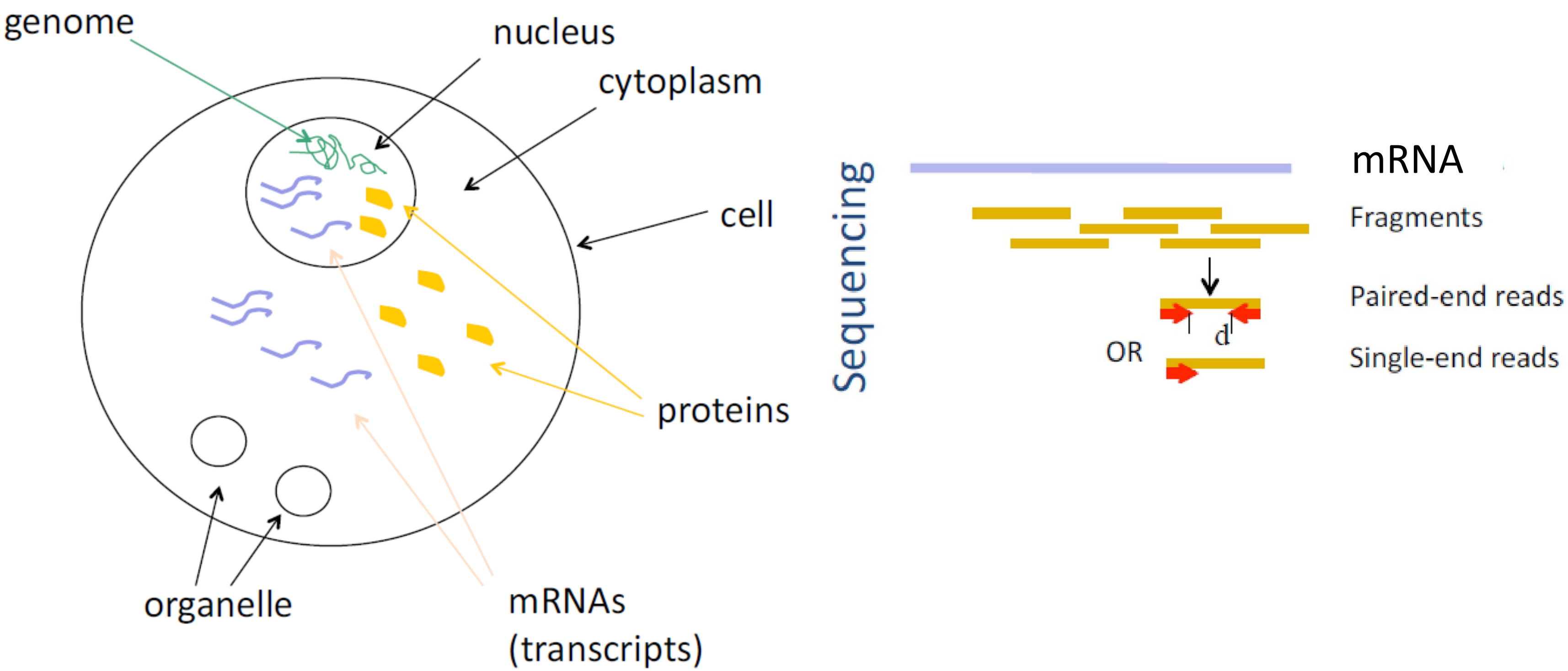
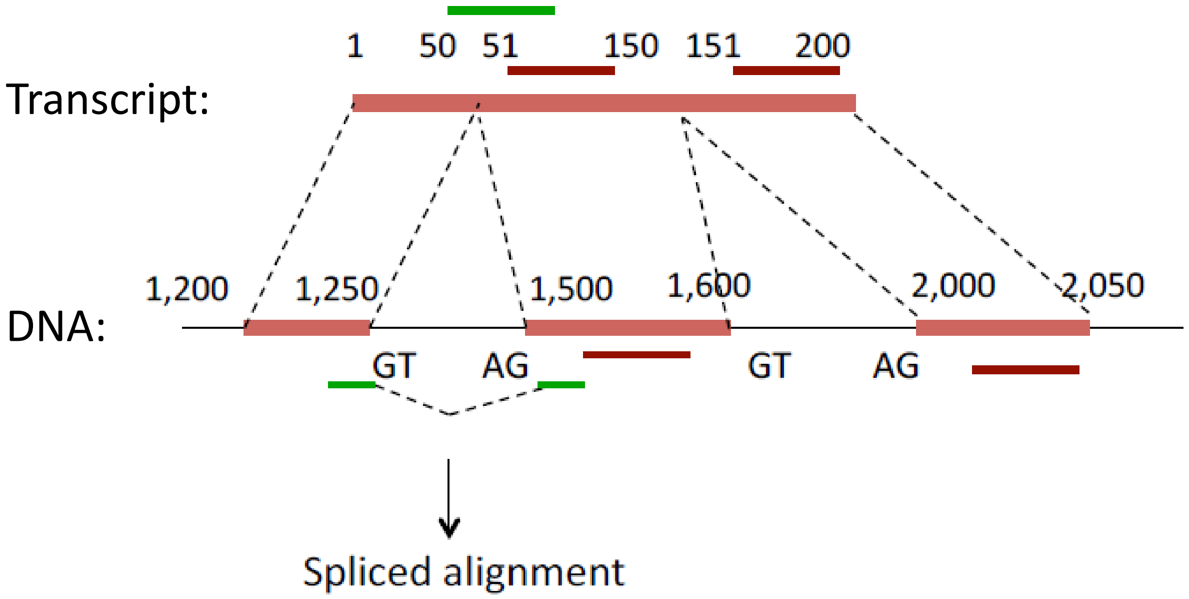
Motivation for analyzing mRNA-Seq data



1. What are the transcript variants of each gene?
2. What genes and transcripts are expressed and at what levels?
3. How do expression levels and transcript usage differ between different conditions?

Credit to https://www.coursera.org, Foivos Gypas

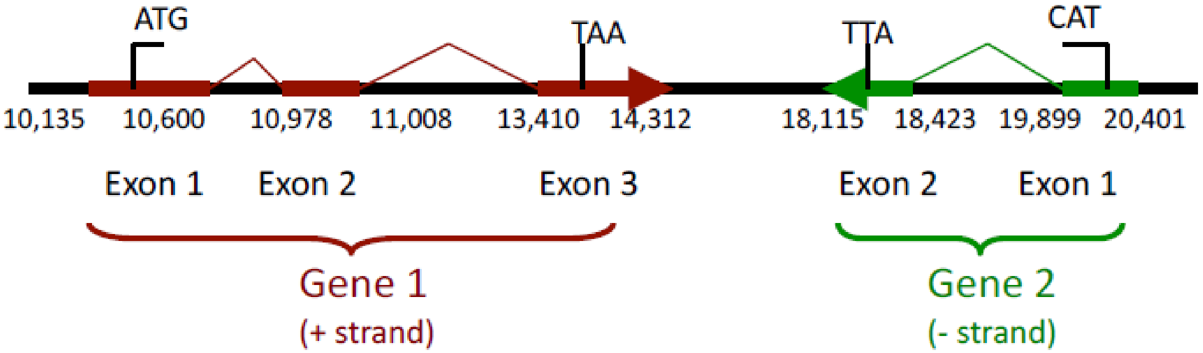
Splicing



https://en.wikipedia.org/wiki/RNA_splicing

Genomic features

- Genome annotation = determine the precise location and structure (intervals, or lists of intervals, and associated biological information) of genomic features along the genome
- Genomic features: genes, promoters, protein binding sites, translation start/stop site, DNaseI sites, etc.
- Example – gene annotations:
  - Exon/intron structure (exon and intron start-end coordinates)
  - Strand (+ or -)
  - Start and end sites for translation (ORF)



Representation: GTF format

```
#chr program feature start end strand frame gene_id; txpt_id

chr7 GF exon 10135 10600 100 + . gene_id “genA”; transcript_id “genA.1”;
chr7 GF exon 10978 11008 100 + . gene_id “genA”; transcript_id “genA.1”;
chr7 GF exon 13410 14312 100 + . gene_id “genA”; transcript_id “genA.1”;
chr7 GF exon 18115 18423 100 - . gene_id “genB”; transcript_id “genB.1”;
chr7 GF exon 19899 20401 100 - . gene_id “genB”; transcript_id “genB.1”;
```

- Each interval feature takes one line
- Columns 1-9 separated by tab ‘\t’; fields within column 9 separated by space ‘ ‘
- Column 9 can have additional attributes
- Coordinates are 1-based

Representation: SAM/BAM format

Header

```
@HD VN:1.0   SO:coordinate
@SQ SN:chr1  LN:248956422
@SQ SN:chr10 LN:133797422
@SQ SN:chr11 LN:135086622
…
@PG ID:TopHat VN:2.0.13 CL:/
data1/igm3/sw/packages/
tophat-2.0.13.Linux_x86_64/
tophat -p 8 -o …
```

Alignments

```
141217_CIDR4_0073_BHCFG7ADXX:2:1111:3128:29074   345
chr1  10021  0  68M  * ACCCTAA...CCCTAAC  @DC?=2...DDDD@?@
AS:i:0 XN:i:0 XM:i:0   XO:i:0 XG:i:0 NM:i:0   MD:Z:68 YT:Z:UU
NH:i:10   CC:Z:chr10   CP:i:10004   XS:A:- HI:i:0

. . .
```

Representation: SAM/BAM format

`141217_CIDR4_0073_BHCFG7ADXX:2:1111:3128:29074`	Read id
`99`	**FLAG**
`chr1`	Chr
`10021`	Start
`0`	Mapping quality
`50M`	**CIGAR** (alignment)
`=`	Mate chr
`10151`	Mate start
`180`	Mate dist
`ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC`	Query seq
`@DC?=2.FFGE@7>C62>BGABGB9HFBAFIIHEGFIIIHFAIIGDA<FC`	Query base quals
`AS:i:0`	Alignment score
`NM:i:0`	Edit distance to reference
`NH:i:10`	Number of hits
`XS:A:-`	Strand
`HI:i:0`	Hit index for this alignment

Tags: [A-Za-z][A-Za-z]:[AifZH]:.*
where A =character; i = integer; f = float; Z=string; H = hex string