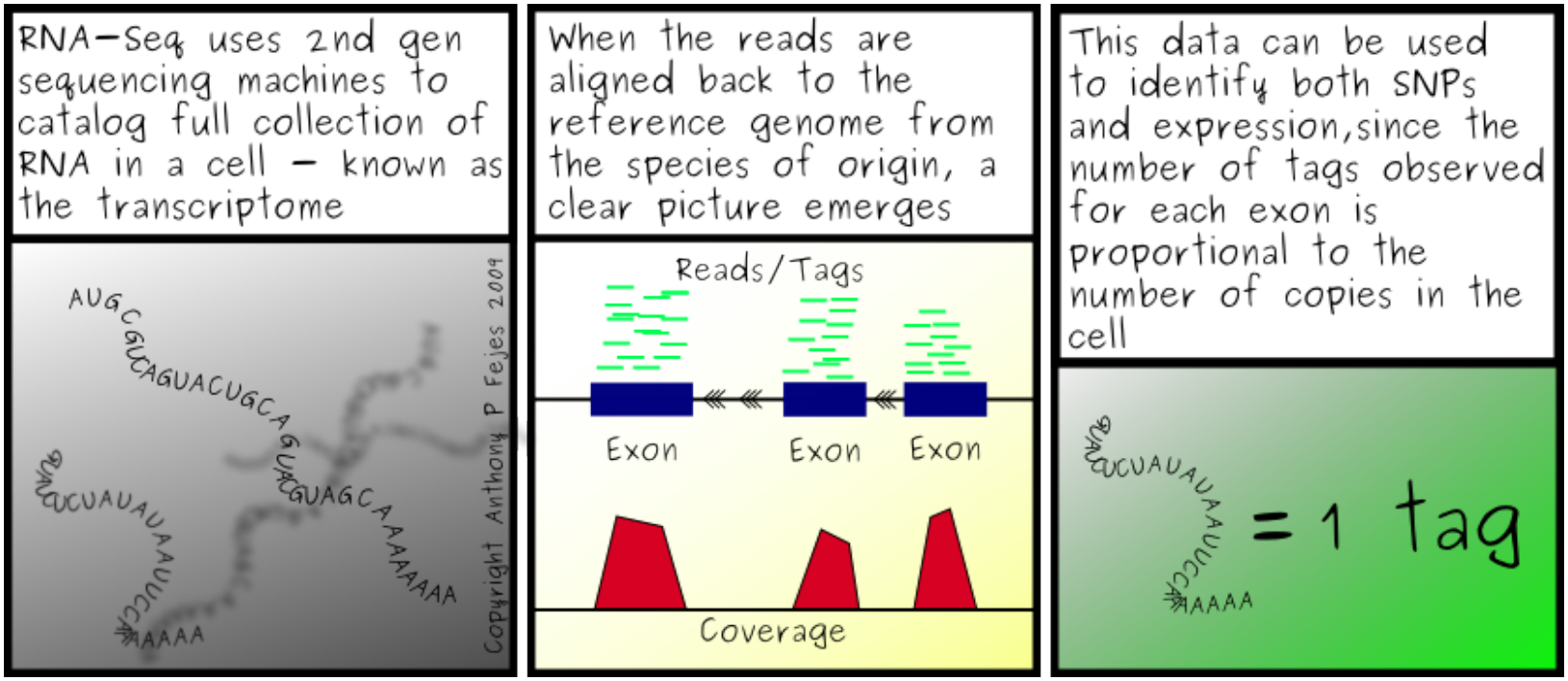
Lectures 13: High throughput sequencing: Beyond the genome

Spring 2017

March 28, 2017



http://www.fejes.ca/2009/06/science-cartoons-5-rna-seq.html

Omics

- Transcriptome - the set of all mRNAs present in a cell
- Proteome – proteins
- Metabolome/physiome - metabolites
- Microbiome – the collection of microbes present in an organism or other location
- Interactome

"In physics… the -*on* suffix has tended to signify an elementary particle: the photon, electron, proton, meson, etc., whereas -*ome* in biology has the opposite intellectual function, of directing attention to a holistic abstraction, an eventual goal…" From: *'Ome Sweet 'Omics*. The Scientist 15(7), 2001

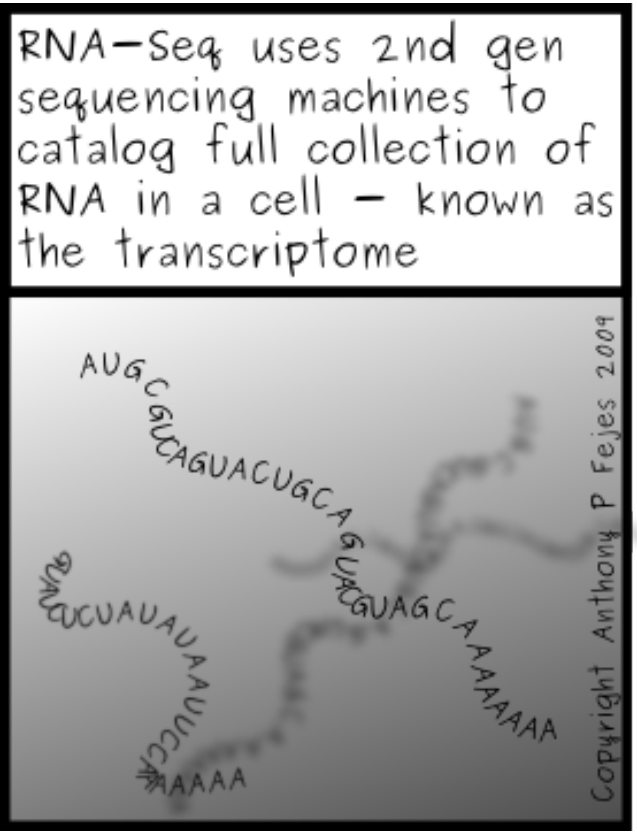
Omics

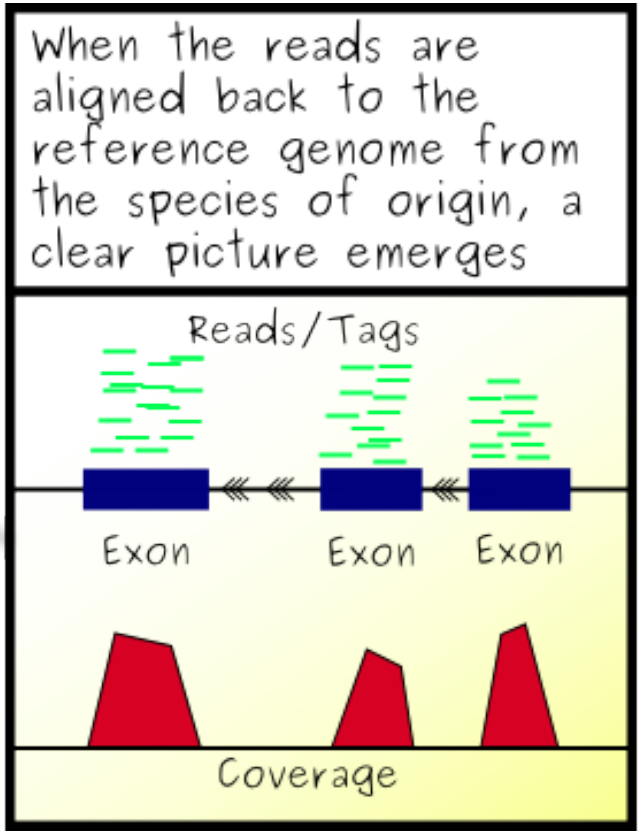
- Biologists have high-throughput methods for probing each -*ome*:
- Transcriptome – RNA-Seq
- Proteome – mass spectrometry, protein arrays
- Microbiome – next generation sequencing
- Interactome – yeast-two-hybrid
- Regulome – ChIP-Seq

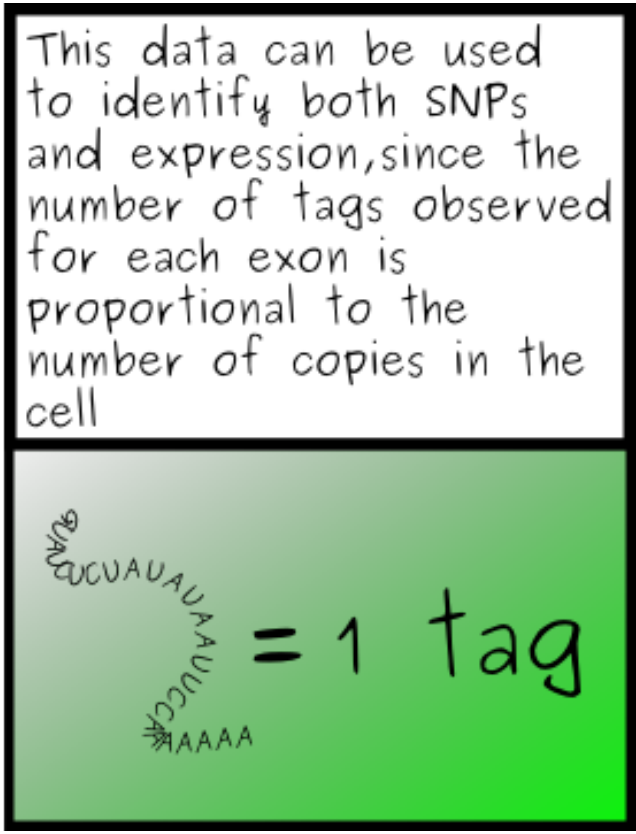
Lots of data for bioinformatics people to analyze!

RNA-seq: profiling the transcriptome

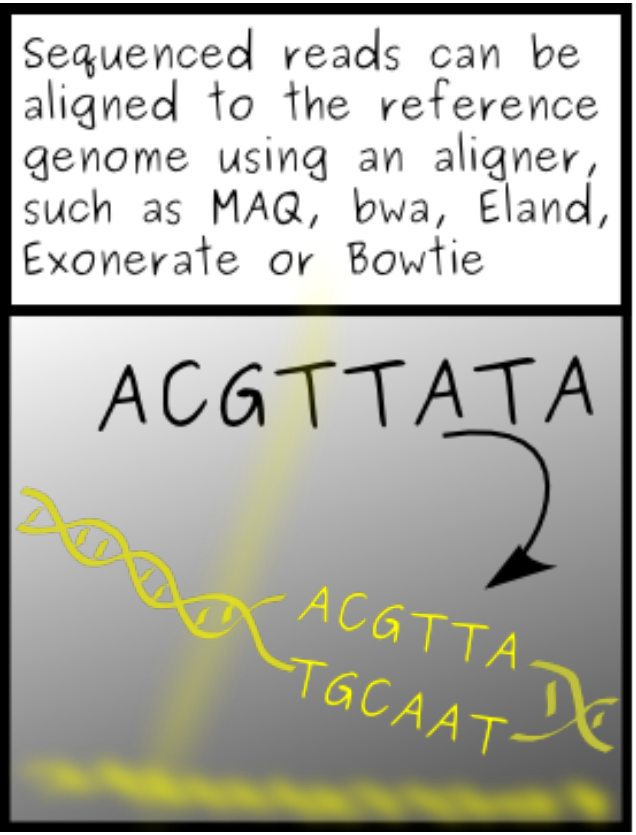
- Technique: sequence the total RNA produced by the cell

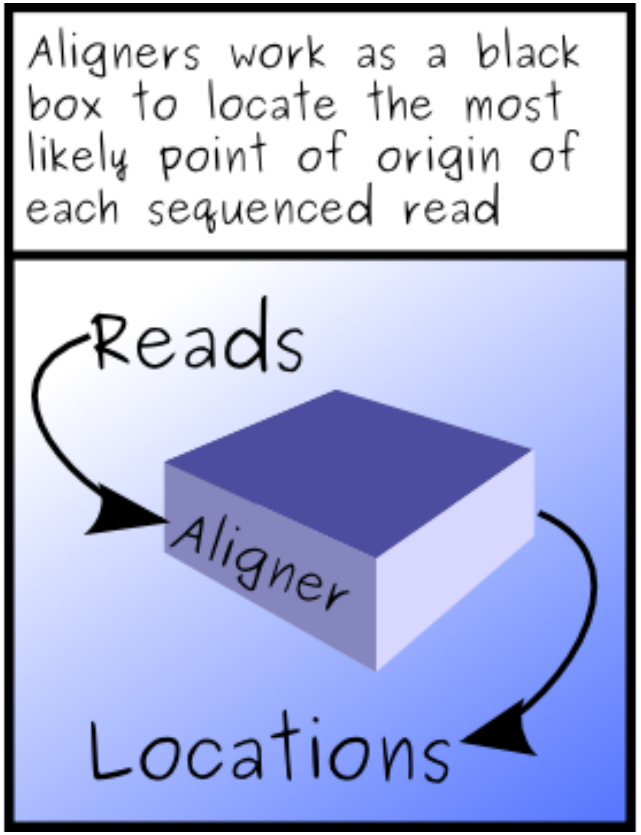


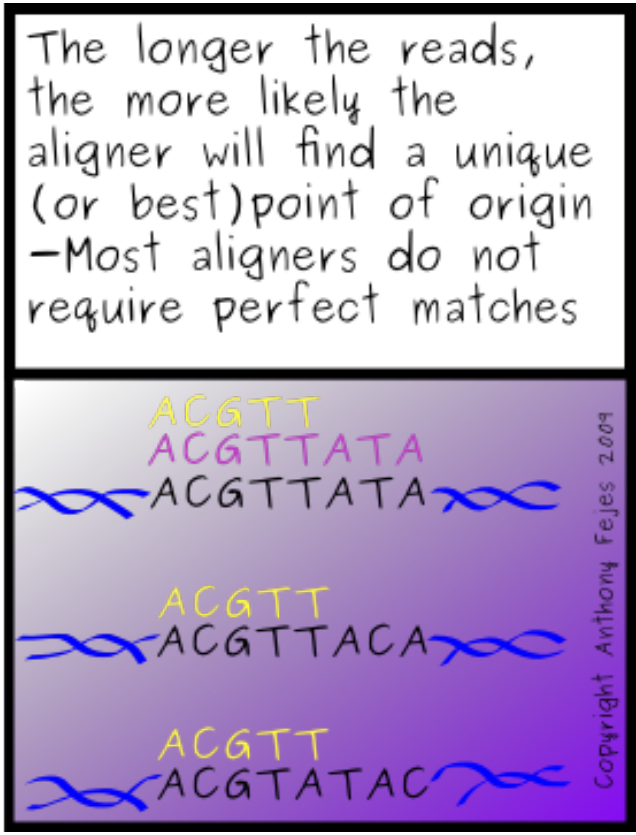




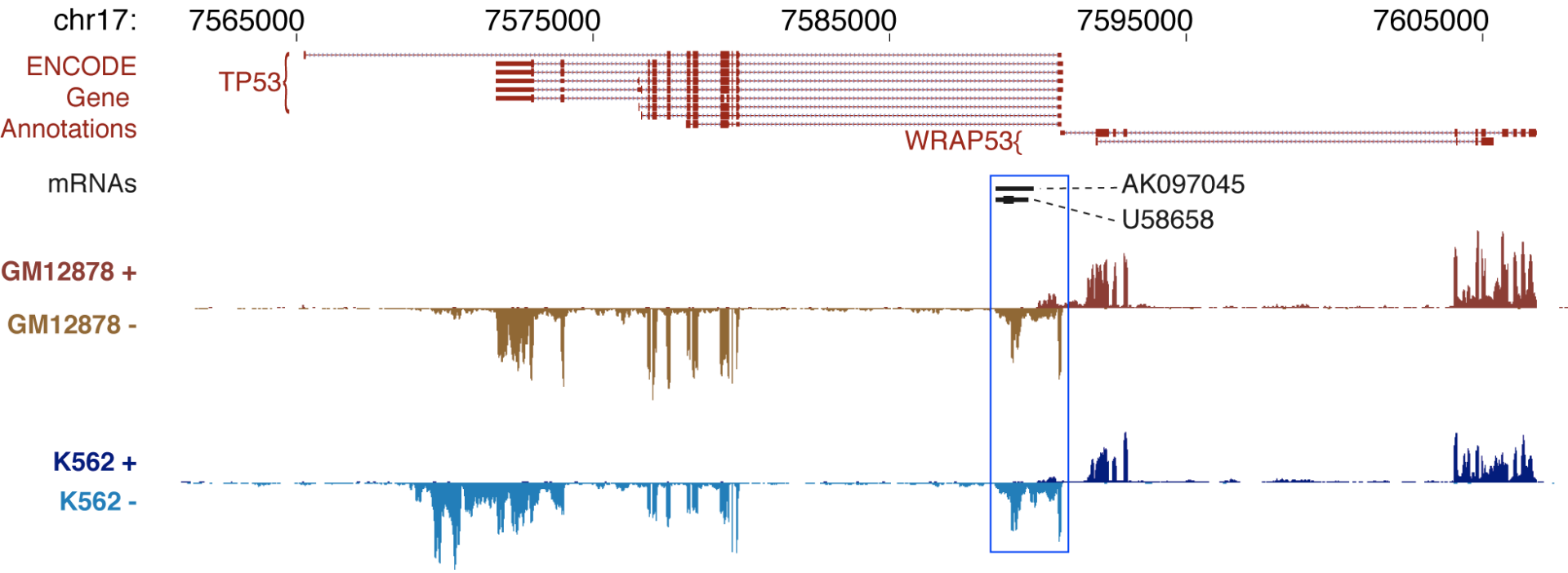
Read mapping





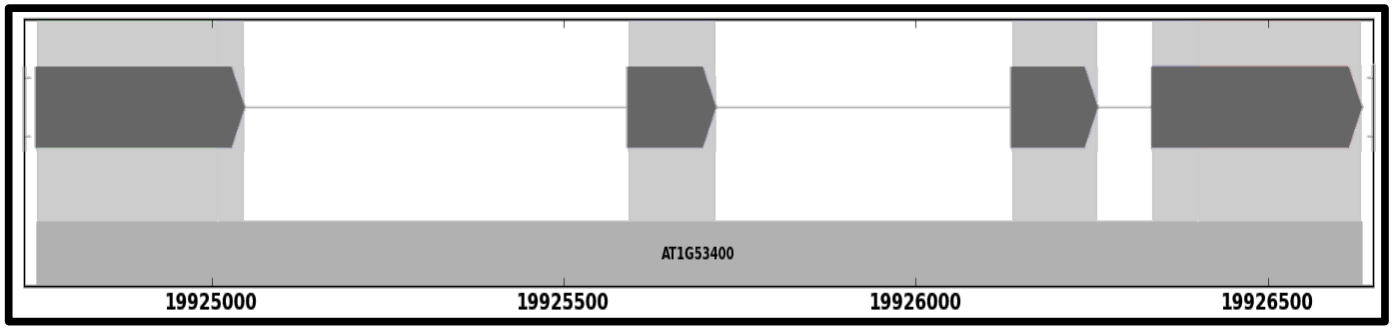


Pile-ups

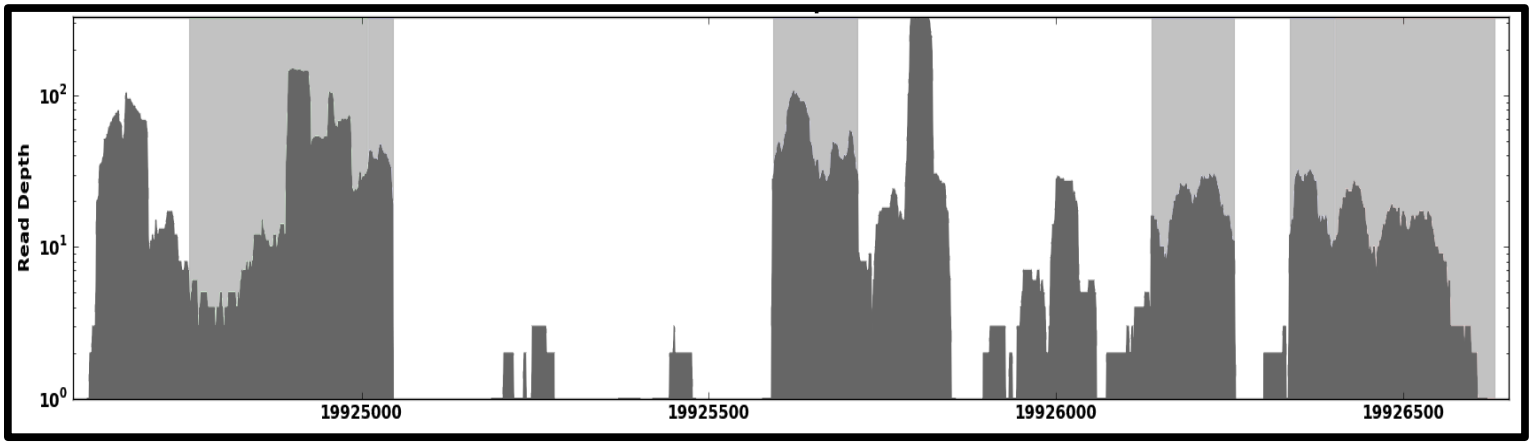


From: The ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol 9(4): e1001046.

Pile-ups



gene model



read depth

Most reads fall into coding exons or UTRs

RNA-seq: profiling the transcriptome

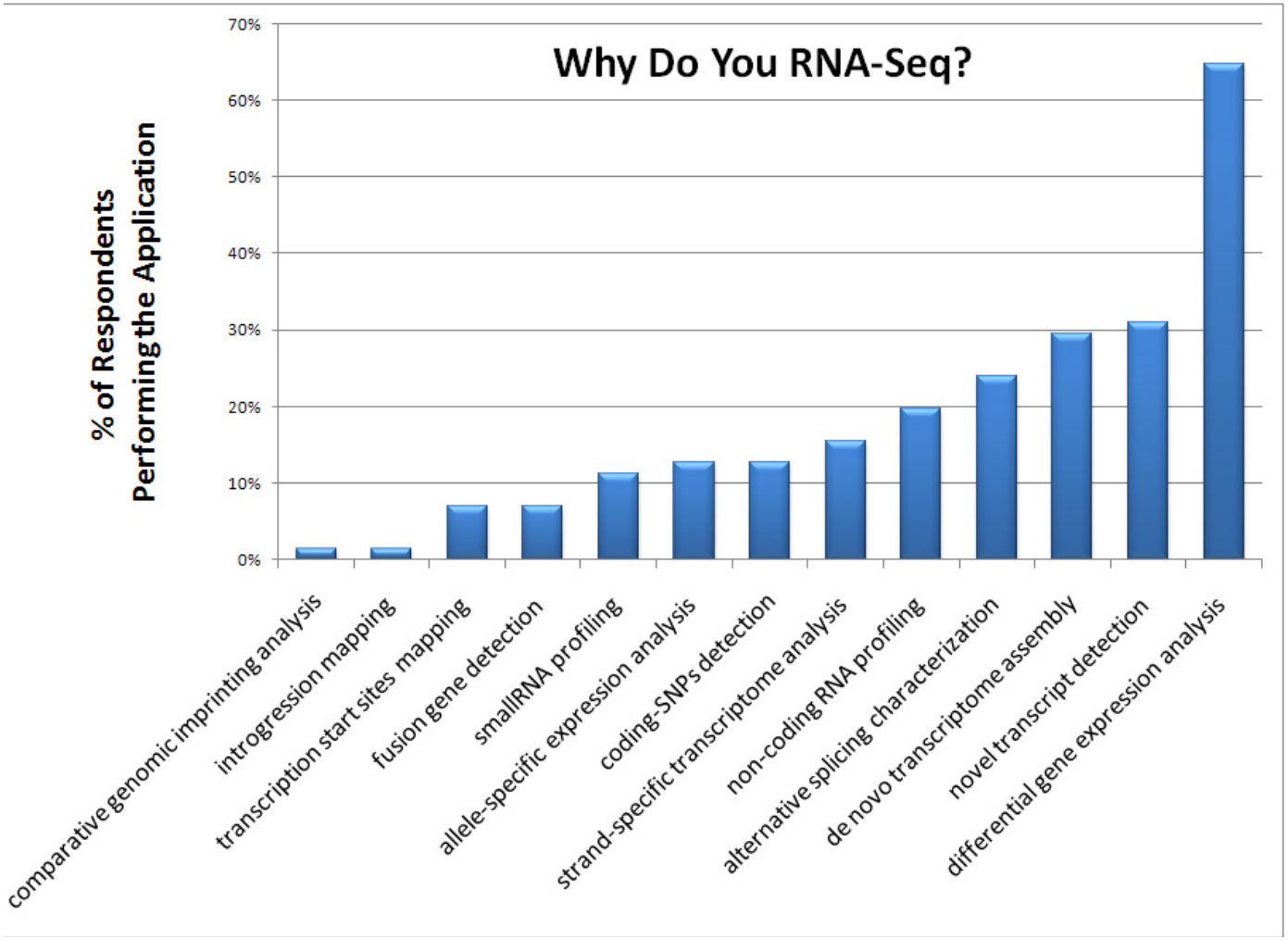
- Technique: sequence the total RNA produced by the cell

- What is this good for?

RNA-seq: profiling the transcriptome

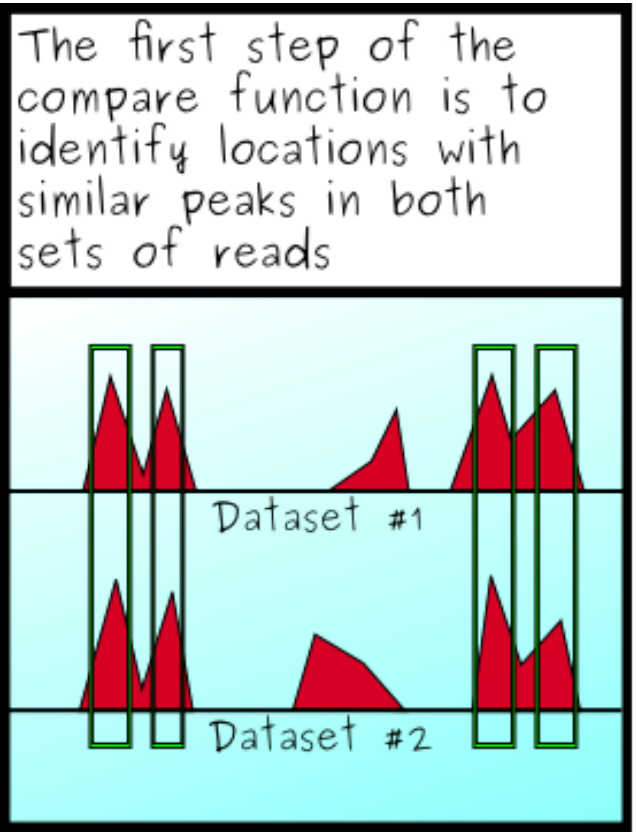
- Genome annotation (transcript assembly)
- Detect alternative splicing
- Obtain gene/transcript expression levels and detection of differential expression
- Allele-specific expression
- Small-RNA transcriptome (different protocol than regular RNA-seq)

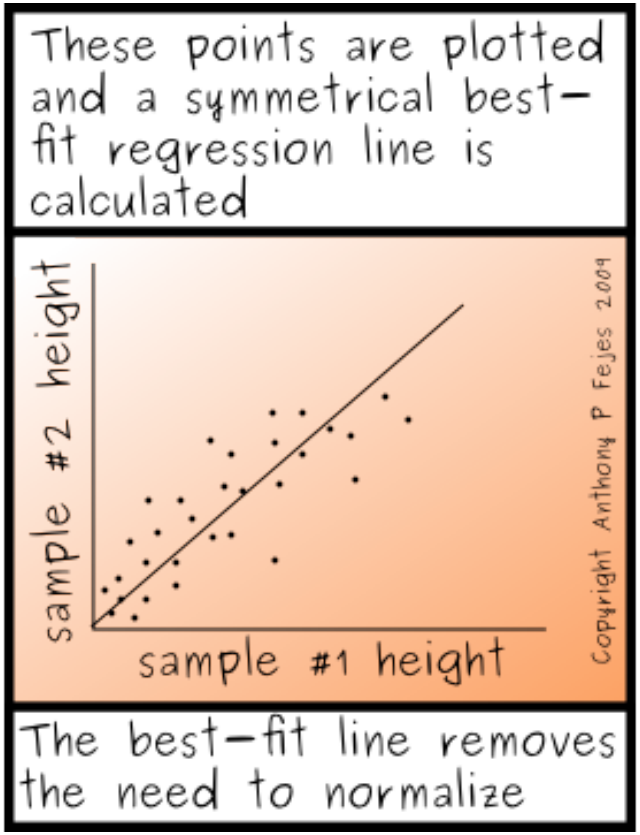
All the uses of RNA-seq

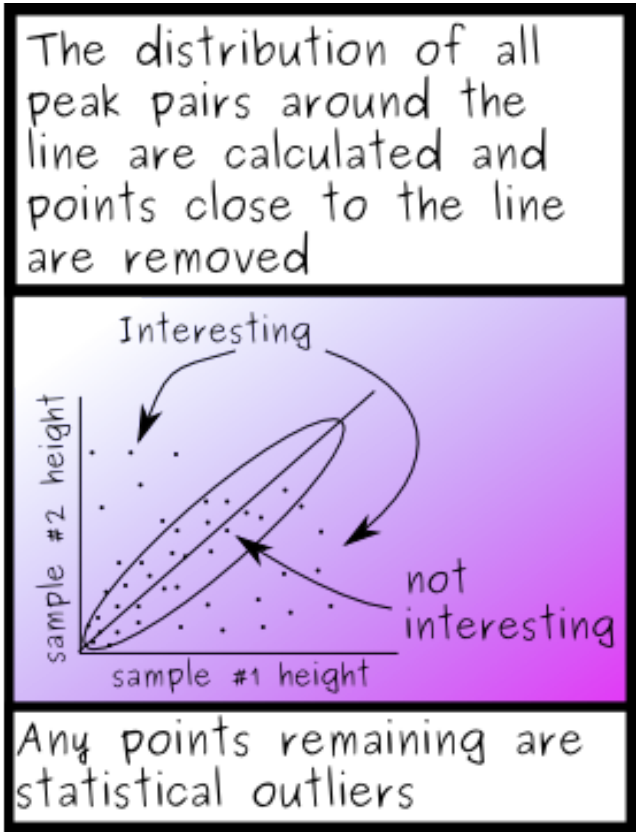


http://www.rna-seqblog.com/news/information/rna-seq-blog-poll-results/

Differential expression

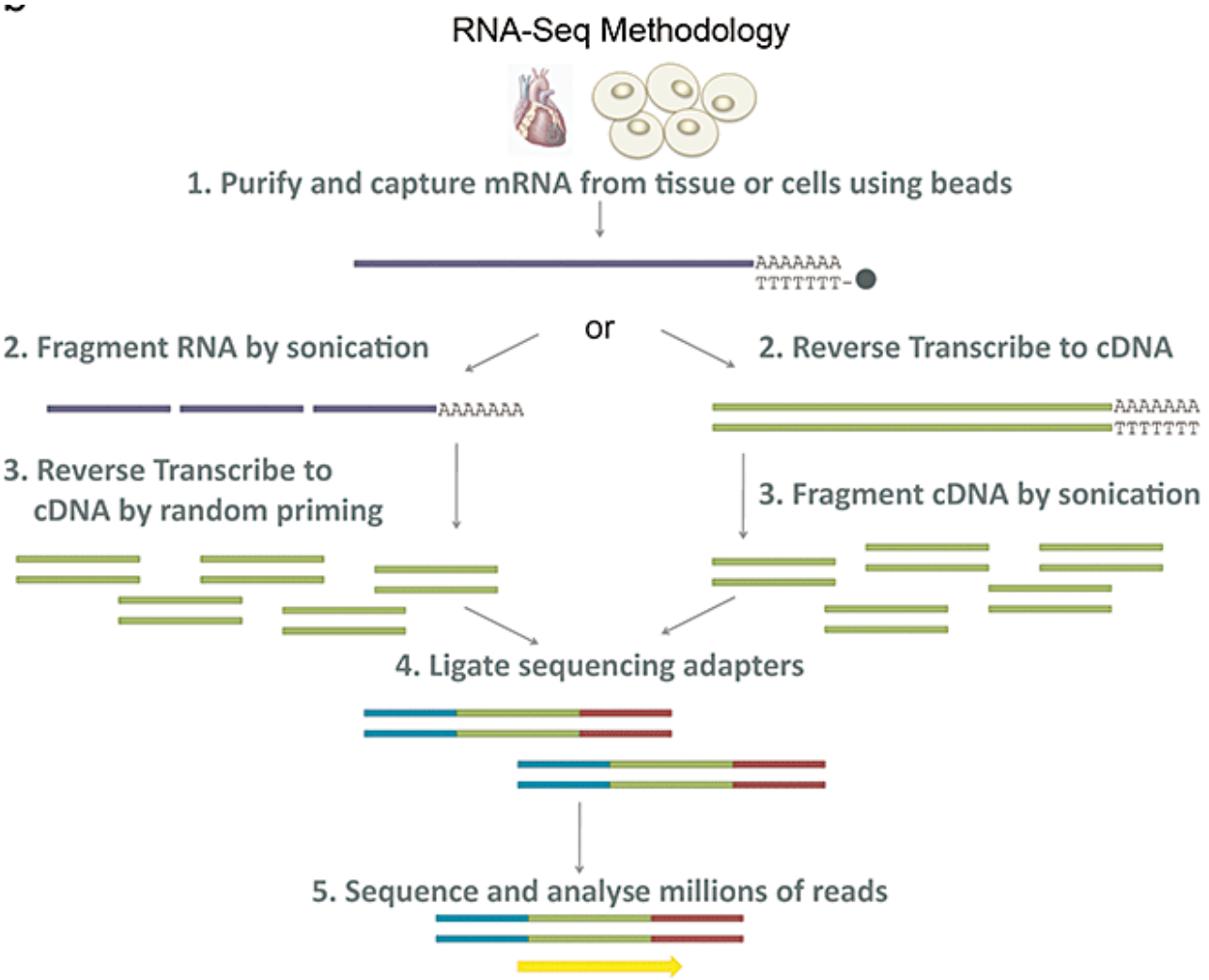






http://www.fejes.ca/labels/figures.html

RNA-seq protocol



Raw and Aligned Reads

- Raw data is a (large) set of sequences
- Typical file format is FASTQ

```
@HWI-EAS255_4_FC2010Y_1_43_110_790
TTAATCTACAGAATAGATAGCTAGCATATATTT
+
hhhhhhhhhhhhhhhdhhhhhhhhhhhdRehdh
```

Read identifier

Bases called

Base quality codes

- Alignment to genome is done by efficient indexing
- Aligned reads in SAM format

```
@HWI-… 163 chr19 9900 10000 16M2I25M
```

Read identifier

Where this read matched

Start and end positions

Codes for match: 16 matches, 2 extra,…

Cataloging the transcriptome

- Transcriptomics involves studying expression at
  - Spatial resolution: tissues, individuals, location
  - Temporal resolution: circadian, seasonal, lifetime

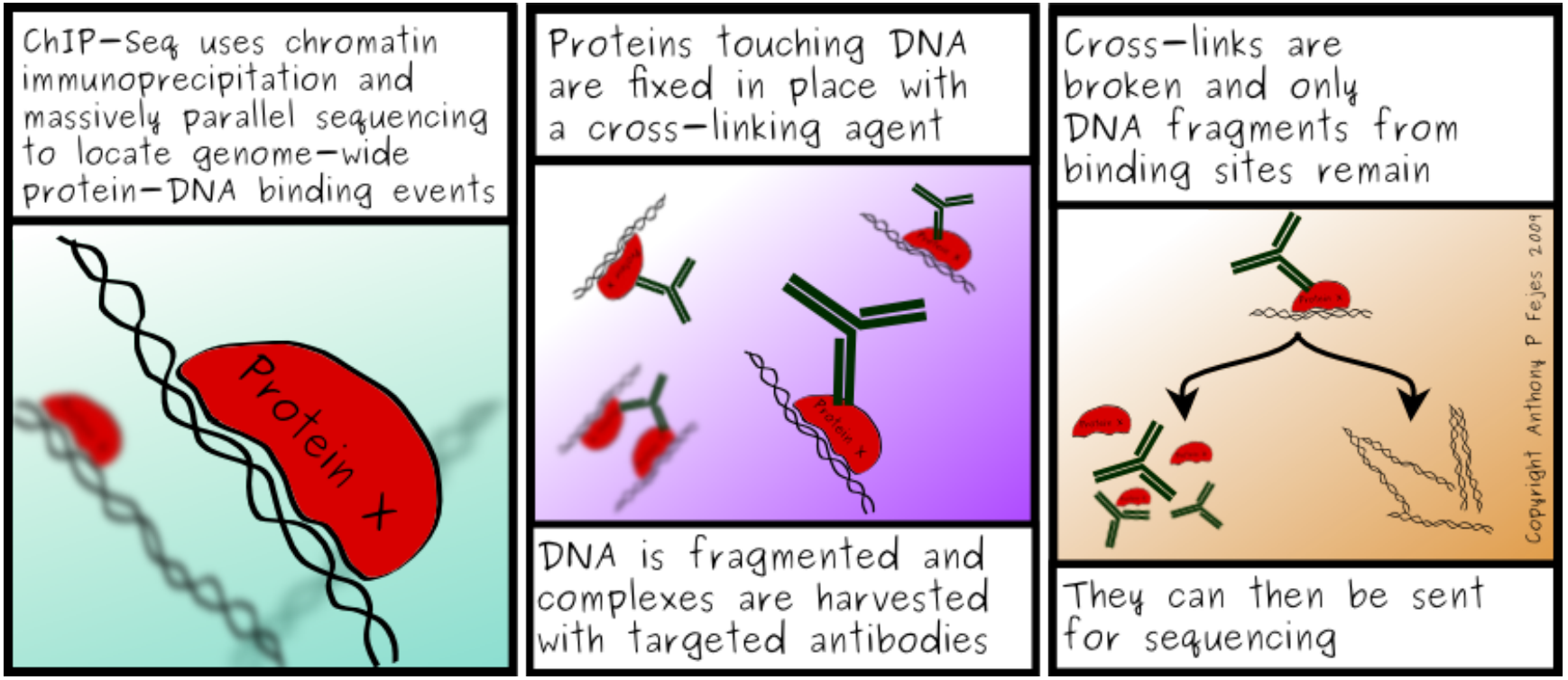
Inter-Genic Reads

- Many reads reflect unannotated genes: opportunity to discover new genes

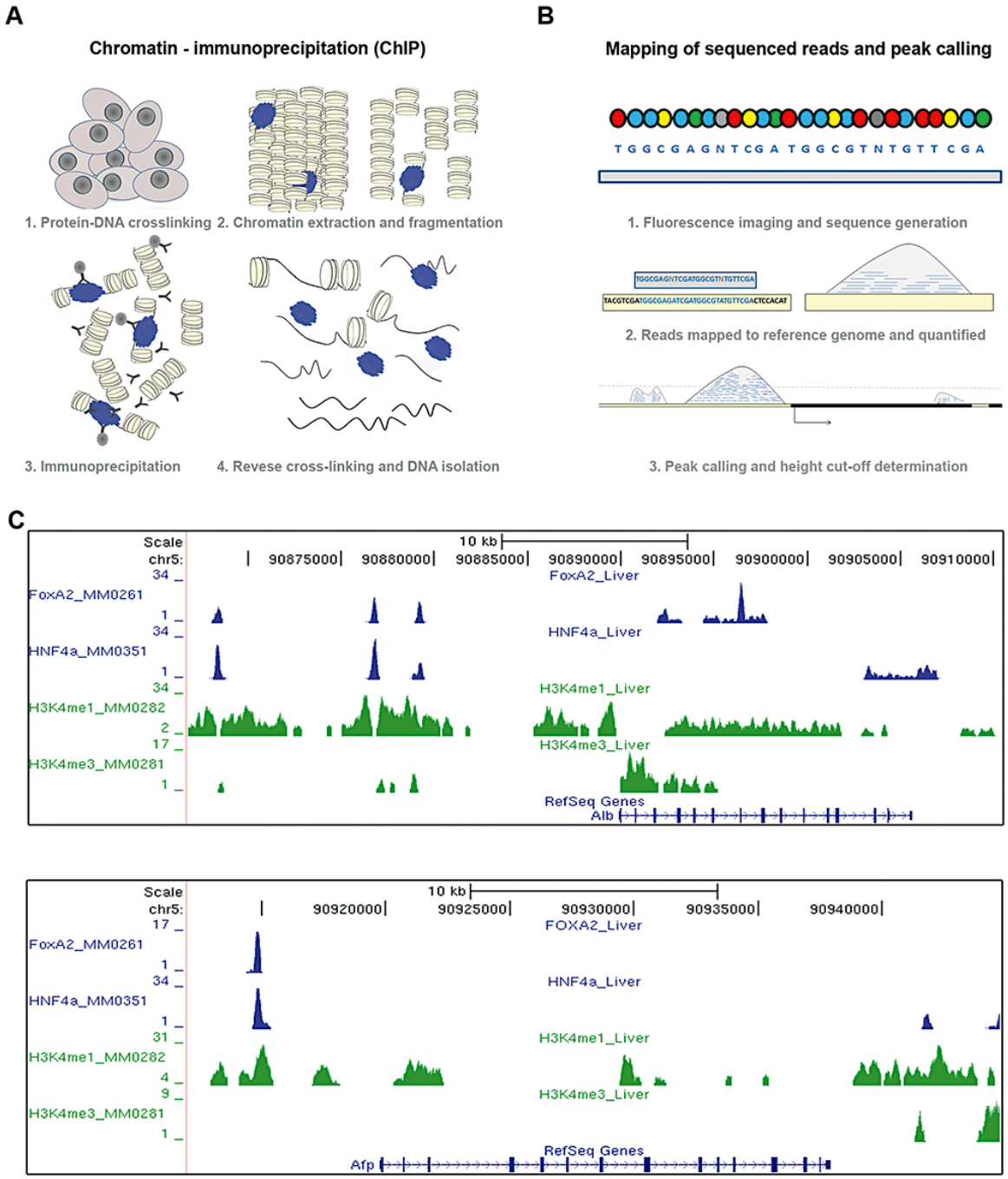
RPKM – A Simple Normalization

- Different numbers of counts per sample (sequencing depth)
- Divide counts in a region of interest (a genomic region or a gene or an exon) by all counts (reads per million reads -RPM)
- Genes have different lengths: divide also by length of gene
- Obtain RPKM (reads per kilobase of exon per million reads)
  - Some use FPKM (fragments/kb/Mr)

ChIP-seq



http://www.fejes.ca/labels/Chip-Seq.html


A
Chromatin - immunoprecipitation (ChIP)
1. Protein-DNA crosslinking
2. Chromatin extraction and fragmentation
3. Immunoprecipitation
4. Revese cross-linking and DNA isolation
B
Mapping of sequenced reads and peak calling
T G G C G A G N T C G A T G G C G T N T G T T C G A
1. Fluorescence imaging and sequence generation
2. Reads mapped to reference genome and quantified
3. Peak calling and height cut-off determination
C
Scale
chr5:
10 kb
90875000
90880000
90885000
90890000
90895000
90900000
90905000
90910000
FoxA2_MM0261
HNF4a_MM0351
H3K4me1_MM0282
H3K4me3_MM0281
FoxA2_Liver
HNF4a_Liver
H3K4me1_Liver
H3K4me3_Liver
RefSeq Genes
Alb
Scale
chr5:
10 kb
90920000
90925000
90930000
90935000
90940000
FoxA2_MM0261
HNF4a_MM0351
H3K4me1_MM0282
H3K4me3_MM0281
FOXA2_Liver
HNF4a_Liver
H3K4me1_Liver
H3K4me3_Liver
RefSeq Genes
Afp

Comments on ChIP-seq

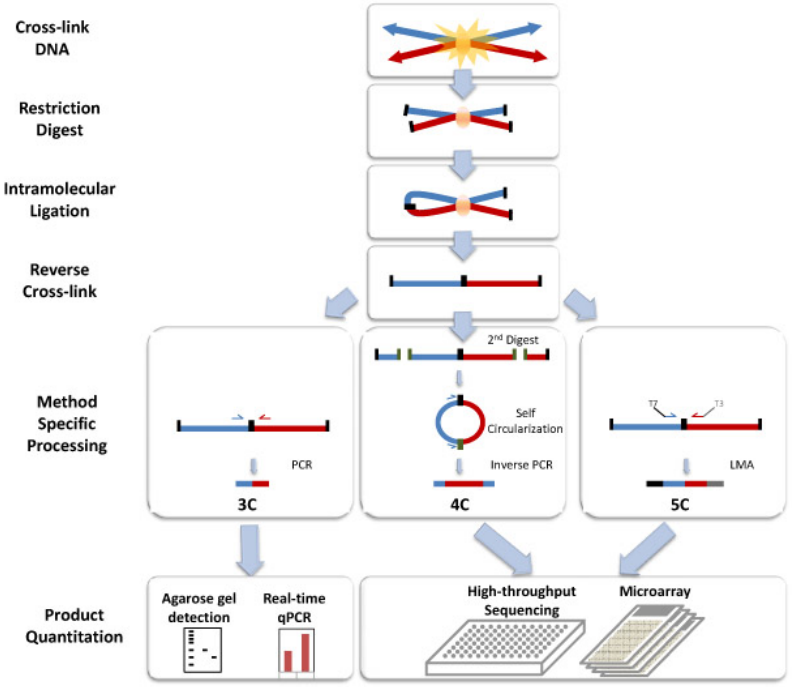
- Genome-wide mapping of transcription factor binding sites
- Computational problems:
  - Peak calling
  - Still need motif finders, but makes the problem easier

Variants

- Apply the methodology to RNA: map RNA-binding sites in mRNA that interact with specific RNA-binding proteins
- CLIP-Seq (cross-linking immunoprecipitation sequencing)
- RIP-Seq (RNA immunoprecipitation sequencing)

Other sequencing-based techniques

- Methyl-seq, BS-seq: methylation
- Chromosome conformation capture (3C-4C-5C-HiC): spatial organization of chromosomes



http://en.wikipedia.org/wiki/Chromosome_conformation_capture

Other sequencing-based techniques

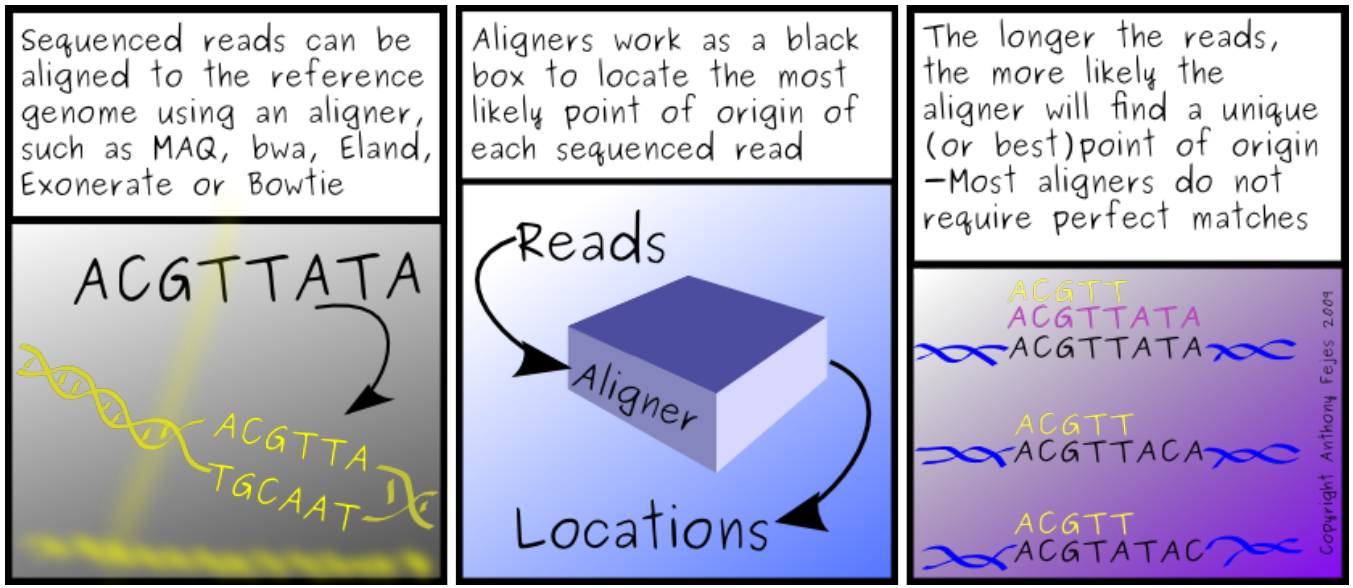
- Methyl-seq, BS-seq: methylation
- Chromosome conformation capture (3C-4C-5C): spatial organization of chromosomes
- seqFold: RNA secondary structure
- DNAase-seq
- And many more!

http://en.wikipedia.org/wiki/Chromosome_conformation_capture

Read mapping

All the sequencing-based techniques require read mapping as a first step.

Existing alignment tools are not fast enough → need new algorithms!



Read mapping

- How is the problem of read mapping different than sequence alignment as we have considered it until now?