Sequencing Transcripts

Overview

I. RNA molecules review

II. ESTs

III. RNASeq

  I. Goals

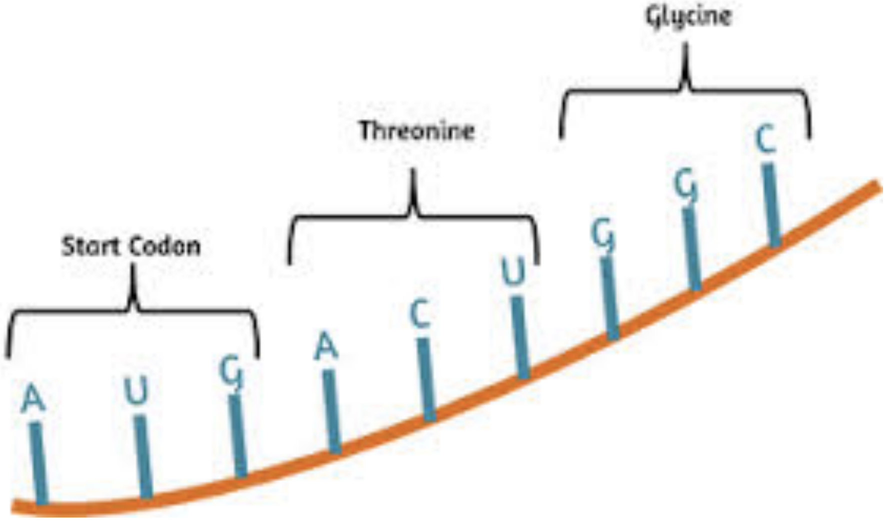
  II. Limitations

IV. Yeast dataset

V. GFF3 file format

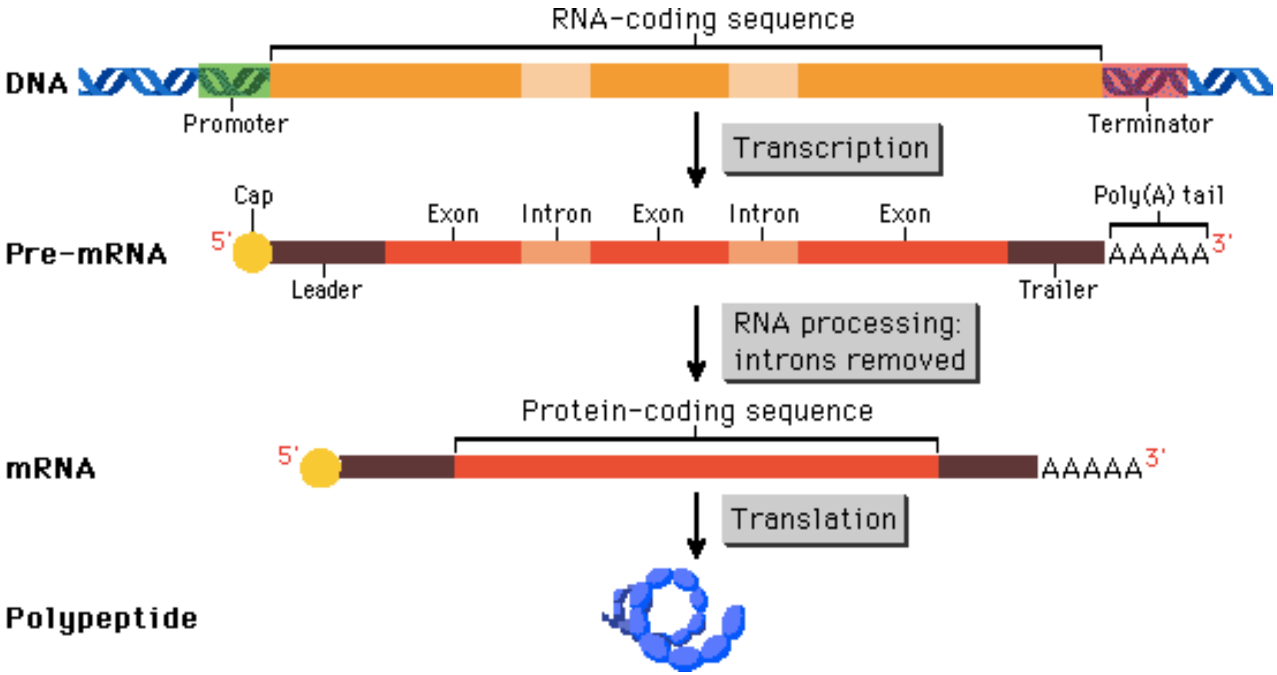
mRNA

- Messenger RNA
- "messenger" because it conveys information from the DNA to the ribosome
- Codons – sets of 3 bases – are translated to amino acids to produce proteins



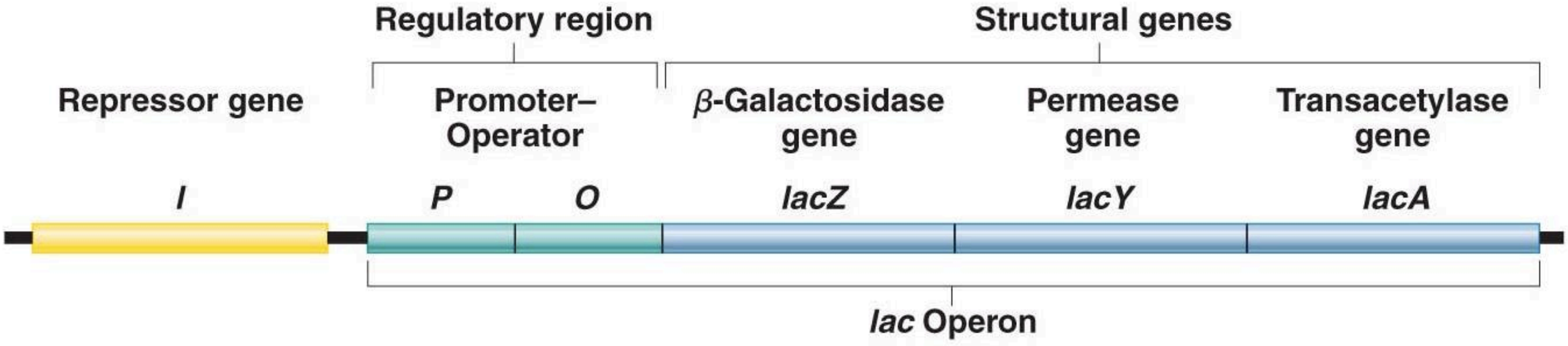
Eukaryotic mRNA

- Transcriptome from DNA to RNA
- Start with a pre-mRNA, then:
  - 5’ cap addition (modified guanine)
  - Splicing - Introns are removed
  - Editing – nucleotides may be altered
  - Polyadenylation – a series of adenine bases is added, called the Poly-A tail
- After processing, it is a mature mRNA



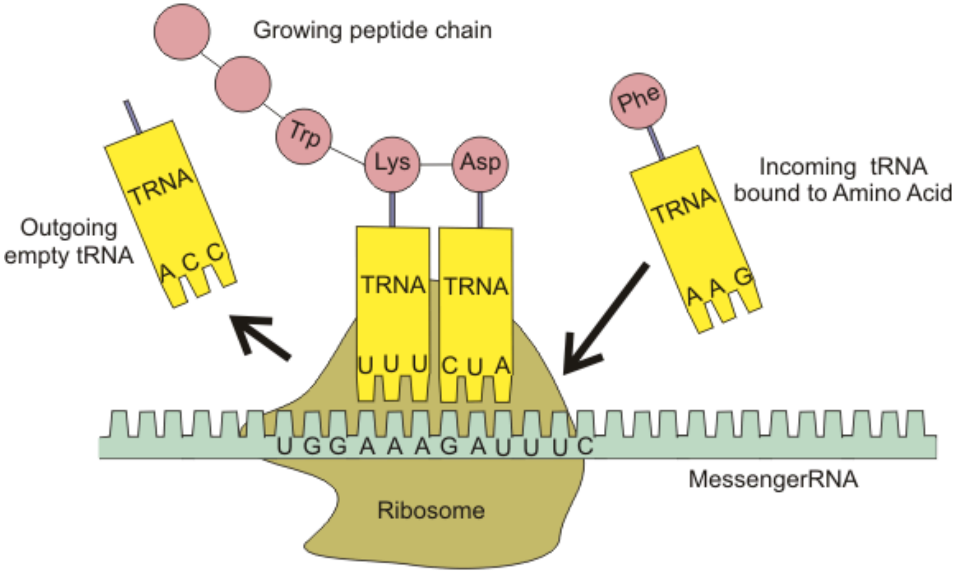
Prokaryotic mRNA

- Polygenic
  - A single mRNA can include several genes
  - Operon
- Less processing – no introns, no cap
- Can have a poly-A tail but not always



Lots of types of RNA

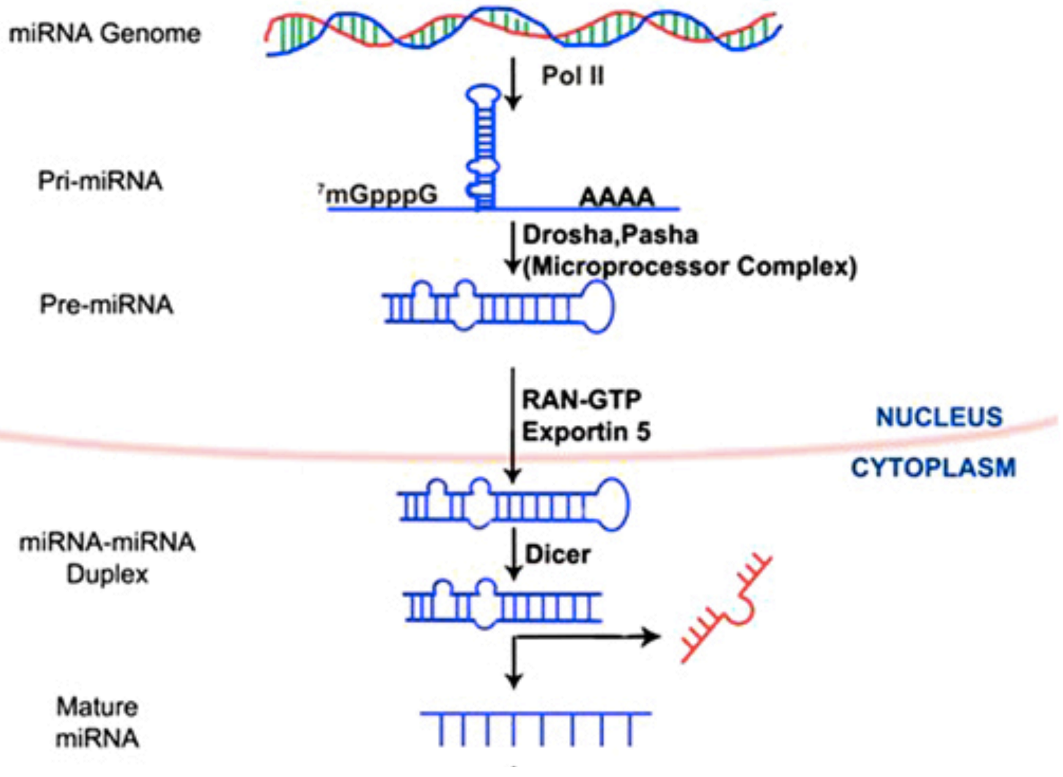
- rRNA
  - Ribosomal RNA
  - Forms the ribosome, which "reads" the mRNA and builds the amino acid chain
- tRNA
  - Transfer RNA
  - Brings the correct amino acid to the ribosome based on the codon being read



Peptide Synthesis

miRNA

- Noncoding
- Functions:
  - RNA silencing
  - post-transcriptional regulation of gene expression
- Mature form is about ~22 nucleotides
- Relatively recent discovery – they were identified as important biological regulators in the early 2000s



Tong and Nemunaitis, 2008

More types of RNA

- siRNAs – short interfering RNAs
- snRNA - small nuclear RNA
- snoRNA - small nucleolar RNA
- lincRNA - long intergenic non-coding RNAs
- Etc.
- (30 types listed on wikipedia)

Greater than 83% of the genome is transcribed (at least in humans) *

-

But its relatively rare

Djebali et al, 2012 Landscape of transcription in human cells. Nature
Hangauer et al., 2013 Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. Plos Genetics

RNASEQ

RNA Sequencing

- Transcriptome shotgun sequencing
- Library prep important. What do you want to sequence?
  - Total RNA (can be up to 90% rRNA)
  - Short RNAs
  - mRNAs (this is the most common)
    - Two methods:
    - Poly-A enrichment
    - Ribosomal RNA Removal

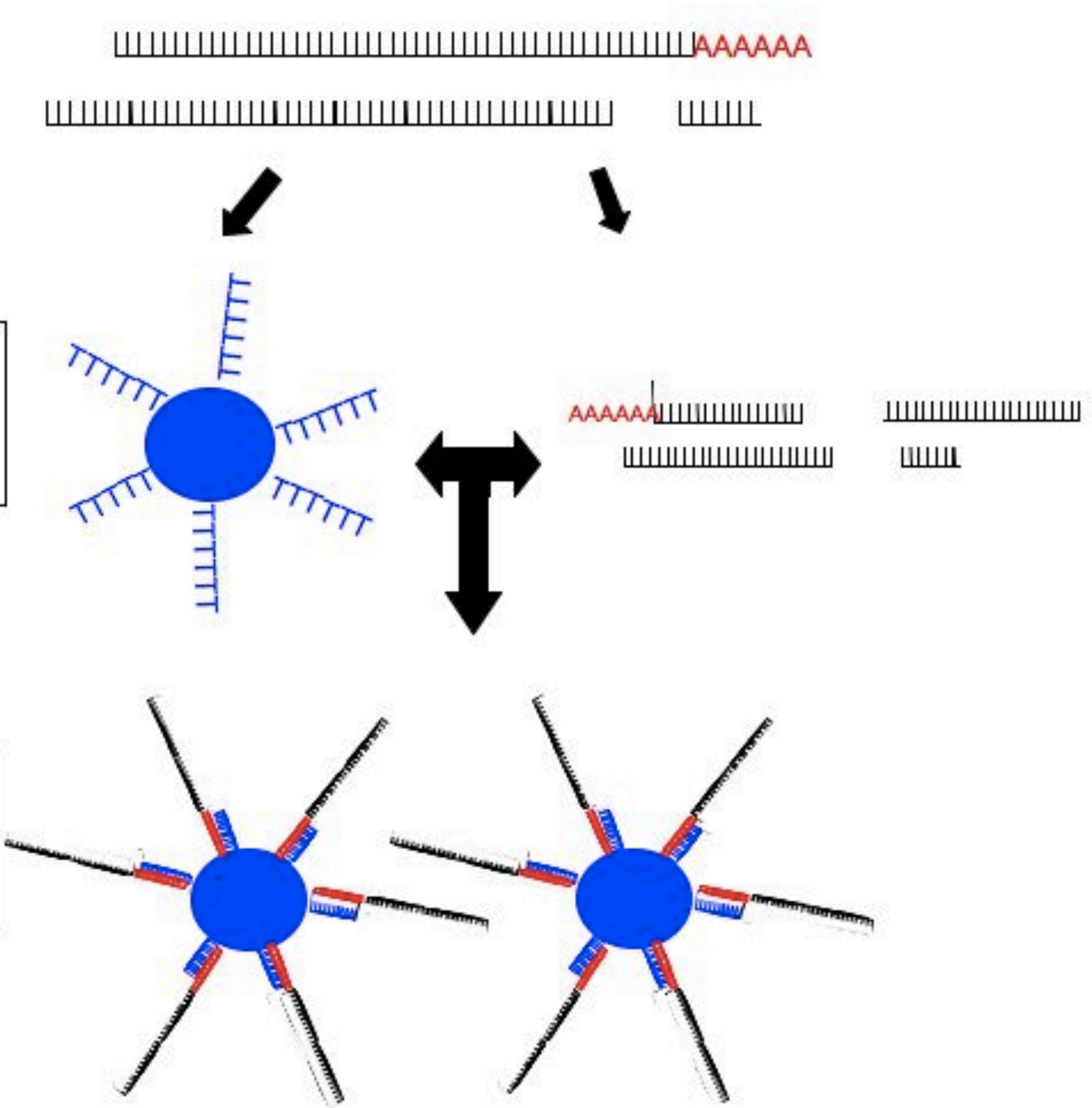
Poly A Enrichment

Isolate Total RNA

Fragmentation and/or Isolation

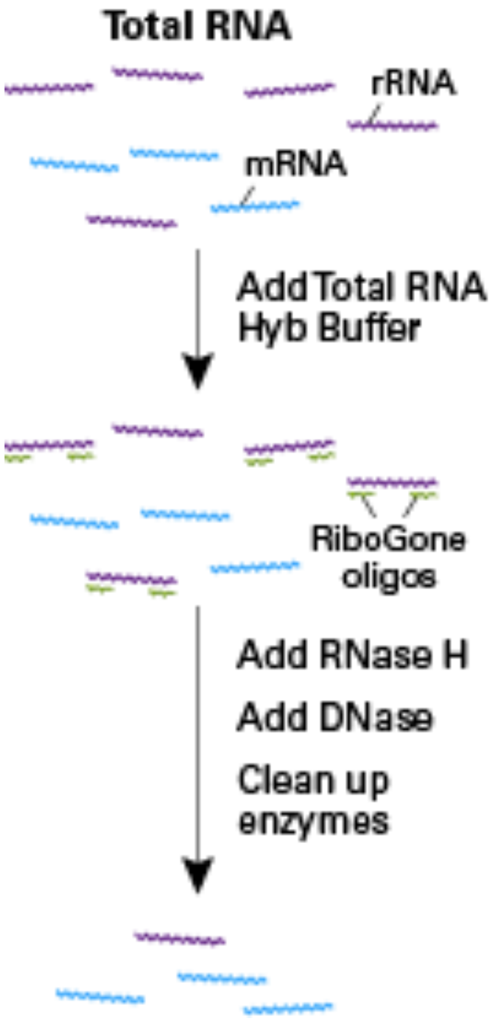
In this case, isolation via Poly(T) coated magnetic beads

Poly(A) RNA molecules bind to the Poly(T) magnetic beads



Remove rRNA

- Goal is to subtract rRNA, thus enriching for mRNA
- Hybridization/bead capture procedure that selectively binds target sequences using biotinylated capture probes
  - This leaves other types of RNA, including non-coding types



Clontech

miRNA-Seq

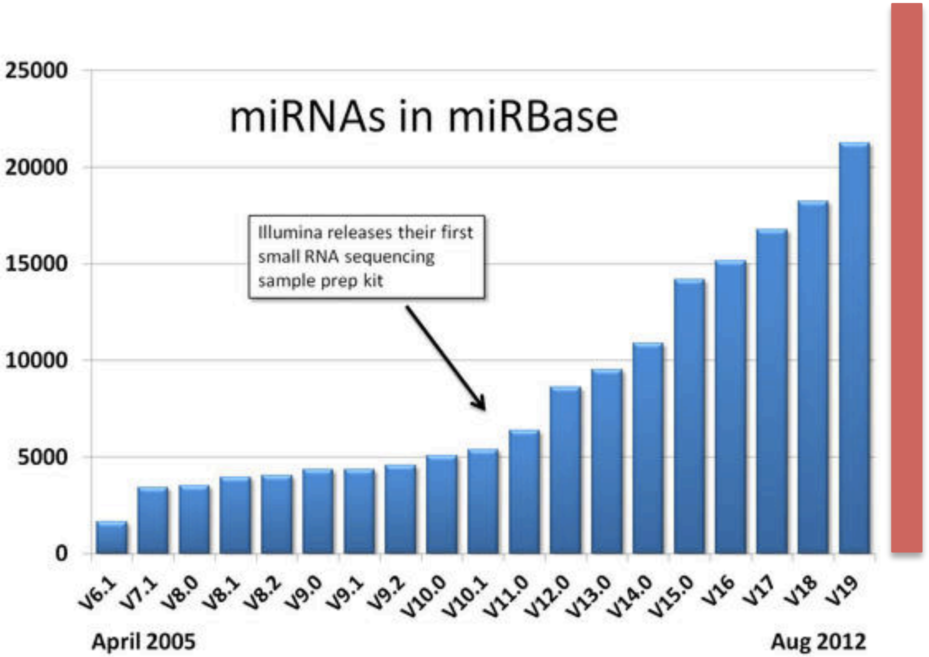
- Target small RNAs from total RNA sample
  - Can use gel electrophoresis to further narrow RNA species by length

miRBase

- 28,645 entries representing hairpin precursor miRNAs
- expressing 35,828 mature miRNA products
- 223
- species

RNASeq Types

- To target small RNA
  - Can start with total RNA or purified small RNA samples
  - Adapters are designed to target miRNAs
    - Small RNAs have a 3' hydroxyl group resulting from enzymatic cleavage by Dicer or other RNA processing enzymes.

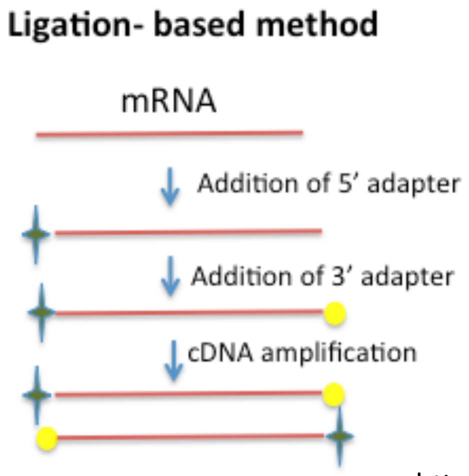

miRNAs in miRBase

Illumina releases their first small RNA sequencing sample prep kit

28,645 entries June 2014

Advantages of Strand Specific Sequencing

- Different protocols take different approaches, but all result in sequencing the original strand only (not the RC)
  - Ligation-based
  - dUTP-based
- Good for assembly and mapping
- Differentiate overlapping genes, psuedogenes, antisense transcripts
- Identifying the transcribed strand for non-coding RNAs



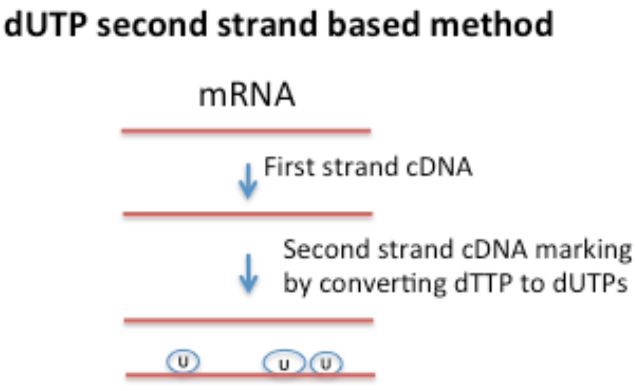
Ligation- based method

mRNA

Addition of 5' adapter

Addition of 3' adapter

cDNA amplification



dUTP second strand based method

mRNA

First strand cDNA

Second strand cDNA marking by converting dTTP to dUTPs

https://cofactorgenomics.com/directional-rna-sequencing/

Public Datasets

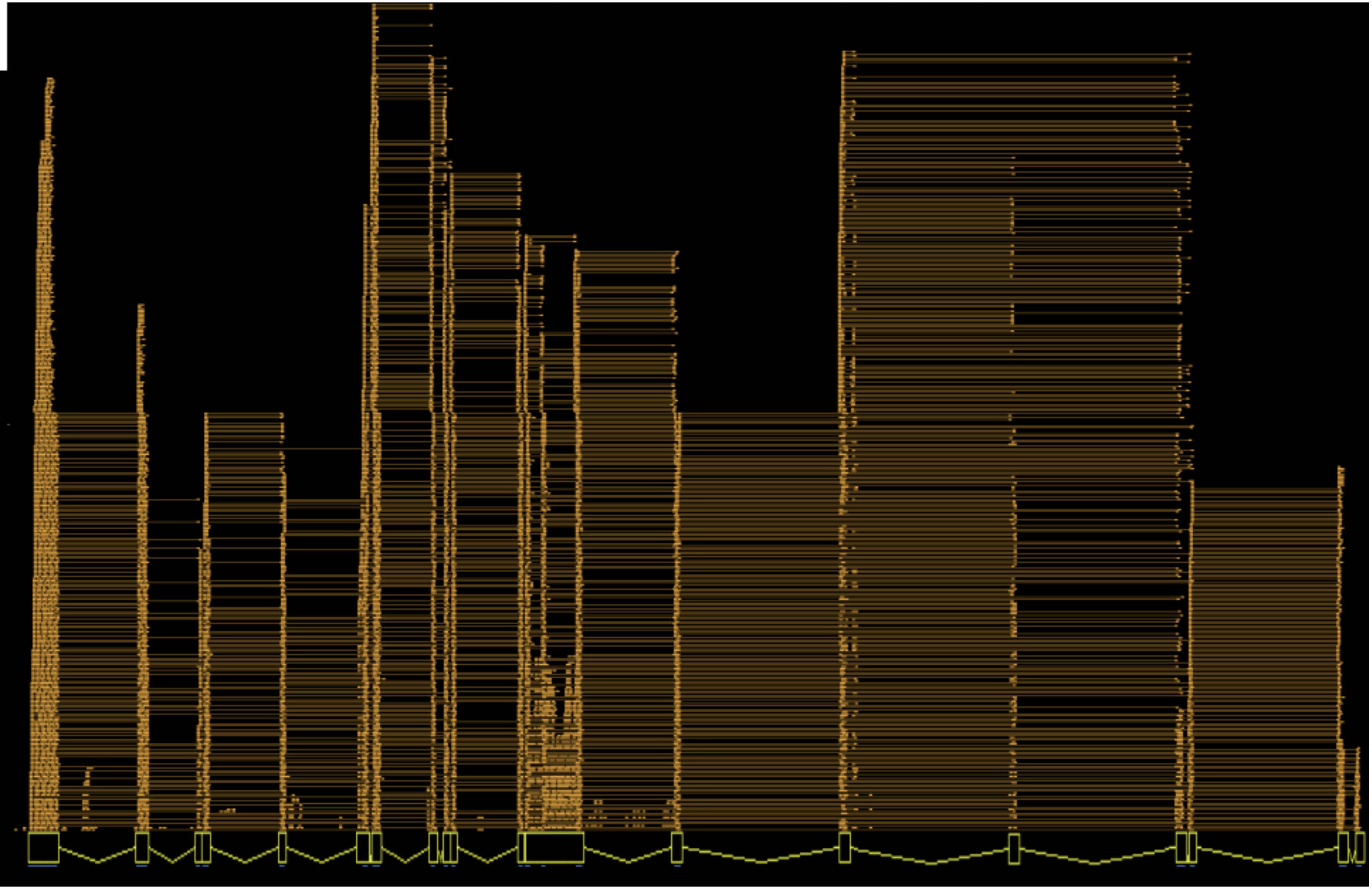
NCBI SRA

- 1,639,024 DNA Seq datasets
- 404,685 RNA Seq datasets
  - 9,898 specify miRNA (~5%)

Experimental Goals for mRNA Seq

- Catalog of genes (What are the genes in this organism?)
- Gene expression levels (What genes are expressed in this tissue or under this condition?)
- Differential gene expression levels (How does gene expression change under different conditions?)
- All of the above for alleles and splice variants
- Annotating the genes in a reference genome
- Variant (Genetic marker) discovery – SNPs, SSRs
- Post-transcriptional modifications, RNA-editing

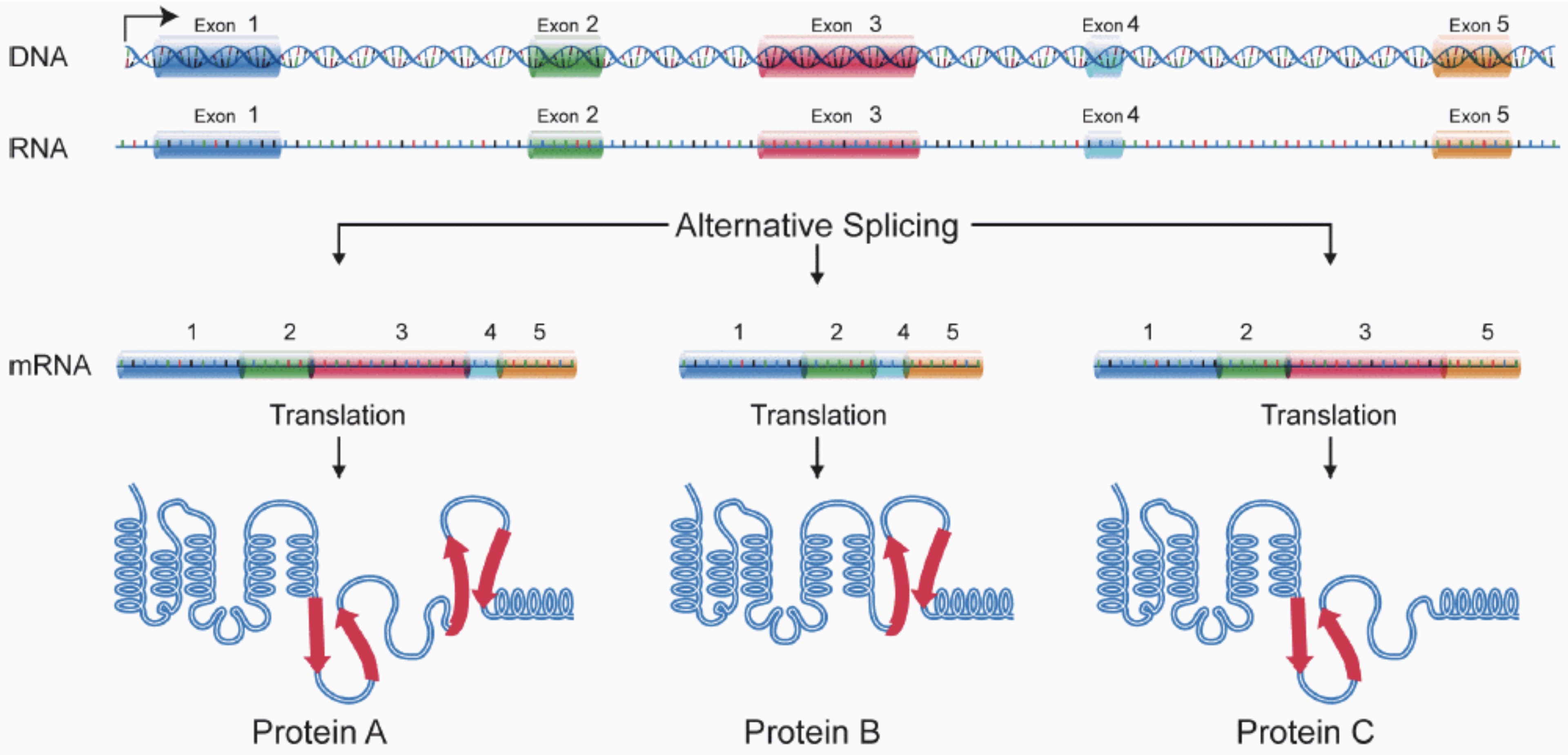
Genome Annotation



Identify exonic structure based on alignments.

Simon White – Ensembl Talk

Alternative Splicing



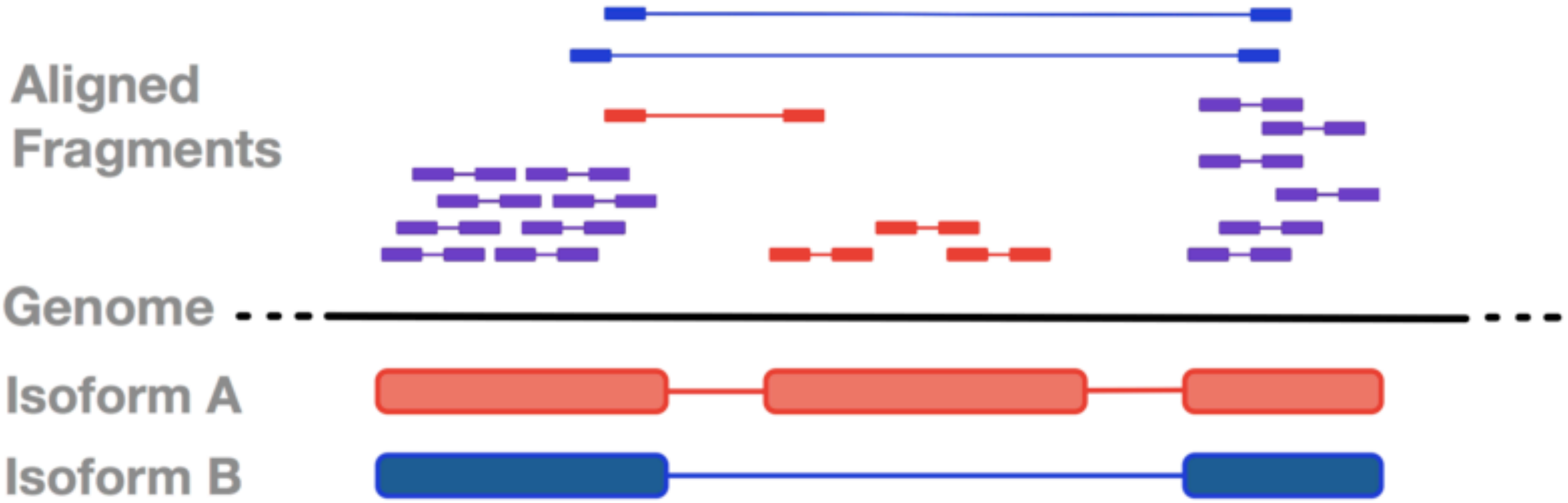
Splice variants are often tissue-specific. In humans, up to 95% of multiexonic genes have multiple splice isoforms.

Detecting Known Isoform Variants

Ambiguous – No information about isoform.

Indicate isoform A.

Indicate isoform B.

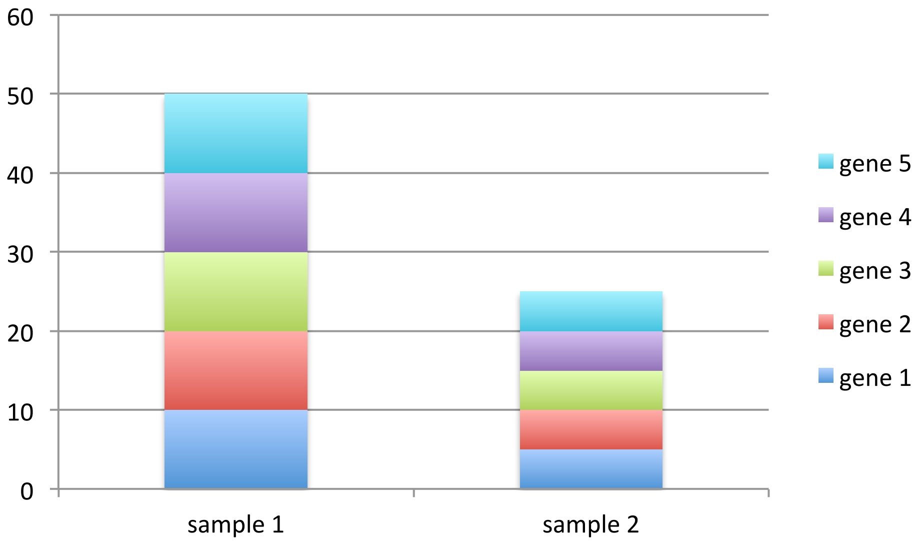


http://cgrlucb.wikispaces.com/Isoform+Deconvolution+and+Unannotated+Species

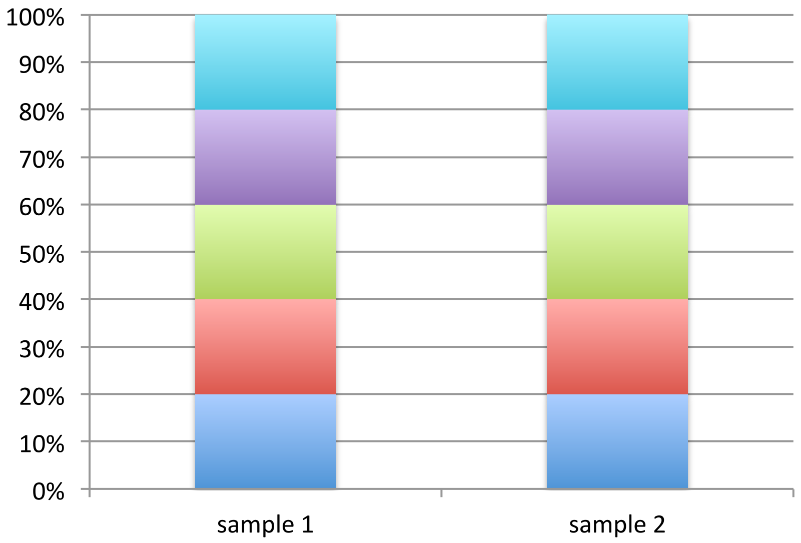
Limitations of RNASeq

RNASeq gives you relative abundance only

Absolute Quantities



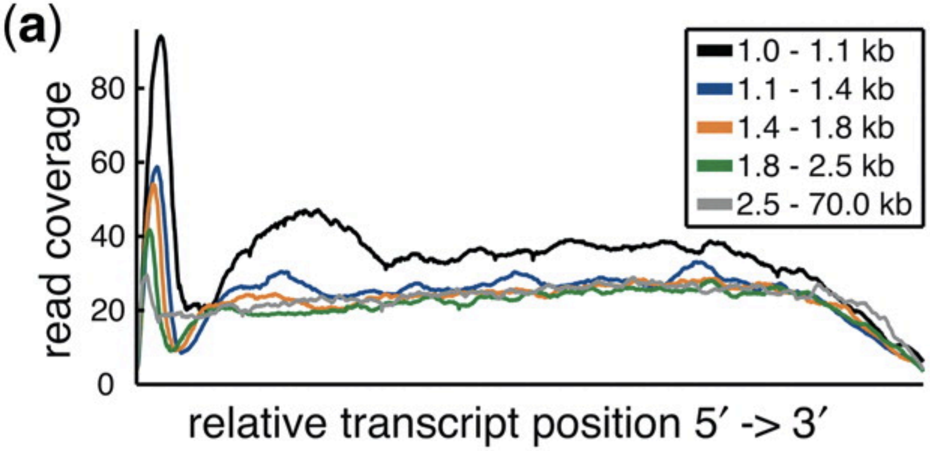
Relative Quantities



Additional information, such as levels of “spike-in” transcripts, are needed for absolute measurements (and these are suspect)

Limitations

- Reverse transcription, PCR and fragmentation steps can introduce biases
    - GC bias, length bias
    - Reads are not uniformly distributed along transcript length
- PCR-free preps are available



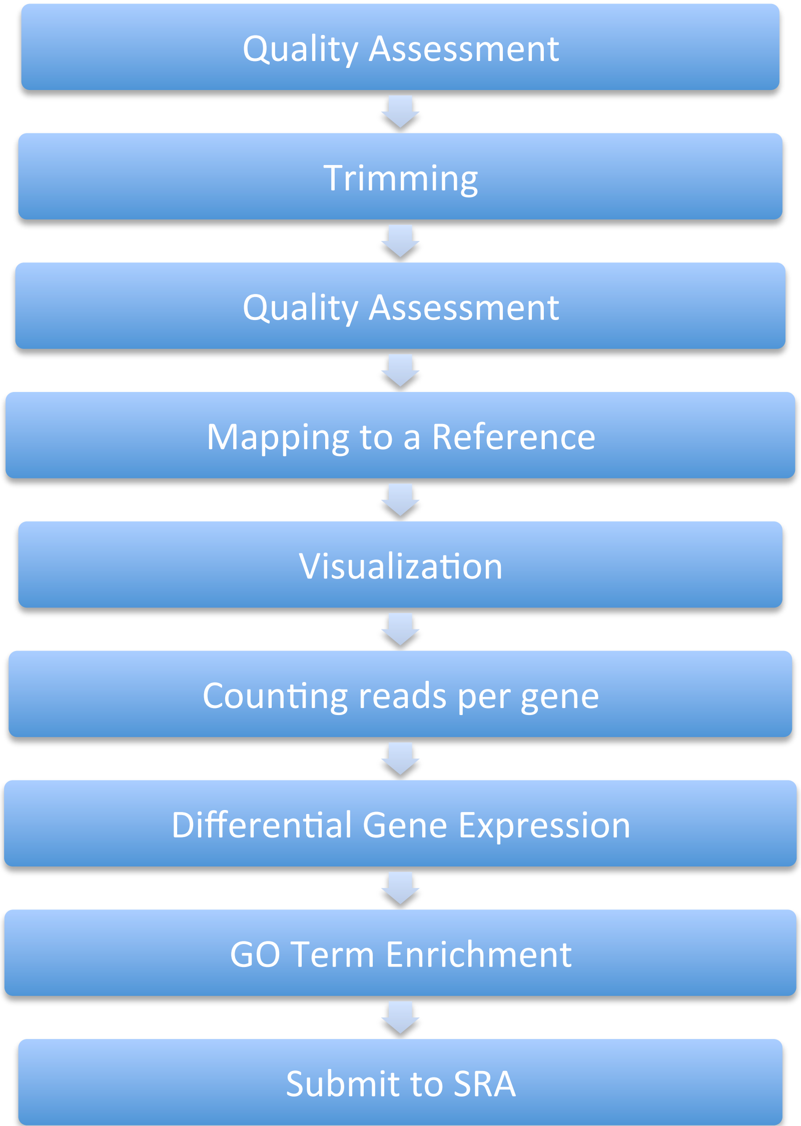
Bohnert and Ratsch, 2010

Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010 Jul;38(12):e131.

Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol.2011;12(3):R22.

Standard protocol for analysis for mRNA sequencing

mRNA Data Analysis Pipeline



Quality Assessment

Trimming

Quality Assessment

Mapping to a Reference

Visualization

Counting reads per gene

Differential Gene Expression

GO Term Enrichment

Submit to SRA

FastQC Babraham Bioinformatics

Trimmomatic

FastQC Babraham Bioinformatics

TopHat
A spliced read mapper for RNA-Seq

igv Integrative Genomics Viewer

HTSeq

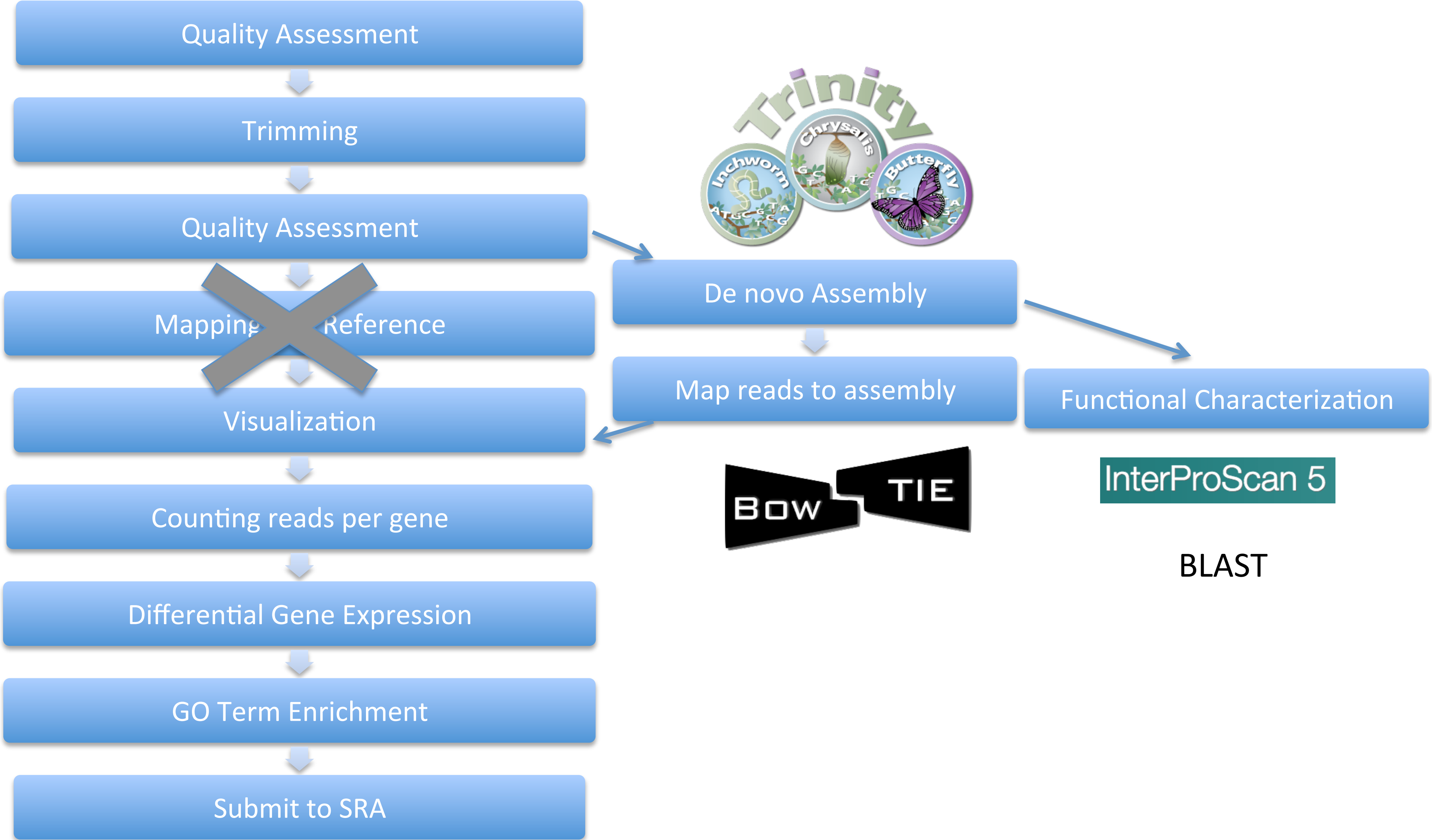
DESeq2

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

go BiNGO

NCBI

What if you don’t have a reference?



Quality Assessment

Trimming

Quality Assessment

Mapping Reference

Visualization

Counting reads per gene

Differential Gene Expression

GO Term Enrichment

Submit to SRA

De novo Assembly

Map reads to assembly

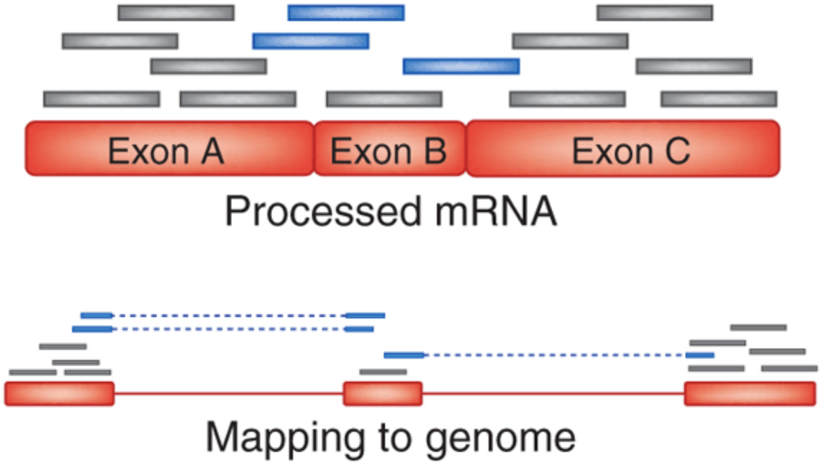
Functional Characterization

BLAST

Mapping to the Reference

- Mapping RNA to a eukaryotic genome is more complicated than mapping DNA
  - Introns
  - Alternative splicing

- Usually, you want to use a mapping software designed for RNASeq
  - The software will use a file (gff3) to know where the genes are located
  - Many RNASeq mapping software packages will also infer gene structures (This is good for identifying novel genes and isoforms)

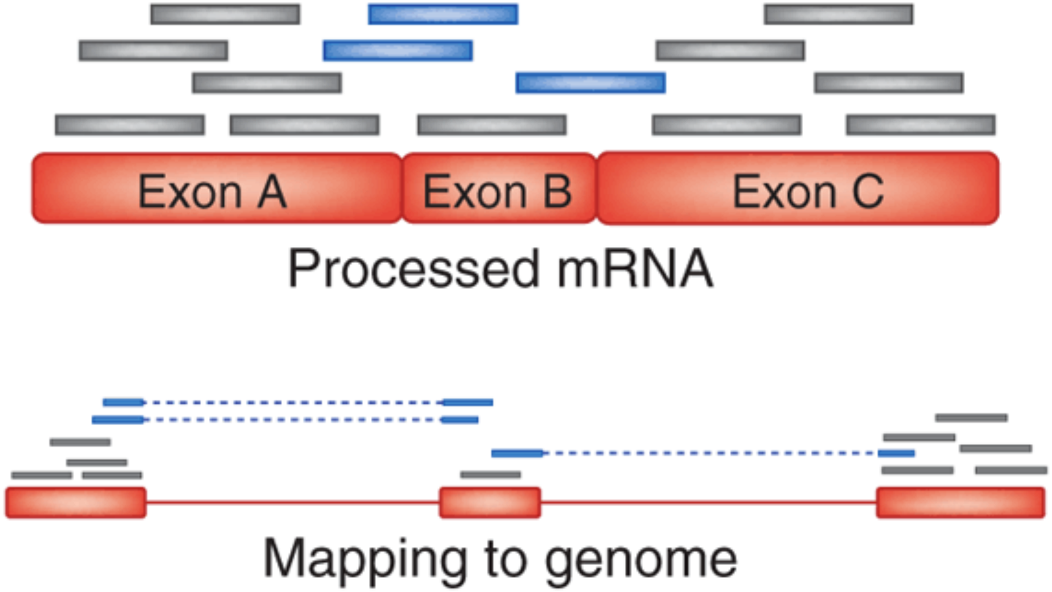


Read Mapping

- Can choose to do un-spliced alignment
- Why would you do this?
  - If you have a prokaryotic system with no introns
  - Align to the reference CDS transcriptome sequences, not the whole genome - Faster. (RapMap)
    - But you miss novel genes/isoforms
  - Some aligners are more sensitive, can align with more differences (MAQ,Stampy)

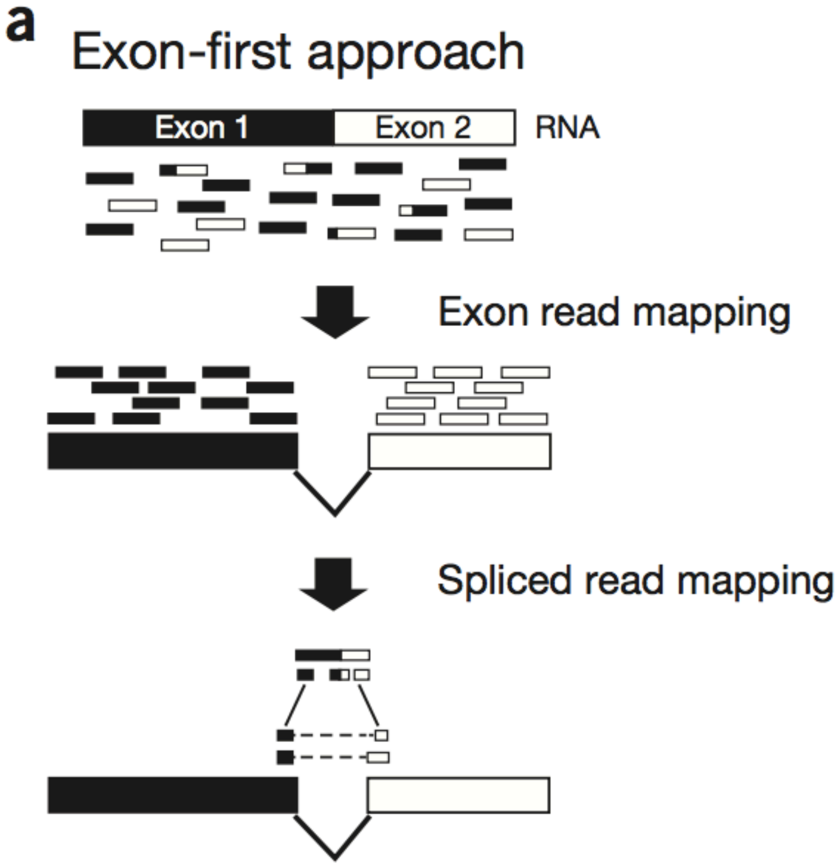
Splice-aware aligners

- Some reads will need to have very large gaps introduced to account for introns
- Intron could occur inside an individual read or between to read pairs
- Two categories
  - Exon-first
  - Seed and extend



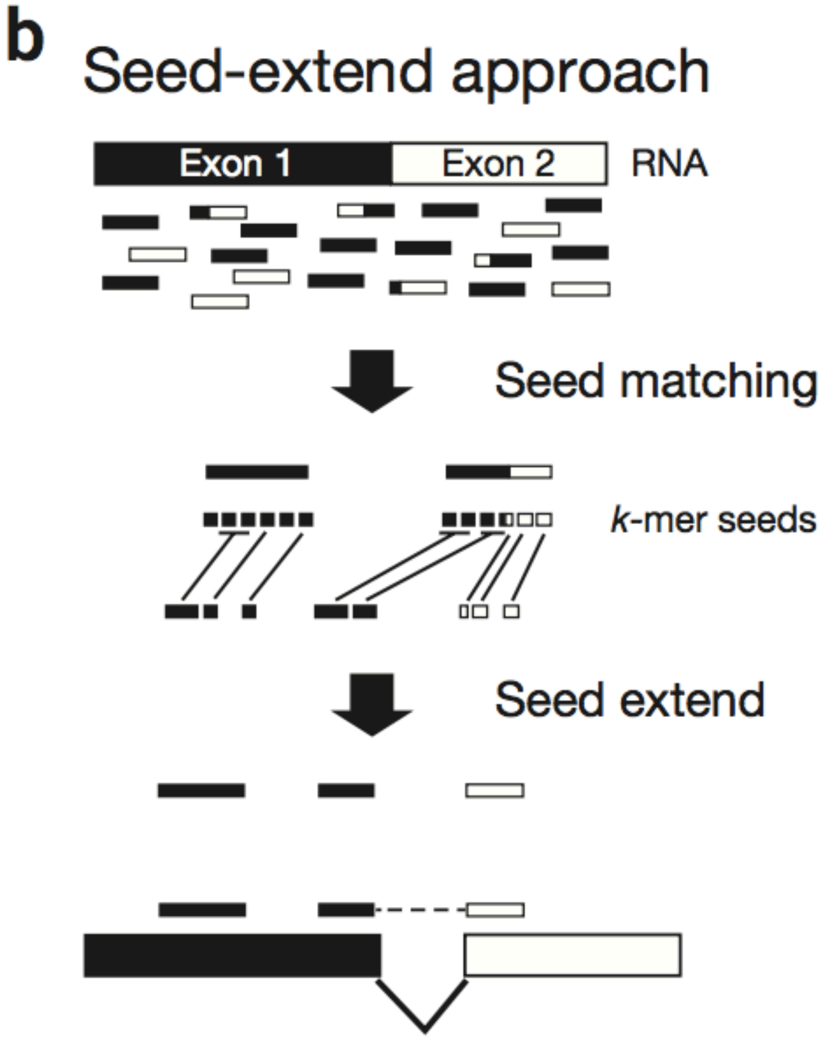
Exon first approach

- First, map reads to the genome contiguously, just like a normal aligner
- For anything that doesn't map, split the read into shorter segments and try to align those
- When a piece aligns, try to find the other piece nearby
- Fast, but biased toward mapping reads to pseudo genes (even if the read would map better to the real gene)



Seed extend approach

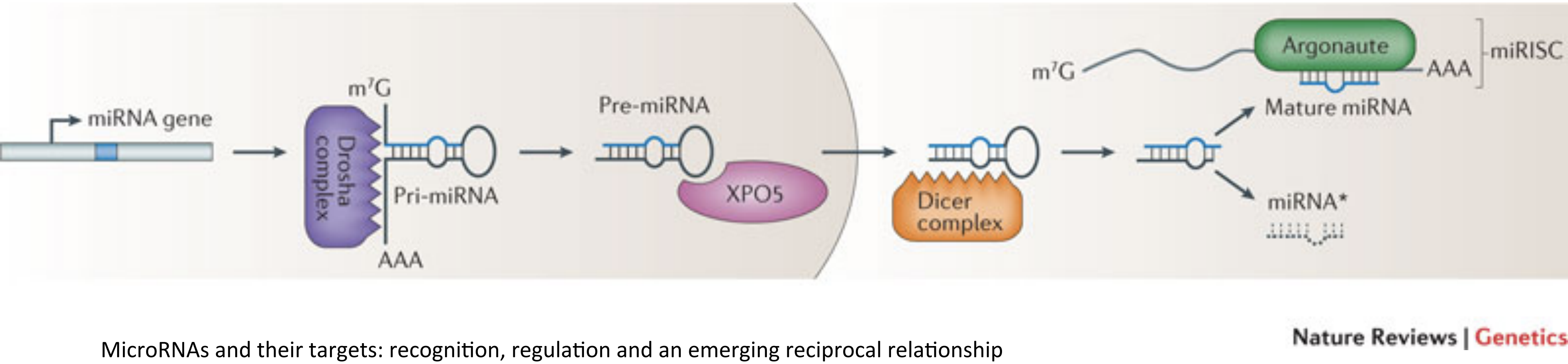
- Break reads into short 'seeds' (e.g. words or kmers)
- Search the genome for the seeds, then localize and optimize the alignment
- Better for polymorphic species
- Usually slower



GSNAP, QPALMA

miRNA Analysis

- Similar to mRNA for differential expression and mapping
- Discovery of novel miRNAs is quite different
  - must examine the RNA structure for folding
  - Look for evidence of the miRNA* (star strand)
  - Conservation of miRNA across species
- Identification of target
  - Find miRNA:mRNA binding pairs
  - Look for conservation across species



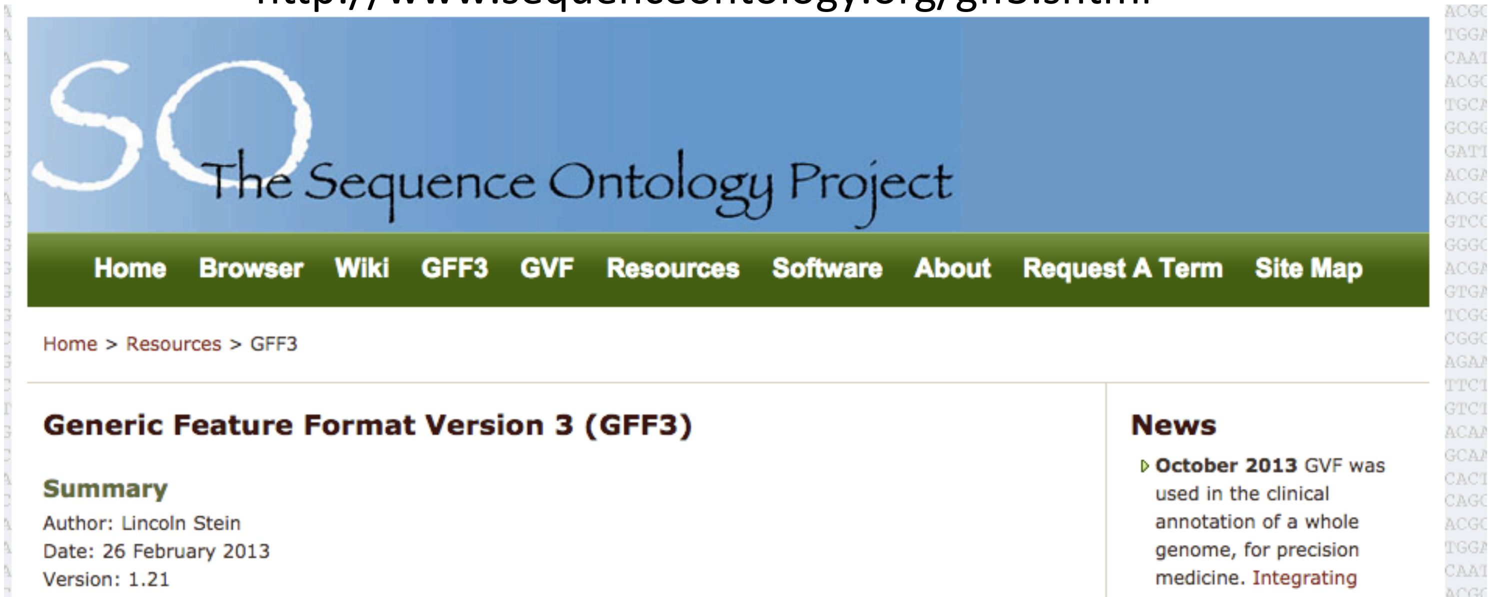
MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship
Amy E. Pasquinelli

GFF3 File Format

GFF- Generic Feature Format

- GFF was the original file format
- Represent genomic features on a sequence
  - gene on a chromosome
- But it did not cover all the use cases needed. Eventually different groups chose to extend it in their own custom ways, and multiple new formats then became common, confusing everyone.

http://www.sequenceontology.org/gff3.shtml



Home > Resources > GFF3

Generic Feature Format Version 3 (GFF3)

Summary

Author: Lincoln Stein
Date: 26 February 2013
Version: 1.21

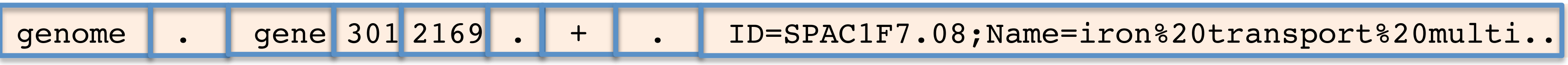
News

- **October 2013** GVF was used in the clinical annotation of a whole genome, for precision medicine. Integrating

GFF3
Generic Feature Format Version 3

- Gff3 format is an attempt to:
  - add and standardize the most common extensions to gff
  - preserve backward compatibility to gff
- Basics:
  - 9 columns
  - Tab delimited
  - Plain text

Backward compatibility - Maintaining compatibility with earlier models or versions of the same product. A new version of a program is said to be backward compatible if it can use files and data created with an older version of the same program.



```
genome  .  gene 301 2169  .  +  .  ID=SPAC1F7.08;Name=iron%20transport%20multi..
```

Column 1: ”seqid”

Column 2: "source”

Column 3: "type”

Columns 4 & 5: "start" and "end”

Column 6: "score”

Column 7: "strand”

Column 8: "phase”

Column 9: "attributes"

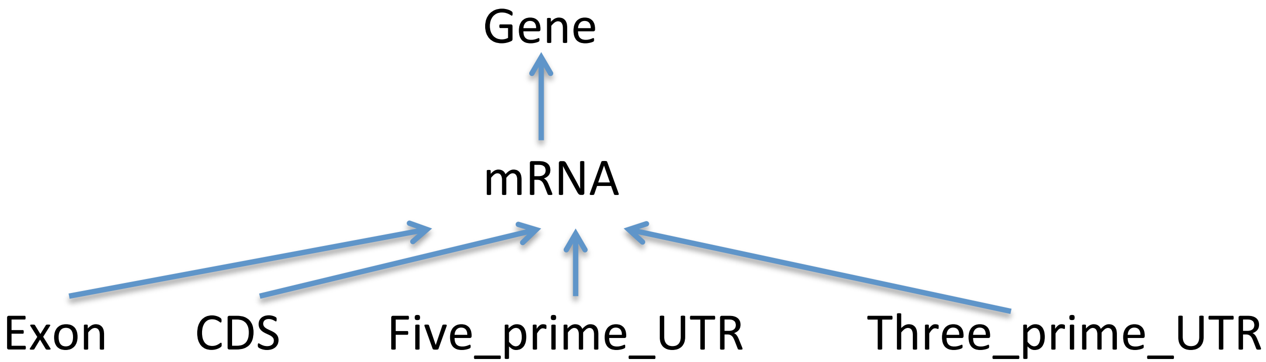
A list of feature attributes in the format tag=value. Multiple tag=value pairs are separated by semicolons

ID= must be unique

```
genome  .  mRNA 301  2169 .  +  .  ID=m.SPAC1F7.08;Parent=SPAC1F7.08;Name=iron…
```

Parent=

Hierarchy of gene pieces



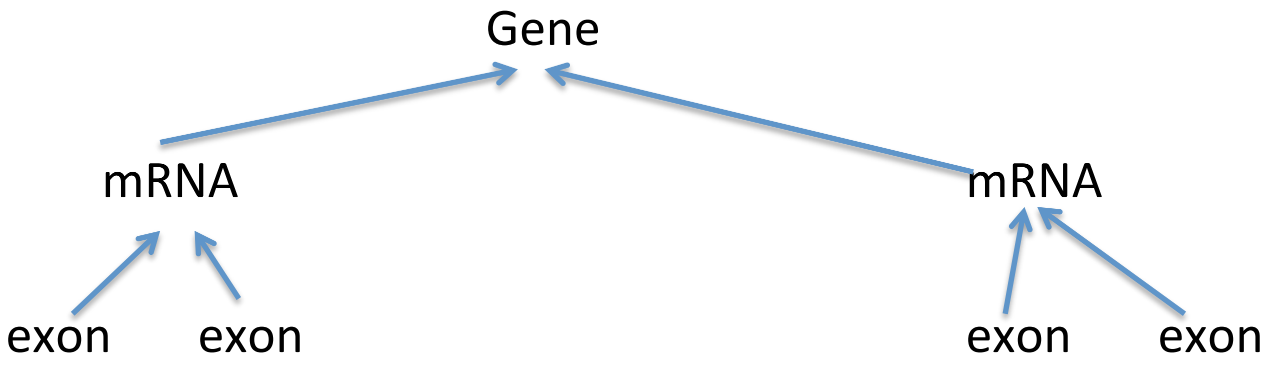
Gene

mRNA

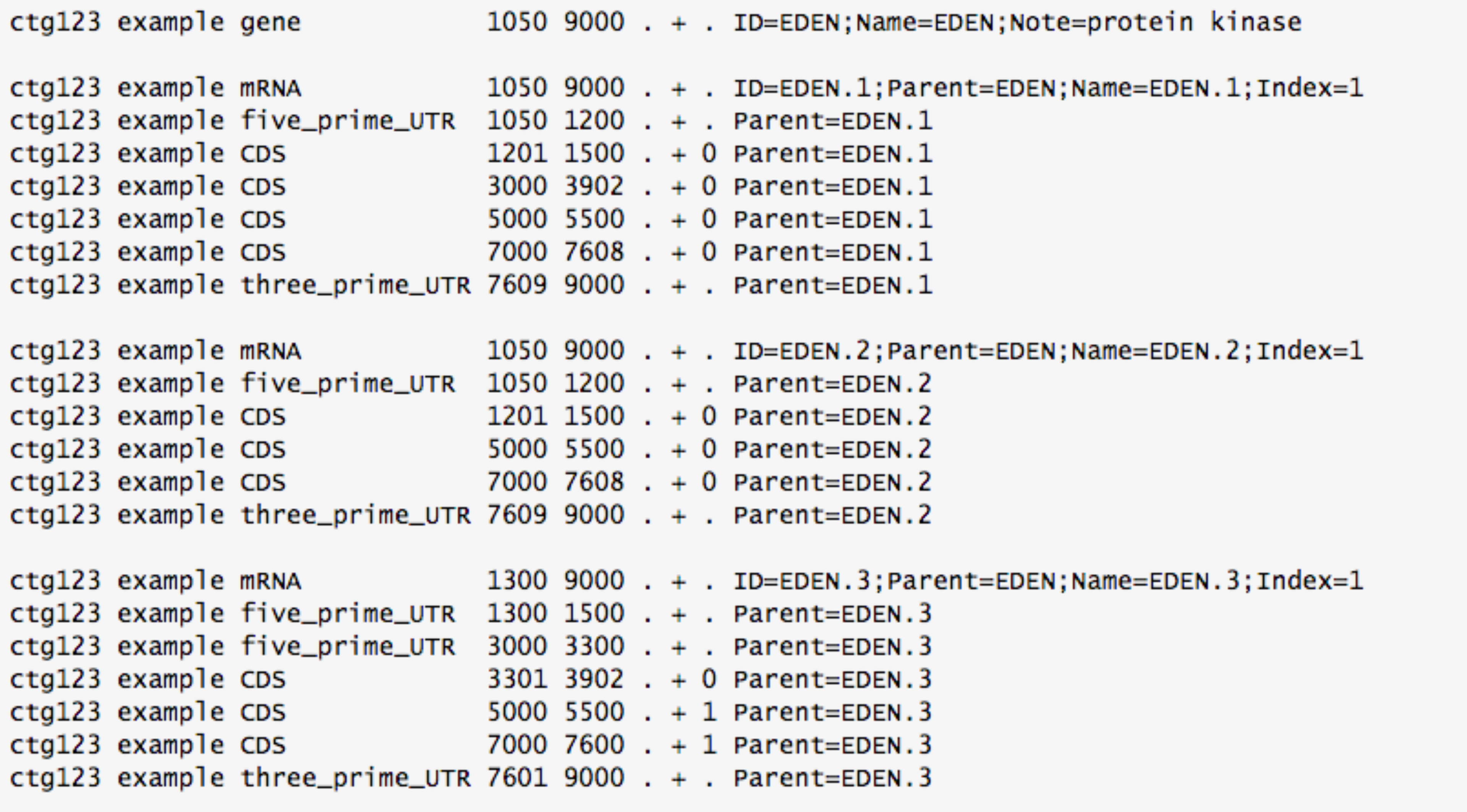
Exon CDS Five_prime_UTR Three_prime_UTR

GFF3
Generic Feature Format Version 3

A feature can have many “children”, allowing for isoforms to be represented as well.



GFF3 – Alternative Isoforms



```
ctg123 example gene            1050 9000 . + . ID=EDEN;Name=EDEN;Note=protein kinase

ctg123 example mRNA            1050 9000 . + . ID=EDEN.1;Parent=EDEN;Name=EDEN.1;Index=1
ctg123 example five_prime_UTR  1050 1200 . + . Parent=EDEN.1
ctg123 example CDS             1201 1500 . + 0 Parent=EDEN.1
ctg123 example CDS             3000 3902 . + 0 Parent=EDEN.1
ctg123 example CDS             5000 5500 . + 0 Parent=EDEN.1
ctg123 example CDS             7000 7608 . + 0 Parent=EDEN.1
ctg123 example three_prime_UTR 7609 9000 . + . Parent=EDEN.1

ctg123 example mRNA            1050 9000 . + . ID=EDEN.2;Parent=EDEN;Name=EDEN.2;Index=1
ctg123 example five_prime_UTR  1050 1200 . + . Parent=EDEN.2
ctg123 example CDS             1201 1500 . + 0 Parent=EDEN.2
ctg123 example CDS             5000 5500 . + 0 Parent=EDEN.2
ctg123 example CDS             7000 7608 . + 0 Parent=EDEN.2
ctg123 example three_prime_UTR 7609 9000 . + . Parent=EDEN.2

ctg123 example mRNA            1300 9000 . + . ID=EDEN.3;Parent=EDEN;Name=EDEN.3;Index=1
ctg123 example five_prime_UTR  1300 1500 . + . Parent=EDEN.3
ctg123 example five_prime_UTR  3000 3300 . + . Parent=EDEN.3
ctg123 example CDS             3301 3902 . + 0 Parent=EDEN.3
ctg123 example CDS             5000 5500 . + 1 Parent=EDEN.3
ctg123 example CDS             7000 7600 . + 1 Parent=EDEN.3
ctg123 example three_prime_UTR 7601 9000 . + . Parent=EDEN.3
```