuebingen.mpg.de/raetsch/lectures/RaetschGenomeInformatics08.pdf`.

M. Bergkessel, G. Wilmes, and C. Guthrie. Snapshot: Formation of mrnps. *Cell*, 136, January 2009.

R. Bohnert, J. Behr, and G Rätsch. Transcript quantification with RNA-Seq data. *BMC Bioinformatics*, 10(S13):P5, 2009. URL `http://www.biomedcentral.com/1471-2105/10/S13/P5`.

D. Campagna, A. Albiero, A. Bilardi, E. Caniato, C. Forcato, S. Manavski, N. Vitulo, and G. Valle. PASS: a program to align short sequences. *Bioinformatics*, 25(7):967, 2009.

References II

RM Clark, G Schweikert, C Toomajian, S Ossowski, G Zeller, P Shinn, N Warthmann, TT Hu, G Fu, DA Hinds, H Chen, KA Frazer, DH Huson, B Schölkopf, M Nordborg, G Rätsch, JR Ecker, and D Weigel. Common sequence polymorphisms shaping genetic diversity in arabidopsis thaliana. *Science*, 317(5836):338–342, 2007. ISSN 1095-9203 (Electronic). doi: 10.1126/science.1138632.

F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch. Qpalma: Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24:i174–i180, 2008.

France Denoeud, Jean-Marc Aury, Corinne Da Silva, Benjamin Noel, Odile Rogier, Massimo Delledonne, Michele Morgante, Giorgio Valle, Patrick Wincker, Claude Scarpelli, Olivier Jaillon, and François Artiguenave. Annotating genomes with massive-scale rna sequencing. *Genome Biol*, 9(12):R175, 2008. doi: 10.1186/gb-2008-9-12-r175.

Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Sharcgs, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res*, 17(11):1697–706, Nov 2007. doi: 10.1101/gr.6435207.

Juliane C. Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105, Sep 2008. doi: 10.1093/nar/gkn425.

L. Florea, G. Hartzell, Z. Zhang, G.M. Rubin, and W. Miller. A computer program for aligning a cdna sequence with a genomic dna sequence. *Genome Research*, 8:967–974, 1998.

References III

M.S. Gelfand, A.A. Mironov, and P.A. Pevzner. Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci.*, 93(17):9061–6, 1996.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, pages 513–520. MIT Press, Cambridge, 2007.

Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, John L Rinn, Eric S Lander, and Aviv Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat Biotechnol*, 28(5): 503–10, May 2010. doi: 10.1038/nbt.1633.

Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, Apr 2010. doi: 10.1093/nar/gkq224.

Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M Sharma, Philipp Khaitovich, Jörg Vogel, Peter F Stadler, and Jörg Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9): e1000502, Sep 2009. doi: 10.1371/journal.pcbi.1000502.

Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, April 2009.

References IV

David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–502, Jun 2007. doi: 10.1126/science.1141319.

J Kärkkäinen. Fast bwt in small space by blockwise suffix sorting. *Theor. Comput. Sci.*, 387: 249–257, 2007.

W.J. Kent. BLAT—the BLAST-like alignment tool. *Genome research*, 12(4):656, 2002.

Vincent Lacroix, Michael Sammeth, Roderic Guigó, and Anne Bergeron. Exact transcriptome reconstruction from short sequence reads. In *WABI*, pages 50–63, 2008.

Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10 (3):R25, 2009. doi: 10.1186/gb-2009-10-3-r25.

Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, February 2010a. doi: 10.1093/bioinformatics/btp692. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/4/493`.

Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, Jul 2009. doi: 10.1093/bioinformatics/btp324.

Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, 26(5):589–95, Mar 2010. doi: 10.1093/bioinformatics/btp698.

References V

Heng Li, Jue Ruan, and Richard Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–8, Nov 2008. doi: 10.1101/gr.078212.108.

Jun Li, Hui Jiang, and Wing Hung Wong. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology*, 11(5):R50, 2010b. doi: 10.1186/gb-2010-11-5-r50.

Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15): 1966–7, Aug 2009. doi: 10.1093/bioinformatics/btp336.

Sam E V Linsen, Elzo de Wit, Georges Janssens, Sheila Heater, Laura Chapman, Rachael K Parkin, Brian Fritz, Stacia K Wyman, Ewart de Bruijn, Emile E Voest, Scott Kuersten, Muneesh Tewari, and Edwin Cuppen. Limitations and possibilities of small RNA digital gene expression profiling. *Nature Methods*, 6(7):474–476, July 2009. ISSN 1548-7091. doi: 10.1038/nmeth0709-474. URL `http://dx.doi.org/10.1038/nmeth0709-474`.

Iain Maccallum, Dariusz Przybylski, Sante Gnerre, Joshua Burton, Ilya Shlyakhter, Andreas Gnirke, Joel Malek, Kevin McKernan, Swati Ranade, Terrance P Shea, Louise Williams, Sarah Young, Chad Nusbaum, and David B Jaffe. Allpaths 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol*, 10(10):R103, 2009. doi: 10.1186/gb-2009-10-10-r103.

S.B. Montgomery, M. Sammeth, and et al. Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, 2010.

References VI

A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, 2008.

Alicia Oshlack and Matthew Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4(1):14, 2009. doi: 10.1186/1745-6150-4-14.

Peter J Park. Chip-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–80, Oct 2009. doi: 10.1038/nrg2641.

Shirley Pepke, Barbara Wold, and Ali Mortazavi. Computation for chip-seq and rna-seq studies. *Nat Methods*, 6(11 Suppl):S22–32, Nov 2009. doi: 10.1038/nmeth.1371.

P A Pevzner, H Tang, and M S Waterman. An eulerian path approach to dna fragment assembly. *Proc Natl Acad Sci U S A*, 98(17):9748–53, Aug 2001. doi: 10.1073/pnas.171285098.

J.K. Pickrell, J.C. Marioni, and et al. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 2010.

G. Rätsch and S. Sonnenburg. Accurate splice site detection for *Caenorhabditis elegans*. In K. Tsuda B. Schoelkopf and J.-P. Vert, editors, *Kernel Methods in Computational Biology*. MIT Press, 2004.

G. Rätsch, S. Sonnenburg, and B. Schölkopf. RASE: recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics*, 21(Suppl. 1):i369–i377, June 2005.

G Rätsch, G Jean, A Kahles, S Sonnenburg, F De Bona, K Schneeberger, J Hagmann, and D Weigel. PALMapper: Fast and accurate alignment of RNA-seq reads. in preparation, 2010.

References VII

Hugues Richard, Marcel H. Schulz, Marc Sultan, Asja Nurnberger, Sabine Schrinner, Daniela Balzereit, Emilie Dagand, Axel Rasche, Hans Lehrach, Martin Vingron, Stefan A. Haas, and Marie-Laure Yaspo. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Research*, page gkq041, February 2010. doi: 10.1093/nar/gkq041. URL `http://nar.oxfordjournals.org/cgi/content/abstract/gkq041v1`.

Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nat Biotechnol*, 27 (1):66–75, Jan 2009. doi: 10.1038/nbt.1518.

Stephen M Rumble, Phil Lacroute, Adrian V Dalca, Marc Fiume, Arend Sidow, and Michael Brudno. Shrimp: accurate mapping of short color-space reads. *PLoS Comput Biol*, 5(5): e1000386, May 2009. doi: 10.1371/journal.pcbi.1000386.

M. Sammeth. The Flux Capacitor. *Website*, 2009a. http://flux.sammeth.net/capacitor.html.

M. Sammeth. The Flux Simulator. *Website*, 2009b. http://flux.sammeth.net/simulator.html.

Korbinian Schneeberger, Jörg Hagmann, Stephan Ossowski, Norman Warthmann, Sandra Gesing, Oliver Kohlbacher, and Detlef Weigel. Simultaneous alignment of short reads against multiple genomes. *Genome Biol*, 10(9):R98, 2009a. doi: 10.1186/gb-2009-10-9-r98.

References VIII

Korbinian Schneeberger, Jörg Hagmann, Stephan Ossowski, Norman Warthmann, Sandra Gesing, Oliver Kohlbacher, and Detlef Weigel. Simultaneous alignment of short reads against multiple genomes. *Genome Biol*, 10(9):R98, Jan 2009b. doi: 10.1186/gb-2009-10-9-r98. URL `http://genomebiology.com/2009/10/9/R98`.

U. Schulze, B. Hepp, C.S. Ong, and G. Ratsch. PALMA: mRNA to genome alignments using large margin algorithms. *Bioinformatics*, 23(15):1892, 2007.

Gabriele Schweikert, Alexander Zien, Georg Zeller, Jonas Behr, Christoph Dieterich, Cheng Soon Ong, Petra Philips, Fabio De Bona, Lisa Hartmann, Anja Bohlen, Nina Krüger, Sören Sonnenburg, and Gunnar Rätsch. mgene: Accurate svm-based gene finding with an application to nematode genomes. *Genome Research*, 2009. URL `http://genome.cshlp.org/content/early/2009/06/29/gr.090597.108.full.pdf+html`. Advance access June 29, 2009.

S. Sonnenburg, G. Rätsch, A. Jagota, and K.-R. Müller. New methods for splice-site recognition. In *Proc. International Conference on Artificial Neural Networks*, 2002.

Sören Sonnenburg, Alexander Zien, and Gunnar Rätsch. ARTS: Accurate Recognition of Transcription Starts in Human. *Bioinformatics*, 22(14):e472–480, 2006.

O. Stegle, P. Drewe, R. Bohnert, K. Borgwardt, and G. Rätsch. Statistical tests for detecting differential rna-transcript expression from read counts. Technical report, Nature Preceedings, 2010.

I. Sutskever. Arachne: A whole genome shotgun assembler. oral presentation, 2008.

References IX

C. Trapnell, L. Pachter, and S.L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105, 2009.

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotech*, advance online publication, May 2010. doi: 10.1038/nbt.1621. URL `http://dx.doi.org/10.1038/nbt.1621`.

J. Usuka, W. Zhu, and V. Brendel. Optimal spliced alignment of homologous cdna to a genomic dna template. *Bioinformatics*, 16(3):203–211, 2000.

Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nat Methods*, 5(9):829–34, Sep 2008. doi: 10.1038/nmeth.1246.

René L Warren, Granger G Sutton, Steven J M Jones, and Robert A Holt. Assembling millions of short dna sequences using ssake. *Bioinformatics*, 23(4):500–1, Feb 2007. doi: 10.1093/bioinformatics/btl629.

Ostell JM. Wheelan SJ, Church DM. Spidey: a tool for mrna-to-genomic alignments. *Genome Research*, 11(11):1952–7, 2001.

References X

G Zeller, RM Clark, K Schneeberger, A Bohlen, D Weigel, and G Ratsch. Detecting polymorphic regions in arabidopsis thaliana with resequencing microarrays. *Genome Res*, 18 (6):918–929, 2008. ISSN 1088-9051 (Print). doi: 10.1101/gr.070169.107.

D.R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18:828–829, 2008.

M. Zhang and W. Gish. Improved spliced alignment from an information theoretic approach. *Bioinformatics*, 22(1):13–20, January 2006.

A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller. Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites. *BioInformatics*, 16(9): 799–807, September 2000.