Gene Architecture & Gene Discovery

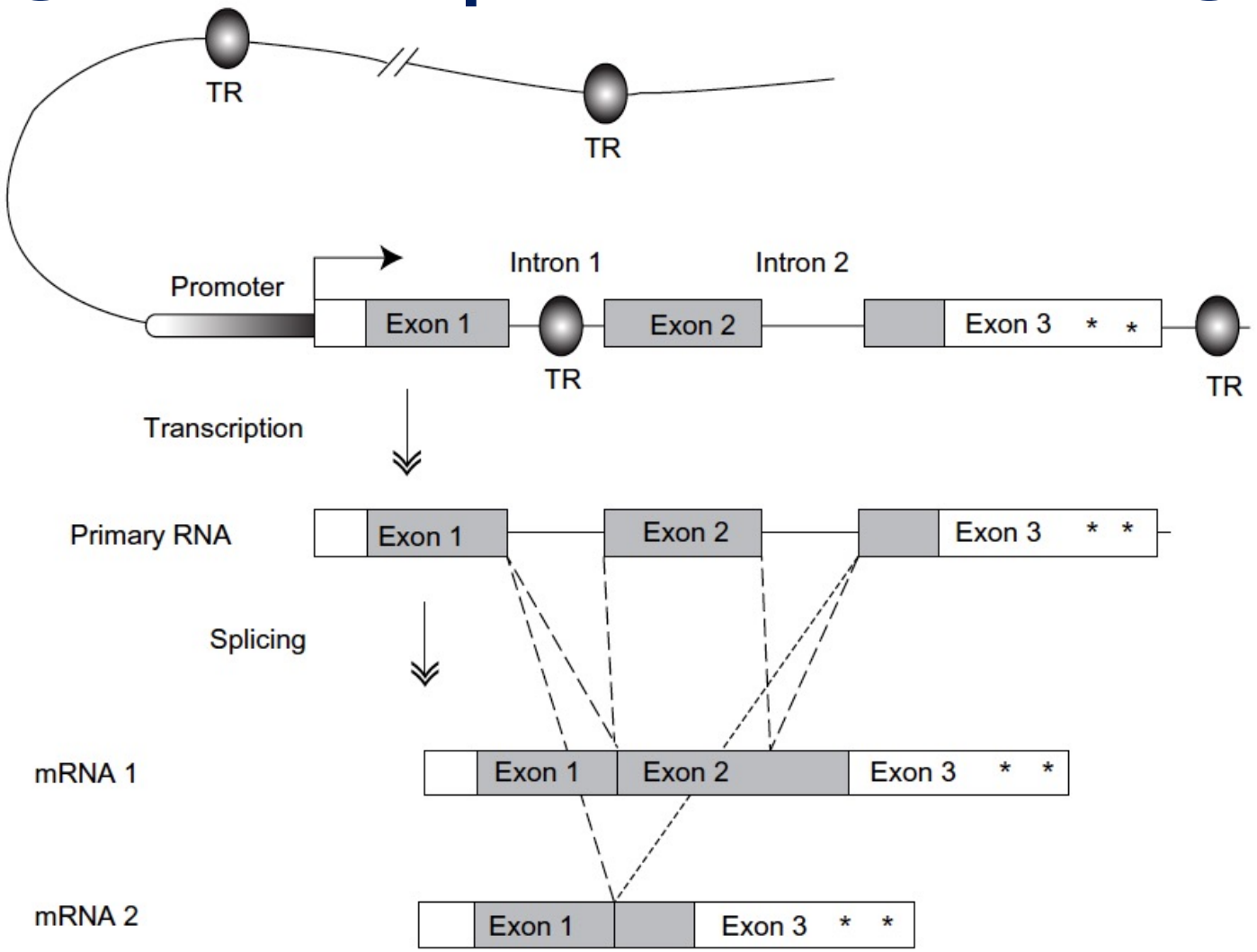
Download Gene Discovery Files from GitHub!

BIOL 435/535:
Bioinformatics

Feb 1, 2022

What is a gene?

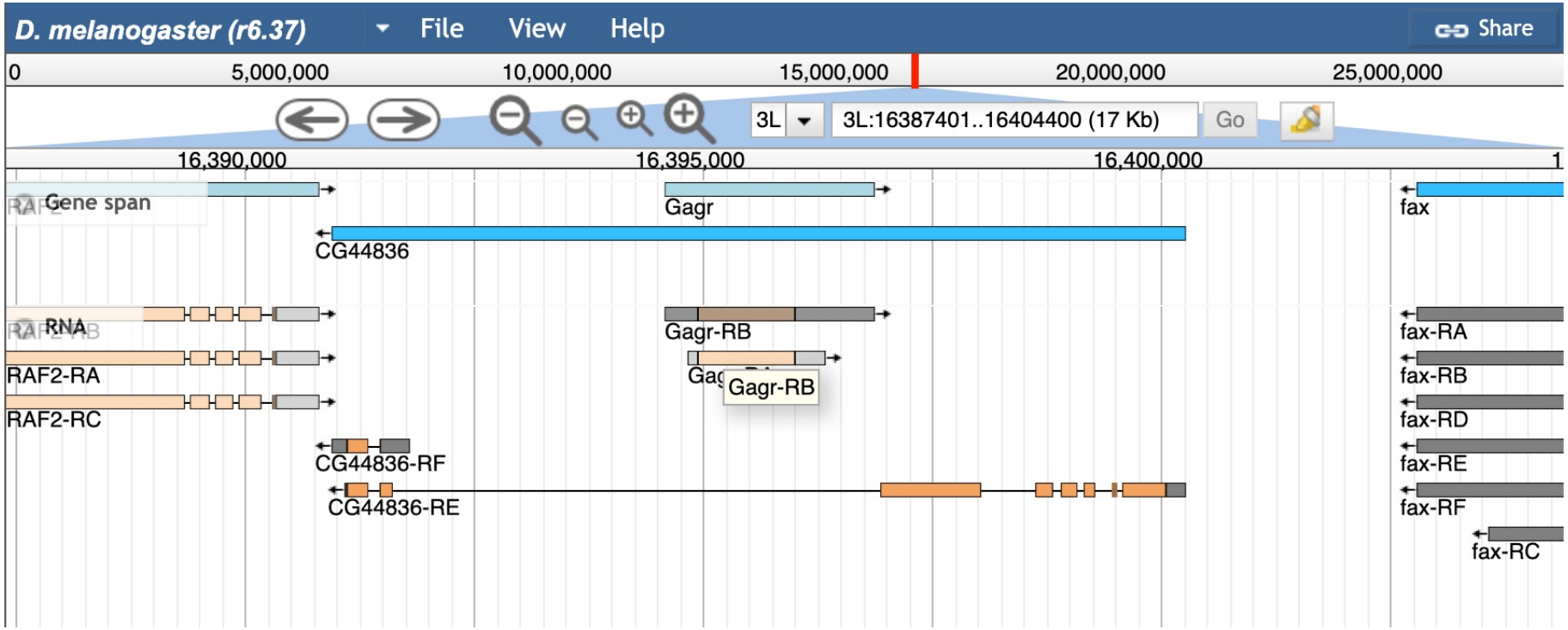
Kinds of genes – protein coding



Sastre 2014 Adv Genomics Genet

Kinds of genes – protein coding

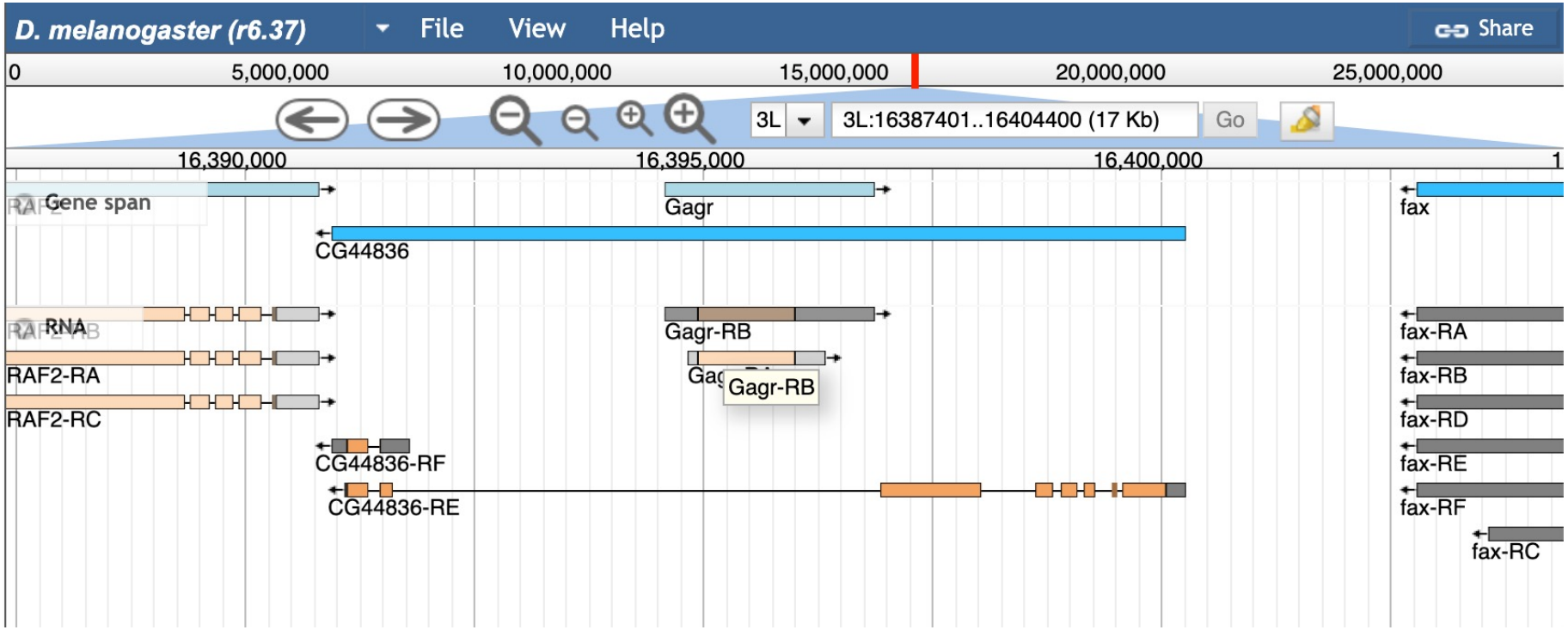
- Genes can be encoded on both strands of DNA (i.e., forward and reverse)



Reverse Complement

Kinds of genes – protein coding

- Genes can be overlapping



Kinds of genes – protein coding

- Redundant, unambiguous genetic code

RNA codon table

1st position	2nd position U	2nd position C	2nd position A	2nd position G	3rd position
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	stop	stop	A
	Leu	Ser	stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G
	Amino Acids				

Ala: Alanine Gln: Glutamine Leu: Leucine Ser: Serine
Arg: Arginine Glu: Glutamic acid Lys: Lysine Thr: Threonine
Asn: Asparagine Gly: Glycine Met: Methionine Trp: Tryptophane
Asp: Aspartic acid His: Histidine Phe: Phenylalanine Tyr: Tyrosine
Cys: Cysteine Ile: Isoleucine Pro: Proline Val: Valine

Expasy

Intron/exon boundaries

- Exons often end with AG
- Introns often begin with GT end with AG

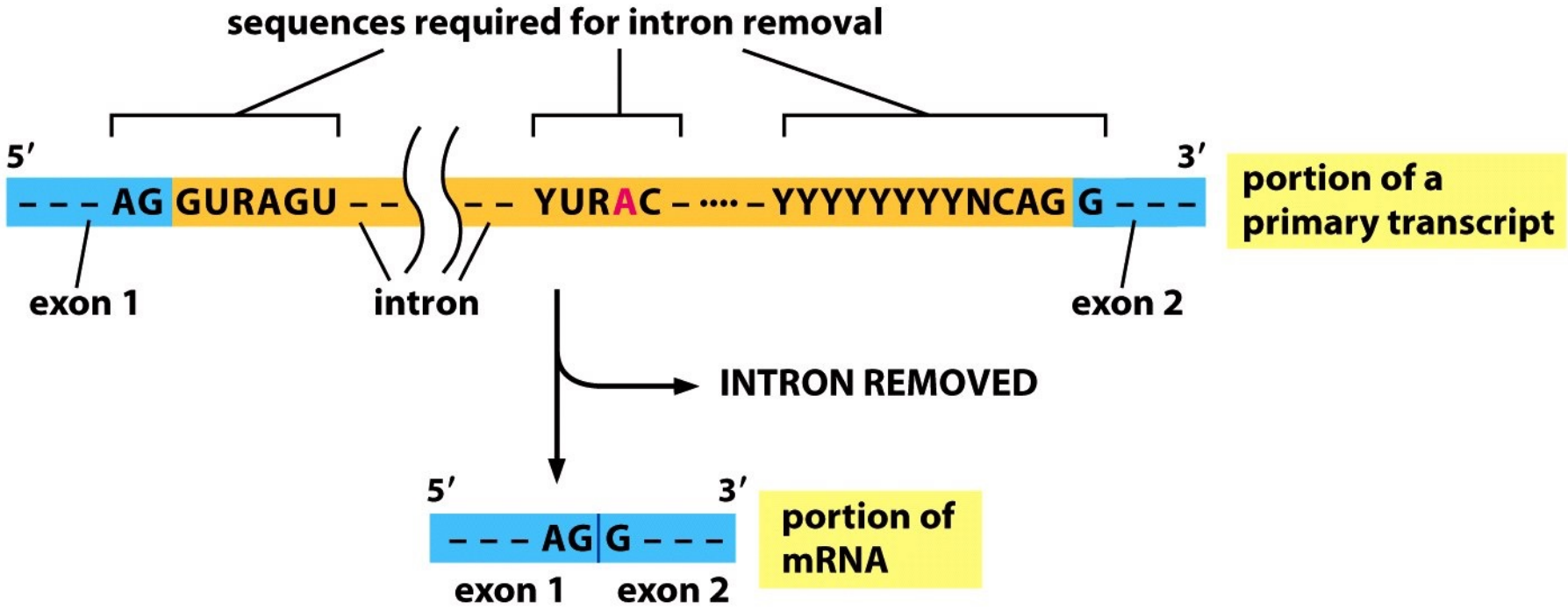
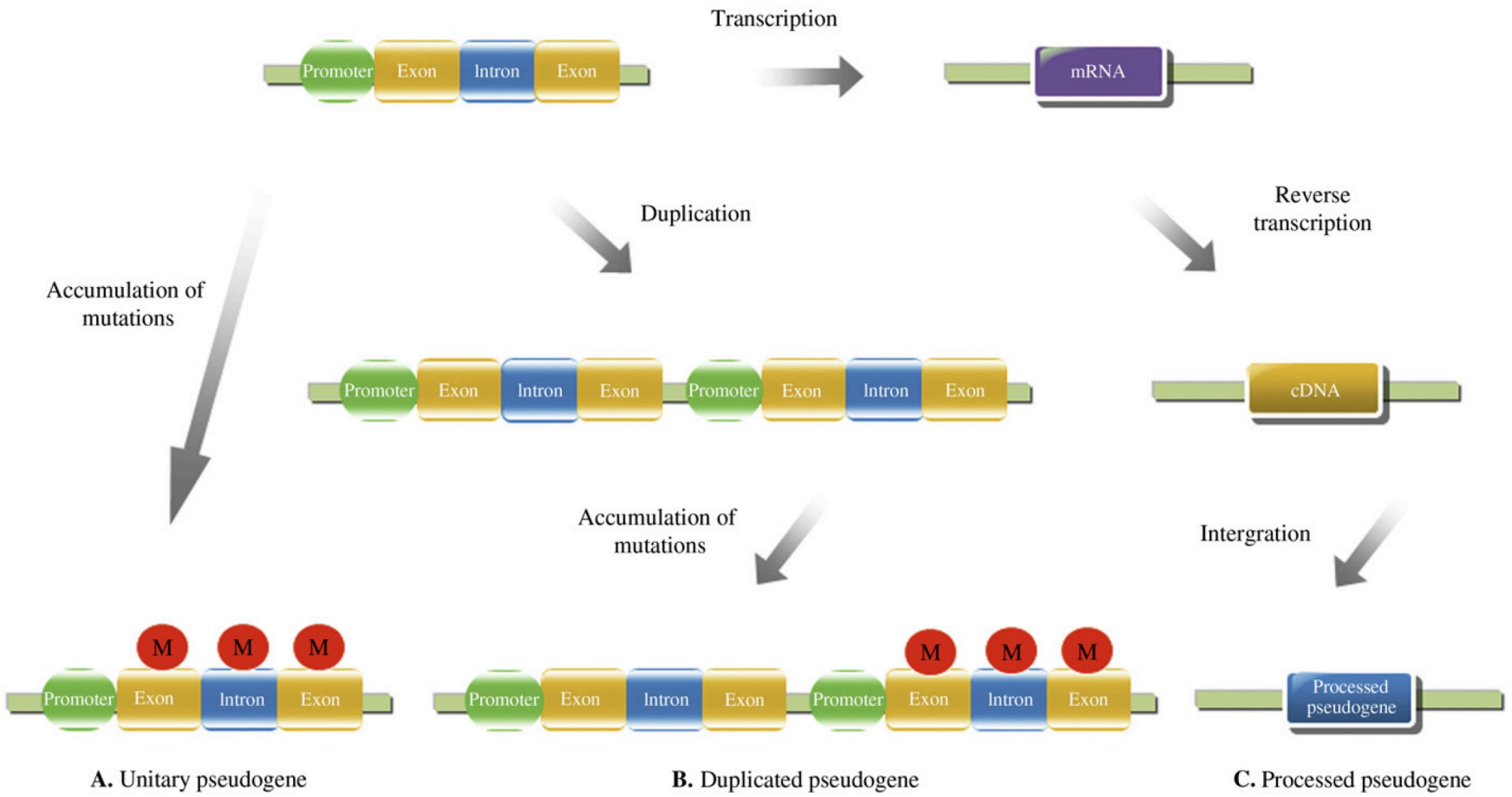


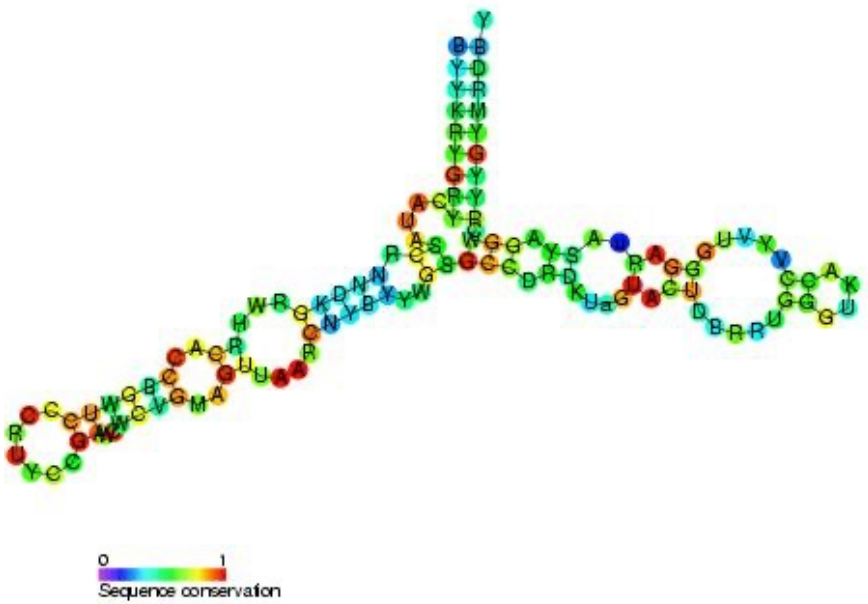
Figure 7-19 Essential Cell Biology 3/e (© Garland Science 2010)

Kinds of genes – pseudogenes

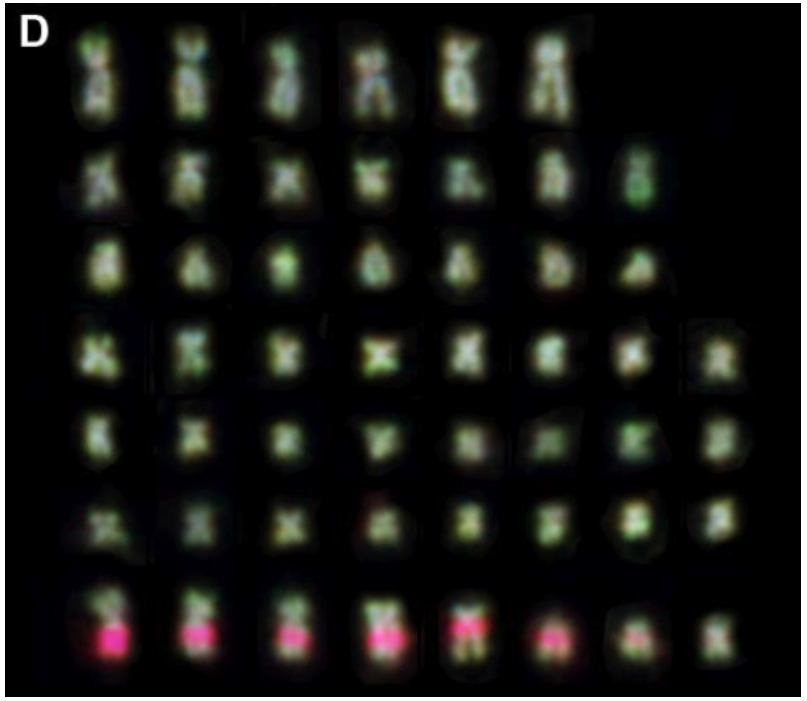


Kinds of genes – rRNAs

- rDNA can represent a substantial fraction of the genome. (many copies)

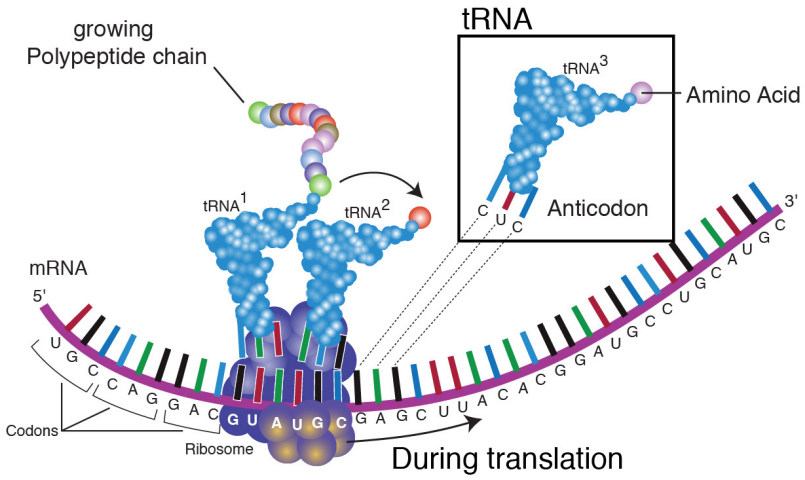


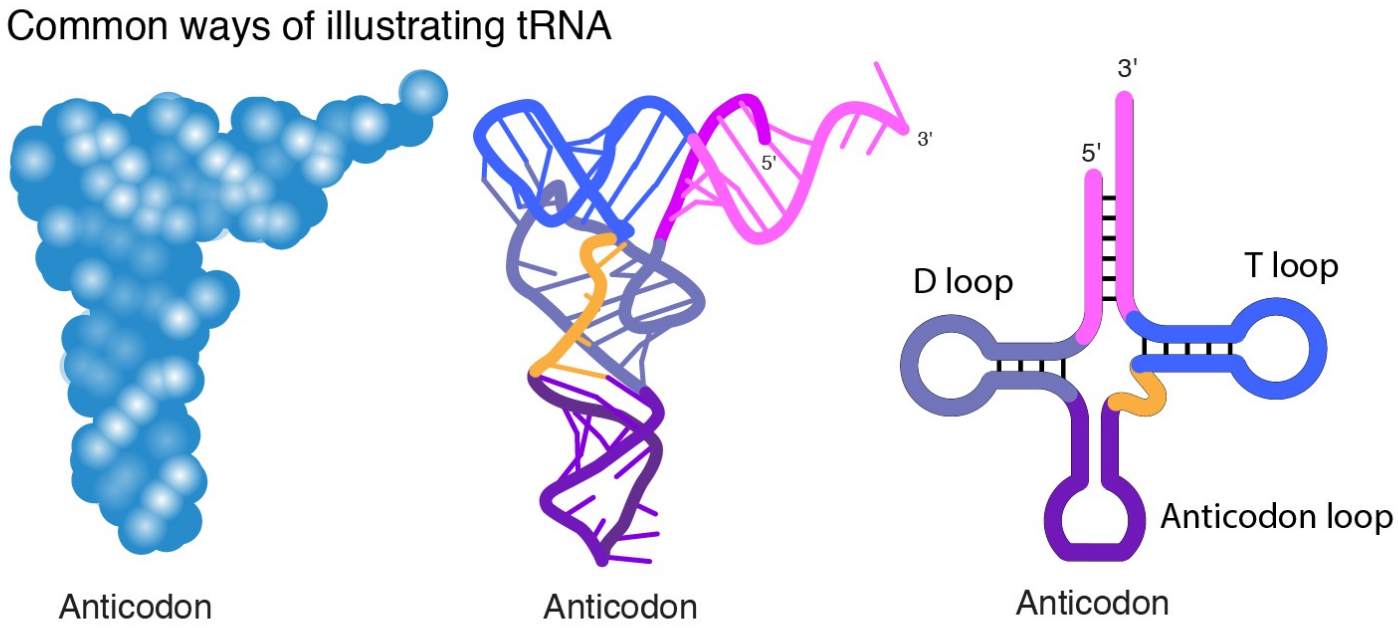
5S Ribosomal subunit
Wikimedia Commons



McElroy et al.,

Kinds of genes – tRNAs





Kinds of genes – other ncRNAs

Table 1: Classification of ncRNAs.

Type	Abbreviation	Full name	Size
Housekeeping ncRNAs	rRNA	ribosomal RNA	120–4,500 nt
	tRNA	tansfer RNA	76–90 nt
	snRNA	small nuclear RNA	100–300 nt
	snoRNA	small nucleolar RNA	60–400
	TERC	telomerase RNA	/
	tRF	tRNA-Derived Fragments	16–28 nt
	tiRNA	tRNA halves	29–50 nt
Regulatory ncRNAs	miRNA	microRNA	21–23 nt
	siRNA	small interfering RNA	20–25 nt
	piRNA	piwi-interacting RNA	26–32 nt
	eRNA	enhancer RNA	50–2,000 nt
	lncRNA	long non-coding RNAs	>200 nt
	circRNA	circular RNA	100–10,000 nt
	Y RNA	Y RNA	/

Genome browsers & Data hubs

- UCSC Genome Browser – hub browser for lots of model genomes
- YeastGenome – *S. cerevisiae* genome browser
- Flybase – *Drosophila* genome browser
- Wormbase – *C. elegans* genome browser
- Xenbase – *Xenopus* genome browser
- TAIR – *Arabidopsis* genome browser
- SolGenomics – Tomato (and relatives) genome browser
- Phytozome – Non-model plant genome repository

General Feature Format (GFF) files

1. **seqid** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. **Important note**: the seq ID must be one used within Ensembl, i.e. a standard chromosome name or an Ensembl identifier such as a scaffold ID, without any additional content such as species or assembly. See the example GFF output below.
2. **source** - name of the program that generated this feature, or the data source (database or project name)
3. **type** - type of feature. Must be a term or accession from the SOFA sequence ontology
4. **start** - Start position of the feature, with sequence numbering starting at 1.
5. **end** - End position of the feature, with sequence numbering starting at 1.
6. **score** - A floating point value.
7. **strand** - defined as + (forward) or - (reverse).
8. **phase** - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
9. **attributes** - A semicolon-separated list of tag-value pairs, providing additional information about each feature. Some of these tags are predefined, e.g. ID, Name, Alias, Parent - see the GFF documentation for more details.

General Feature Format (GFF) files

chromosome
gene
mRNA
five_prime_UTR
exon
CDS
three_prime_UTR
ncRNA_gene
lnc_RNA
miRNA
pre_miRNA
tRNA
ncRNA
snoRNA
snRNA
SRP_RNA
rRNA
RNase_MRP_RNA

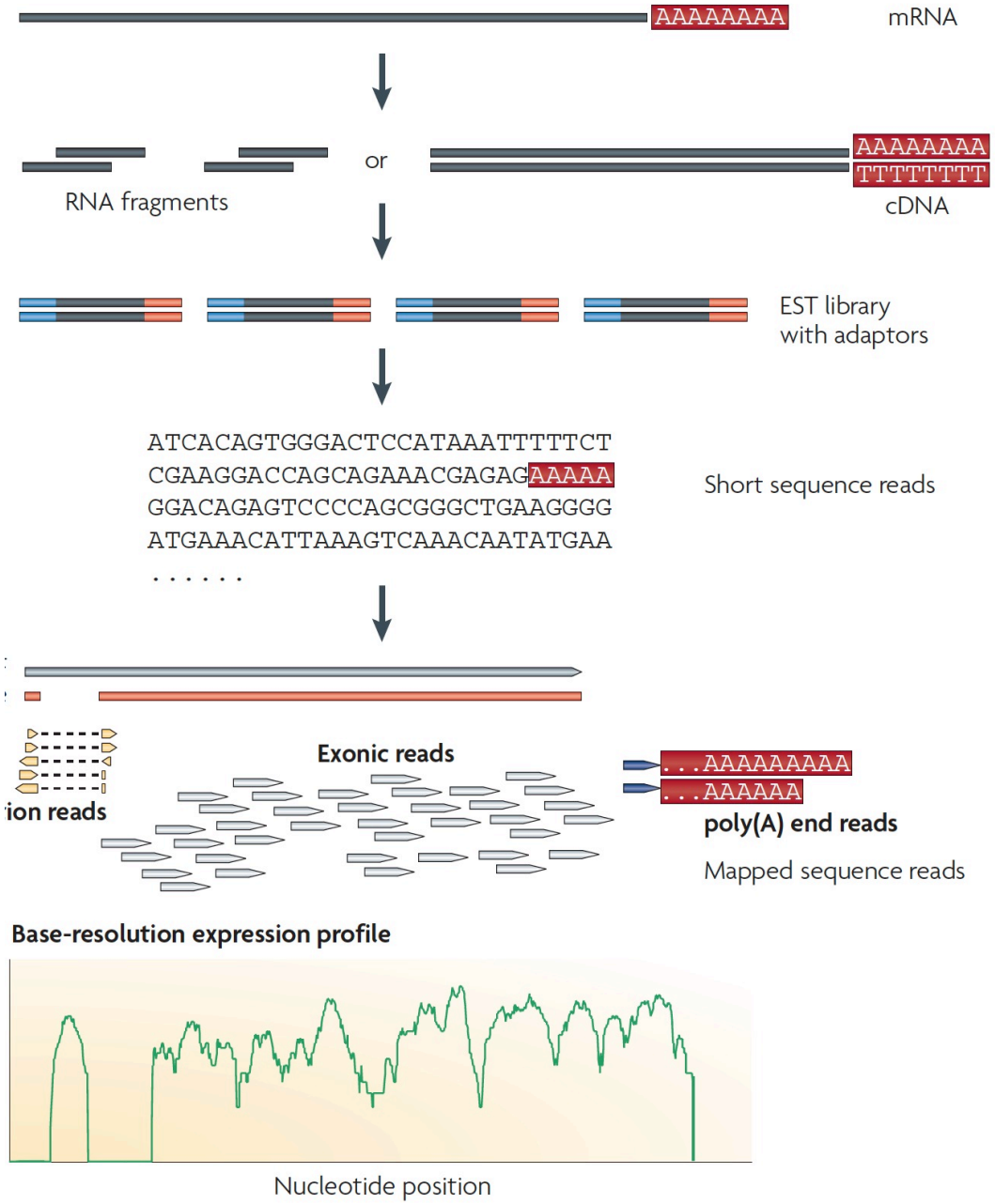
Arabidopsis_thaliana.TAIR10.41.gff3

(might have to use gunzip to de-compress)

RNA sequencing is the best way to find genes

Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project

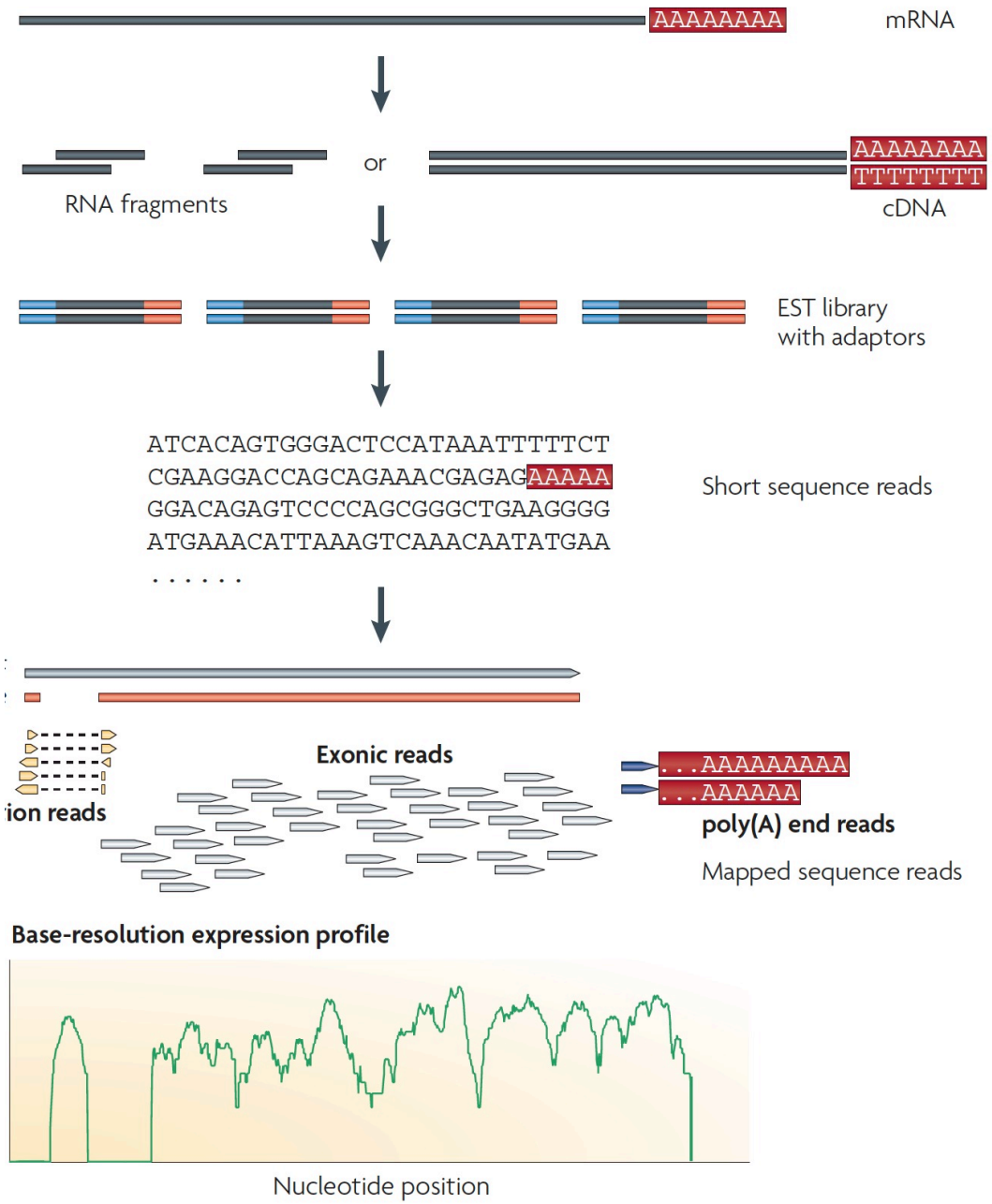
MARK D. ADAMS, JENNY M. KELLEY, JEANNINE D. GOCAYNE, MARK DUBNICK, MIHAEL H. POLYMEROPOULOS, HONG XIAO, CARL R. MERRIL, ANDREW WU, BJORN OLDE, RUBEN F. MORENO, ANTHONY R. KERLAVAGE, W. RICHARD MCCOMBIE, J. CRAIG VENTER*



Adams et al., 1991 Science; Wang et al., 2009 Nat Rev

RNA sequencing is the best way to find genes

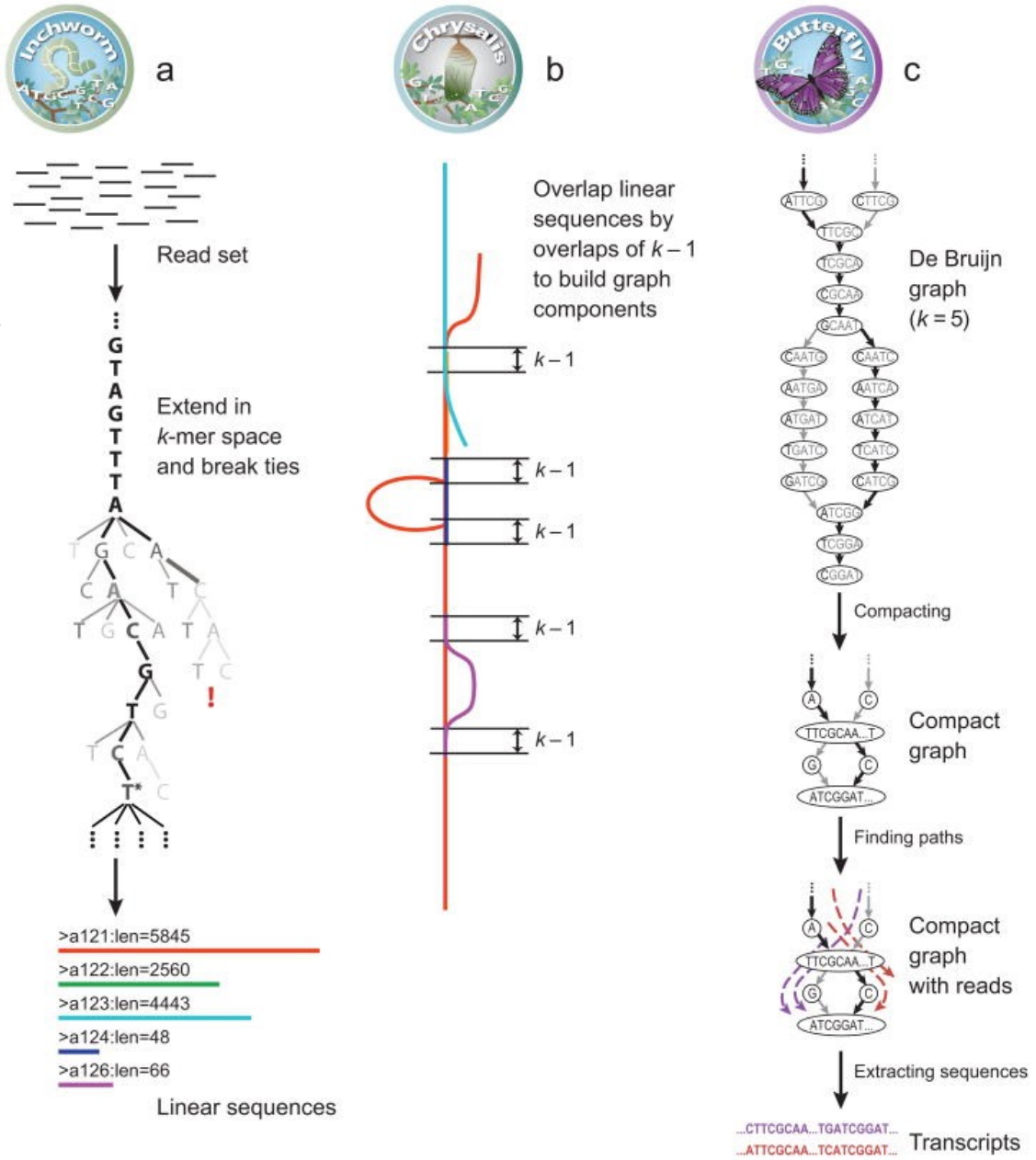
- Generate/sequence cDNA library



Wang et al., 2009 Nat Rev Genet

RNA sequencing is the best way to find genes

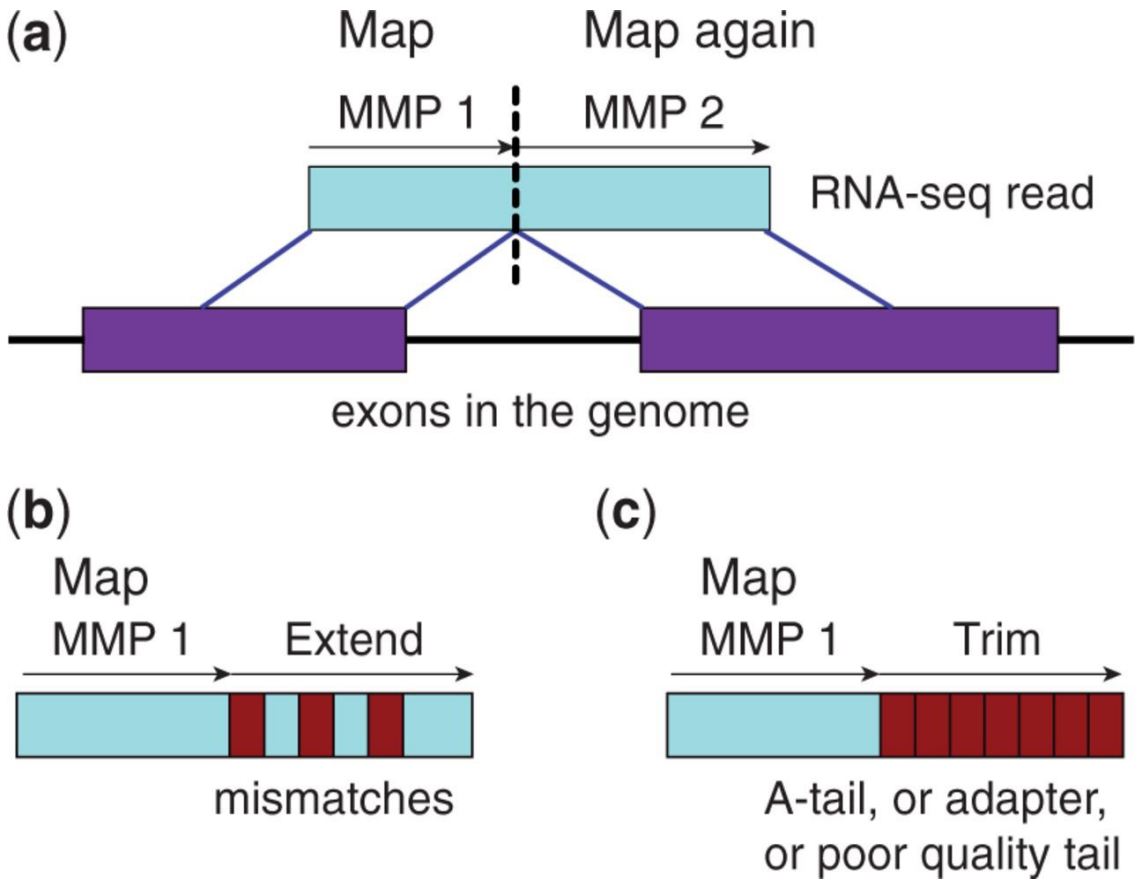
- Generate/sequence cDNA library
- Assemble RNAseq reads (optional and genome-free)



Trinity – Grabherr et al., 2011 Nat Biotechnol.

RNA sequencing is the best way to find genes

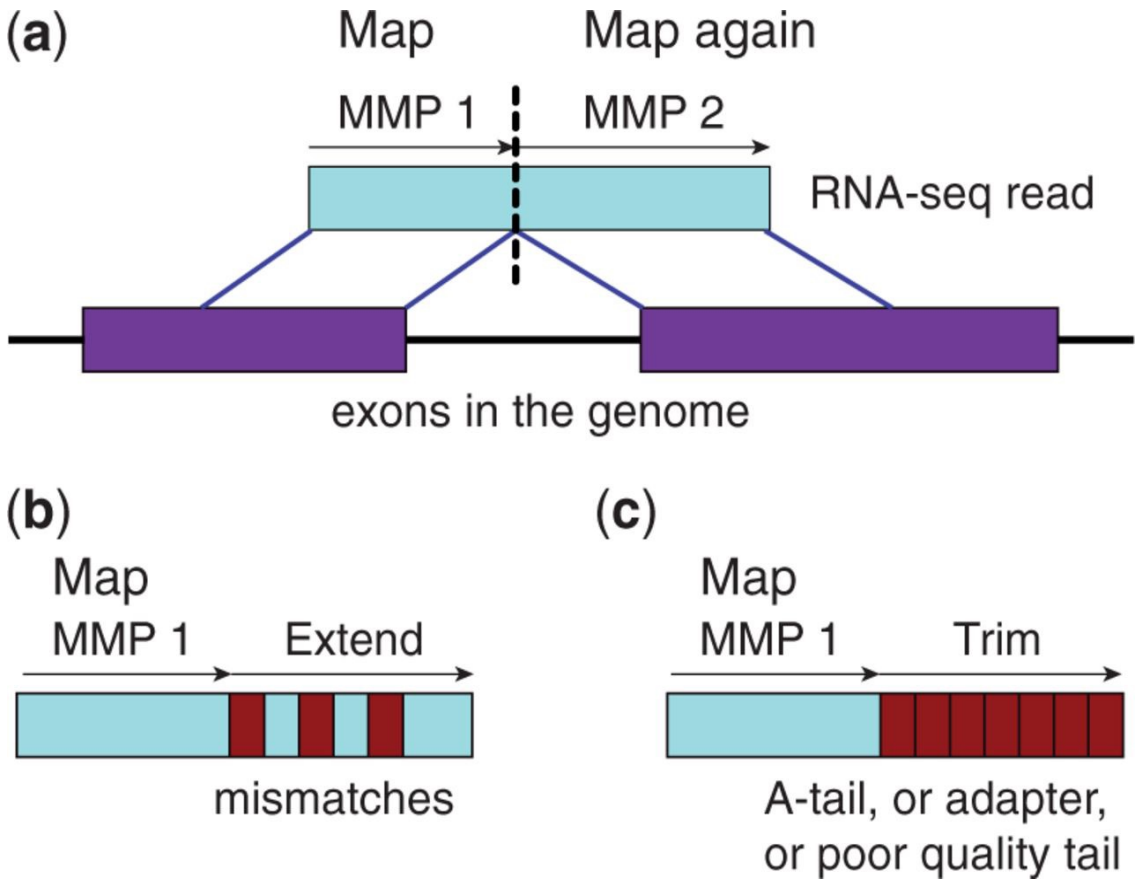
- Generate/sequence cDNA library
- Assemble RNAseq reads (optional and genome-free)
- Align RNAseq reads to genome sequence



Dobin et al., 2013 Bioinformat

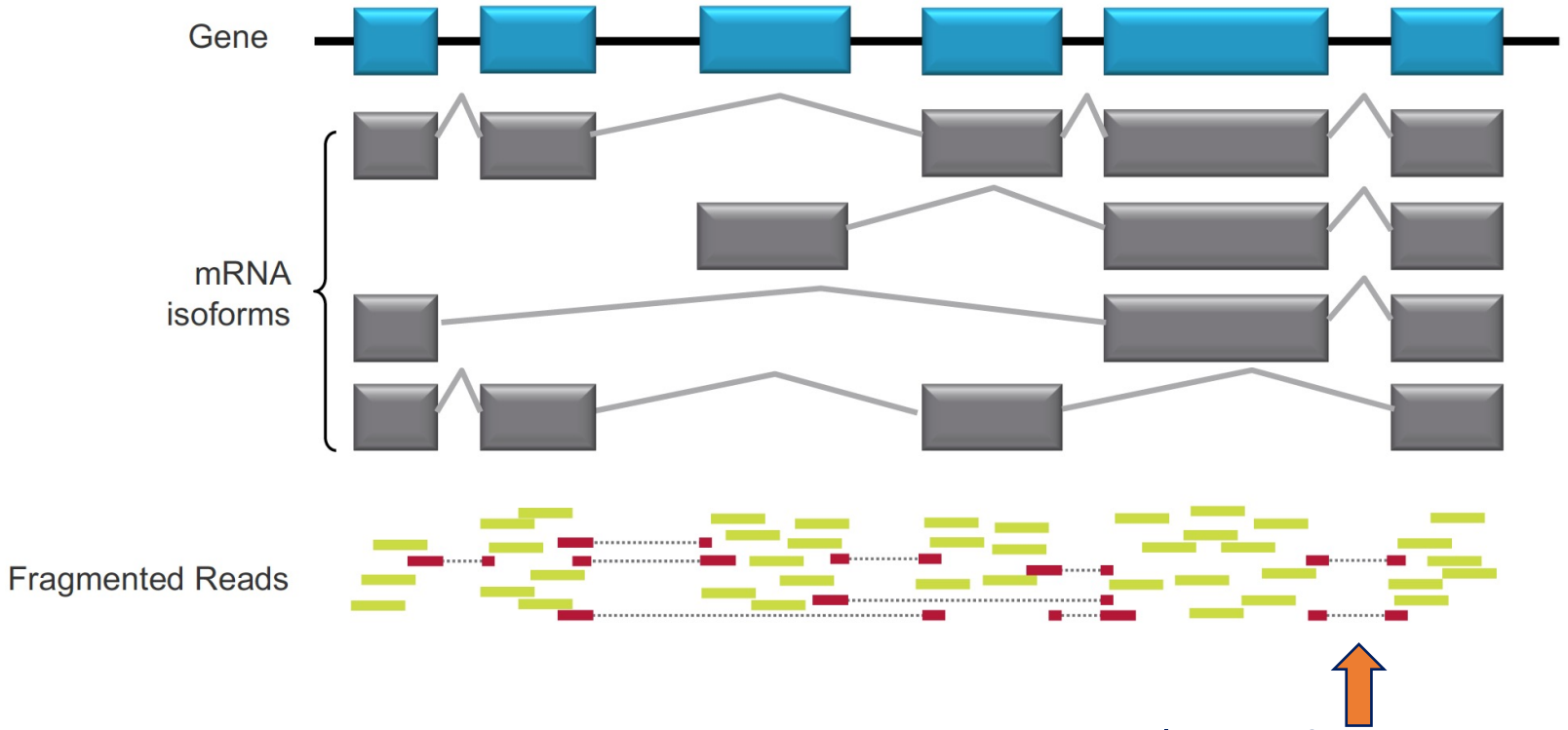
RNA sequencing is the best way to find genes

- Generate/sequence cDNA library
- Assemble RNAseq reads (optional and genome-free)
- Align RNAseq reads to genome sequence
- Use gaps to identify intron/exon boundaries, look for ORFs



Dobin et al., 2013 Bioinformat

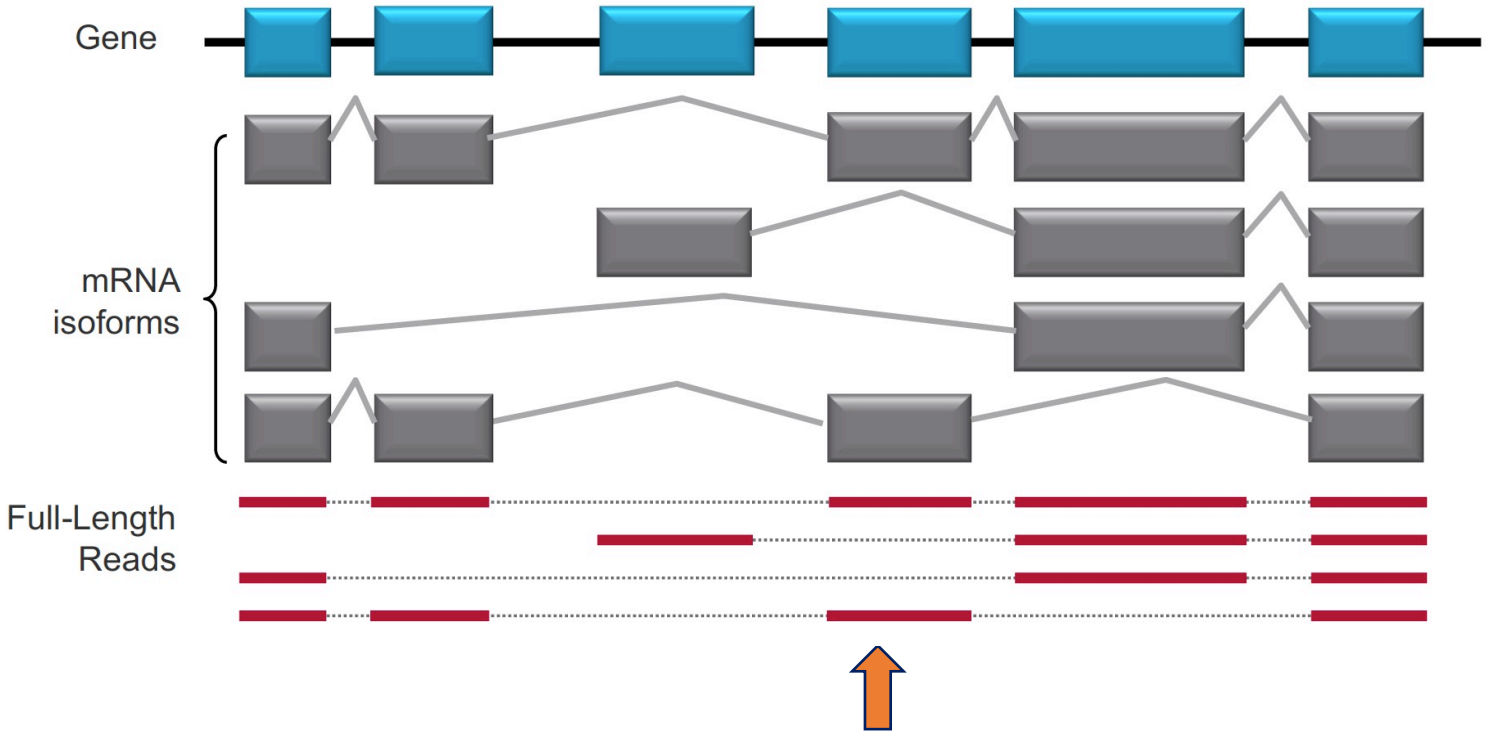
Developing high-quality gene models from short reads is **challenging!**



Have to rely on deep sequencing to ensure you capture enough splice-junction-spanning reads

Slides by Liz Zheng @MagDoll

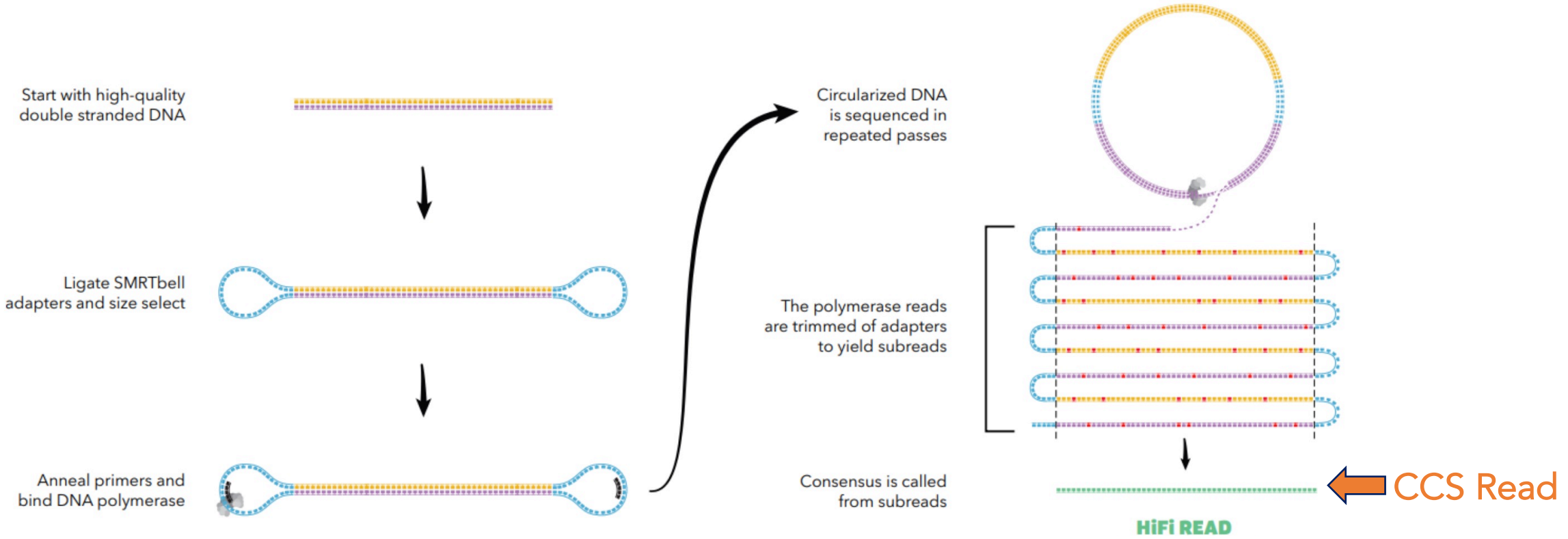
Sequencing cDNAs with long reads mostly alleviates that problem



Directly sequence entire transcripts with IsoSeq

Slides by Liz Zheng @MagDoll

IsoSeq: Uses single-molecule, real-time sequencing, plus Circular Consensus Sequencing to produce high quality, full-length transcript reads

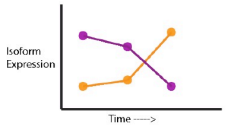


Slides by Liz Zheng @MagDoll

Bioinformatic tools associated with IsoSeq now coming online

		Input	Output
Sequencing			`subreads.bam`
Iso-Seq Analysis		`ccs.bam`	`Collapsed unique transcripts (GFF, FASTA)`

Transcript Classification		`Unique transcripts`
		`Reference genome`
		`Annnotation (GTF)`
		`CAGE Peak`
		`Junction data...`
			`Transcript classification`
			`Junction classification`
			`Figures`

Functional Annotation		`SQANTI output`	`Annotated GTF`
Differential Analysis		`Experimental design` `Annotated GTFs`	

Potential Project option

Supporting Tools

- collapse redundant transcripts
- merge multi-sample output
- saturation curve
- file format conversion
- single cell analysis

CO ding GEN ome reconstruction Tool

- gene family finding
- genome reconstruction
- evaluate assembly

TAMA

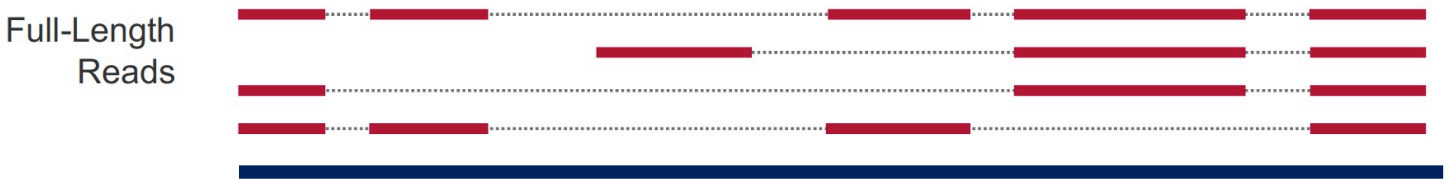
- collapse redundant transcripts
- merge multi-sample output
- NMD/ORF prediction
- transcript filtering

TALON

- long read processing & annotation pipeline developed independently by ENCODE4

Slides by Liz Zheng @MagDoll

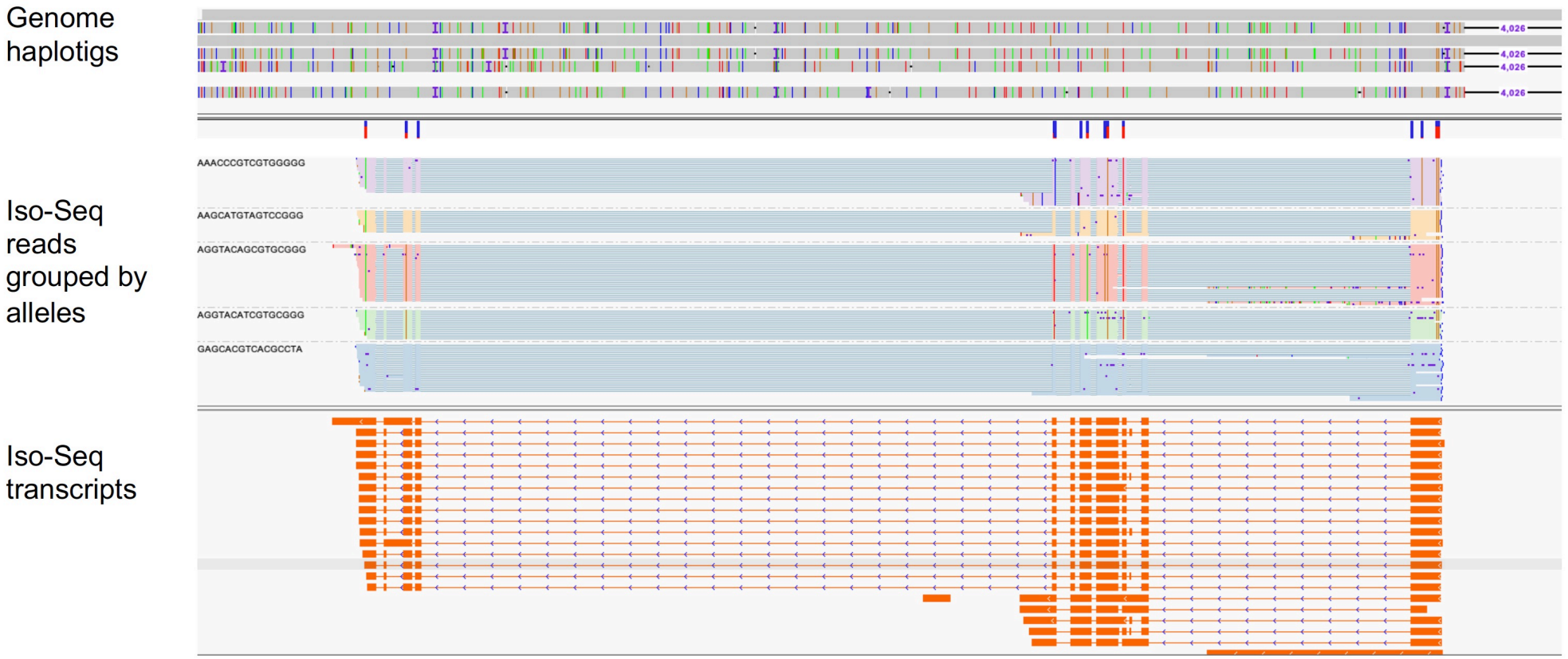
Gene models are no longer computational inferences, but direct biological observations



Full length transcripts mapped to genome

Slides by Liz Zheng @MagDoll

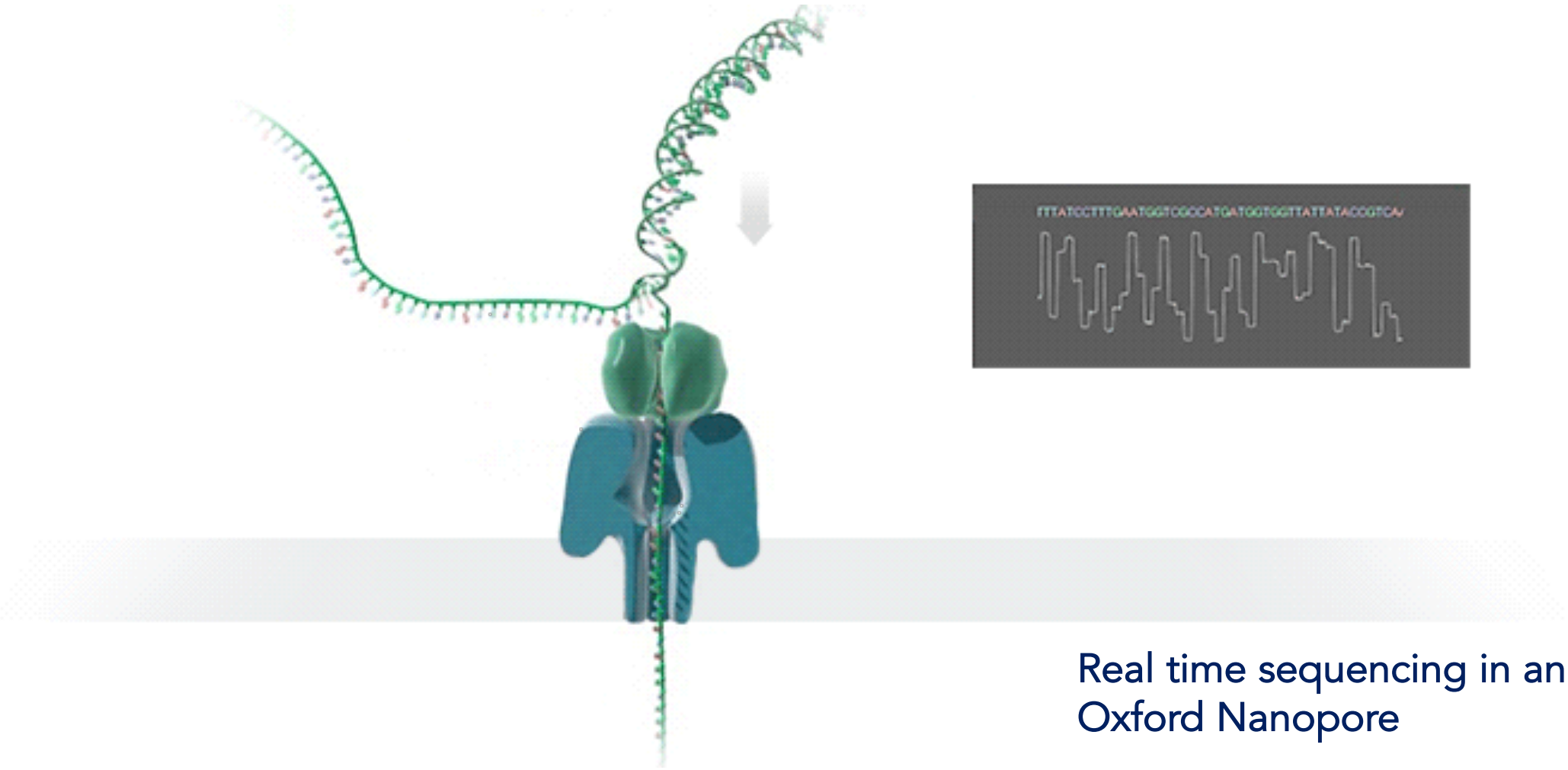
Clustering full-length transcripts with IsoPhase



IsoPhase claims to be able to distinguish IsoForms/Alleles/Paralogs

Slides by Liz Zheng @MagDoll

Other long-read technologies can also be used for RNAseq



Real time sequencing in an Oxford Nanopore

IsoSeq is awesome, still doesn't solve everything

What IsoSeq can give you:

- **8 Million** reads, **4 Million** CCS reads, **500k** Full length transcripts
- High quality, direct observations of tens of thousands of expressed genes
- Immediate, assembly-free genome annotation

What IsoSeq cannot give you:

- The full transcriptome
- Quantitative estimates of expression levels
- Analysis of reads produced on sequencing machines using older chemistry

IsoSeq is awesome, still doesn't solve everything

How to differentiate transcripts originating from similar but non-identical genes?

- Paralogs, isoforms, alleles (IsoPhase?)

PacBio-sourced bioinformatic tools are designed to highlight the PacBio technology, not intended for every use

- Need to be aware of the limitations of the technology
- Implement **your own workflow**, not someone else's!

De novo gene discovery – how to find genes *without* RNAseq

- Identifying genes in the absence of RNAseq data is difficult!
  - RNA can't be reliably sequenced from some organisms

- How might one discover genes without RNA?

De novo gene discovery – how to find genes *without* RNAseq

LilyPad.fasta

Open Reading Frames

GeneMarkS – coding potential

Glimmer

Next up: Primer Design & ENCODE discussion

Please Read: ENCODE 2011

Homework #3 Posted on Canvas/GitHub