Genome annotation and short read assembly

Lucile Soler
SciLifeLab RNAseq workshop

November 2018

Based on Manfred Grabherr presentation

1. Introduction to annotation

What is annotation ?

Structural annotation:

Find out where the regions of interest (usually genes) are in the sequence data and what they look like.

VS

functional annotation:

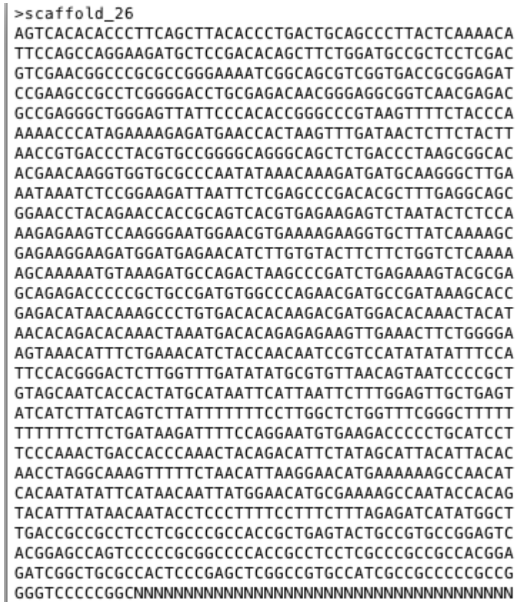
Find out what the regions do. What do they code for?

*It is the **annotation** that bridges the gap from the sequence to the biology of the organism*

Introduction to annotation

From a genome…

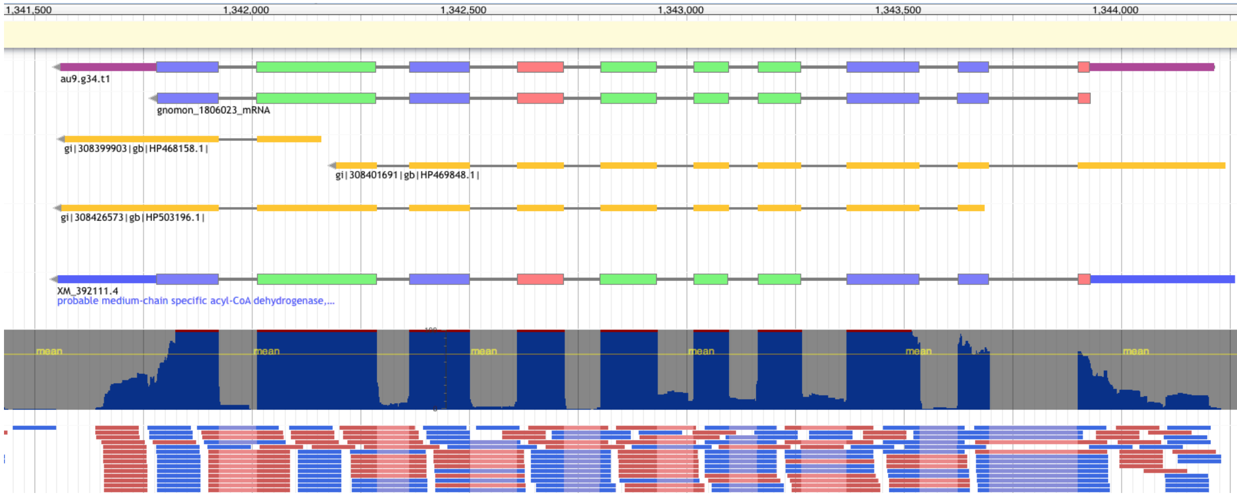
FASTA



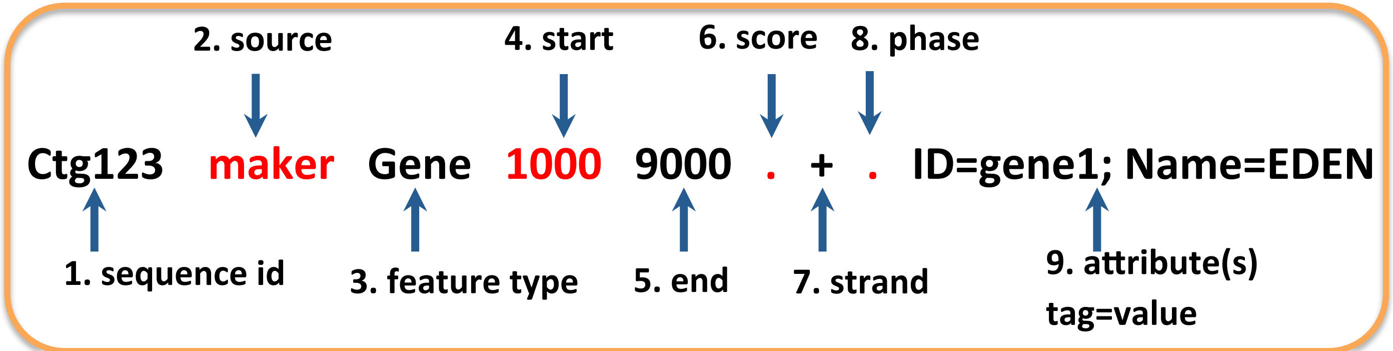
```
>scaffold_26
AGTCACACACCCTTCAGCTTACACCCTGACTGCAGCCCTTACTCAAAACA
TTCCAGCCAGGAAGATGCTCCGACACAGCTTCTGGATGCCGCTCCTCGAC
GTCGAACGGCCCGCGCCGGGAAAATCGGCAGCGTCGGTGACCGCGGAGAT
CCGAAGCCGCCTCGGGGACCTGCGAGACAACGGGAGGCGGTCAACGAGAC
GCCGAGGGCTGGGAGTTATTCCCACACCGGGCCCGTAAGTTTTCTACCCA
AAAACCCATAGAAAAGAGATGAACCACTAAGTTTGATAACTCTTCTACTT
AACCGTGACCCTACGTGCCGGGGCAGGGCAGCTCTGACCCTAAGCGGCAC
ACGAACAAGGTGGTGCGCCCAATATAAACAAAGATGATGCAAGGGCTTGA
AATAAATCTCCGGAAGATTAATTCTCGAGCCCGACACGCTTTGAGGCAGC
GGAACCTACAGAACCACCGCAGTCACGTGAGAAGAGTCTAATACTCTCCA
AAGAGAAGTCCAAGGGAATGGAACGTGAAAAGAAGGTGCTTATCAAAAGC
GAGAAGGAAGATGGATGAGAACATCTTGTGTACTTCTTCTGGTCTCAAAA
AGCAAAAATGTAAAGATGCCAGACTAAGCCCGATCTGAGAAAGTACGCGA
GCAGAGACCCCCGCTGCCGATGTGGCCCAGAACGATGCCGATAAAGCACC
GAGACATAACAAAGCCCTGTGACACACAAGACGATGGACACAAACTACAT
AACACAGACACAAACTAAATGACACAGAGAGAAGTTGAAACTTCTGGGGA
AGTAAACATTTCTGAAACATCTACCAACAATCCGTCCATATATATTTCCA
TTCCACGGGACTCTTGGTTTGATATATGCGTGTTAACAGTAATCCCCGCT
GTAGCAATCACCACTATGCATAATTCATTAATTCTTTGGAGTTGCTGAGT
ATCATCTTATCAGTCTTATTTTTTTCCTTGGCTCTGGTTTCGGGCTTTTT
TTTTTTCTTCTGATAAGATTTTCCAGGAATGTGAAGACCCCCTGCATCCT
TCCCAAACTGACCACCCAAACTACAGACATTCTATAGCATTACATTACAC
AACCTAGGCAAAGTTTTTCTAACATTAAGGAACATGAAAAAAGCCAACAT
CACAATATATTCATAACAATTATGGAACATGCGAAAAGCCAATACCACAG
TACATTTATAACAATACCTCCCTTTTCCTTTCTTTAGAGATCATATGGCT
TGACCGCCGCCTCCTCGCCCGCCACCGCTGAGTACTGCCGTGCCGGAGTC
ACGGAGCCAGTCCCCCGCGGCCCCACCGCCTCCTCGCCCGCCGCCACGGA
GATCGGCTGCGCCACTCCCGAGCTCGGCCGTGCCATCGCCGCCCCCGCCG
GGGTCCCCCGGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

…to an annotated gene

GFF


1,341,500
1,342,000
1,342,500
1,343,000
1,343,500
1,344,000
au9.g34.t1
gnomon_1806023_mRNA
gi|308399903|gb|HP468158.1|
gi|308401691|gb|HP469848.1|
gi|308426573|gb|HP503196.1|
XM_392111.4
probable medium-chain specific acyl-CoA dehydrogenase,...
mean
mean
mean
mean
mean
mean

- 9 columns
- 1 feature = 1 line



2. source 4. start 6. score 8. phase

Ctg123 maker Gene 1000 9000 . + . ID=gene1; Name=EDEN

1. sequence id 3. feature type 5. end 7. strand 9. attribute(s) tag=value

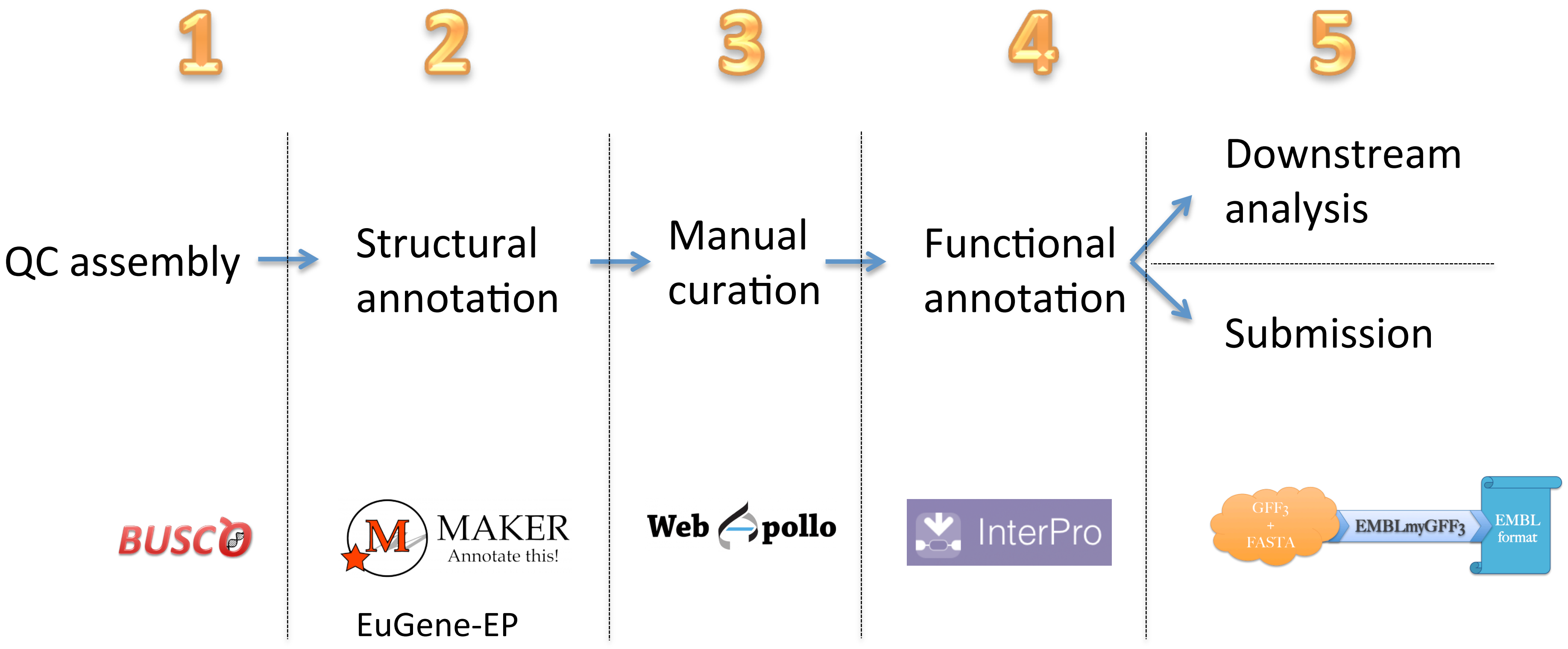
One gene in GFF3 format:

```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . Gene   1000  9000 . + . ID=gene1;Name=EDEN
ctg123 . mRNA  1050  9000 . + . ID=mRNA1;Parent=gene1
ctg123 . exon   1050  1500 . + . ID=exon1;Parent=mRNA1
ctg123 . exon   7000  9000 . + . ID=exon2;Parent=mRNA1
ctg123 . CDS   1201  1500 . + 0 ID=cds1;Parent=mRNA1;Name=edenprotein.1
ctg123 . CDS   7000  7600 . + 0 ID=cds1;Parent=mRNA1;Name=edenprotein.1
```

/!\ different version 1, 2, 2.5, 3

GTF = GFF version 2

The main steps in genome annotation

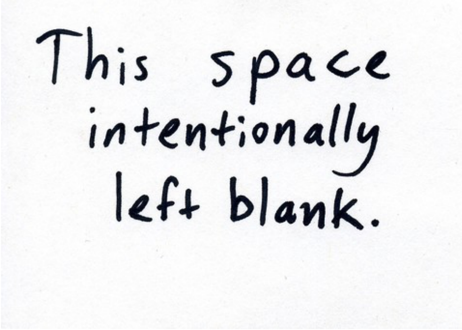


1 QC assembly → 2 Structural annotation → 3 Manual curation → 4 Functional annotation → 5 Downstream analysis / Submission

EuGene-EP

Types of external data used

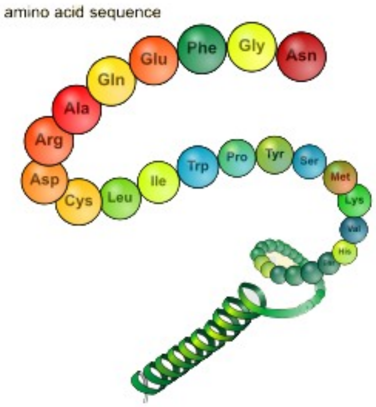
Ø



This space intentionally left blank.

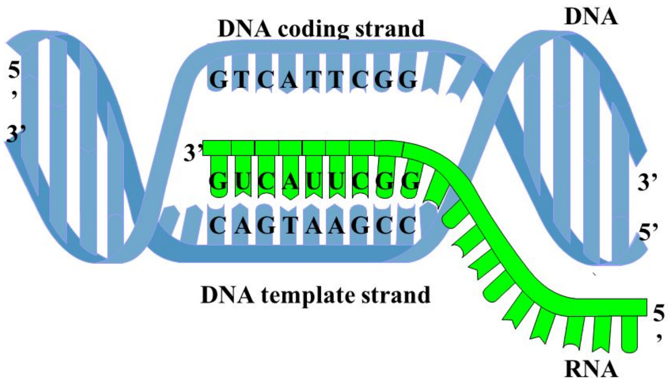
Proteins

- Known amino acid sequences from other organisms



Transcripts

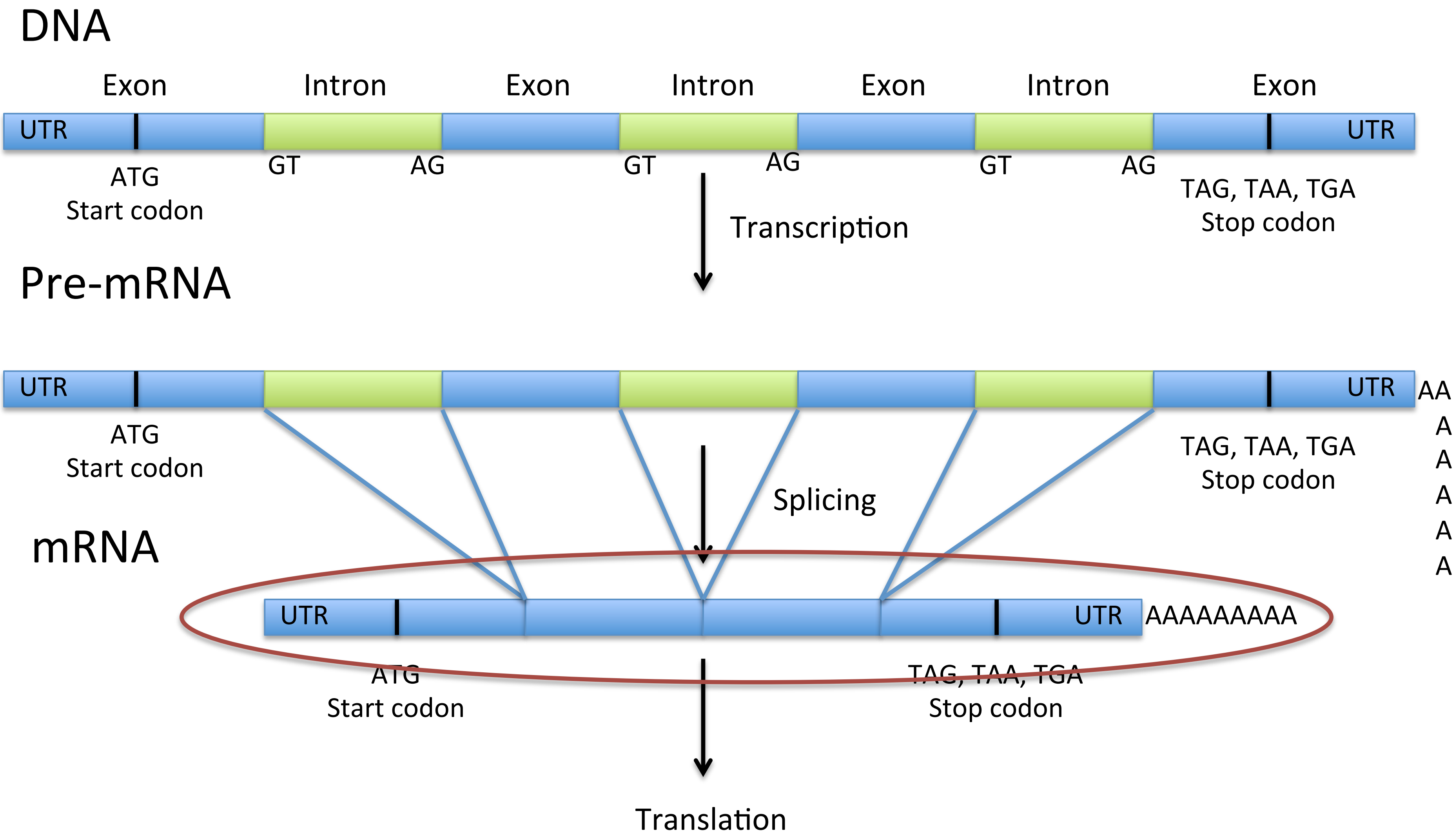
- Assembled from RNA-seq or downloaded ESTs



Types of data used: RNA-seq

- Should always be included in an annotation project
- From the same organism as the genomic data => unbiased
- /!\ Can be very noisy (tissue/species dependent), can include pre-mRNA
- Sample different tissues or life stages if possible
- Avoid gonads; muscle or liver is good

Types of data used: RNA-seq



2. Assembly of transcripts

RNA-seq

RNA-seq (short-reads) need to be assembled first

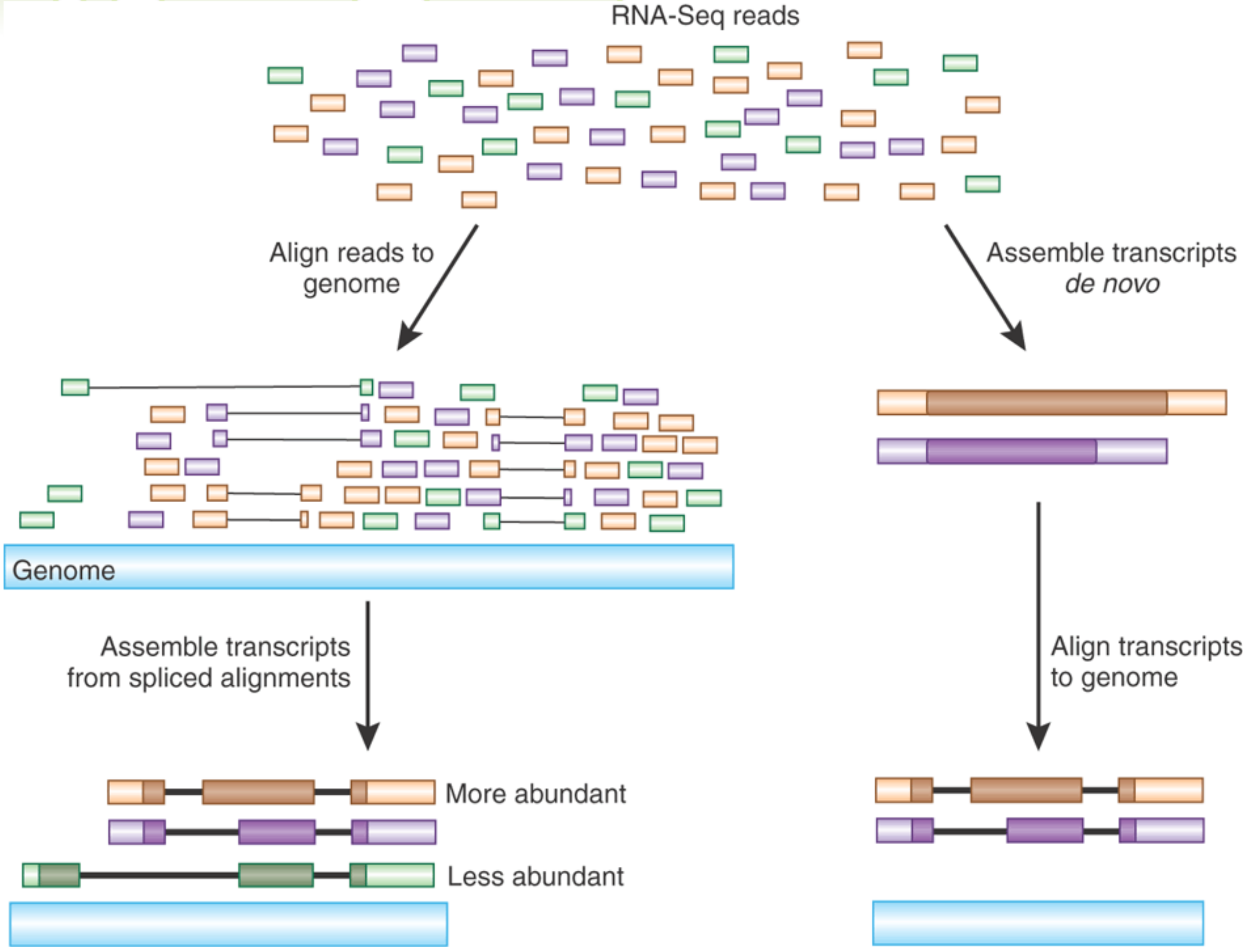
- Genome guided assembly

=> Cufflinks/Stringtie/…: mapped reads -> transcripts

- *De novo*

=> Trinity: assembles transcripts without a genome

Assembly of transcript

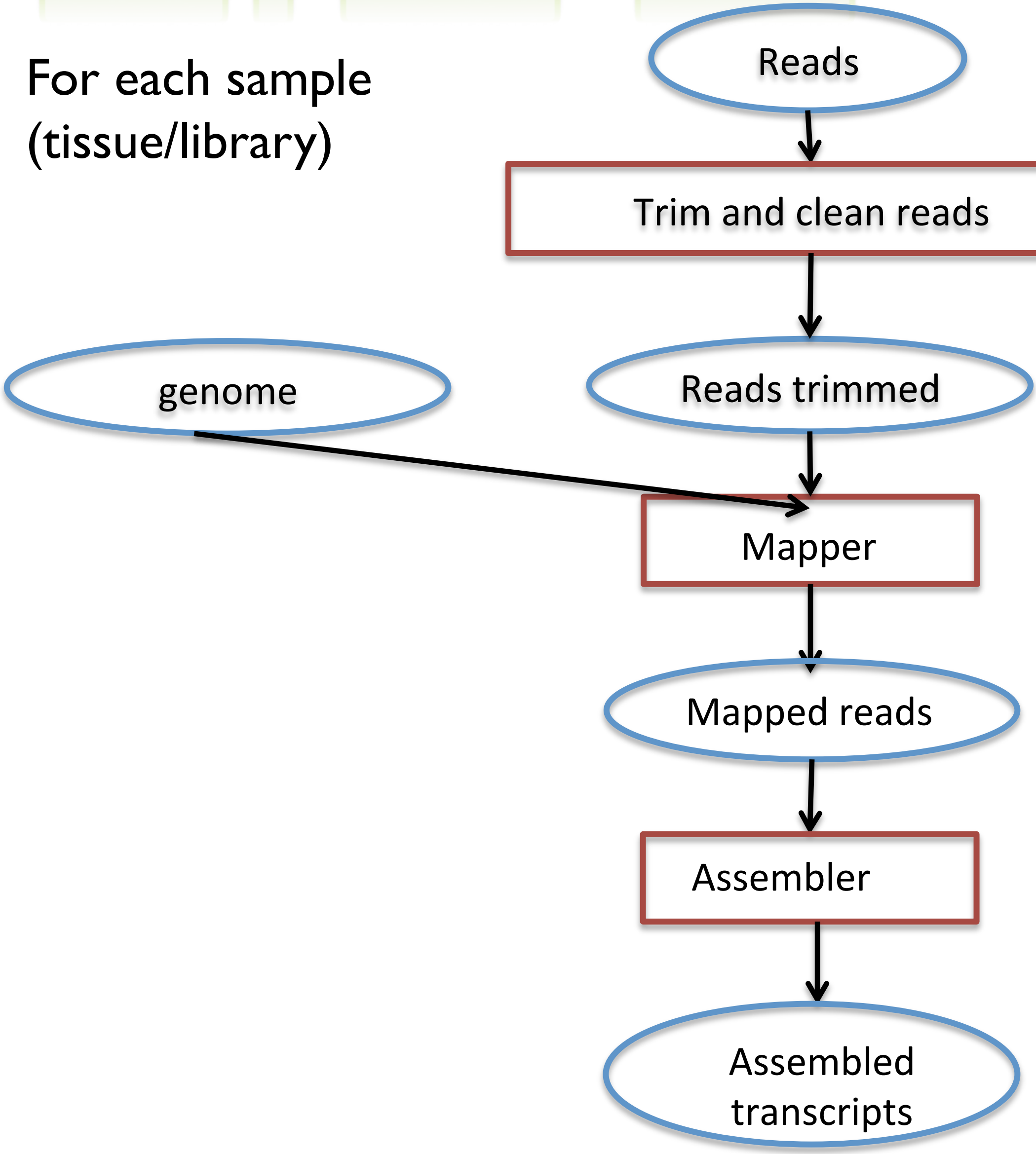

RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Haas and Zody, Nature Biotechnology 28, 421–423 (2010)

Genome guided transcriptome assembly

Genome guided transcriptome assembly

For each sample (tissue/library)



E.g : trimmomatic

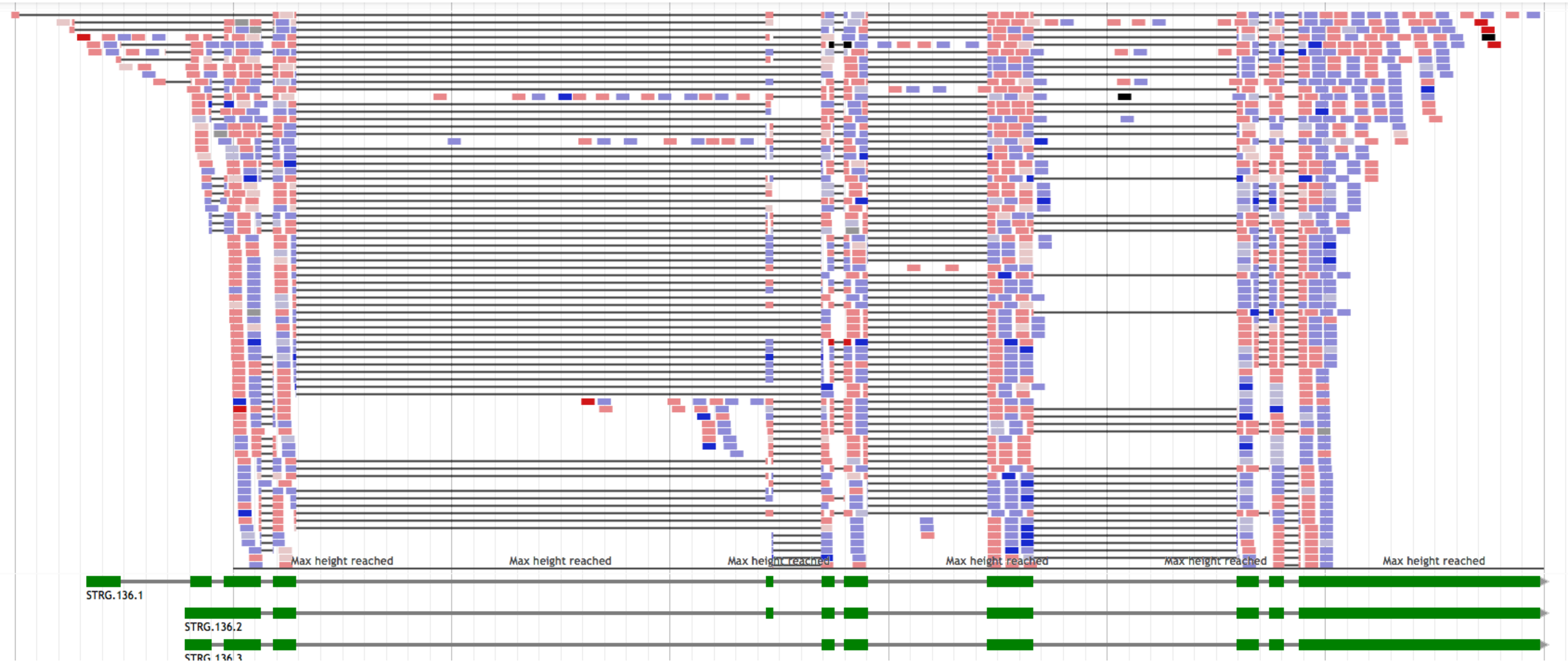
E.g : hisat2, star

E.g : stringtie, cufflinks

Genome-guided transcriptome assembly

- Need a very good reference (genome most of the time)
- Can use existing annotation (GTF/GFF file) (in option for stringtie)
- Can detect novel transcripts

RNA-seq - Spliced reads



De-novo transcriptome assembly

De-novo transcriptome assembly

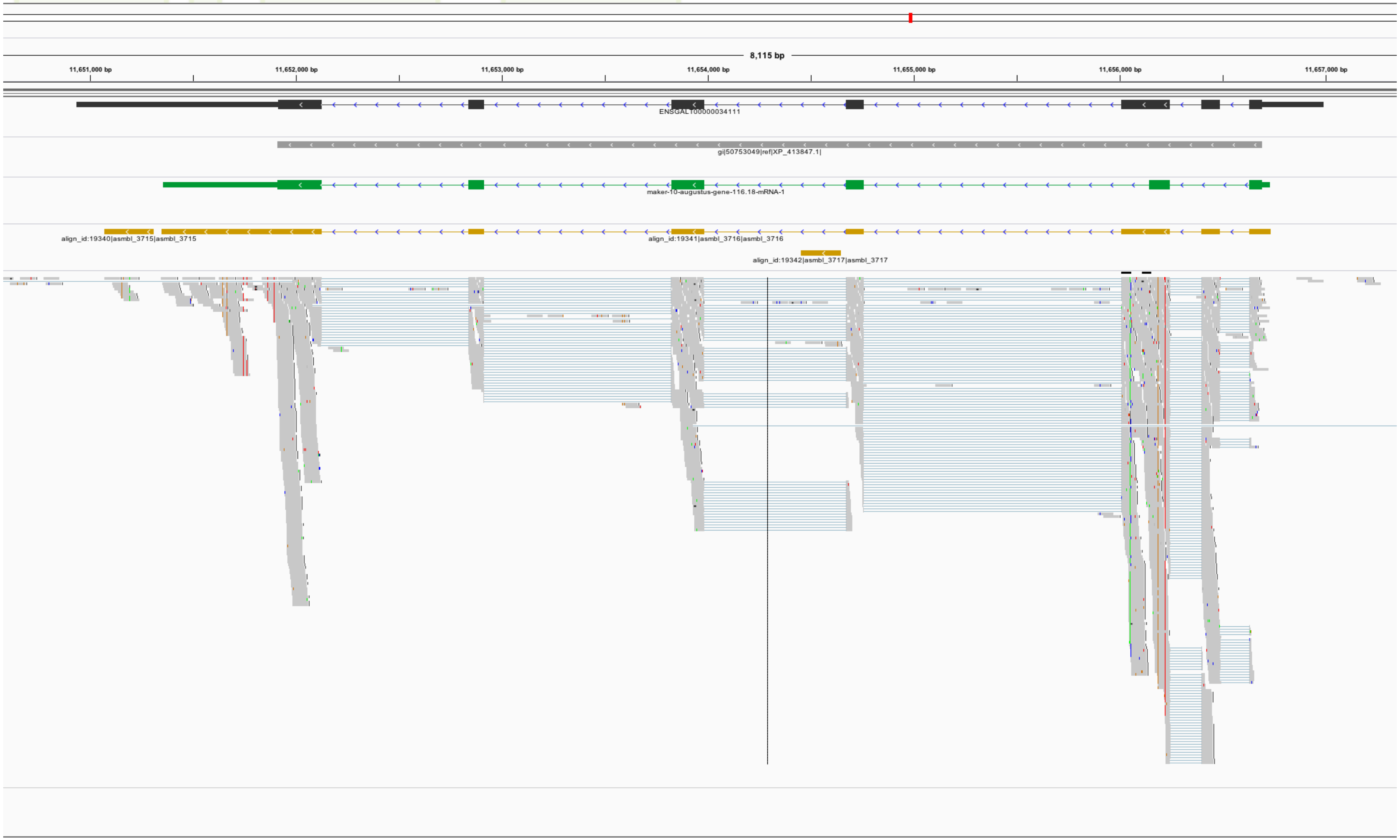
- Most used programs (latest release date):
  - Trinity (Sept 2018)
  - SOAPdenovo-Trans (July 2013)
  - Trans-ABySS (Feb 2018)
  - Velvet+Oases (March 2015)
- Originally SOAPdenovo, ABySS and Velvet for de novo genome assembly
- "SOAPdenovo-Trans incorporates the error-removal model from Trinity and the robust heuristic graph traversal method from Oases."

Trinity, Grabherr et al. 2011

De-novo transcriptome assembly

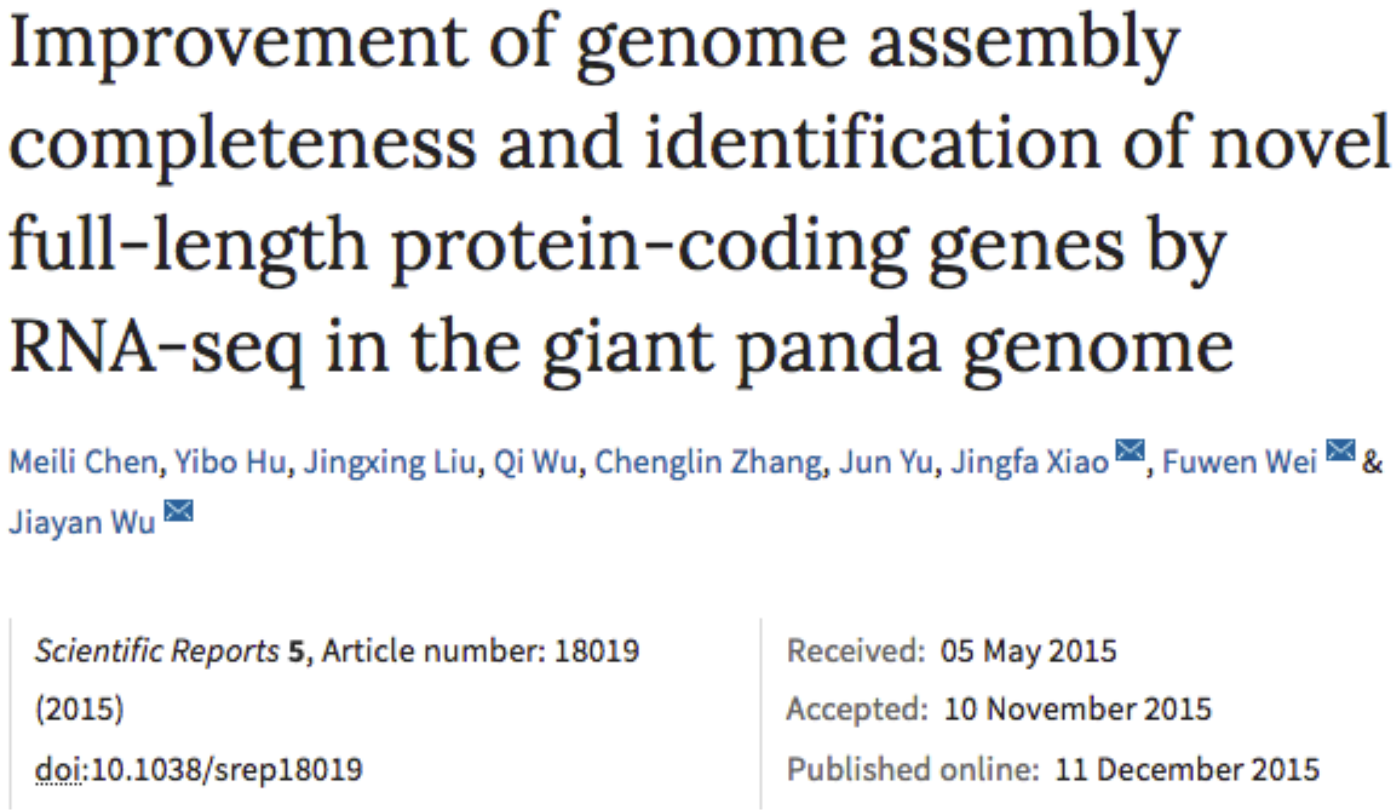
- No reference needed
- Many programs available
- Lots of potential transcripts. Filter!

Mapped Trinity-assembled transcripts


8,115 bp
11,651,000 bp
11,652,000 bp
11,653,000 bp
11,654,000 bp
11,655,000 bp
11,656,000 bp
11,657,000 bp
ENSGALT00000034111
gi|50753049|ref|XP_413847.1|
maker-10-augustus-gene-116.18-mRNA-1
align_id:19340|asmbl_3715|asmbl_3715
align_id:19341|asmbl_3716|asmbl_3716
align_id:19342|asmbl_3717|asmbl_3717

Combining method

Combining both method



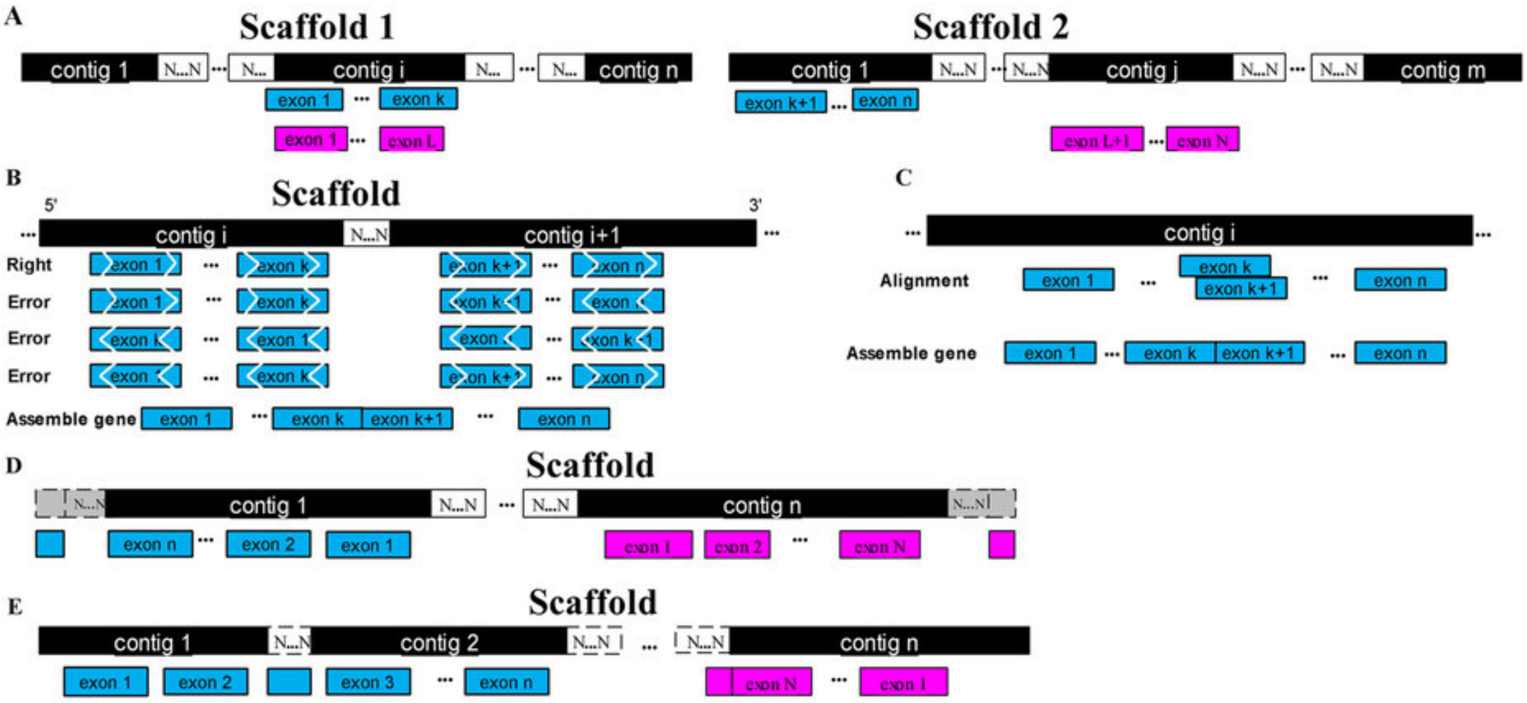
Improvement of genome assembly completeness and identification of novel full-length protein-coding genes by RNA-seq in the giant panda genome

Meili Chen, Yibo Hu, Jingxing Liu, Qi Wu, Chenglin Zhang, Jun Yu, Jingfa Xiao ✉, Fuwen Wei ✉ & Jiayan Wu ✉

Scientific Reports **5**, Article number: 18019 (2015)
doi:10.1038/srep18019

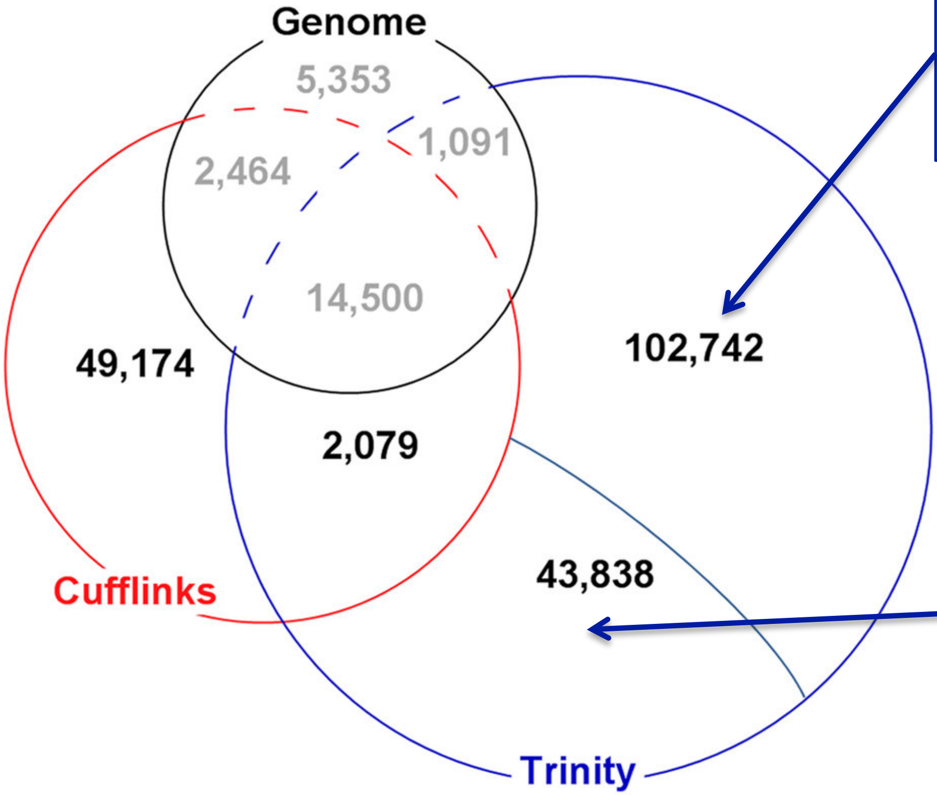
Received: 05 May 2015
Accepted: 10 November 2015
Published online: 11 December 2015

Improvement of genome assembly



(A) Scaffolding improvement; (B) Scaffolding inconsistencies; (C) Nest assembly errors; (D) Boundary extensions; (E) Gap closure

Transcriptome reconstruction



Transcripts located to scaffolds that did not cover any known gene models

Transcripts unaligned back to the giant panda draft genome

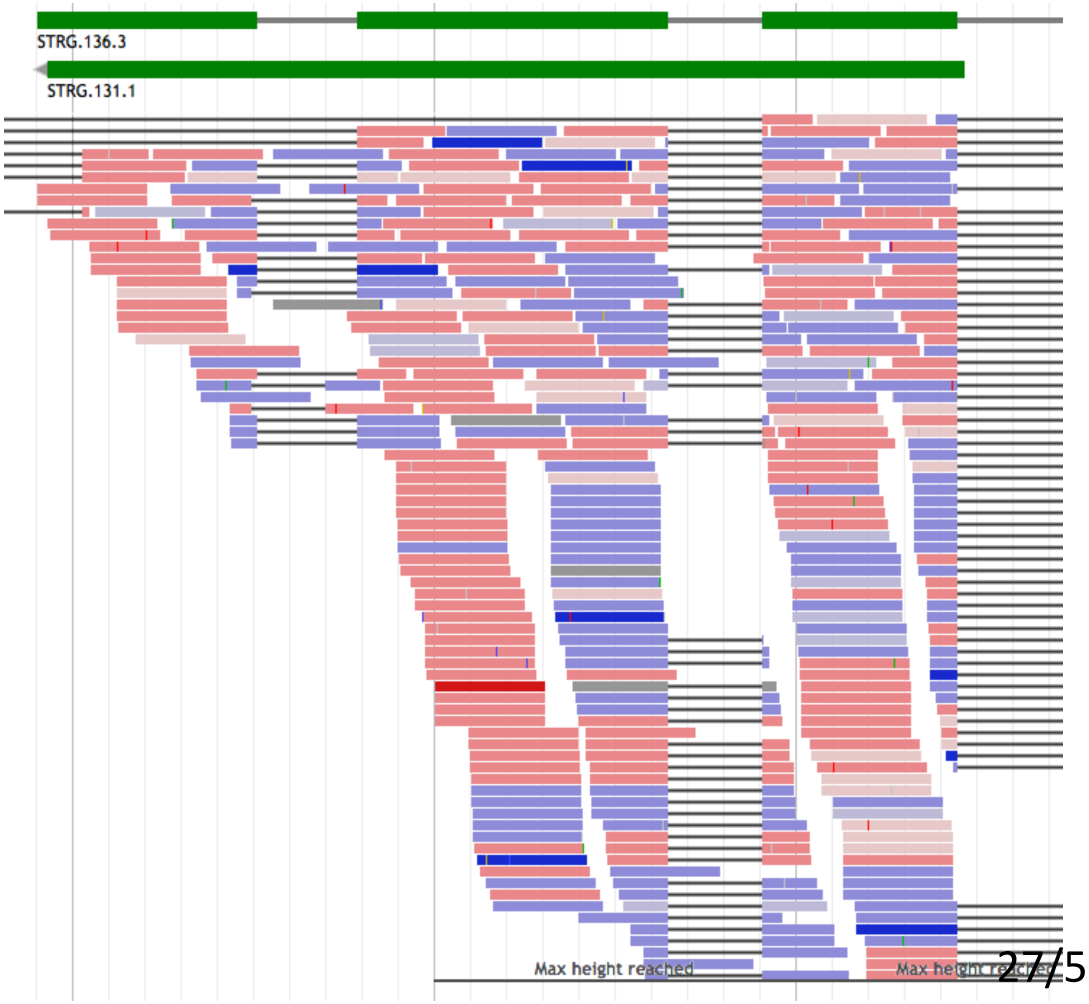
49,174 + 2,079 + 43,838 + 102,742 = 197,833 potential novel transcripts!

Combining both method

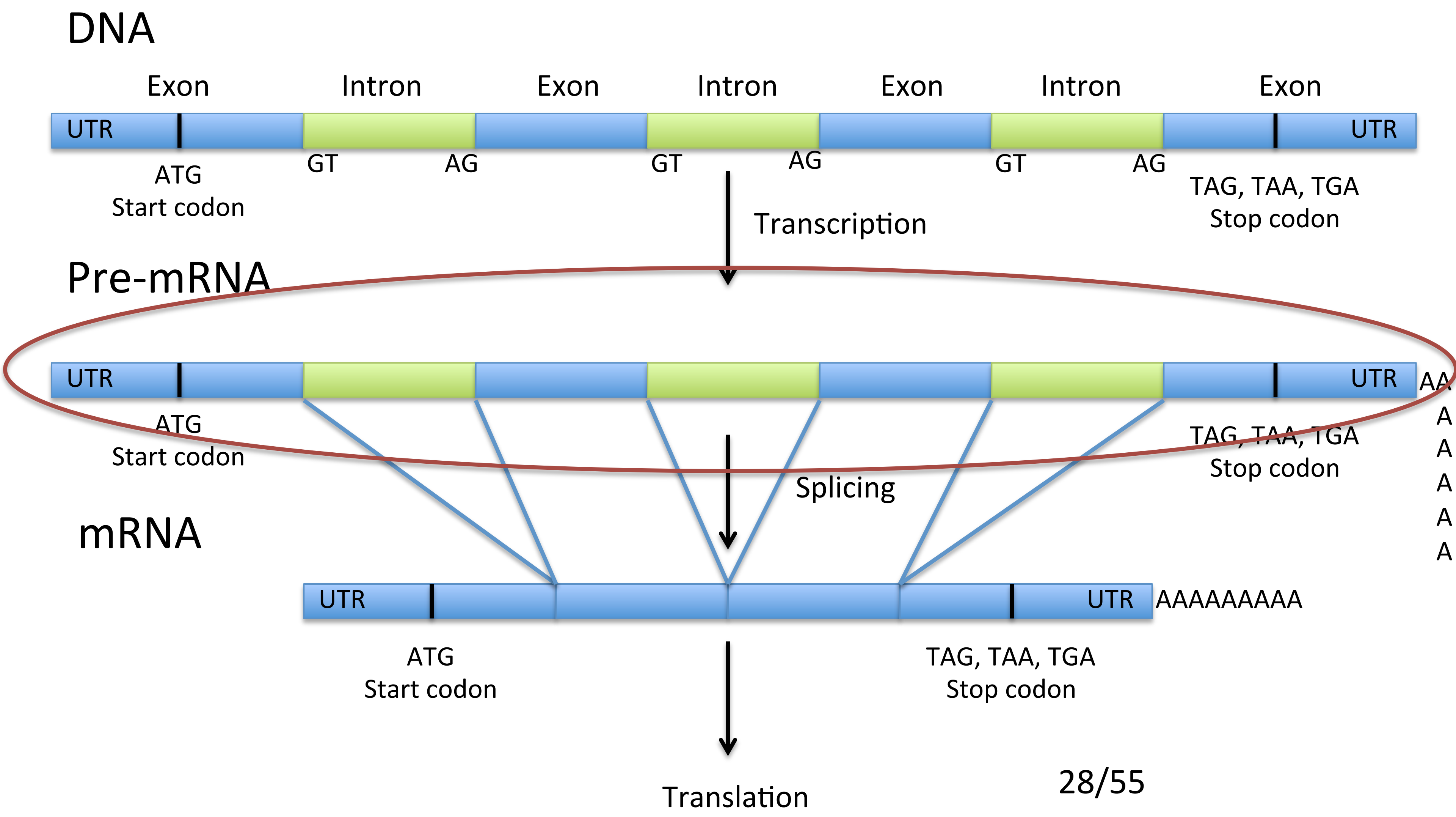
- Useful if the reference is incomplete
- Can help improving the reference
- Can help annotating the reference
- Need to filter the results!

3. How does it look when it does not look good?

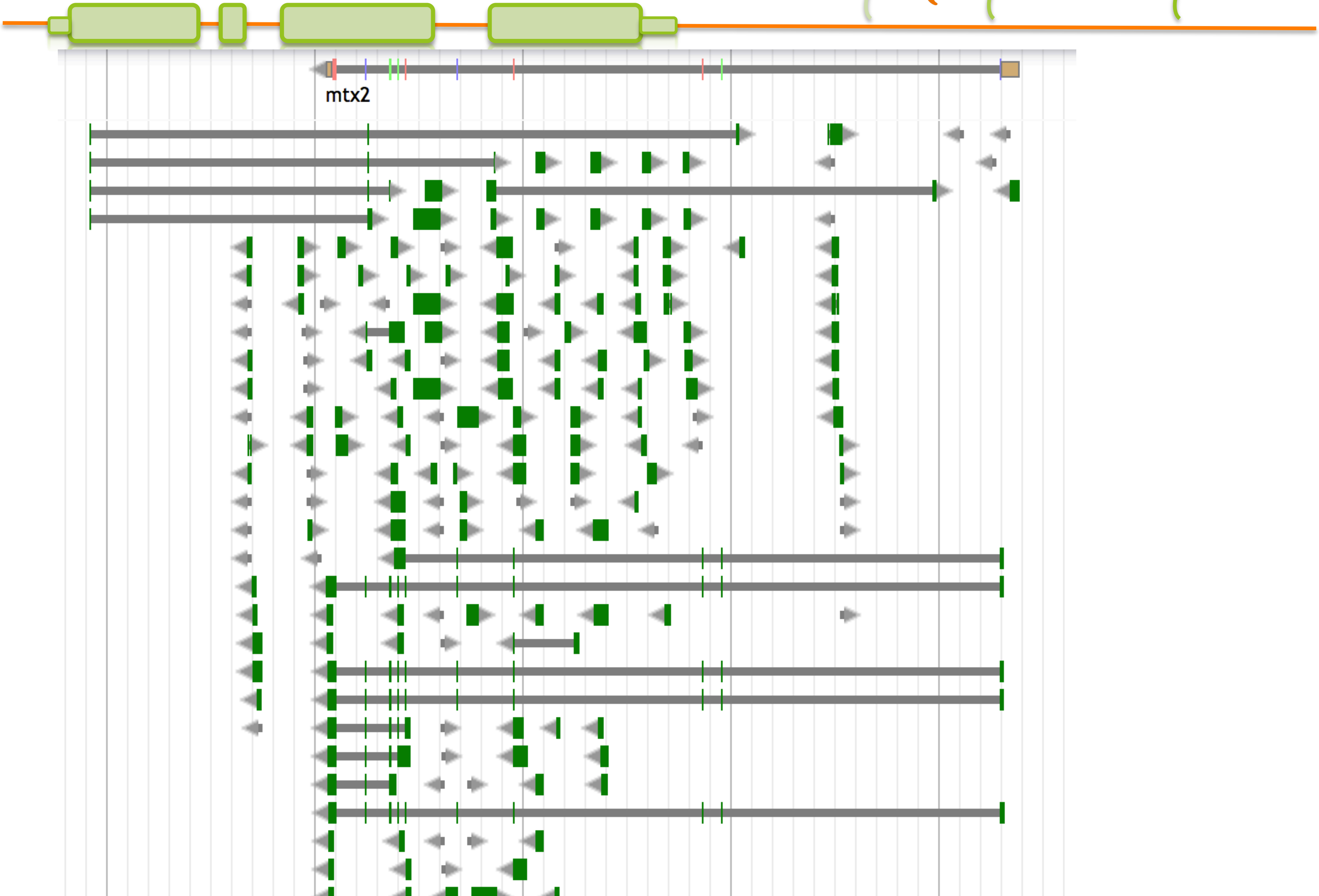
RNA-seq – pre-mRNA noise



Types of data used: RNA-seq



Trinity noise



4. Conclusion/summary

Conclusion/summary

- RNAseq data should always be included in an annotation project
- From the same organism as the genomic data => unbiased
- Can be used before annotation or after to improve an annotation already existing
- Sample different tissues or life stages if possible
- Avoid gonads; muscle or liver is good

- /!\ Can be very noisy (tissue/species dependent), can include pre-mRNA

- Combining method is best if possible

THE END