Introduction to Sequencing & Functional Genomics

Gunnar Rätsch

Computational Biology Center
Memorial Sloan Kettering Cancer Center

Memorial Sloan-Kettering
Cancer Center

cBio@MSKCC

Introduction

Part I: Sequencing Basics

- The Rise of the Sequencers
- From Genome sequencing to Counting assays
- Biological question, sequencing, analysis

Part II: RNA-seq Basics

- Read Mapping and common analysis steps
- Gene and transcript quantification, Caveats

Part III: Chip-Seq Basics

- Enhancers, Promoters
- ChIP, DNase, ATAC, & friends

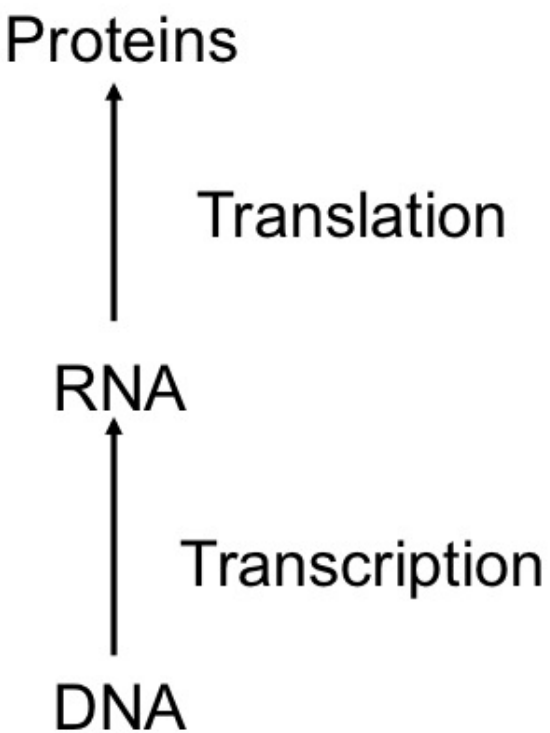
Part IV: Integration Approaches

- From DNA to Phenotype
- Integration into QTLs

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."

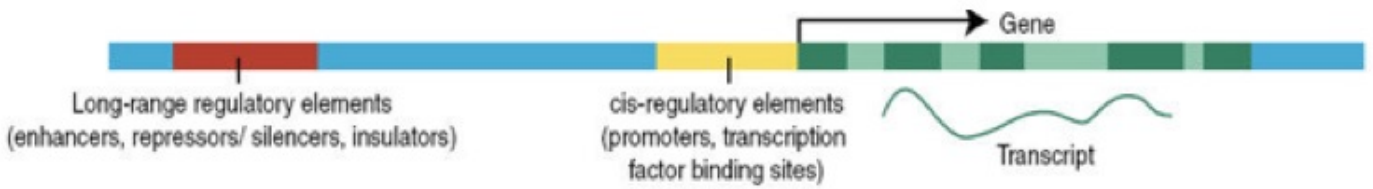
The Central Dogma and Transcriptional Regulation



- 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”.

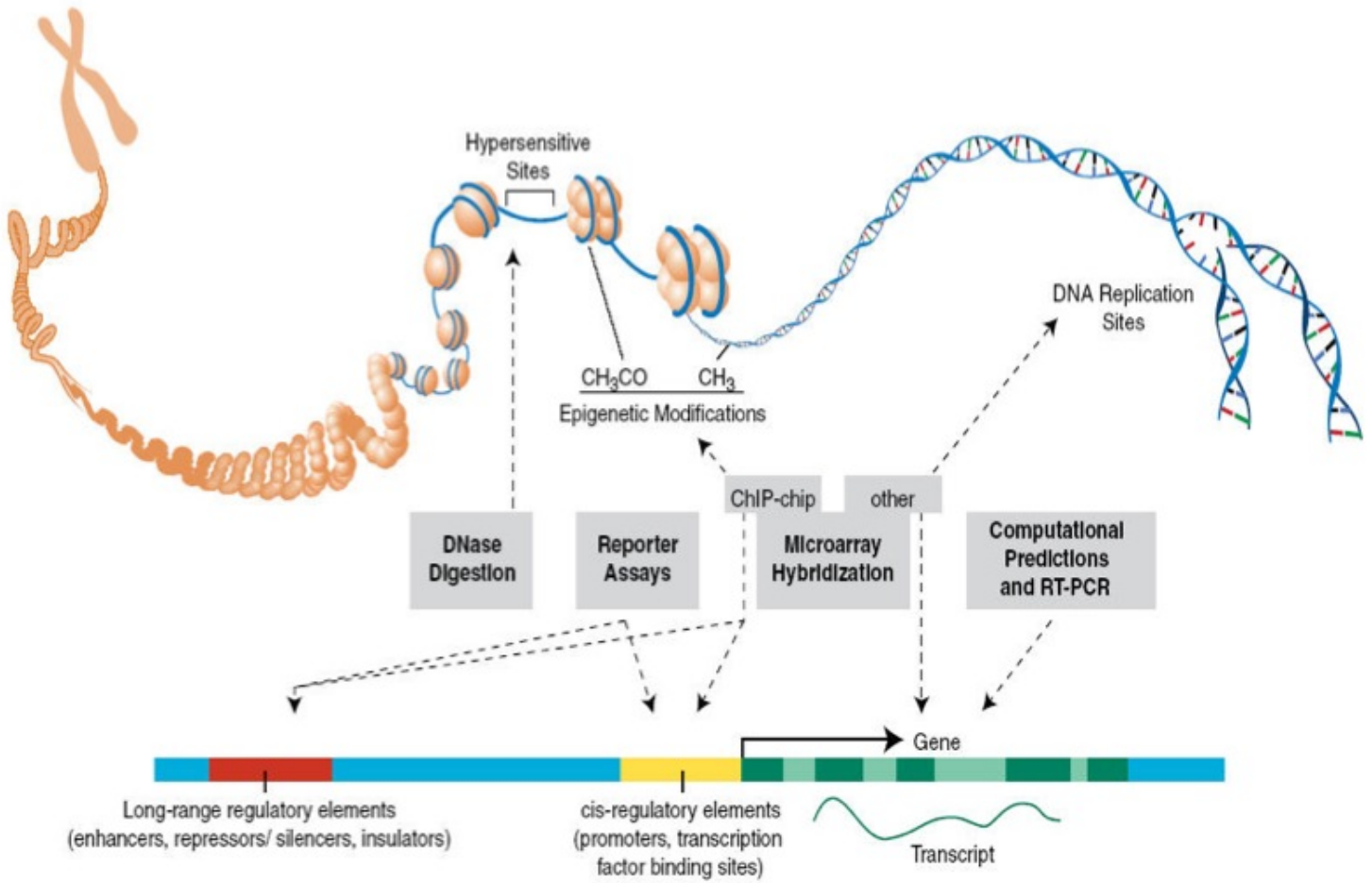
Sequencing Genomes

- 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

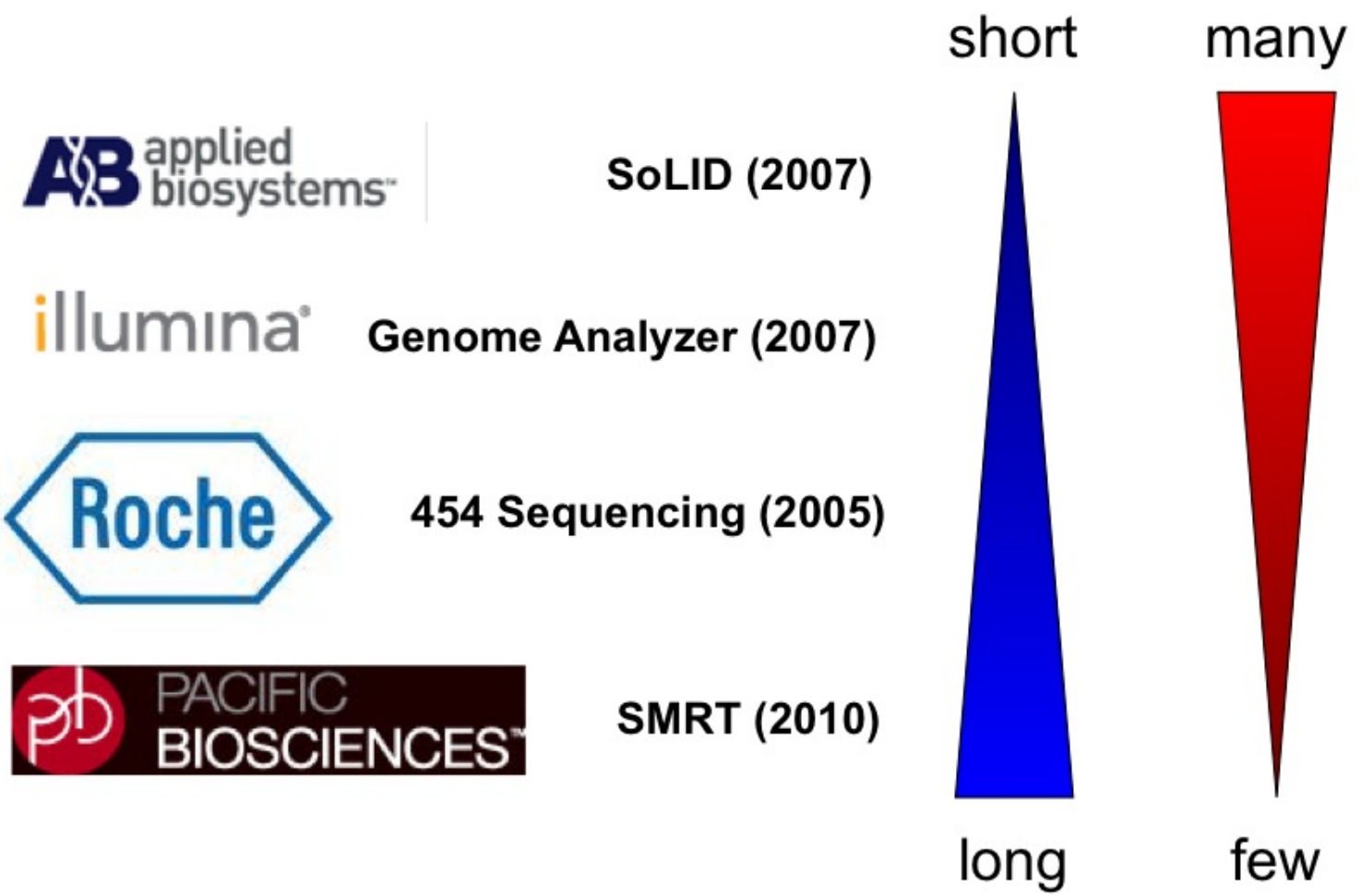


ENCODE Project Consortium (2004). Science 306: 636.

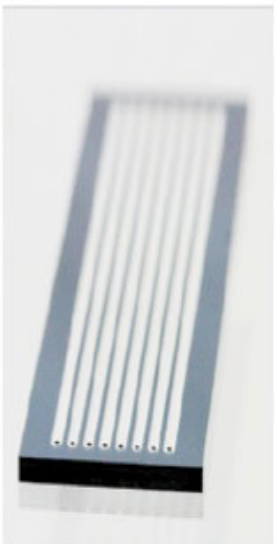
An Investment: the $1000 human genome

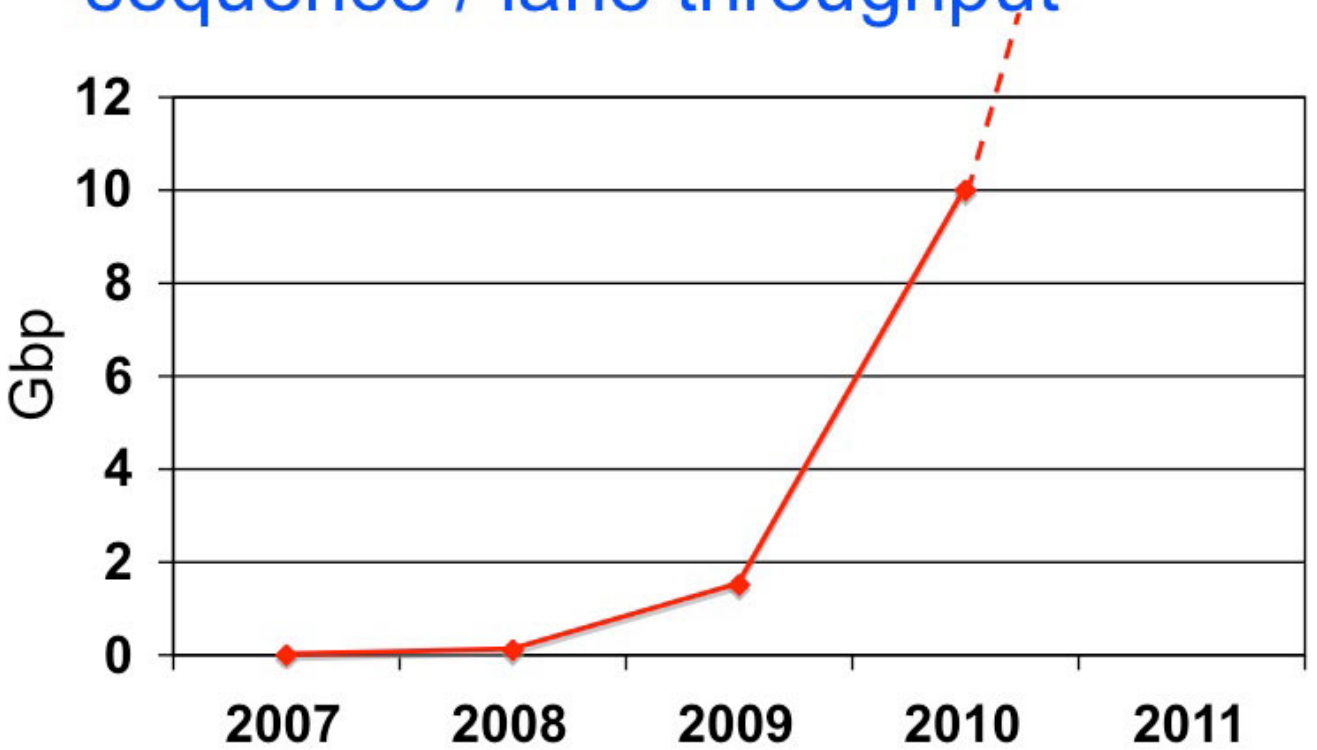
- The situation by the turn of the century:
  - Cost of a human-size draft genome (8x) in 2003: $50M
  - 4 main publicly supported genome centers in the US received the bulk of the money set aside for sequencing:
    - MIT (Broad)
    - Washington University
    - Baylor
    - DOE
- In 2003, NHGRI committed 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

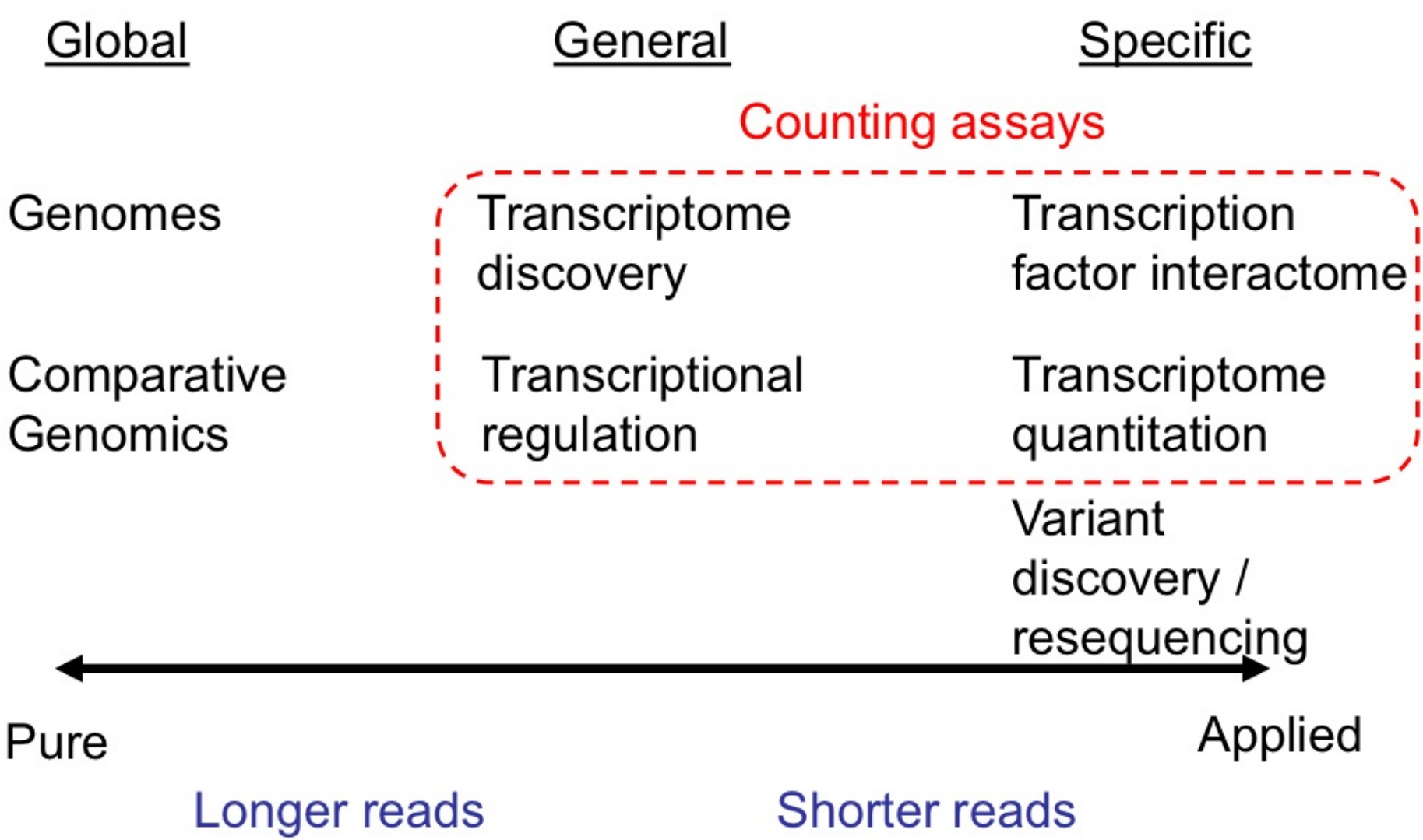


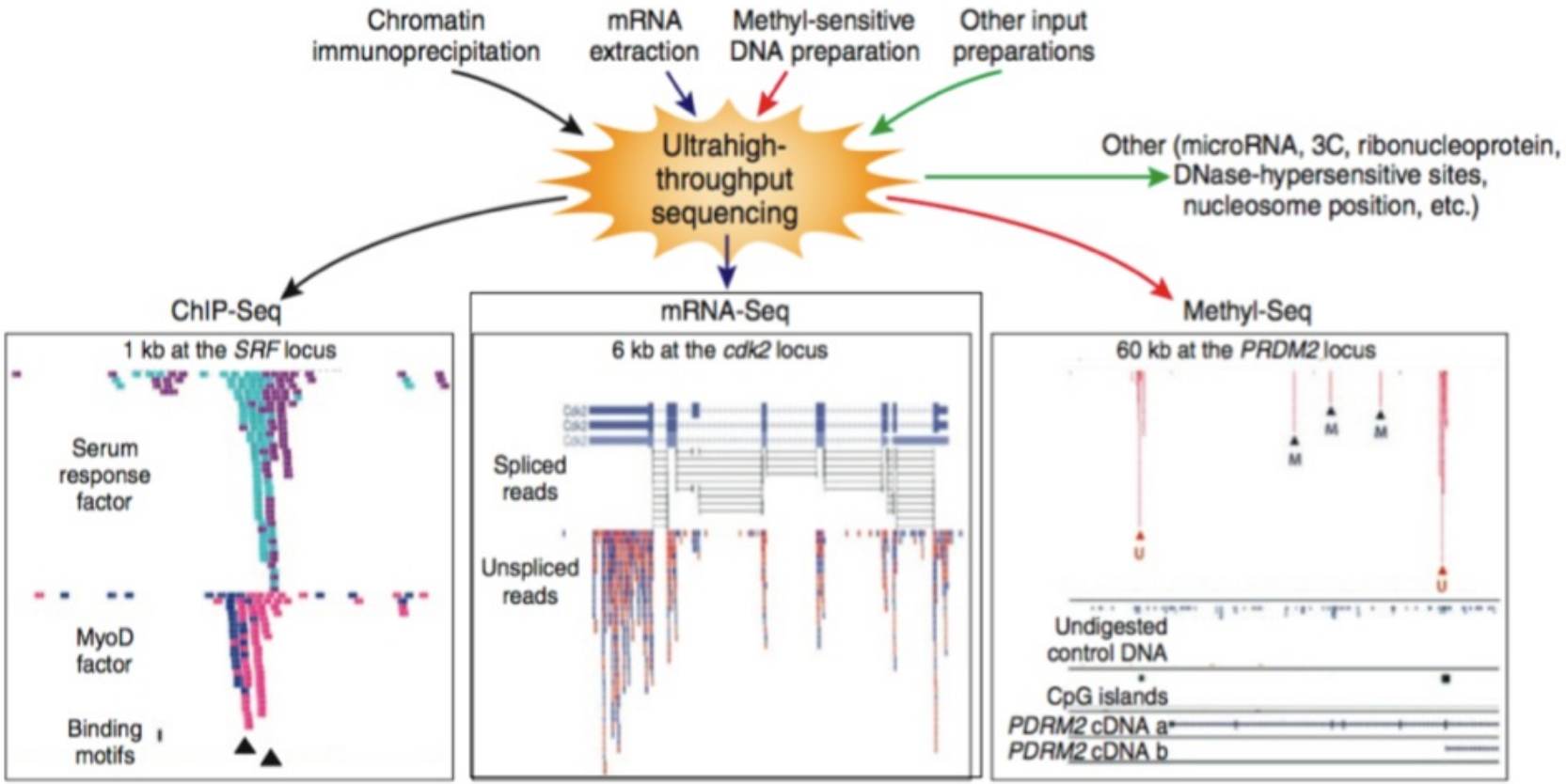


	Jan-07	Jan-08	Jan-09	Jan-10	Jan-11	Jan-13
Gbp / lane	0.025	0.128	1.5	10	30-50	≈75
Read type:	1x25	1x36	2x75	2x100	2x150	2x150

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

Genomics: a maturing field





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.

Biological Question, Sequencing & Analysis

- New genome — qualitative
  - Sequence DNA ⇒ assemble
- Genomic variation — qualitative
  - Sequence DNA ⇒ assemble ⇒ align ⇒ detect
  - Sequence ⇒ align to DNA ⇒ detect
- New gene/transcript — qualitative
  - Sequence RNA ⇒ assemble
  - Sequence RNA ⇒ align to DNA ⇒ detect
- Gene/transcript expression — quantitative
  - Sequence RNA ⇒ align to known RNAs ⇒ count
  - Sequence RNA ⇒ align to DNA ⇒ count
- DNA/RNA-Protein binding — quantitative
  - Binding assay ⇒ Sequence ⇒ align to DNA ⇒ identify peaks
- Methylation — quant- & qualitative
  - Bisulfit treatment ⇒ Sequence DNA ⇒ align to DNA⇒ count
- ...

Biological Question, Sequencing & Analysis

- New genome — qualitative
  - Sequence DNA ⇒ assemble
- Genomic variation — qualitative
  - Sequence DNA ⇒ assemble ⇒ align ⇒ detect
  - Sequence ⇒ align to DNA ⇒ detect
- New gene/transcript — qualitative
  - Sequence RNA ⇒ assemble
  - Sequence RNA ⇒ align to DNA ⇒ detect
- Gene/transcript expression — quantitative
  - Sequence RNA ⇒ align to known RNAs ⇒ count
  - Sequence RNA ⇒ align to DNA ⇒ count
- DNA/RNA-Protein binding — quantitative
  - Binding assay ⇒ Sequence ⇒ align to DNA ⇒ identify peaks
- Methylation — quant- & qualitative
  - Bisulfit treatment ⇒ Sequence DNA ⇒ align to DNA⇒ count
- ...

Biological Question, Sequencing & Analysis

- New genome — qualitative
  - Sequence DNA ⇒ assemble
- Genomic variation — qualitative
  - Sequence DNA ⇒ assemble ⇒ align ⇒ detect
  - Sequence ⇒ align to DNA ⇒ detect
- New gene/transcript — qualitative
  - Sequence RNA ⇒ assemble
  - Sequence RNA ⇒ align to DNA ⇒ detect
- Gene/transcript expression — quantitative
  - Sequence RNA ⇒ align to known RNAs ⇒ count
  - Sequence RNA ⇒ align to DNA ⇒ count
- DNA/RNA-Protein binding — quantitative
  - Binding assay ⇒ Sequence ⇒ align to DNA ⇒ identify peaks
- Methylation — quant- & qualitative
  - Bisulfit treatment ⇒ Sequence DNA ⇒ align to DNA⇒ count
- ...

Biological Question, Sequencing & Analysis

- New genome — qualitative
  - Sequence DNA ⇒ assemble
- Genomic variation — qualitative
  - Sequence DNA ⇒ assemble ⇒ align ⇒ detect
  - Sequence ⇒ align to DNA ⇒ detect
- New gene/transcript — qualitative
  - Sequence RNA ⇒ assemble
  - Sequence RNA ⇒ align to DNA ⇒ detect
- Gene/transcript expression — quantitative
  - Sequence RNA ⇒ align to known RNAs ⇒ count
  - Sequence RNA ⇒ align to DNA ⇒ count
- DNA/RNA-Protein binding — quantitative
  - Binding assay ⇒ Sequence ⇒ align to DNA ⇒ identify peaks
- Methylation — quant- & qualitative
  - Bisulfit treatment ⇒ Sequence DNA ⇒ align to DNA⇒ count
- ...

Biological Question, Sequencing & Analysis

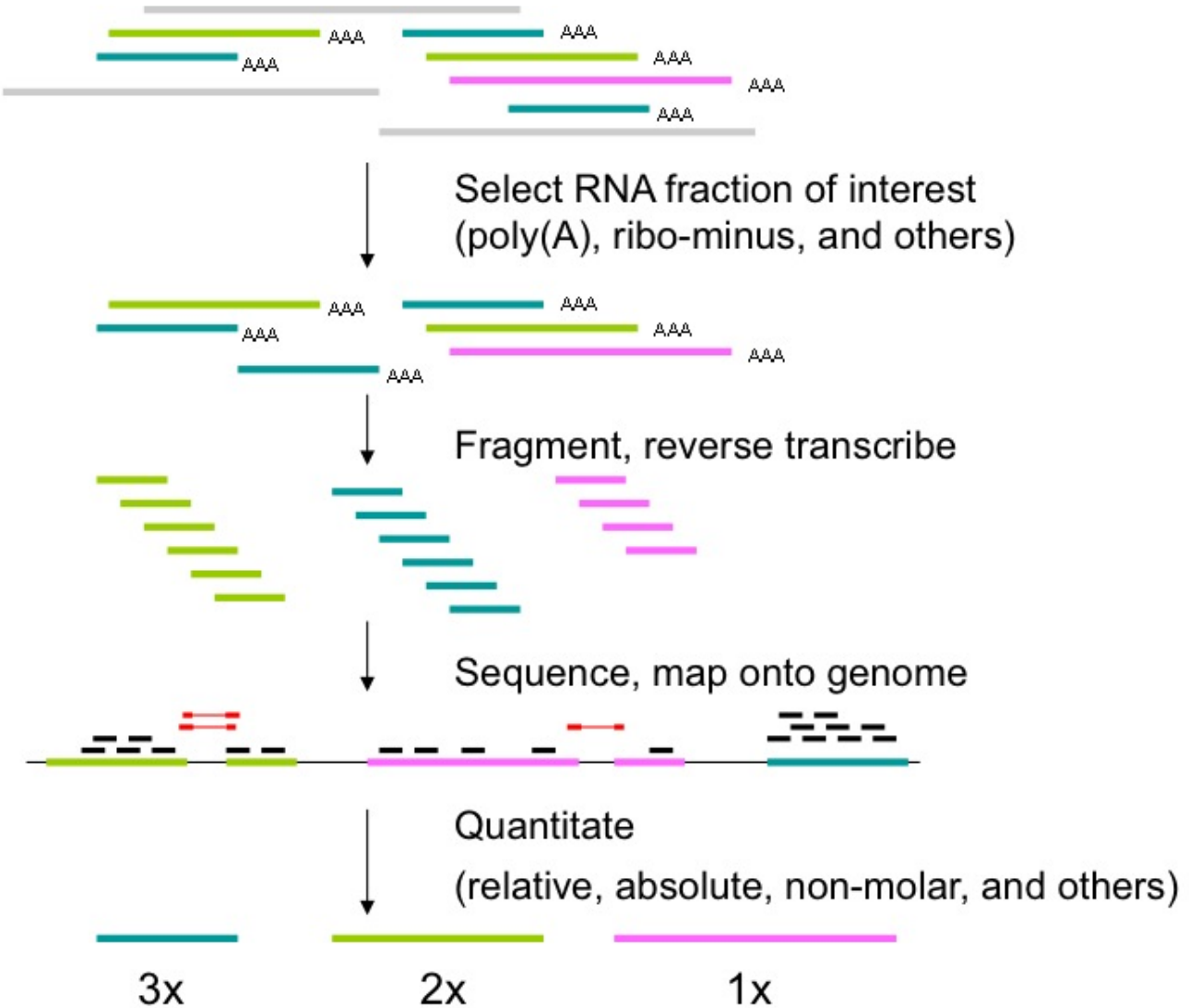
- New genome — qualitative
  - Sequence DNA ⇒ assemble
- Genomic variation — qualitative
  - Sequence DNA ⇒ assemble ⇒ align ⇒ detect
  - Sequence ⇒ align to DNA ⇒ detect
- New gene/transcript — qualitative
  - Sequence RNA ⇒ assemble
  - Sequence RNA ⇒ align to DNA ⇒ detect
- Gene/transcript expression — quantitative
  - Sequence RNA ⇒ align to known RNAs ⇒ count
  - Sequence RNA ⇒ align to DNA ⇒ count
- DNA/RNA-Protein binding — quantitative
  - Binding assay ⇒ Sequence ⇒ align to DNA ⇒ identify peaks
- Methylation — quant- & qualitative
  - Bisulfit treatment ⇒ Sequence DNA ⇒ align to DNA⇒ count
- ...

Biological Question, Sequencing & Analysis

- New genome — qualitative
  - Sequence DNA ⇒ assemble
- Genomic variation — qualitative
  - Sequence DNA ⇒ assemble ⇒ align ⇒ detect
  - Sequence ⇒ align to DNA ⇒ detect
- New gene/transcript — qualitative
  - Sequence RNA ⇒ assemble
  - Sequence RNA ⇒ align to DNA ⇒ detect
- Gene/transcript expression — quantitative
  - Sequence RNA ⇒ align to known RNAs ⇒ count
  - Sequence RNA ⇒ align to DNA ⇒ count
- DNA/RNA-Protein binding — quantitative
  - Binding assay ⇒ Sequence ⇒ align to DNA ⇒ identify peaks
- Methylation — quant- & qualitative
  - Bisulfit treatment ⇒ Sequence DNA ⇒ align to DNA⇒ count
- ...

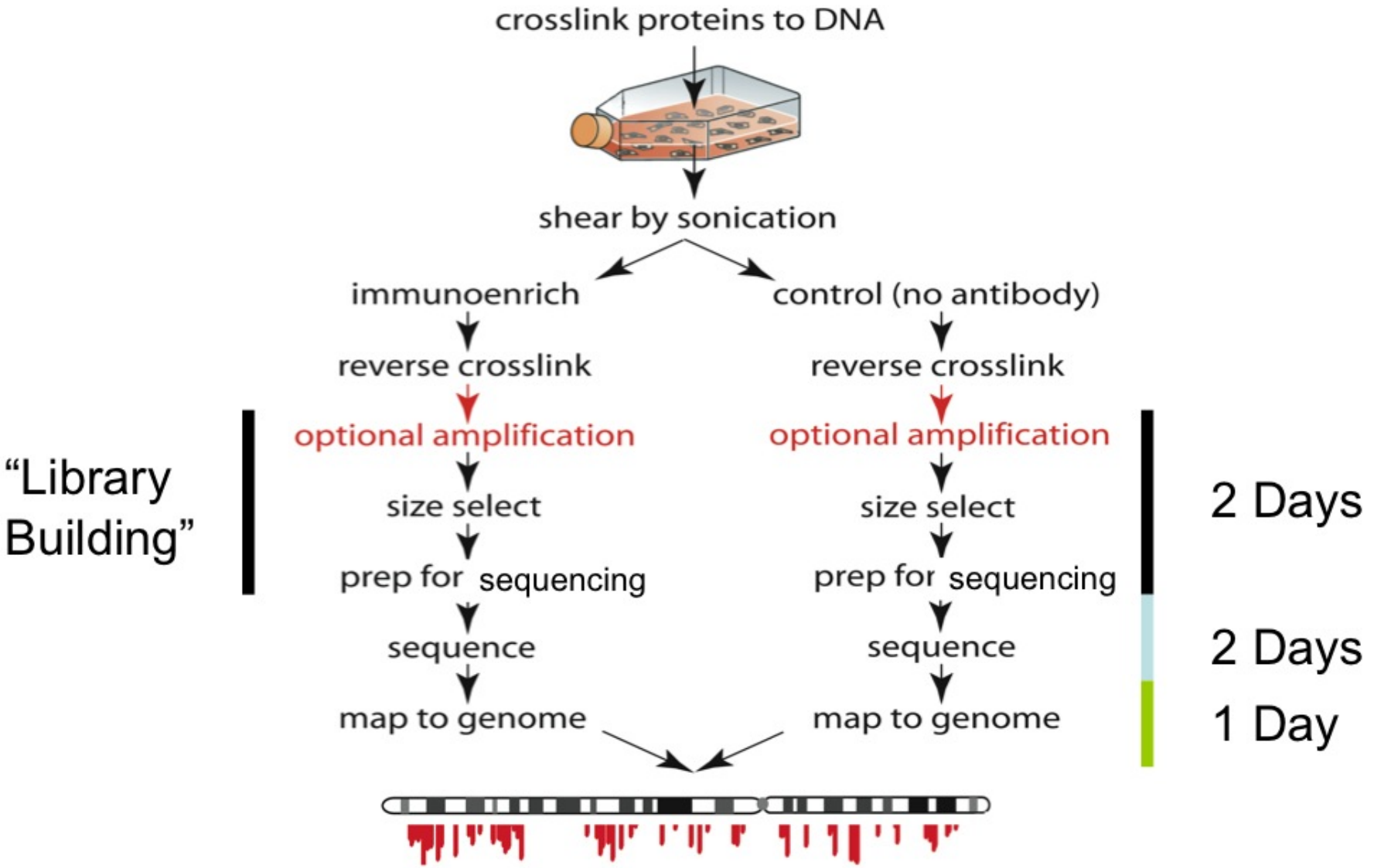
RNA-seq

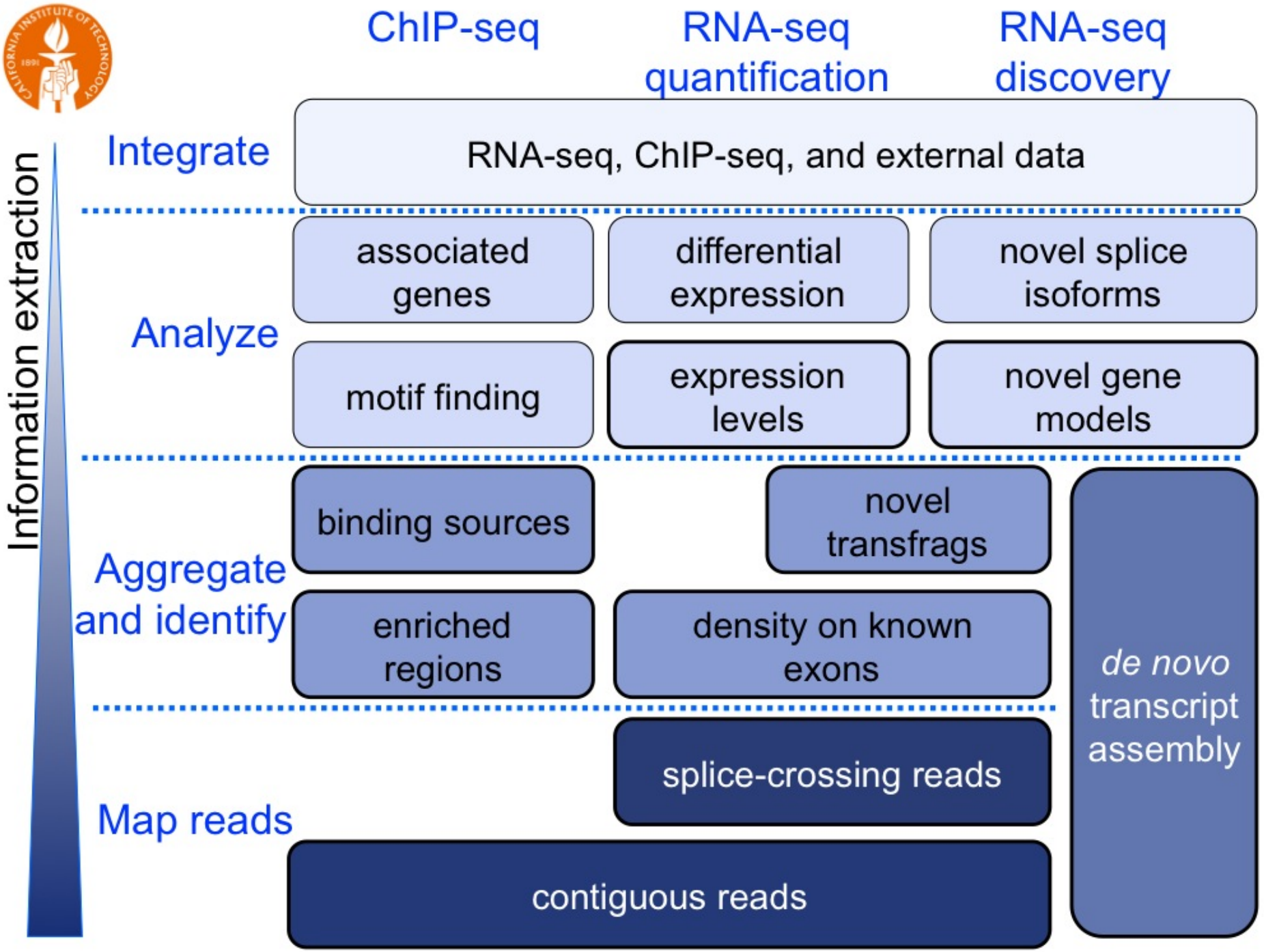
A digital counting method for transcriptome discovery and quantification



ChIP-seq

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




ChIP-seq
RNA-seq quantification
RNA-seq discovery
Information extraction
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
de novo transcript assembly
Map reads
splice-crossing reads
contiguous reads

RNA-seq: 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.

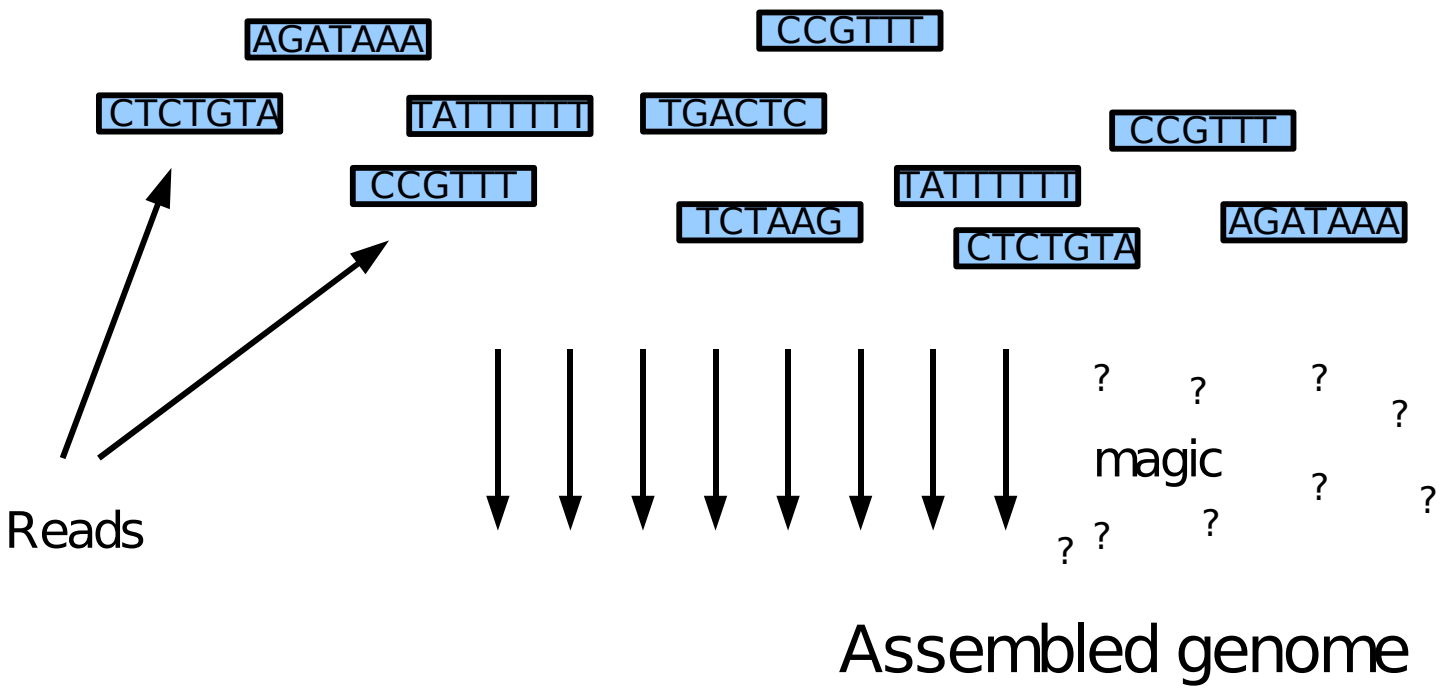
Read Analysis I

- Assembly

  $\Rightarrow$ generate contigs
- Mapping/Alignments

  $\Rightarrow$ map/align reads back to a known genome
- Quantification

  $\Rightarrow$ Estimate abundances of transcripts/binding . . .



Read Analysis I

- Assembly
  ⇒ generate contigs
- Mapping/Alignments
  ⇒ map/align reads back to a known genome
- Quantification
  ⇒ Estimate abundances of transcripts/binding . . .

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

Read Analysis I

- Assembly
  $\Rightarrow$ generate contigs
- Mapping/Alignments
  $\Rightarrow$ map/align reads back to a known genome
- Quantification
  $\Rightarrow$ Estimate abundances of transcripts/binding ...

Problem: hundreds of millions of reads of short length

$\Rightarrow$ Big computational challenge

Read Analysis - Mapping

Read 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

Example: 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
```

Example: 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
```

Example: 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
```

Example: 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
```

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:

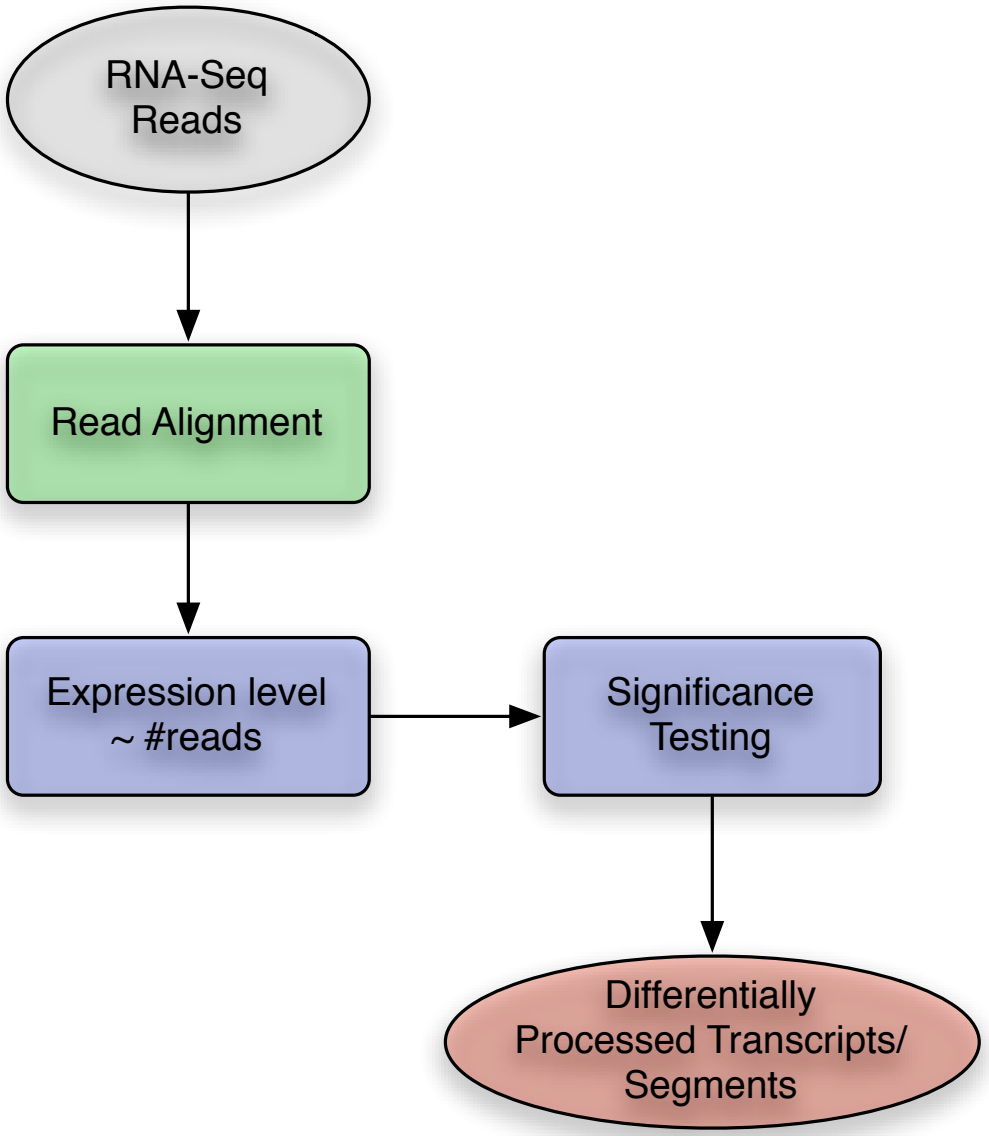
- *blast* [Altschul et al., 1990], *spliced alignments* [Gelfand et al., 1996], *sim4* [Florea et al., 1998], *GeneSeqer* [Usuka et al., 2000], *Spidey* [Wheelan SJ, 2001], *blat* [Kent, 2002], *exalin* [Zhang and Gish, 2006], *Palma* [Schulze et al., 2007]

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

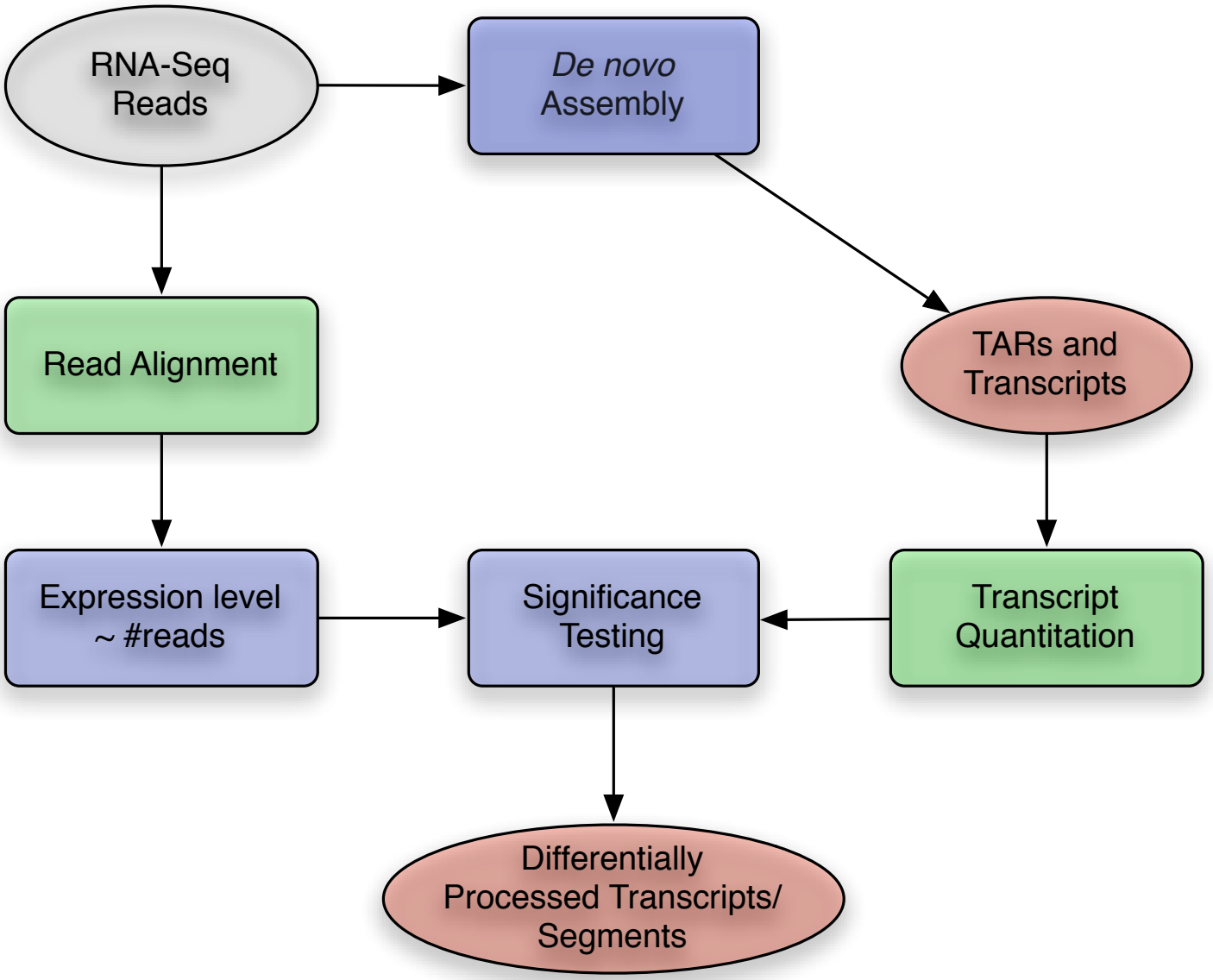
specific for spliced NGS read alignment:

- 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)

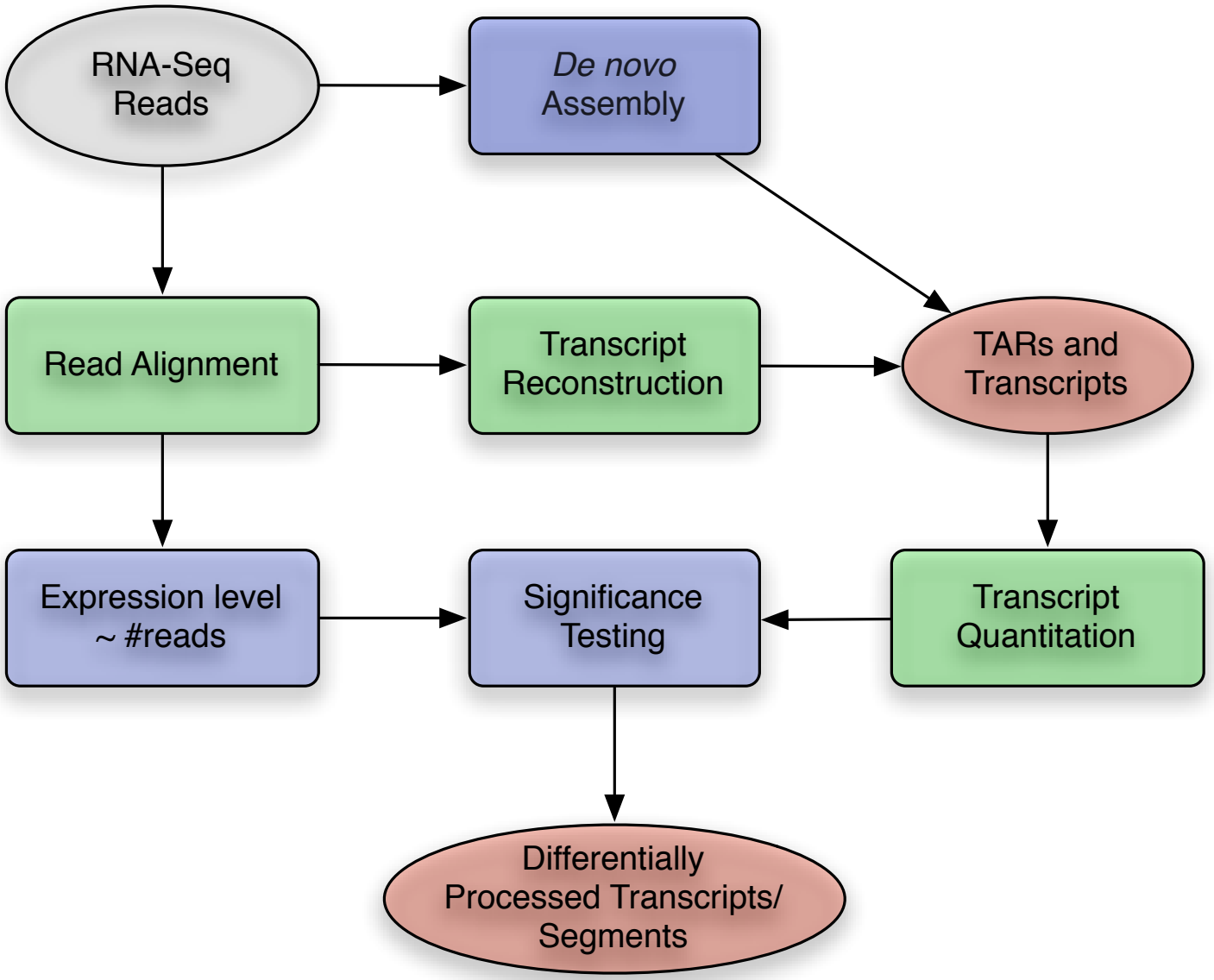
Common RNA-Seq Analysis Steps



Common RNA-Seq Analysis Steps



Common RNA-Seq Analysis Steps



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)**

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)**

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):

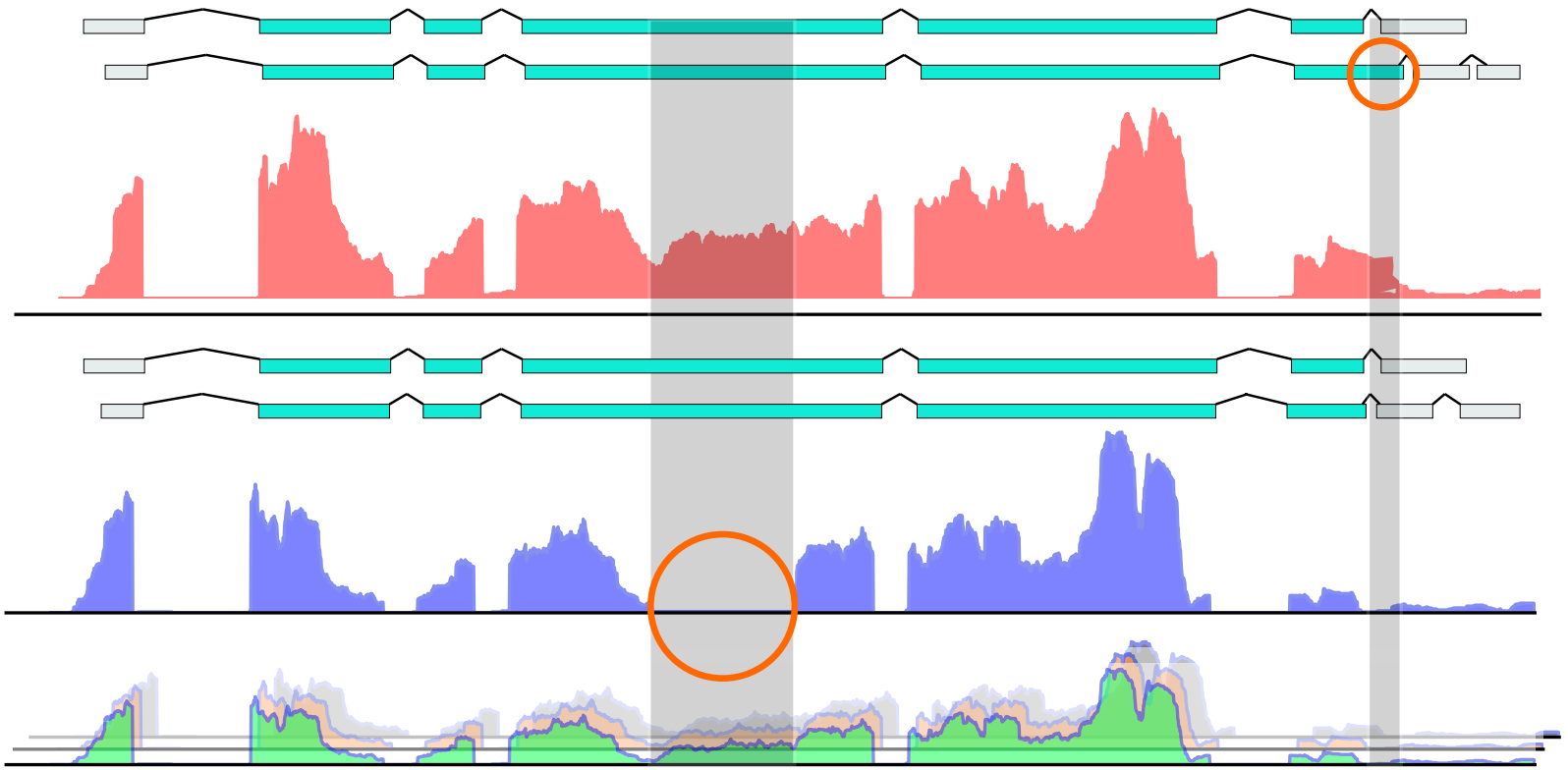
$\Rightarrow$ **Fragments per kilobase per million mapped reads (FPKM)**

Estimate Gene Expression: Caveats

- RPKM/FPKM values are strongly dependent on the expression level of the highest expressed genes (largest fraction of reads, e.g. rRNA contamination)
- Effect of genomic variation
- Alternative transcripts/RNA-processing may lead to differential read counts

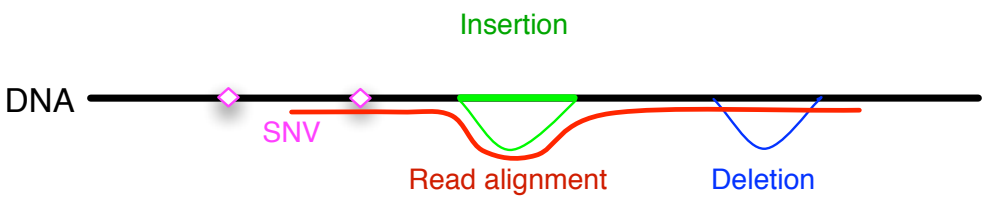
Estimate Gene Expression: Caveats

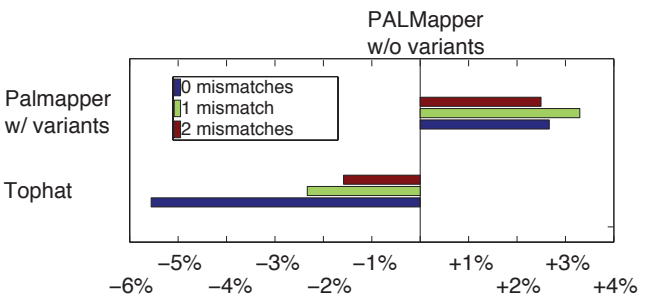
- RPKM/FPKM values are strongly dependent on the expression level of the highest expressed genes (largest fraction of reads, e.g. rRNA contamination)
- Effect of genomic variation



Estimate Gene Expression: Caveats

- RPKM/FPKM values are strongly dependent on the expression level of the highest expressed genes (largest fraction of reads, e.g. rRNA contamination)
- Effect of genomic variation

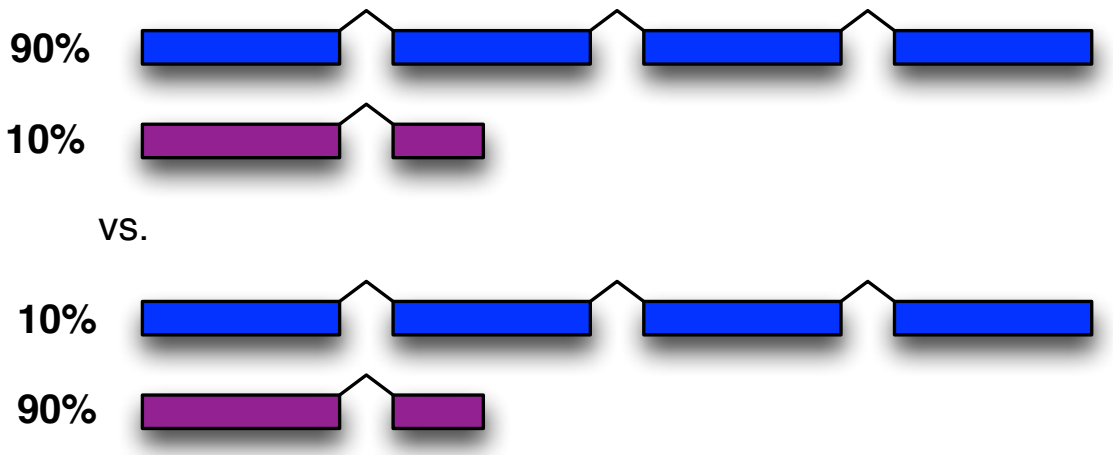




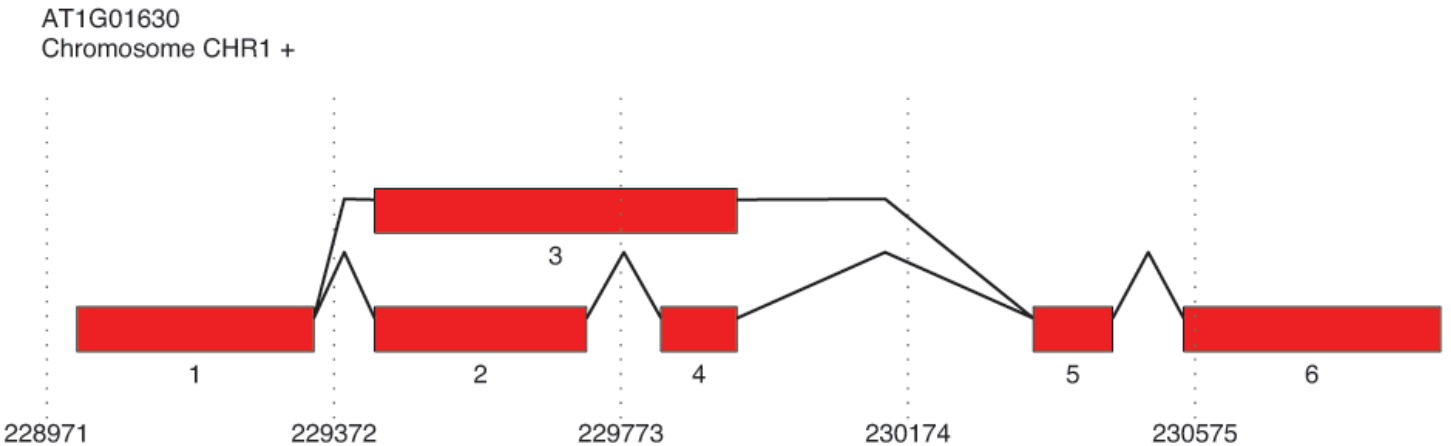
- Alternative transcripts/RNA-processing may lead to differential read counts

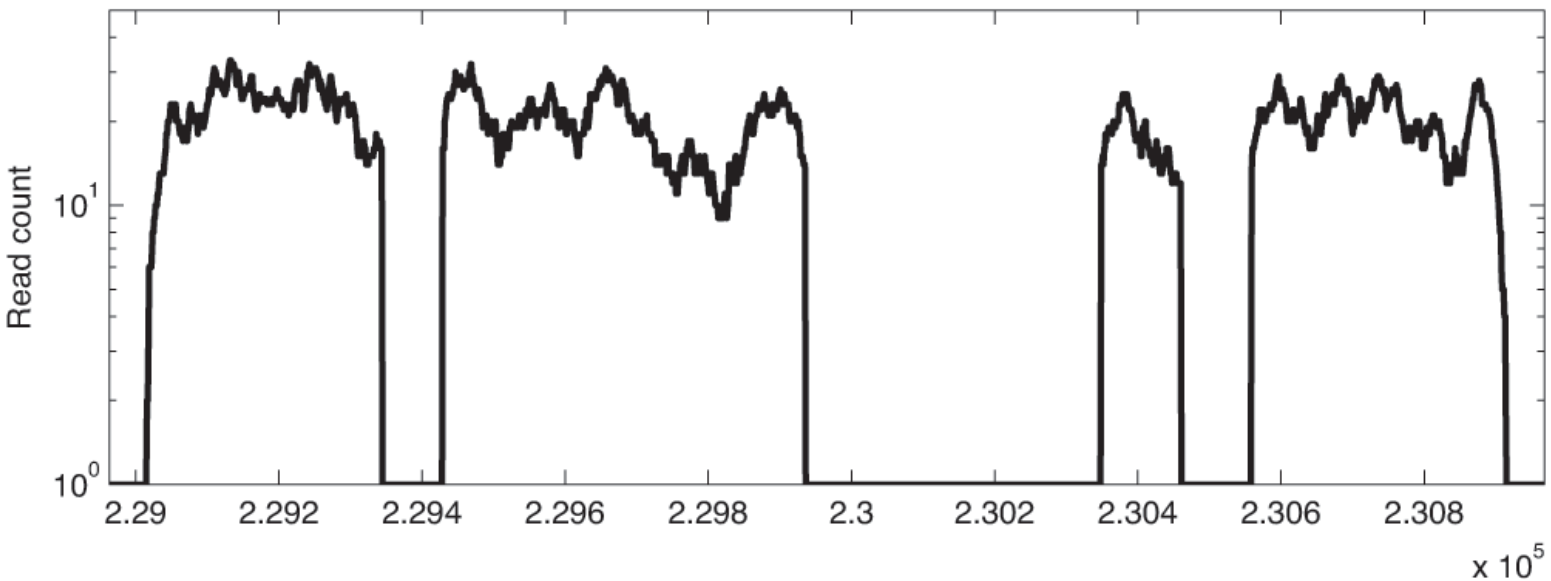
Estimate Gene Expression: Caveats

- RPKM/FPKM values are strongly dependent on the expression level of the highest expressed genes (largest fraction of reads, e.g. rRNA contamination)
- Effect of genomic variation
- Alternative transcripts/RNA-processing may lead to differential read counts

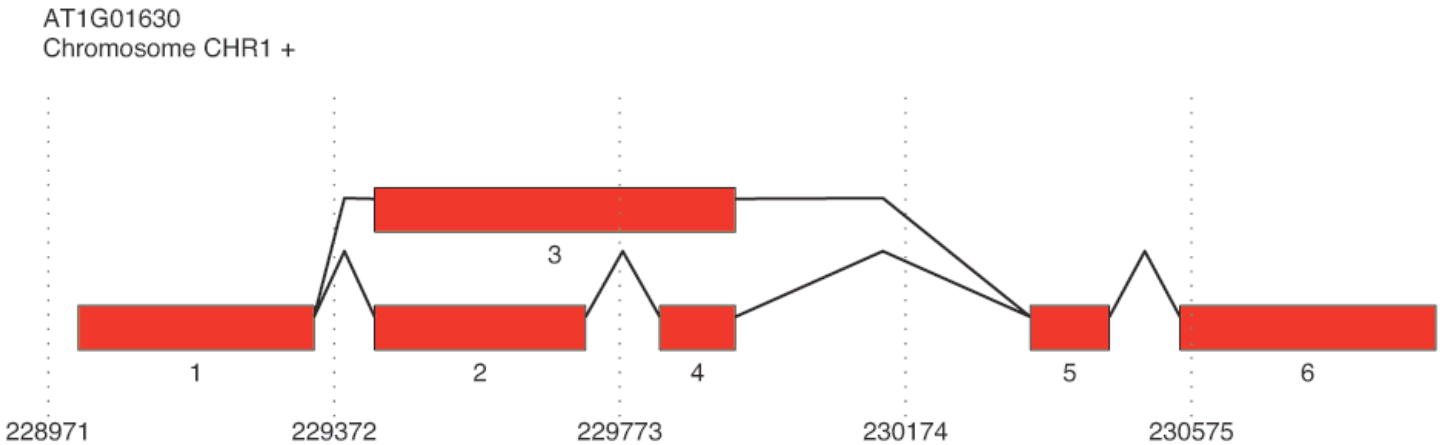


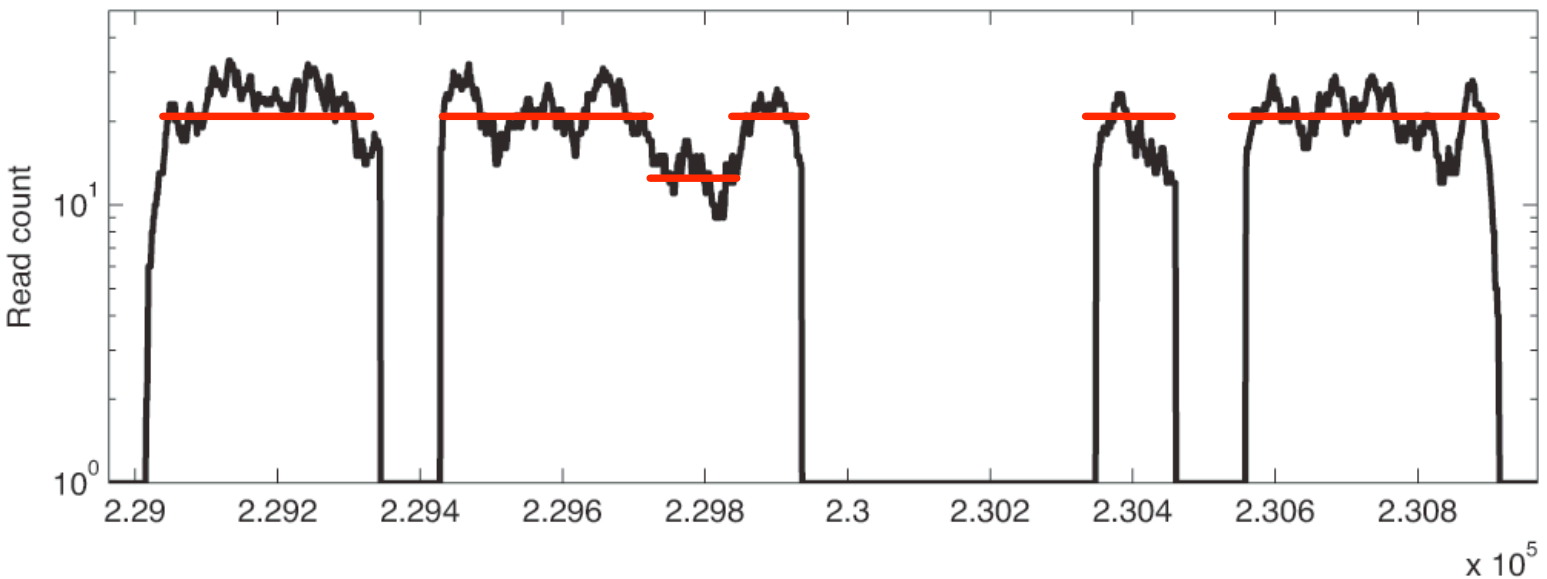
Quantitation of Transcripts





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, 2009]
- Squared differences using quadratic programming [Bohnert et al., 2009]
- (approximate) Negative Binomial distribution [Behr et al., 2013]

Other methods: [Li et al., 2010], [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:

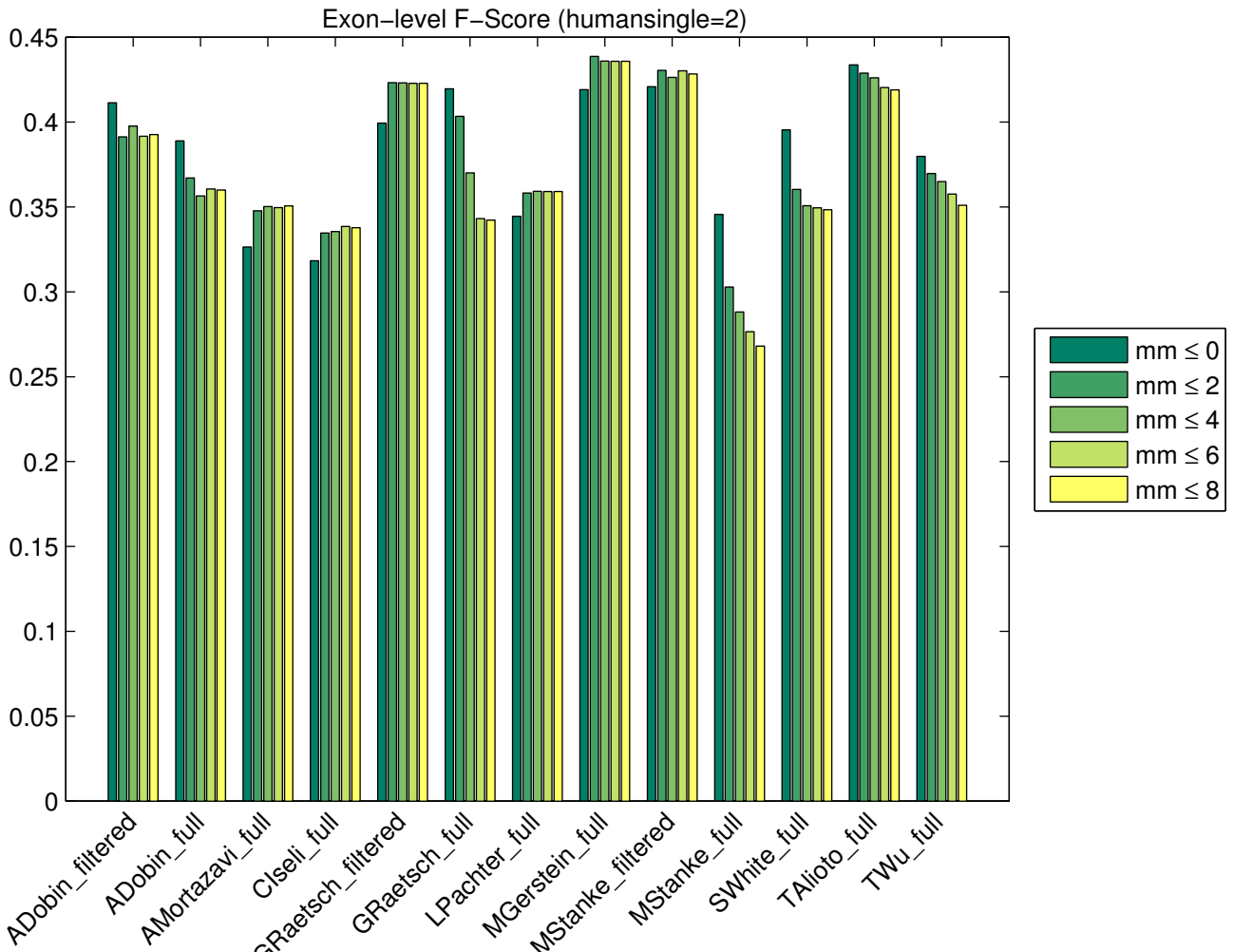
- Poisson distributions [Jiang and Wong, 2009]
- Absolute differences using a flow-network [Sammeth, 2009]
- Squared differences using quadratic programming [Bohnert et al., 2009]
- (approximate) Negative Binomial distribution [Behr et al., 2013]

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

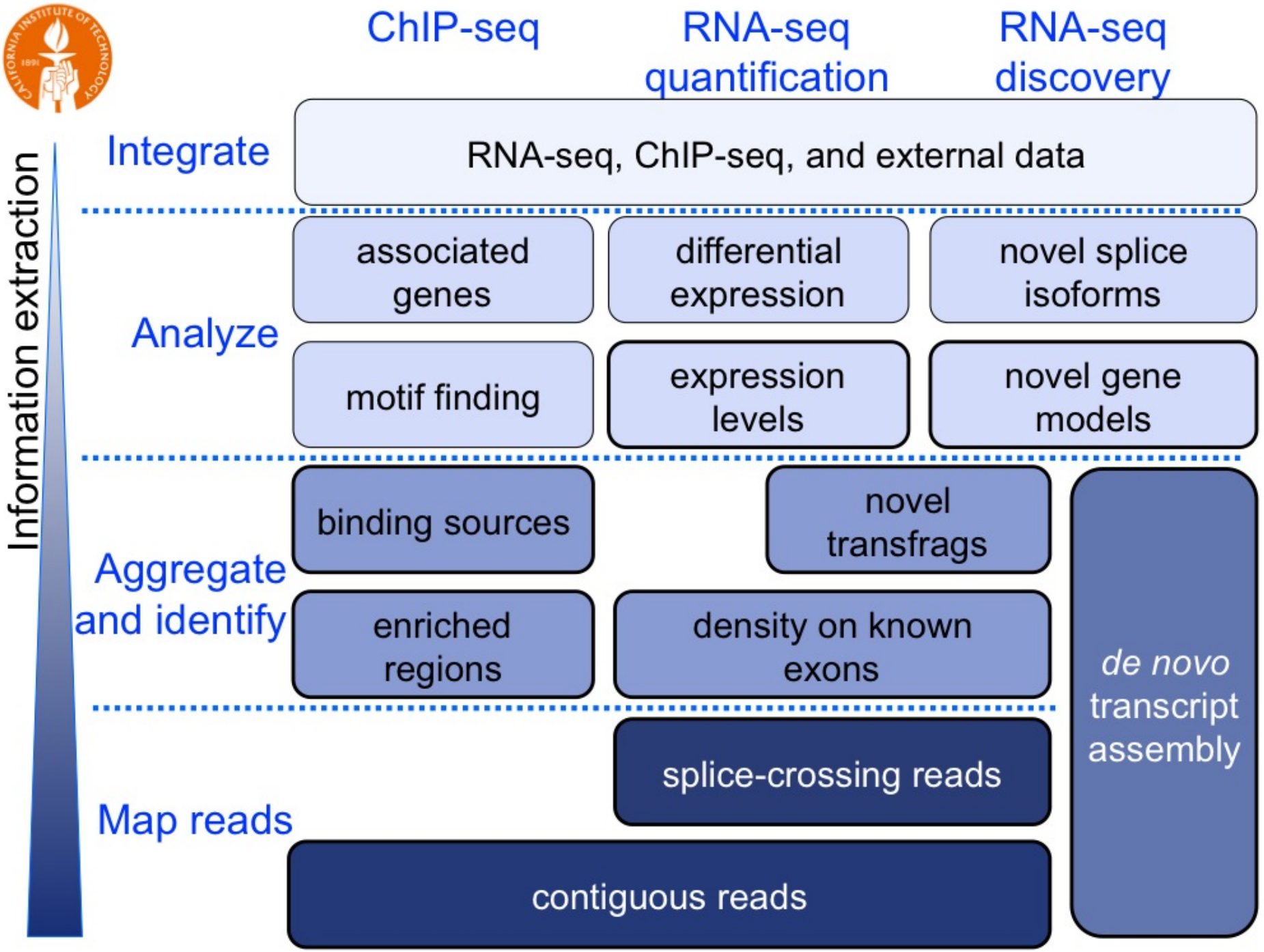
Problems:

- 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

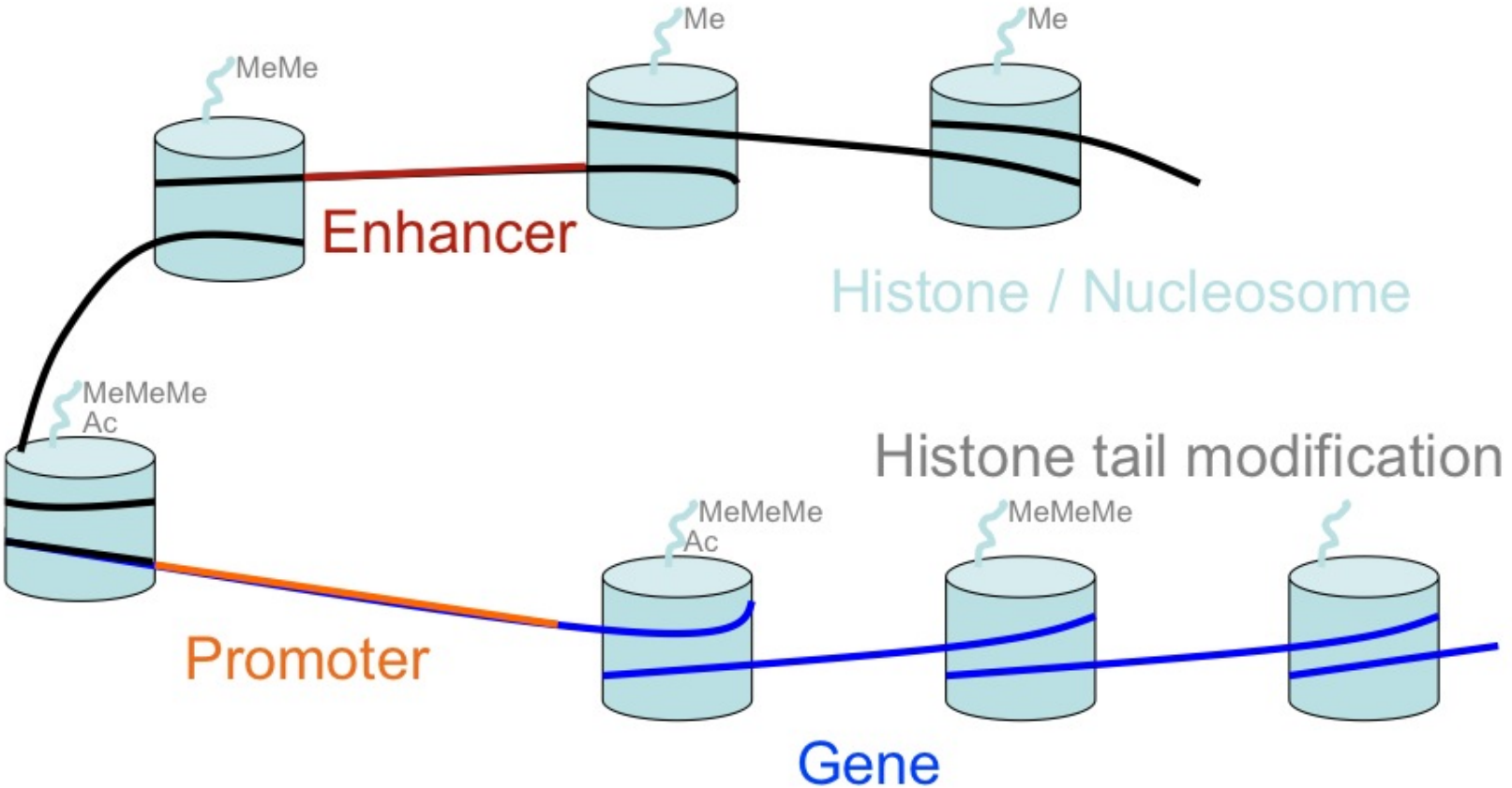
Effects of Alignents on Downstream Analysis
(Cufflinks: Human - Exon F-score)



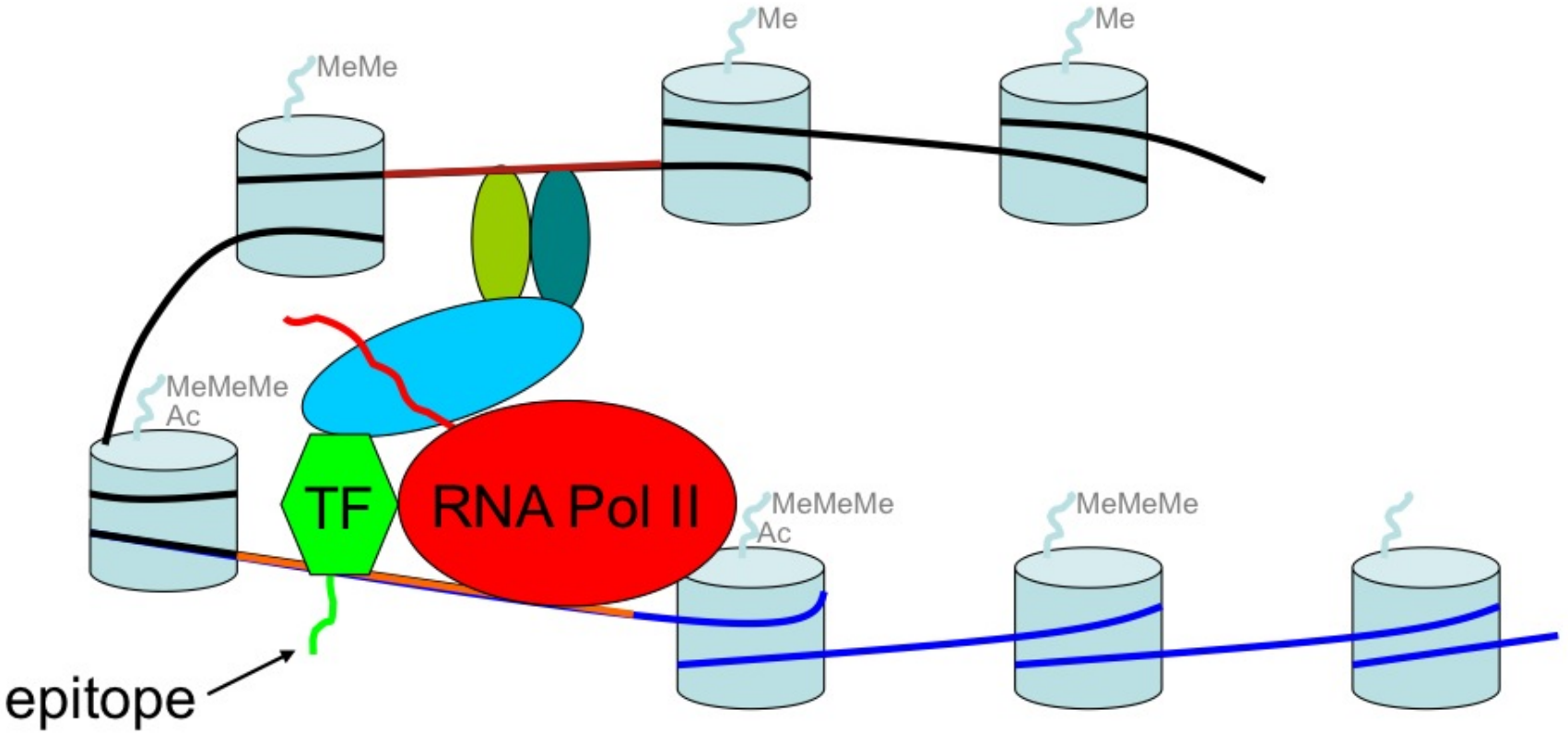
Filter: by max edit ops (0 – 8); prediction F-Score (exon level)


ChIP-seq
RNA-seq quantification
RNA-seq discovery
Information extraction
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
de novo transcript assembly
Map reads
splice-crossing reads
contiguous reads

Chromatin Immunoprecipitation (ChIP)

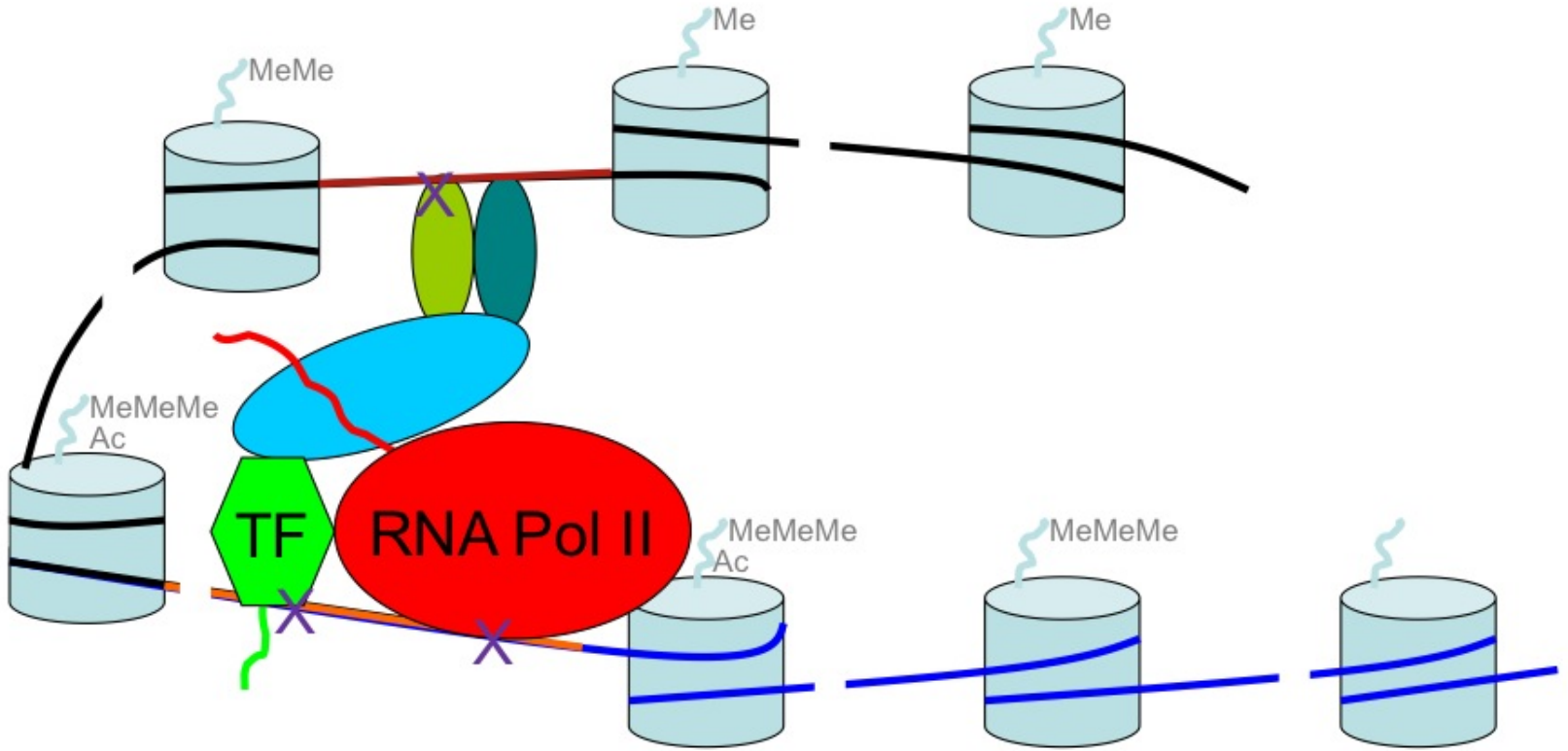


Chromatin Immunoprecipitation (ChIP)



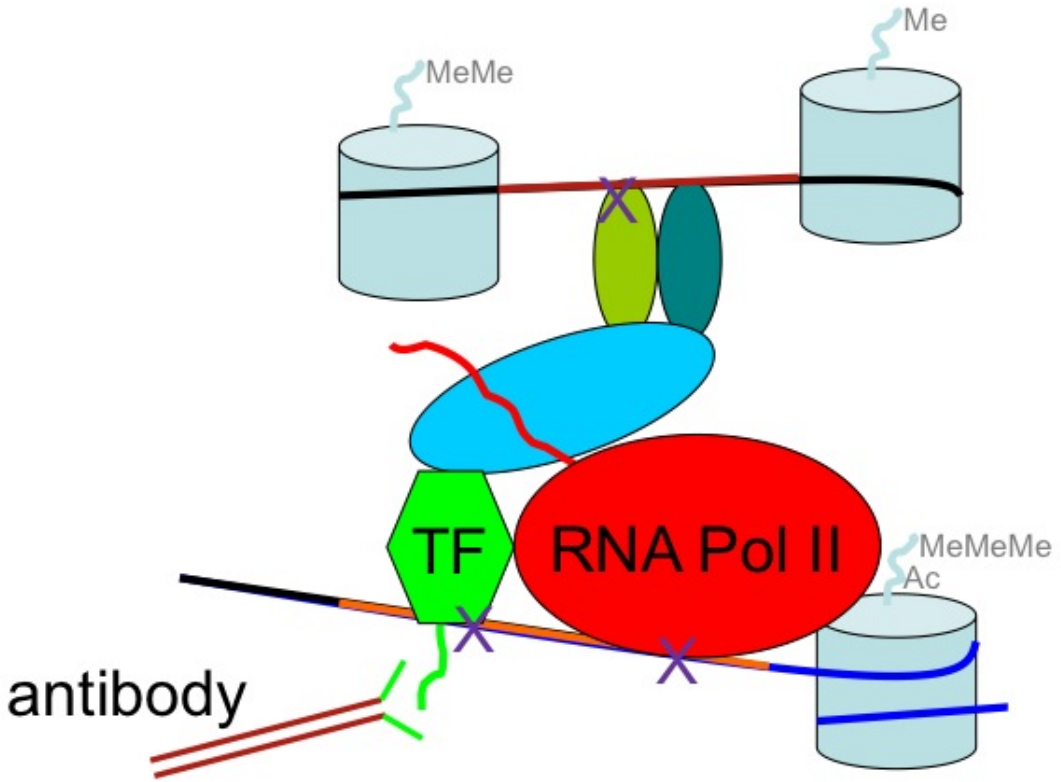
Transcription Factors + histones + DNA = chromatin

Chromatin Immunoprecipitation (ChIP)



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

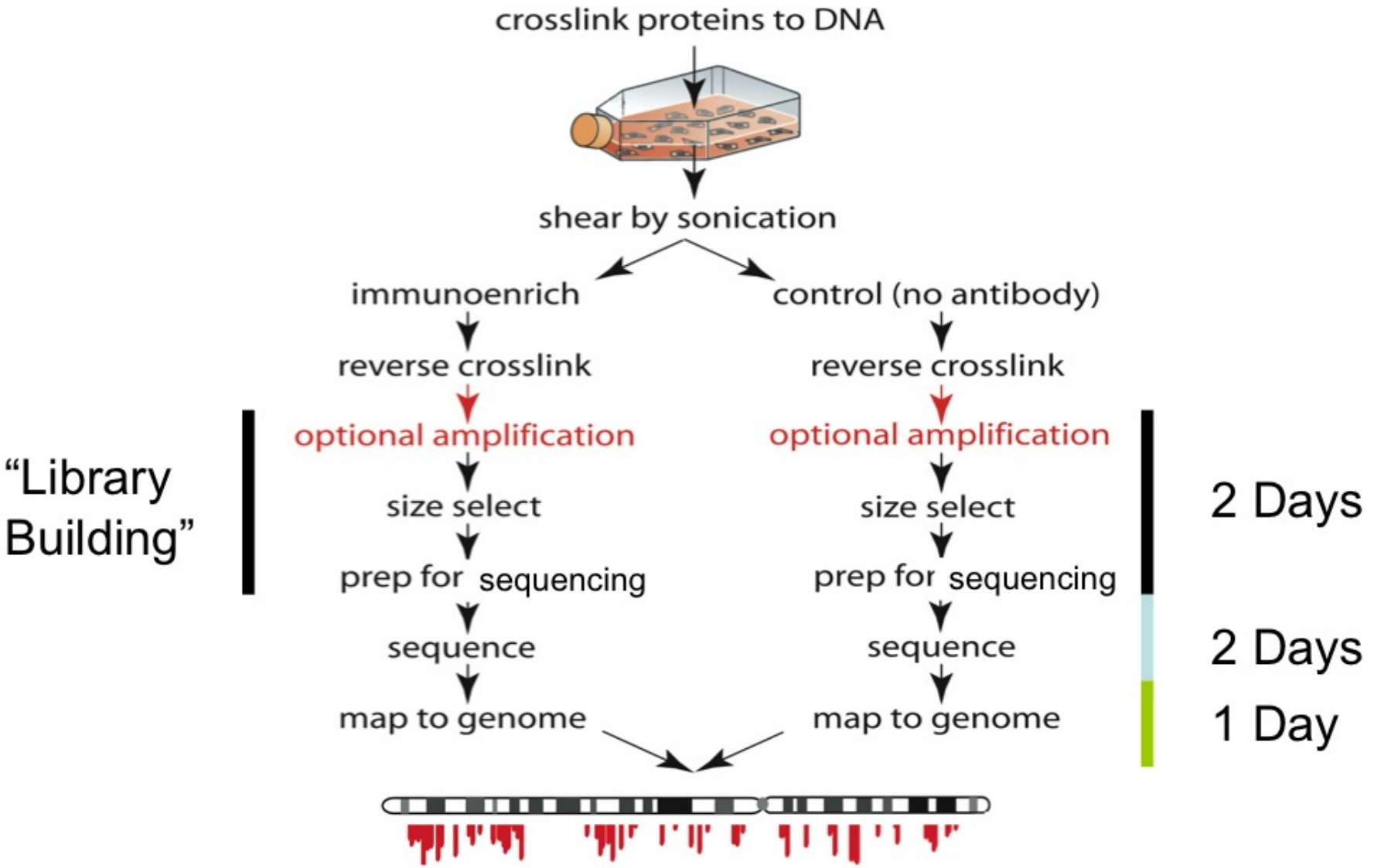
Chromatin Immunoprecipitation (ChIP)

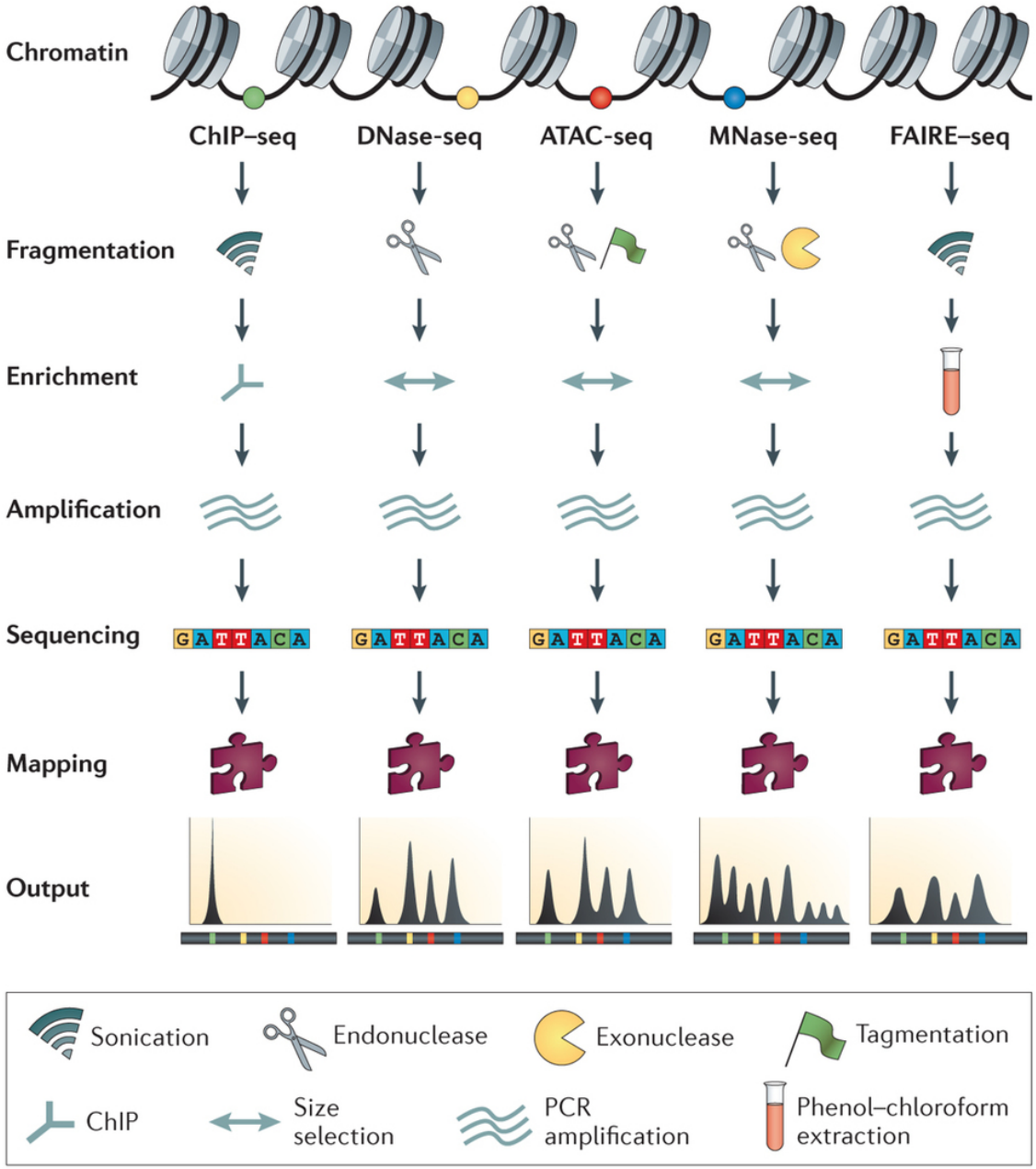


3. Use antibody specific to a factor to retrieve DNA fragments that are (not necessarily directly) bound.
4. Reverse crosslinks and sequence ends of fragments.

ChIP-seq

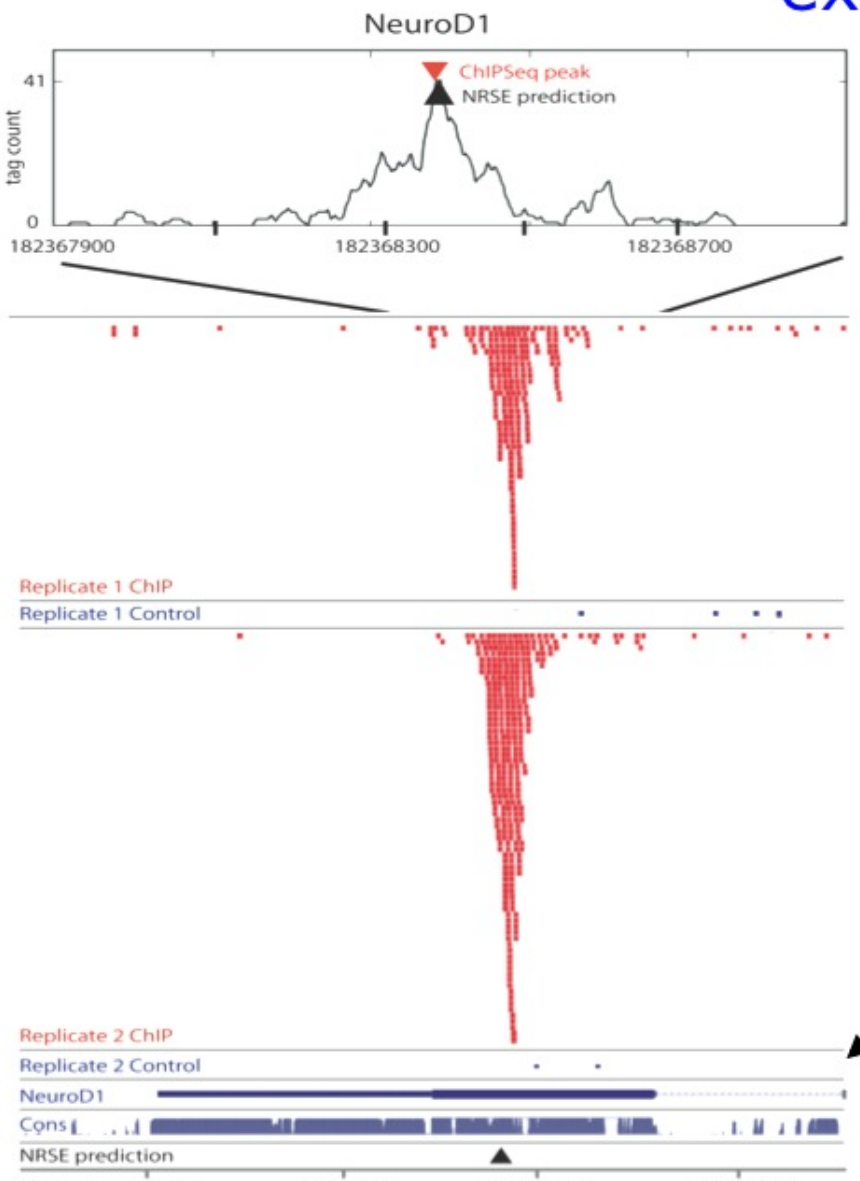
A digital counting method to score site occupancy by DNA binding proteins

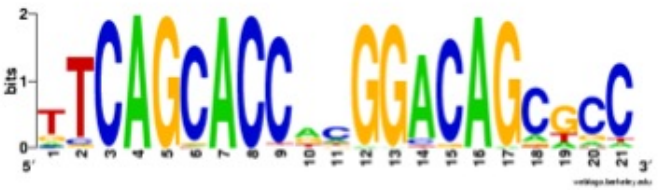




Nature Reviews | Genetics

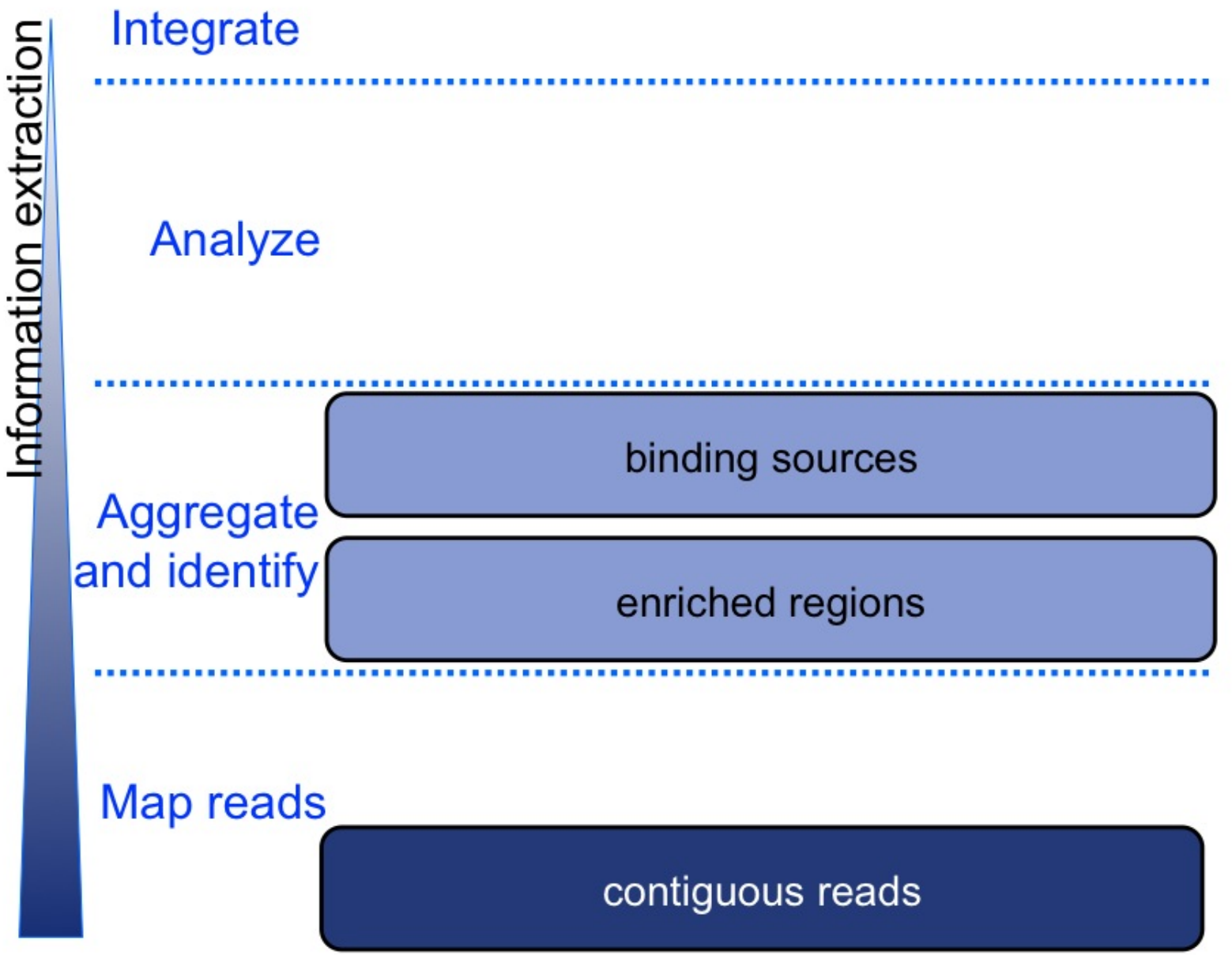
ChIP-Seq identifies NRSF occupancy in NeuroD1 exon





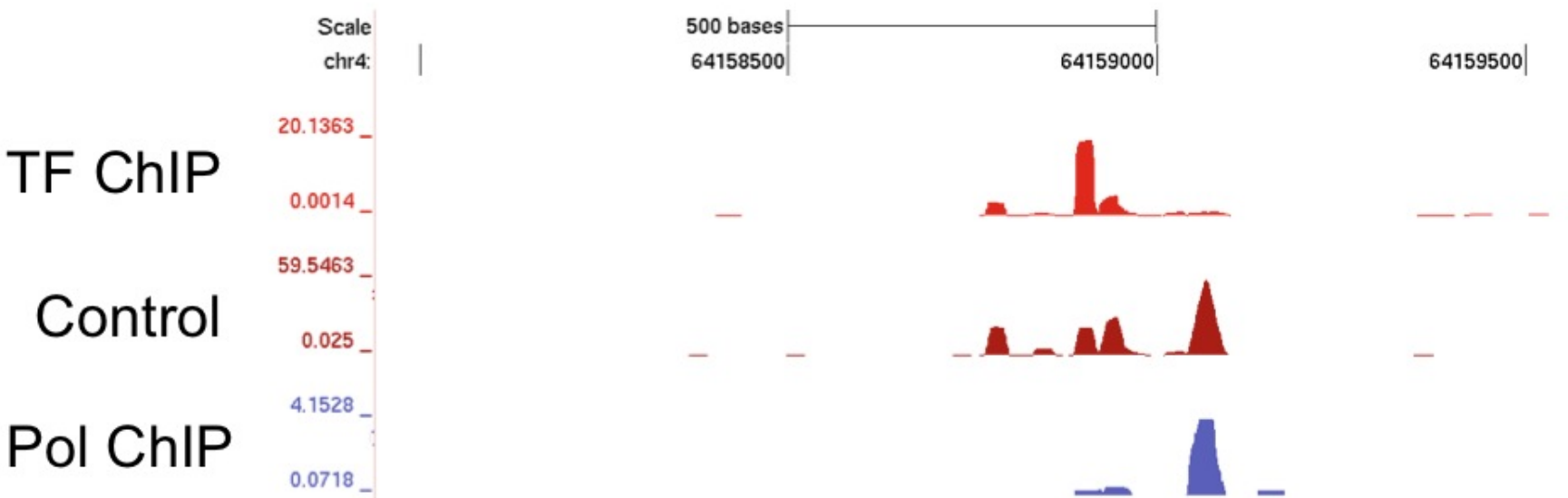
Controls

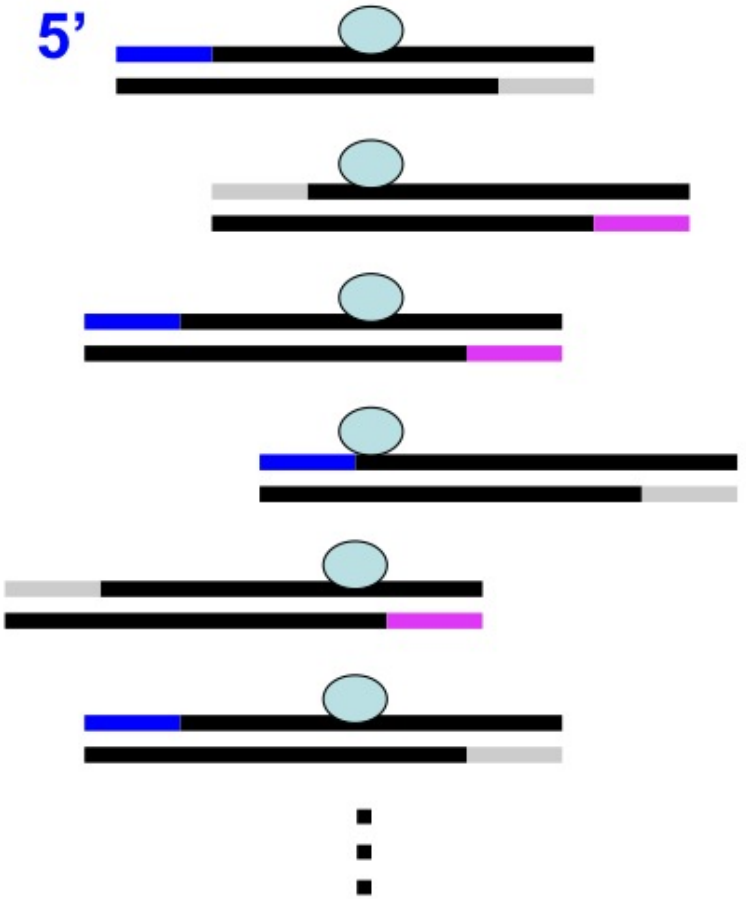
(Johnson et al, 2007)


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

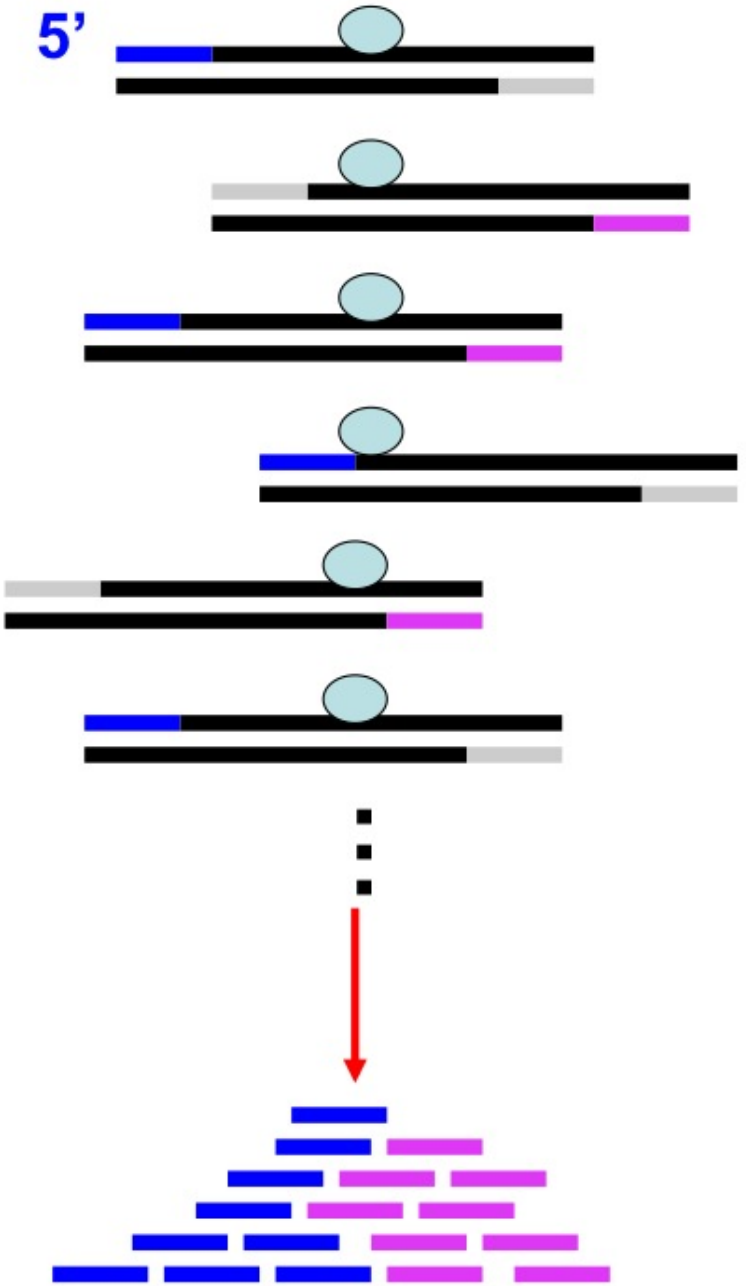
Why do we need a control ?

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



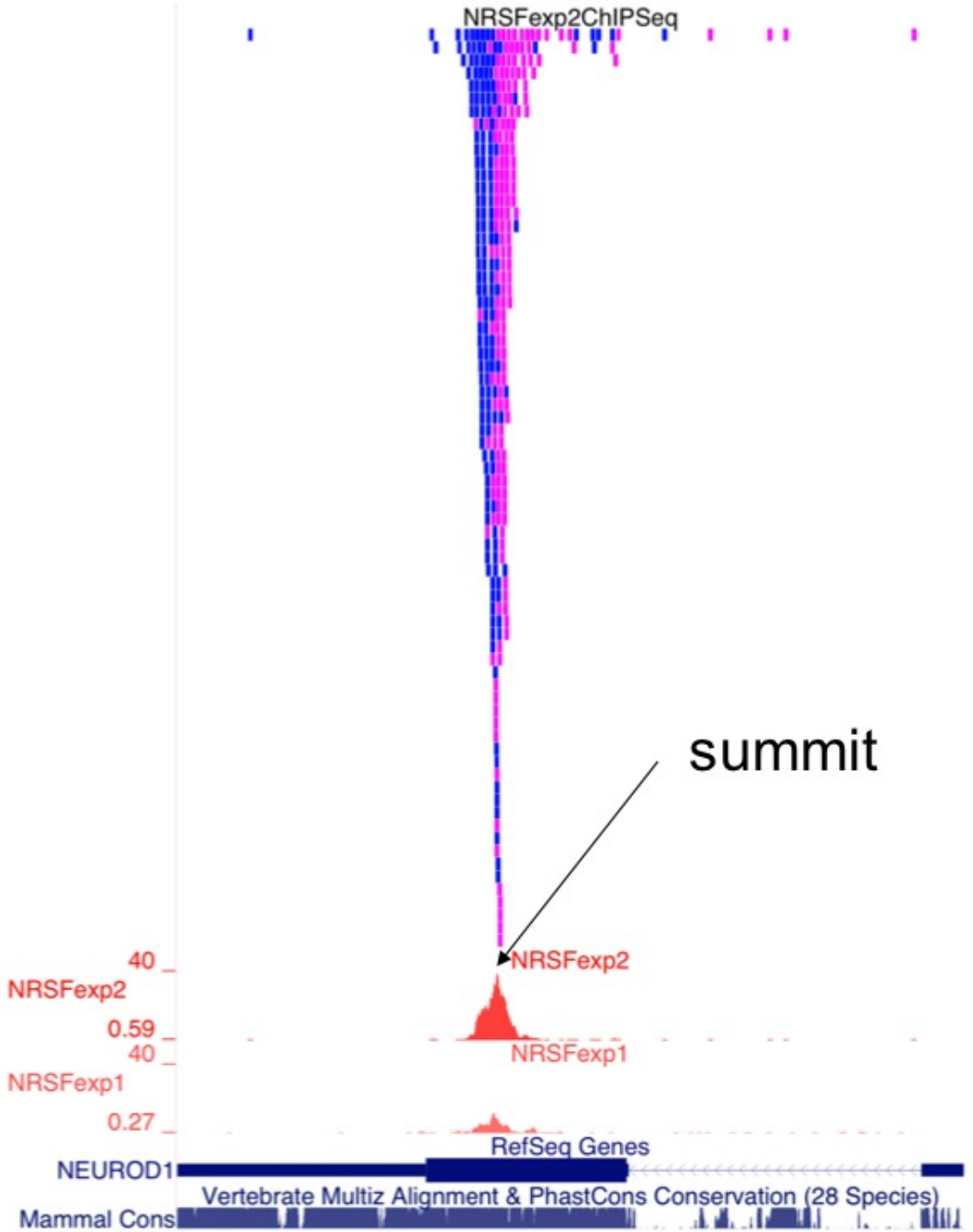


Since Illumina uses polymerase, we are always sequencing from 5’ end of fragment (blue or purple), typically observing one end of the fragment.



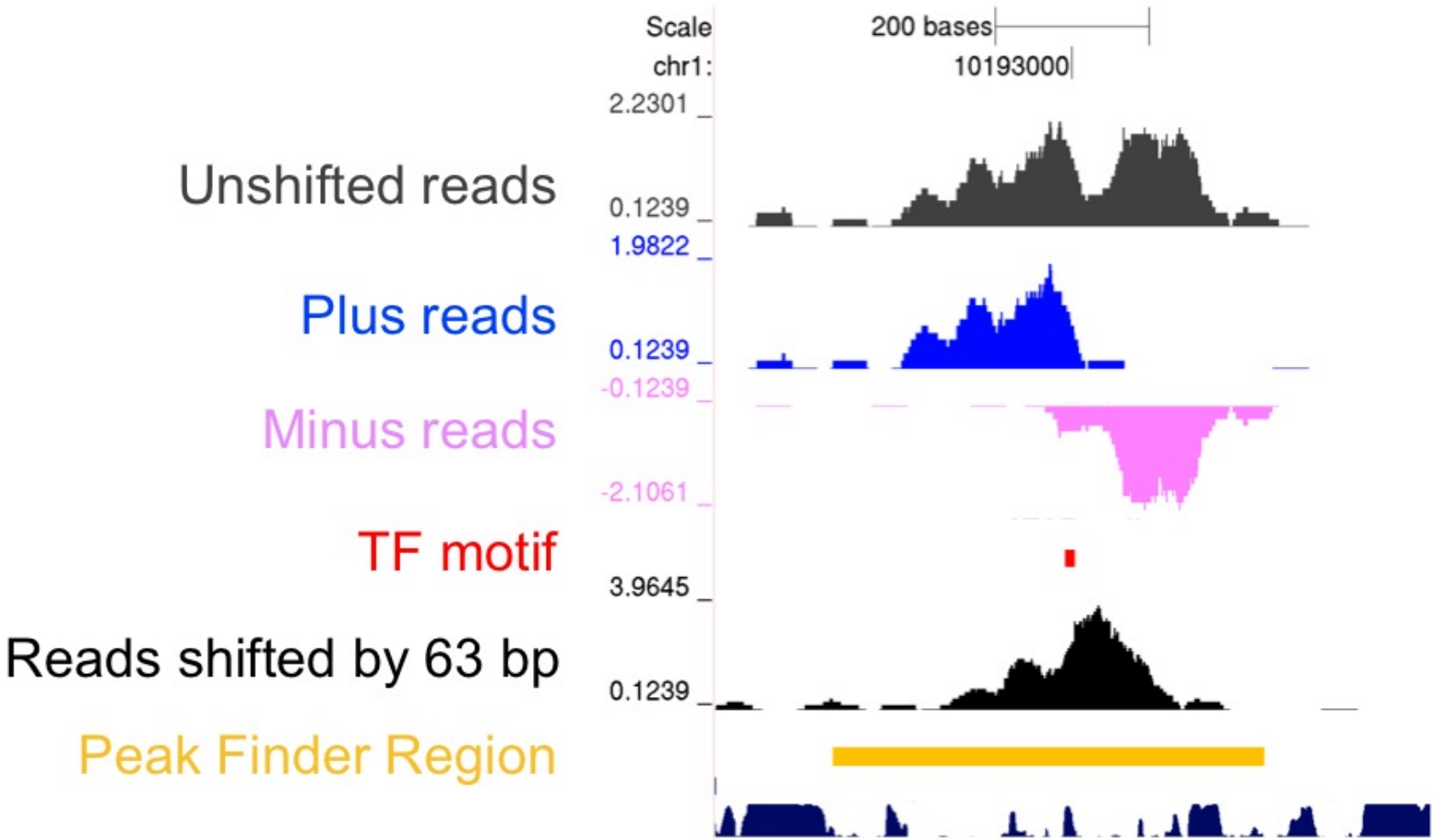
Since Illumina uses polymerase, we are always sequencing from 5’ end of fragment (blue or purple), typically observing one end of the fragment.

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

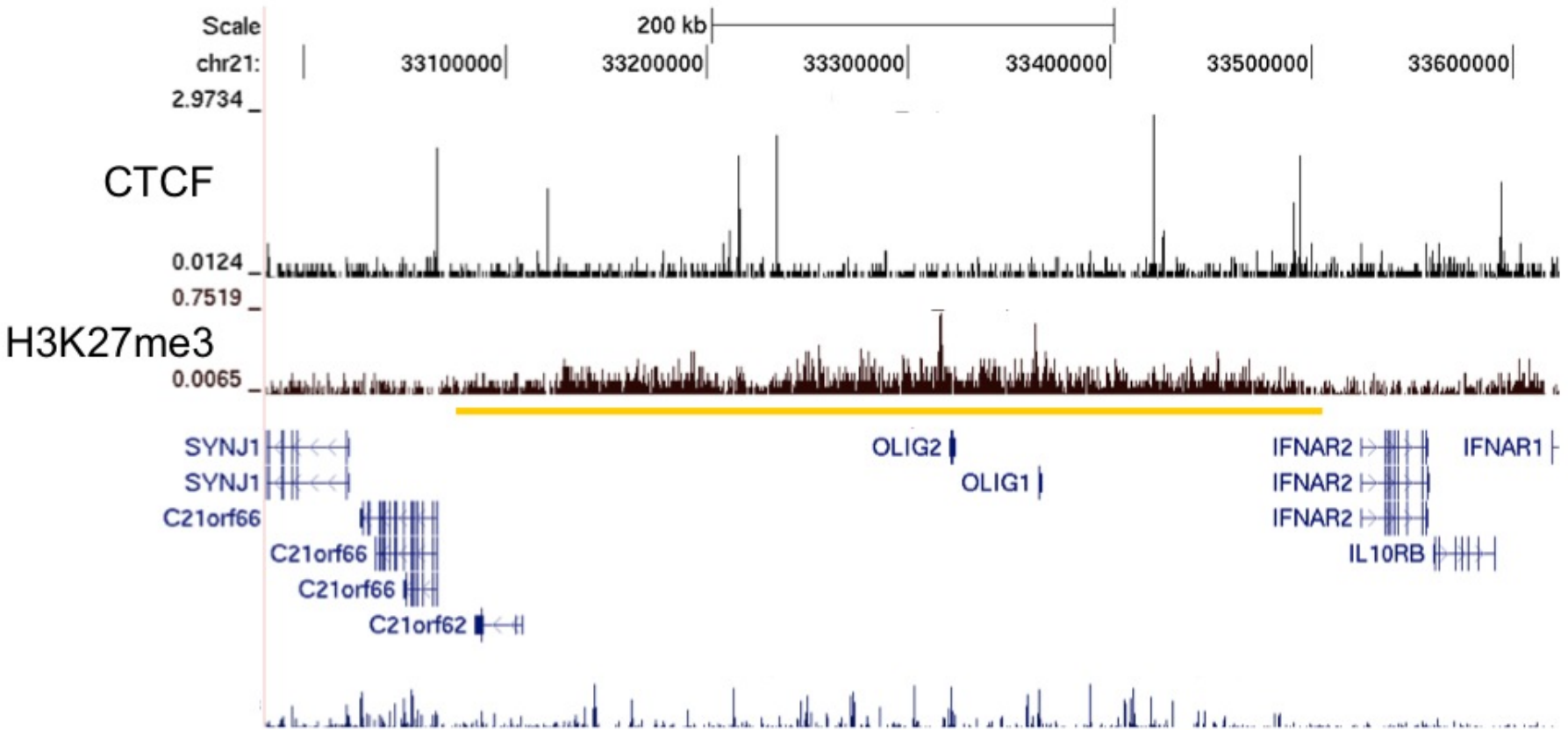
Narrow ChIP-seq peaks

Single source such as a Transcription Factor binding site

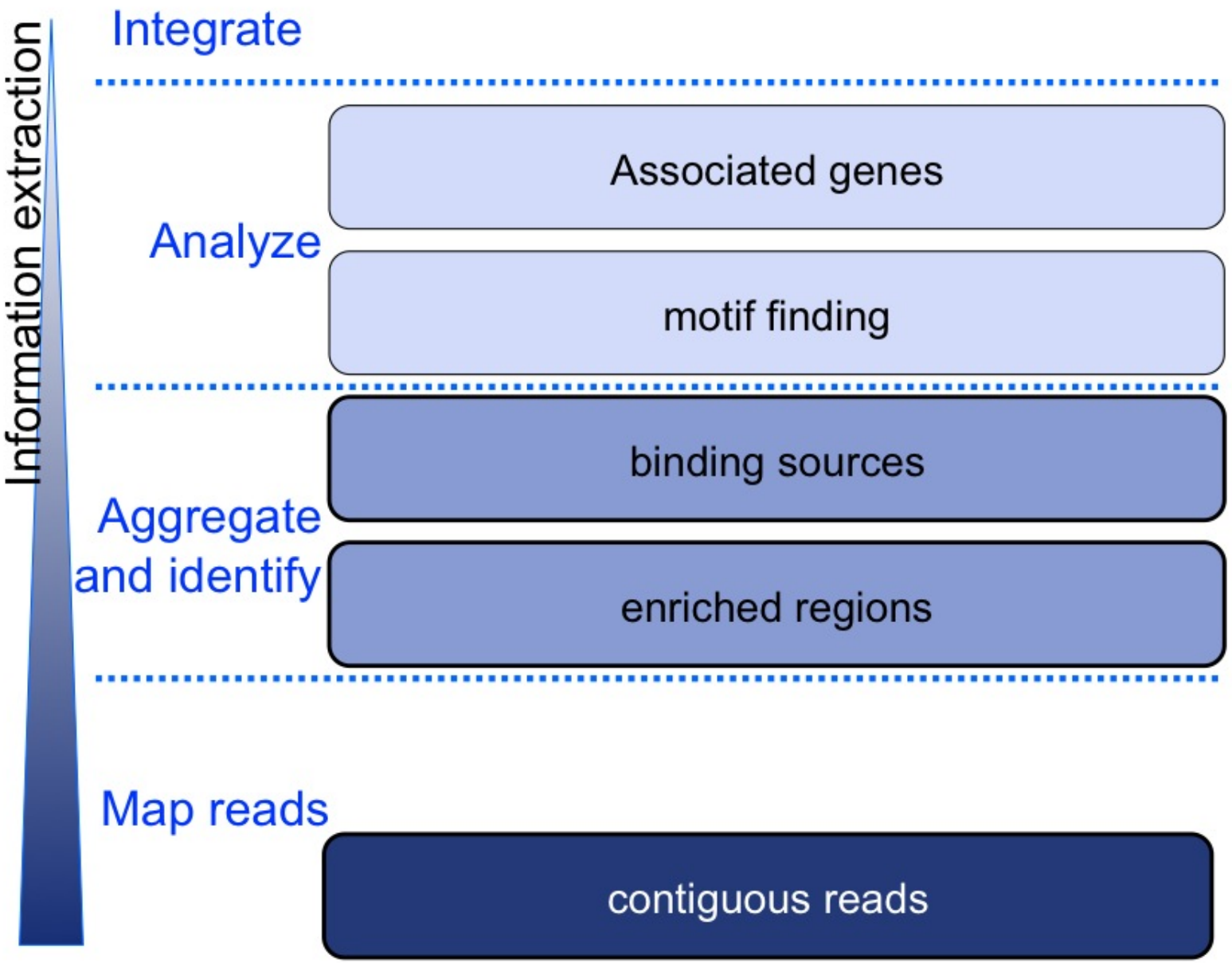


Broad ChIP-seq peaks

Some repressive chromatin histone modification marks



No benefit to shifting or extending - use a sliding window


Information extraction
Integrate
Associated genes
Analyze
motif finding
binding sources
Aggregate and identify
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

Introduction

Part I: Sequencing Basics

- The Rise of the Sequencers
- From Genome sequencing to Counting assays
- Biological question, sequencing, analysis

Part II: RNA-seq Basics

- Read Mapping and common analysis steps
- Gene and transcript quantification, Caveats

Part III: Chip-Seq Basics

- Enhancers, Promoters
- ChIP, DNase, ATAC, & friends

Part IV: Integration Approaches

- From DNA to Phenotype
- Integration into QTLs

The combinatorial problem

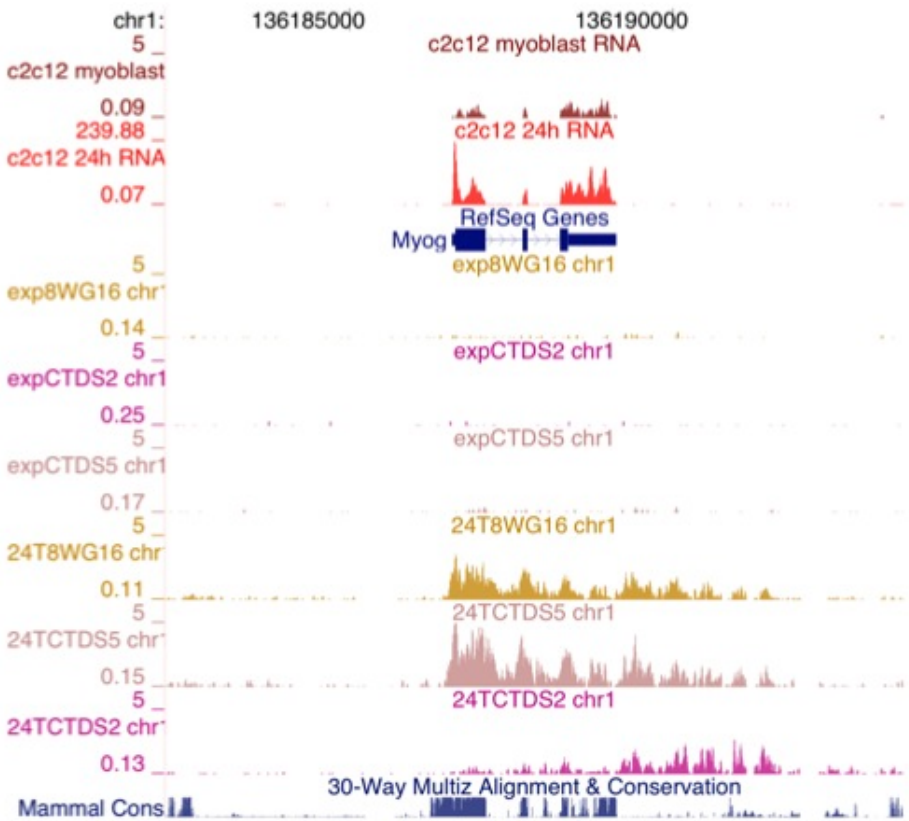
•It is now relatively easy to generate dozens of ChIP-seq and/or RNA-seq for a biological sample of interest and to analyze them singly.

•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.

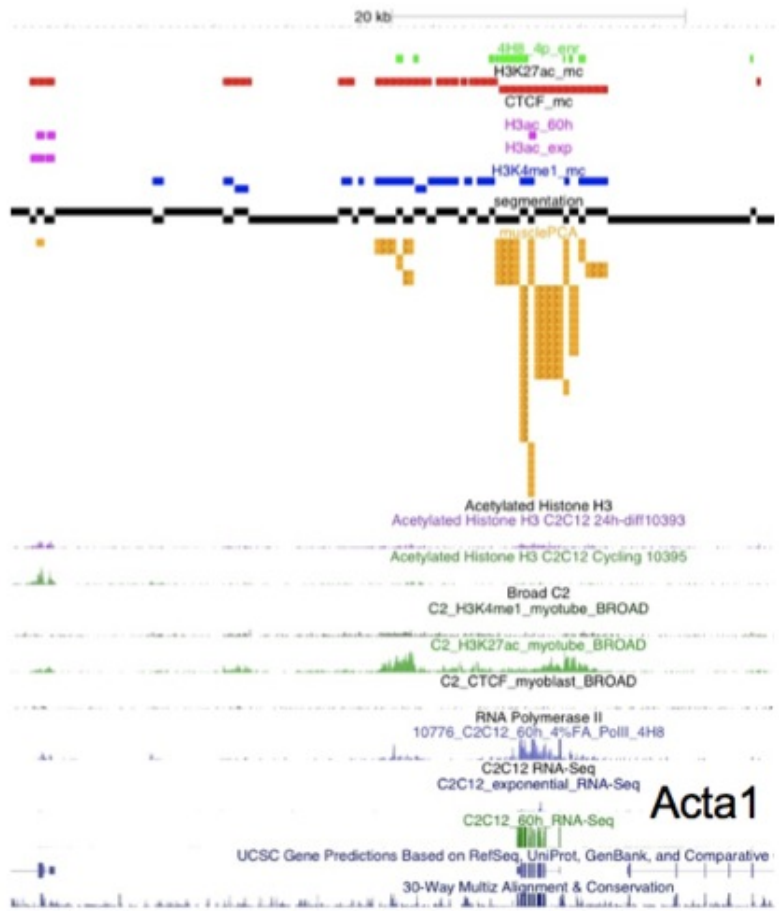
- Which are the interesting ones ?
- What are the region boundaries ?



Analyzing multiple ChIP-seq datasets jointly

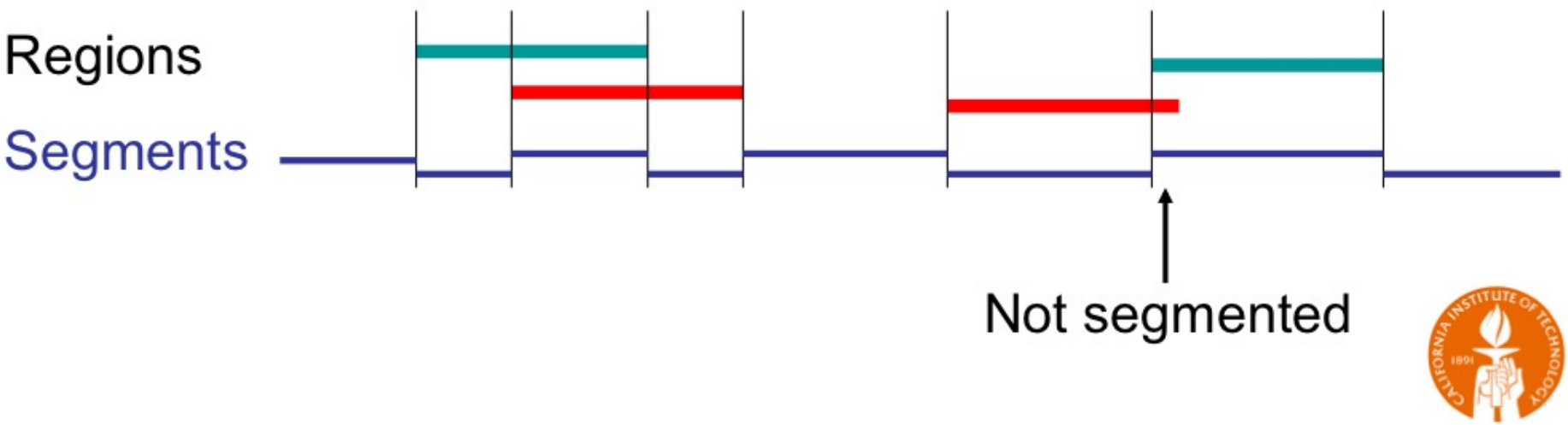
Most integrative analyses boil down to:

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



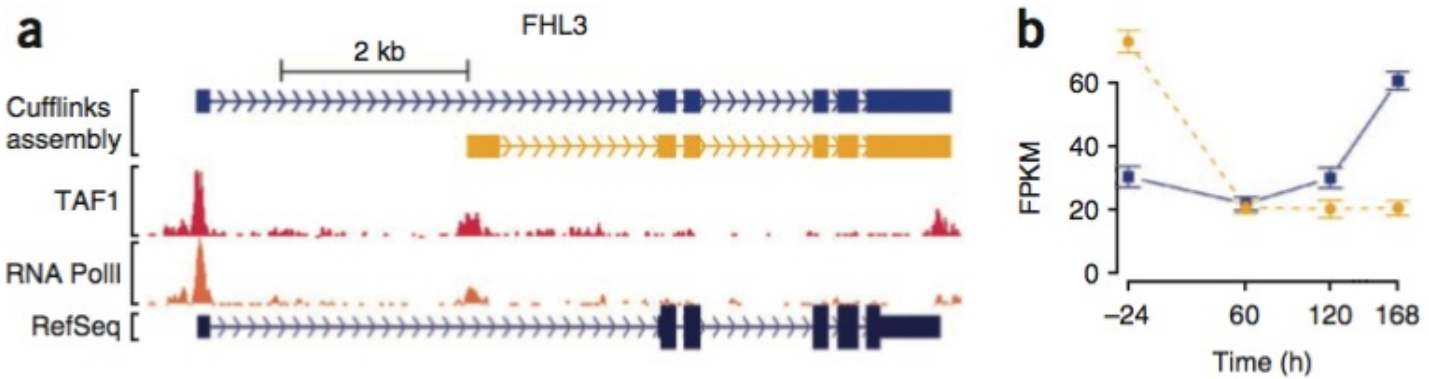
Segmenting the genome

- 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)



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 ?



Integration in QTLs

Few approaches:

- *A posteriori* for "validation"
  - Identify QTLs
  - Match with known functional annotations to find overlap
- *A priori* for "filtering"
  - Filter variants down to those that have a relevant functional annotation
  - Perform QTL analysis on subset with increased power (on subset)
  - Useful for small datasets and rare variants
- *In situ* during inference
  - Learn weighting of functional annotation types
  - ... while performing the associations

Integration in QTLs

Few approaches:

- *A posteriori* for "validation"
  - Identify QTLs
  - Match with known functional annotations to find overlap
- *A priori* for "filtering"
  - Filter variants down to those that have a relevant functional annotation
  - Perform QTL analysis on subset with increased power (on subset)
  - Useful for small datasets and rare variants
- *In situ* during inference
  - Learn weighting of functional annotation types
  - ... while performing the associations

Integration in QTLs

Few approaches:

- *A posteriori* for "validation"
  - Identify QTLs
  - Match with known functional annotations to find overlap
- *A priori* for "filtering"
  - Filter variants down to those that have a relevant functional annotation
  - Perform QTL analysis on subset with increased power (on subset)
  - Useful for small datasets and rare variants
- *In situ* during inference
  - Learn weighting of functional annotation types
  - ... while performing the associations

Acknowledgments

We gratefully acknowledge help with the material:

- Anna Abelin, Caltech
- Jonas Behr, FML & Sloan Kettering Institute
- Regina Bohnert, FML
- Philipp Drewe, FML & Sloan Kettering Institute
- Fabio De Bona, FML & Google Research
- André Kahles, FML & Sloan Kettering Institute
- 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://raetschlab.org/lectures/mlpm-ngs-lecture.pdf`

News and Opportunities

the current view

soon the new view

Current topics: ML for phenotyping from medical records, cancer, large-scale genomics, decision support systems, gene regulation

(come and talk to me if you'd like to learn more and look for opportunities)

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.

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`.

J. Behr, A. Kahles, Y. Zhong, and G. Rätsch. Mitie: Accurate transcript prediction with integer programming. *Bioinformatics*, 2013. under revision.

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.

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.

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.

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.

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.

References III

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

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.

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 2010. 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 IV

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.

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.

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.

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.

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 V

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*, 2009. http://flux.sammeth.net/capacitor.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.

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`.

References VI

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.

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

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`.

References VII

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.

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

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.

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.