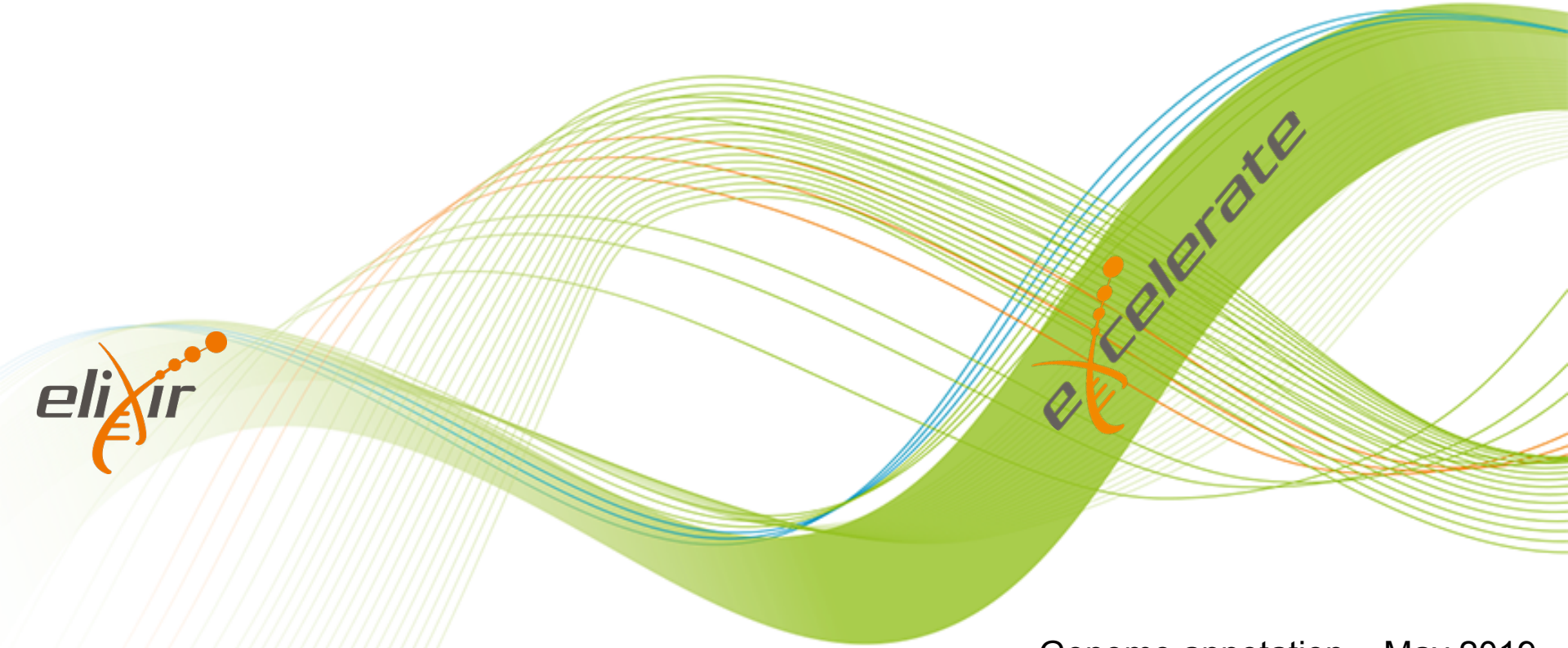
SciLifeLab

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Henrik Lantz

Introduction to genome annotation



SciLifeLab vision

To be an internationally leading center that develops, uses and provides access to advanced technologies for molecular biosciences with focus on health and environment.

Solna

Uppsala

National organisation

SciLifeLab Platforms

SciLifeLab

Bioinformatics

- Compute and Storage U
- Long-term Support G, Li, Lu, S, U, Um
- Support and Infrastructure G, Li, Lu, S, U, Um
- Systems Biology G

Cellular and Molecular Imaging

- Advanced Light Microscopy S
- BioImage Informatics U
- Cell Profiling S
- Cryo-EM S, Um
- Protein Science Facility S
- Swedish NMR Centre G, Um

Chemical Biology and Genome Engineering

- Chemical Biology Consortium Sweden S, Um
- Genome Engineering Zebrafish U
- High Throughput Genome Engineering S

Diagnostics Development

- Clinical Genomics Göteborg G
- Clinical Genomics Lund Lu
- Clinical Genomics Stockholm S
- Clinical Genomics Uppsala U

Drug Discovery and Development

- ADME (Absorption, Distribution, Metabolism, Excretion) of Therapeutics U
- Biochemical and Cellular Assay S
- Biophysical Screening and Characterization U
- Human Antibody Therapeutics Lu, S
- In Vitro and Systems Pharmacology U
- Medicinal Chemistry – Hit2Lead S
- Medicinal Chemistry – Lead Identification U
- Protein Expression and Characterization S

Genomics

- National Genomics Infrastructure S, U
- Ancient DNA U

Proteomics and Metabolomics

- Autoimmunity Profiling S
- Chemical Proteomics and Proteogenomics S
- Clinical Biomarkers U
- PLA Proteomics U
- Plasma Profiling S
- Swedish Metabolomics Centre Um

Single Cell Biology

- Eukaryotic Single Cell Genomics S
- Mass Cytometry Li, S
- Microbial Single Cell Genomics U
- Single Cell Proteomics U

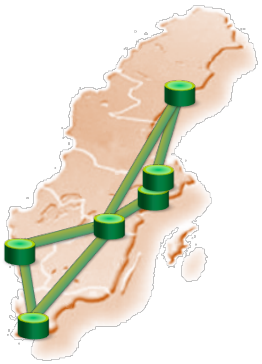
G - Göteborg
Li - Linköping
Lu - Lund
S - Stockholm
U - Uppsala
Um - Umeå

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

4 facilities, 80 FTEs

- **Short-term support and infrastructure**
  Wide competence in bioinformatics, Assembly/Annotation, SysDev
- **Long-term support**
  Large collaborative projects selected by scientific ranking
- **Systems biology**
  Network analyses and Integrative bioinformatics
- **Compute and storage**
  Computational and storage resources for bioinformatics, especially next-generation sequencing



Support, Infrastructure and Training

400 consultations

Future compute infrastructure

800 compute projects
200 software and databases

35 training events
500 PhD/post-docs

Data publishing and open science
Secure sharing of sensitive data

Efficient tools and
workflows

200 research projects

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:

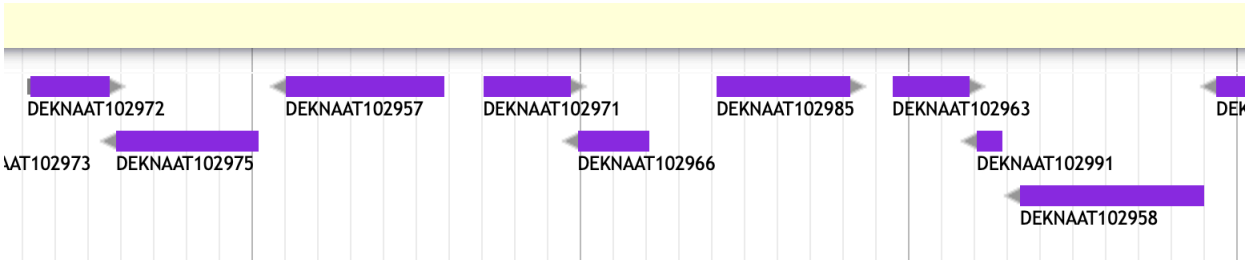
Find out what the regions do. What do they code for?

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

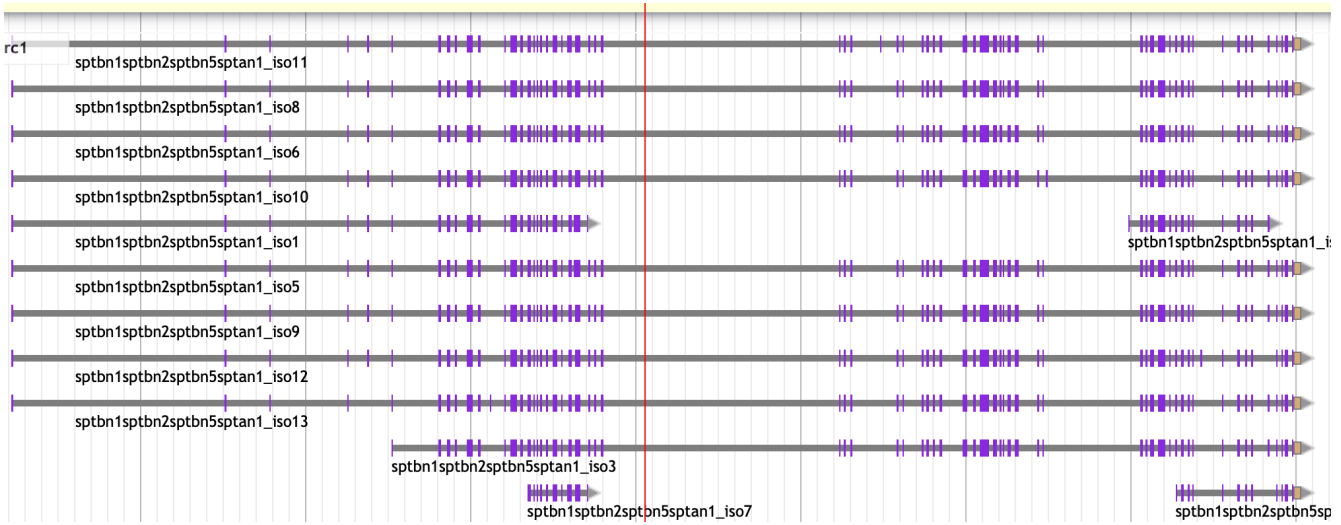
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Organisms differ in genomic complexity

A yeast



A crustacean



Zoomed in

Introduction to annotation

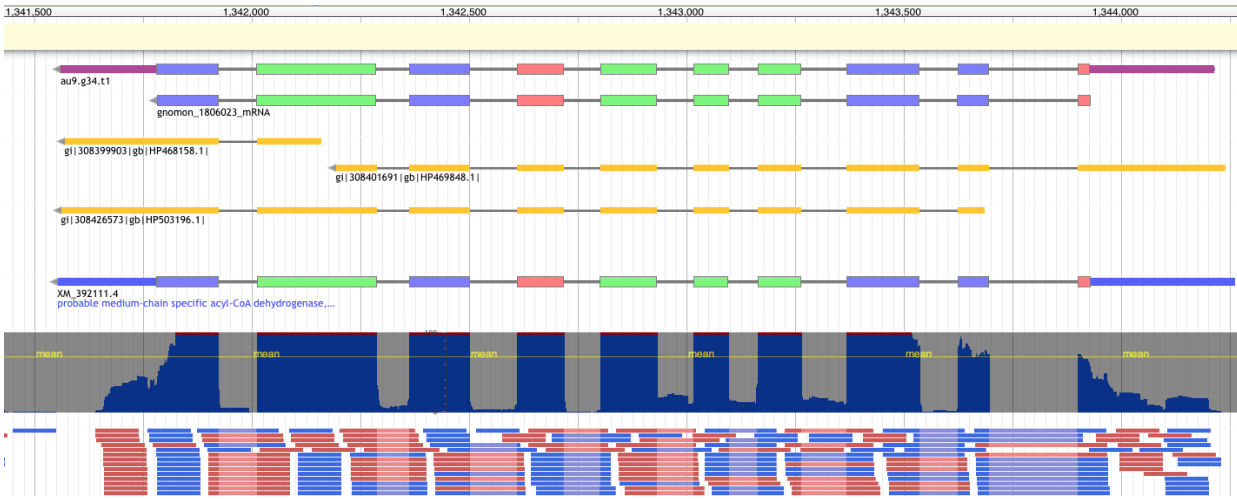
From a genome…

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
GGGTCCCCCGGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

…to an annotated gene

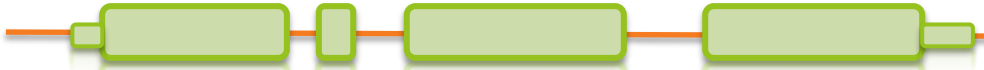
GTF/GFF



- 9 columns
- 1 feature = 1 line

1. sequence id	2. source	3. feature type	4. start	5. end	6. score	7. strand	8. phase	9. attribute(s) tag=value
Ctg123	maker	Gene	1000	9000	.	+	.	ID=gene1; Name=EDEN

Introduction to annotation: GFF3



```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
```

Header

- 9 columns
- 1 feature = 1 line

```
##gff-version 3
scaffold_7      maker   gene    133848  144662  .       -       .       ID=C55462A8A38E2878A71E2ABDAFB34661;Name=maker-scaffold_7-augustus-gene-0.11
scaffold_7      maker   mRNA    133848  144662  .       -       .       ID=A649E923246BADE2184E579FA9124ABD;Parent=C55462A8A38E2878A71E2ABDAFB34661;Name=1:cornix-all_
scaffold_7      maker   exon    138974  139077  .       -       .       ID=A649E923246BADE2184E579FA9124ABD:exon:7;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   exon    135098  135281  .       -       .       ID=A649E923246BADE2184E579FA9124ABD:exon:6;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   exon    139616  139836  .       -       .       ID=A649E923246BADE2184E579FA9124ABD:exon:5;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   exon    144511  144662  .       -       .       ID=A649E923246BADE2184E579FA9124ABD:exon:4;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   exon    136342  136437  .       -       .       ID=A649E923246BADE2184E579FA9124ABD:exon:3;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   exon    133848  134338  .       -       .       ID=A649E923246BADE2184E579FA9124ABD:exon:2;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   exon    141262  141383  .       -       .       ID=A649E923246BADE2184E579FA9124ABD:exon:1;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   exon    144138  144296  .       -       .       ID=A649E923246BADE2184E579FA9124ABD:exon:0;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   five_prime_UTR  144592  144662  .       -       .       ID=A649E923246BADE2184E579FA9124ABD:five_prime_utr;Parent=A649E923246BADE2184E579FA91
scaffold_7      maker   CDS     144511  144591  .       -       0       ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   CDS     144138  144296  .       -       0       ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   CDS     141262  141383  .       -       0       ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   CDS     139616  139836  .       -       1       ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   CDS     138974  139077  .       -       2       ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   CDS     136342  136437  .       -       0       ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   CDS     135098  135281  .       -       0       ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   CDS     134262  134338  .       -       2       ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7      maker   three_prime_UTR 133848  134261  .       -       .       ID=A649E923246BADE2184E579FA9124ABD:three_prime_utr;Parent=A649E923246BADE2184E579FA91
scaffold_7      maker   gene    83101   117593  .       +       .       ID=D3BD9A5F27797F56A844A2E890FF6B99;Name=maker-scaffold_7-augustus-gene-0.6
scaffold_7      maker   mRNA    83101   117593  .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26;Parent=D3BD9A5F27797F56A844A2E890FF6B99;Name=maker-scaffol
scaffold_7      maker   exon    95748   95871   .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26:exon:8;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7      maker   exon    99113   99137   .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26:exon:9;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7      maker   exon    90664   90748   .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26:exon:10;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7      maker   exon    110231  110356  .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26:exon:11;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7      maker   exon    113609  113679  .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26:exon:12;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7      maker   exon    94057   94117   .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26:exon:13;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7      maker   exon    84578   84670   .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26:exon:14;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7      maker   exon    115452  115536  .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26:exon:15;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7      maker   exon    111579  111669  .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26:exon:16;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7      maker   exon    102917  103016  .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26:exon:17;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7      maker   exon    96766   96849   .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26:exon:18;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7      maker   exon    86666   86750   .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26:exon:19;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7      maker   exon    99944   100109  .       +       .       ID=CFF5DDA190832937C45A0D2E674C9C26:exon:20;Parent=CFF5DDA190832937C45A0D2E674C9C26
```

1) sequence id

2) source

3) feature type

(SO term = 2278 possibilities)

4) start

5) end

6) score

7) strand

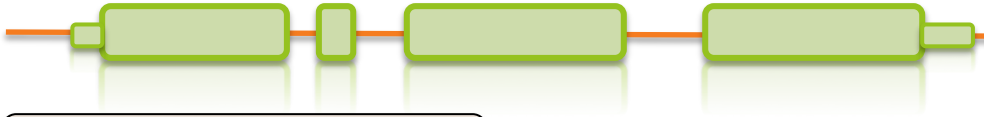
8) phase

9) attributes

tag=value

! Features are grouped by **parent** relationship

Introduction to annotation: GTF2.X



- 9 columns
- 1 feature = 1 line

```
#!genome-build GRCz11
#!genome-date 2017-05
```

Header

Ctg123	.	Gene	1000	9000	.	+	.	gene_id gene1; name EDEN;
ctg123	.	Transcript	1050	9000	.	+	.	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	Transcript	1050	9000	.	+	.	gene_id gene1; transcript_id=t2; name EDEN;
ctg123	.	exon	1300	1500	.	+	.	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	exon	1050	1500	.	+	.	gene_id gene1; transcript_id=t1; name EDEN;
tg123	.	exon	1050	1500	.	+	.	gene_id gene1; transcript_id=t2; name EDEN;
ctg123	.	exon	3000	3902	.	+	.	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	exon	5000	5500	.	+	.	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	exon	5000	5500	.	+	.	gene_id gene1; transcript_id=t2; name EDEN;
ctg123	.	exon	7000	9000	.	+	.	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	exon	7000	9000	.	+	.	gene_id gene1; transcript_id=t2; name EDEN;
ctg123	.	CDS	1201	1500	.	+	0	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	CDS	3000	3902	.	+	0	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	CDS	5000	5500	.	+	0	gene_id gene1; transcript_id=t1; name EDEN;
ctg123	.	CDS	7000	7600	.	+	0	gene_id gene1; transcript_id=t1; name EDEN;
Ctg123	.	CDS	1201	1500	.	+	0	gene_id gene1; transcript_id=t2; name EDEN;
ctg123	.	CDS	5000	5500	.	+	0	gene_id gene1; transcript_id=t2; name EDEN;
Ctg123	.	CDS	7000	7600	.	+	0	gene_id gene1; transcript_id=t2; name EDEN;

1) sequence id

2) source

3) feature type
(9 possibilities)

4) start

5) end

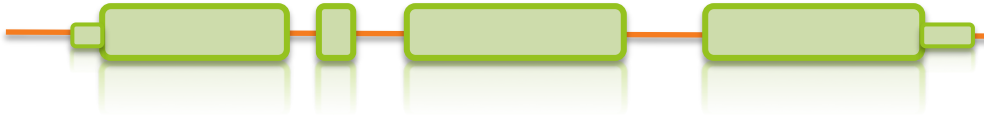
6) score

7) strand

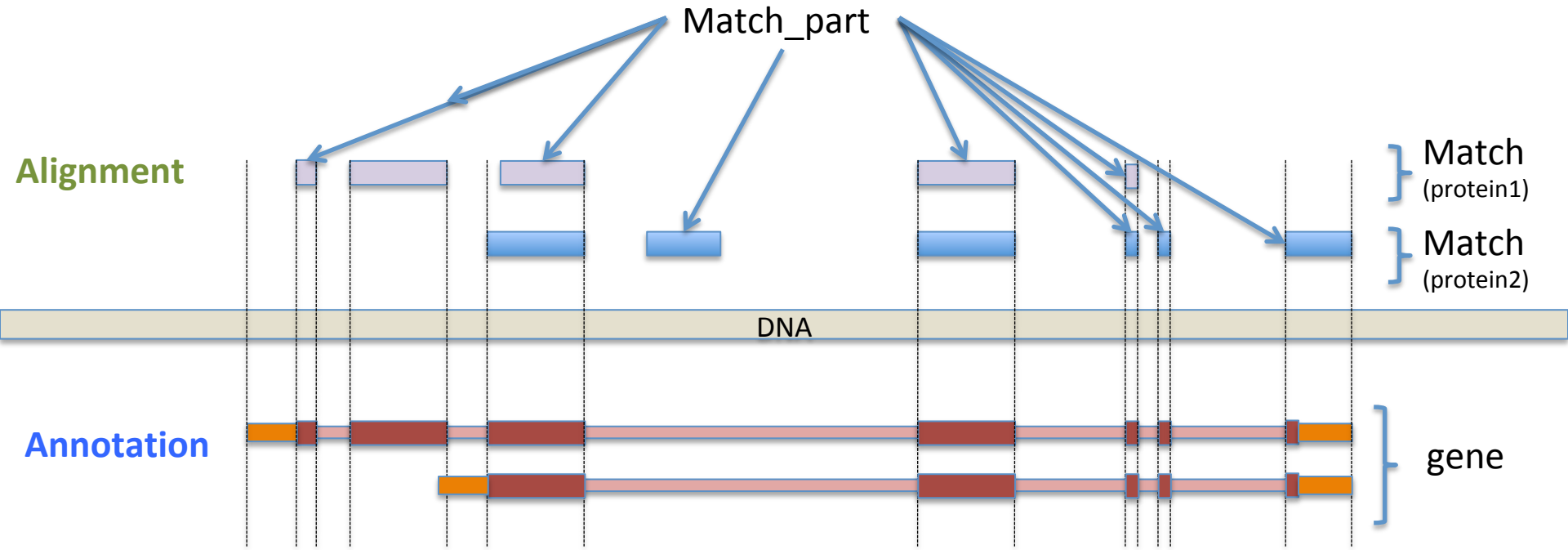
8) phase

9) attributes
tag value;

! Features grouped by a **common attribute** (gene_id / transcript_id)



/!\ different type of gff: **annotation** / **alignment** / other



Intron, exon, CDS,splice site, UTR, mRNA, isoforms

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

The main steps in genome annotation

1 2 3 4 5

QC assembly → Structural annotation → Manual curation → Functional annotation → Downstream analysis / Submission

BUSCO

MAKER Annotate this!

EuGene-EP

Web Apollo

InterPro

GFF3 + FASTA → EMBLmyGFF3 → EMBL format

Before annotation – check assembly quality

- The quality of the assembly will heavily influence the quality of the annotation
  - ❑ SNP-errors can change start/stop-codons
  - ❑ Indels can cause frame-shifts
  - ❑ High fragmentation could break loci
  - ❑ missing loci cannot be annotated

=> Annotation tools have difficulties to deal with those problems

Assembly check and preparation

- Fragmentation (N50, number of sequences, how many small contigs)
- Sanity of the fasta file (Ns, IUPAC, lowercase nucleotides)
- Completeness / duplication / fragmentation BUSCO
- Presence of Organelles
- Other (GC content, how distant from other species)

BUSCO output

```
# BUSCO version is: 3.0.2
# The lineage dataset is: fungi_odb9 (Creation date: 2016-02-13,
number of species: 85, number of BUSCOs: 290)
#
# Summarized benchmarking in BUSCO notation for file genome.fa
# BUSCO was run in mode: genome

    C:98.6%[S:97.9%,D:0.7%],F:0.0%,M:1.4%,n:290

    286 Complete BUSCOs (C)
    284 Complete and single-copy BUSCOs (S)
    2   Complete and duplicated BUSCOs (D)
    0   Fragmented BUSCOs (F)
    4   Missing BUSCOs (M)
    290 Total BUSCO groups searched
```

Repeat Masking

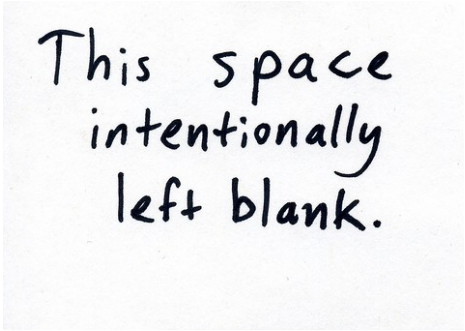
- Repeatmodeler to find new repeats
  http://www.repeatmasker.org/RepeatModeler/

- Repeatmasker to mask known repeats
  http://www.repeatmasker.org

+ Save time
+ Increase quality of the annotation

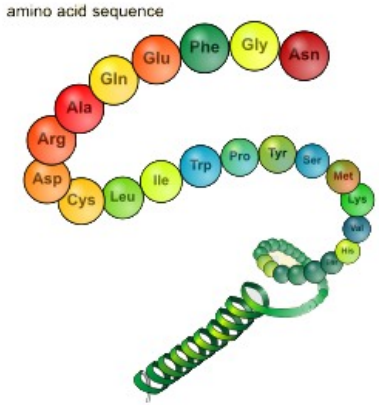
Types of external data used

Ø



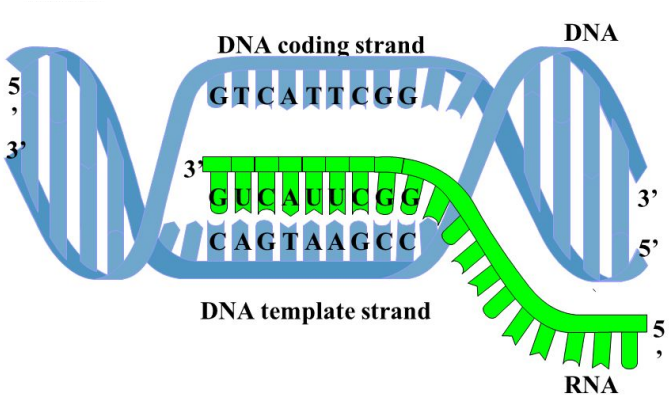
Proteins

- Known amino acid sequences from other organisms



Transcripts

- Assembled from RNA-seq or downloaded ESTs



NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

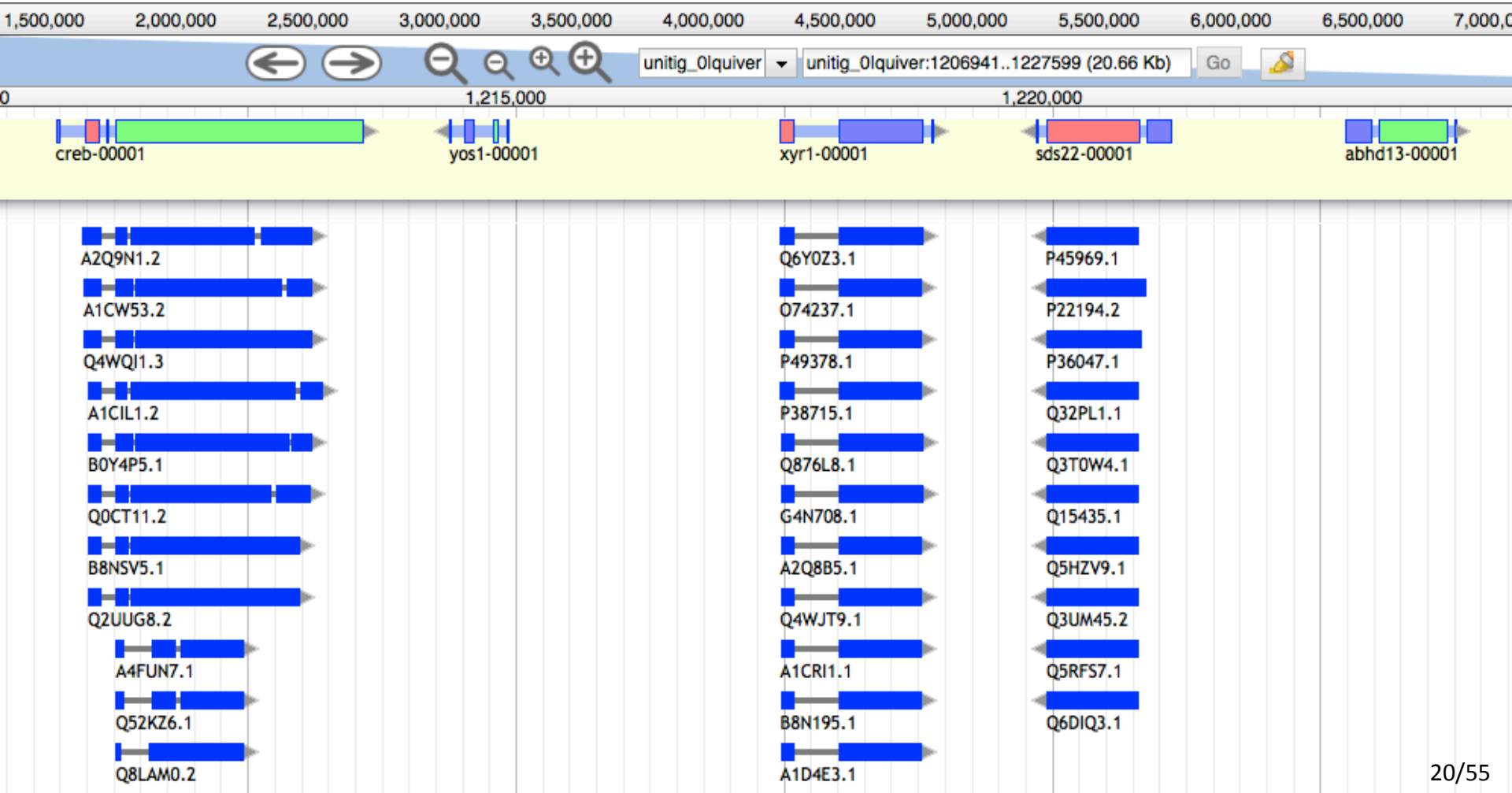
Types of data used: Proteins

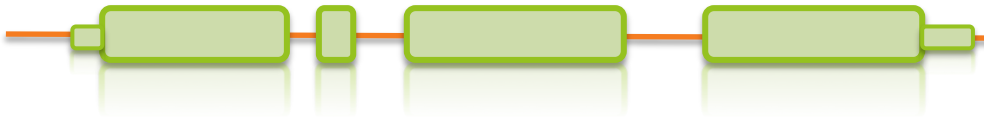
- Conserved in sequence => conserved annotation with little noise
- Proteins from model organisms often used => bias?
- Proteins can be incomplete => problems as many annotation procedures are heavily dependent on protein alignments

>ENSTGUP00000017616 pep:novel chromosome:taeGut3.2.4:8_random:2849599:2959678:-1 gene:ENSTGUG00000017338 transcript:ENSTGUT00000018018 g
RSPNATEYNWHHLRYPKIPERLNPPAAAGPALSTAEGWMLPWGNGQHPLLARAPGKGRER
DGKELIKKPKTFKFTFLKKKKKKKKTFK
>ENSTGUP00000017615 pep:novel chromosome:taeGut3.2.4:23_random:205321:209117:1 gene:ENSTGUG00000017337 transcript:ENSTGUT00000018017 ger
PDLRELVLMFEHLHRVRNGGFRNSEVKKWPDRSPPPYHSFTPAQKSFSLAGCSGESTKMG
IKERMRLSSSQRQGSRGRQQHLGPPLHRSPSPEDVAEATSPTKVQKSWSFNDRTRFRASL
RLKPRIPAEGDCPPEDSGEERSSPCDLTFEDIMPAVKTLIRAVRILKFLVAKRKFKETLR
PYDVKDVIEQYSAGHLDMLGRIKSLQTRVEQIVGRDRALPADKKVREKGEKPALEAELVD
ELSMMGRVVKVERQVQSIEHKLDLLLGLYSRCLRKGSANSLVLAAVRVPPGEPDVTSDYQ
SPVEHEDISTSAQSLSISRLASTNMD

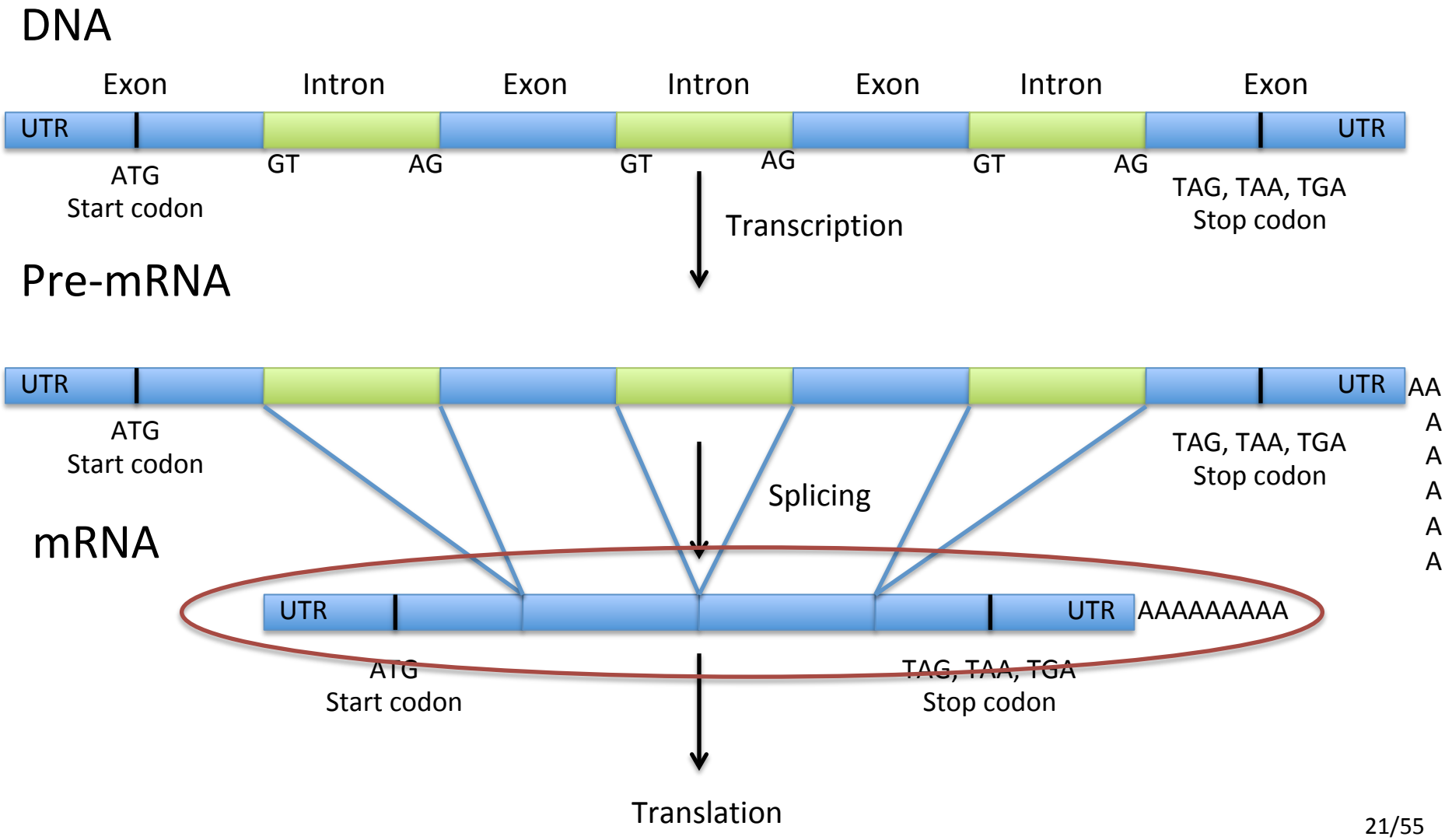
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Protein sequences are aligned to the genome



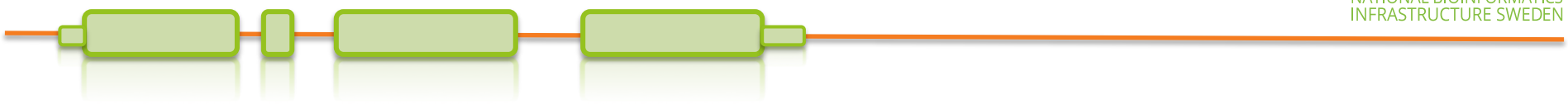


Types of data used: RNA-seq

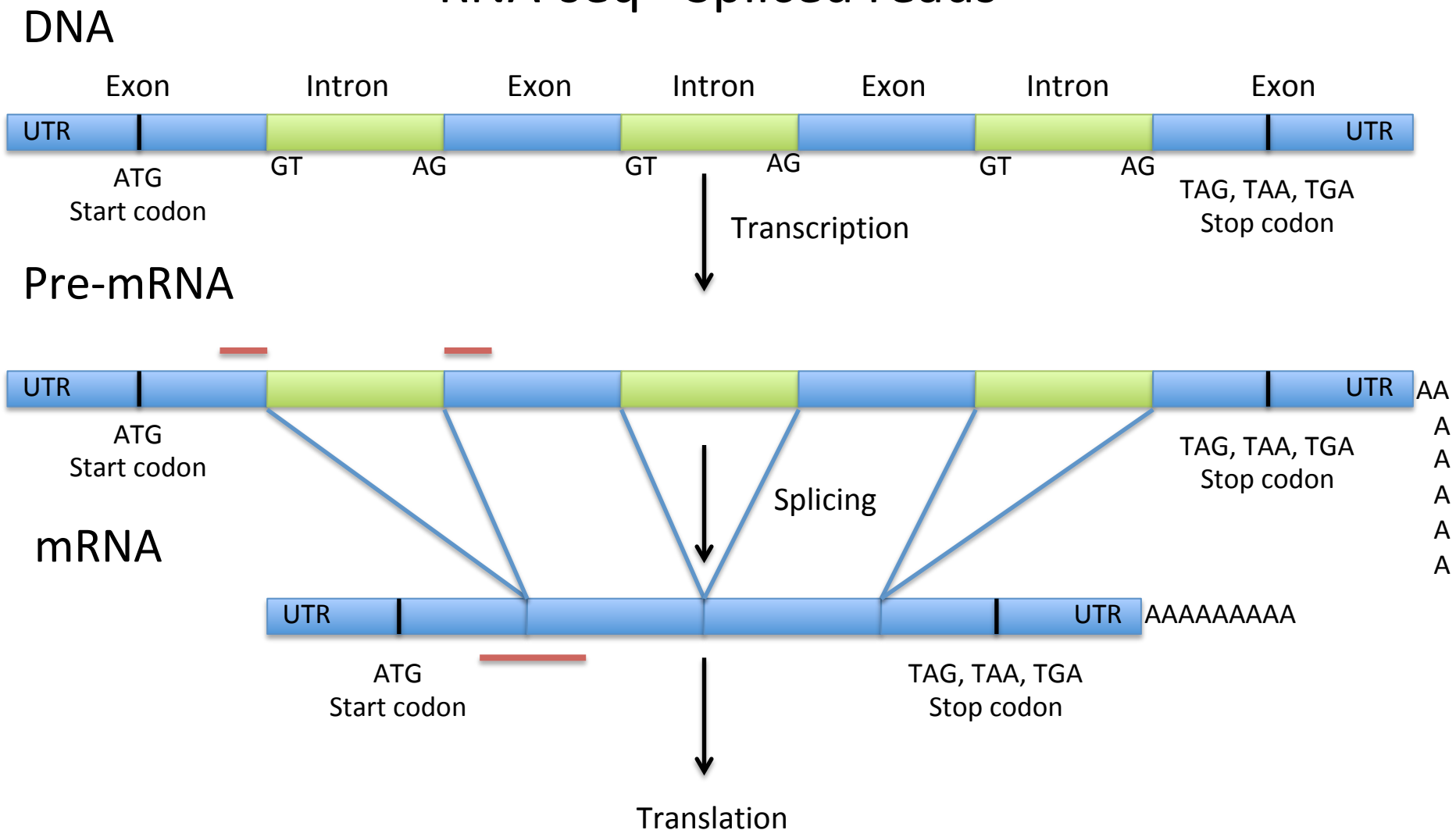


Types of data used: RNA-seq

- 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

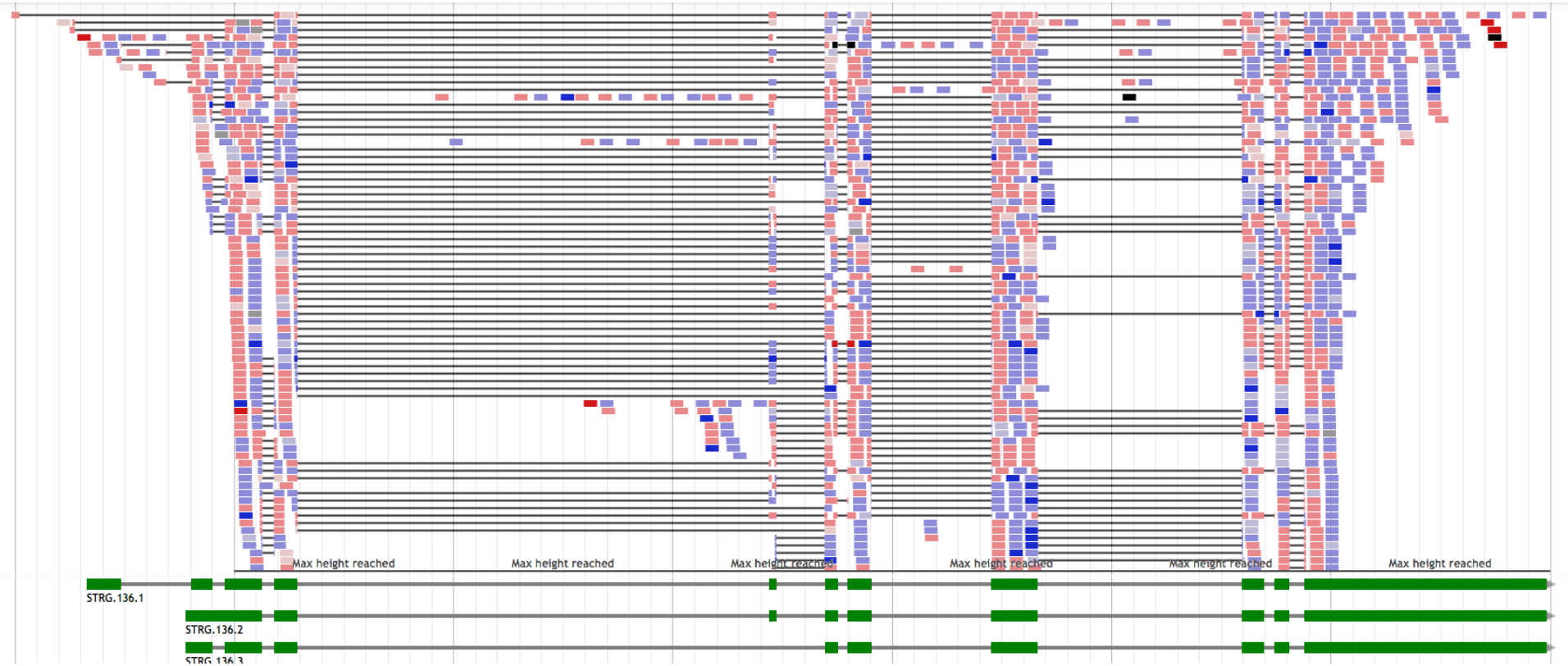


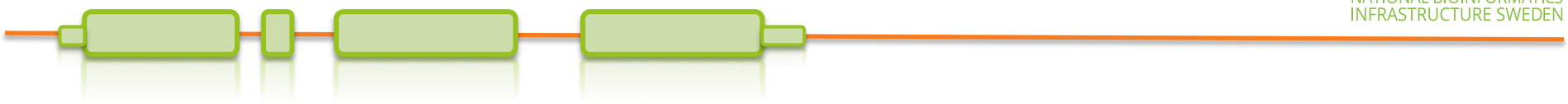
RNA-seq - Spliced reads



NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

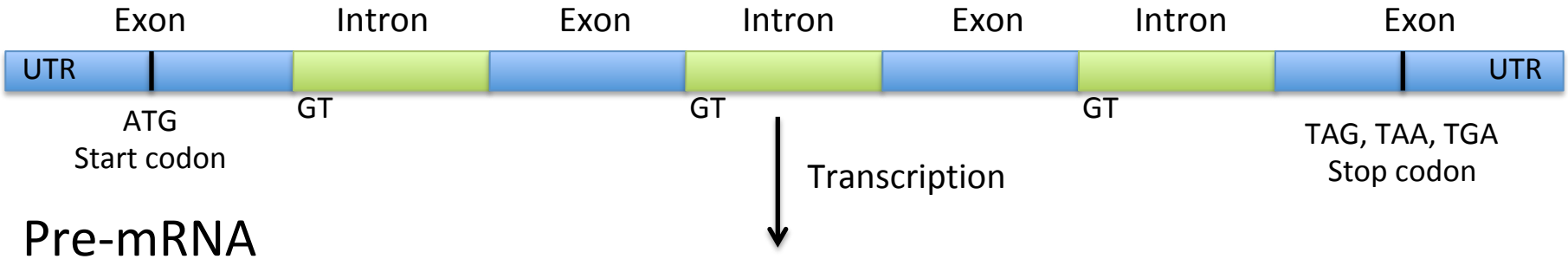
RNA-seq - Spliced reads



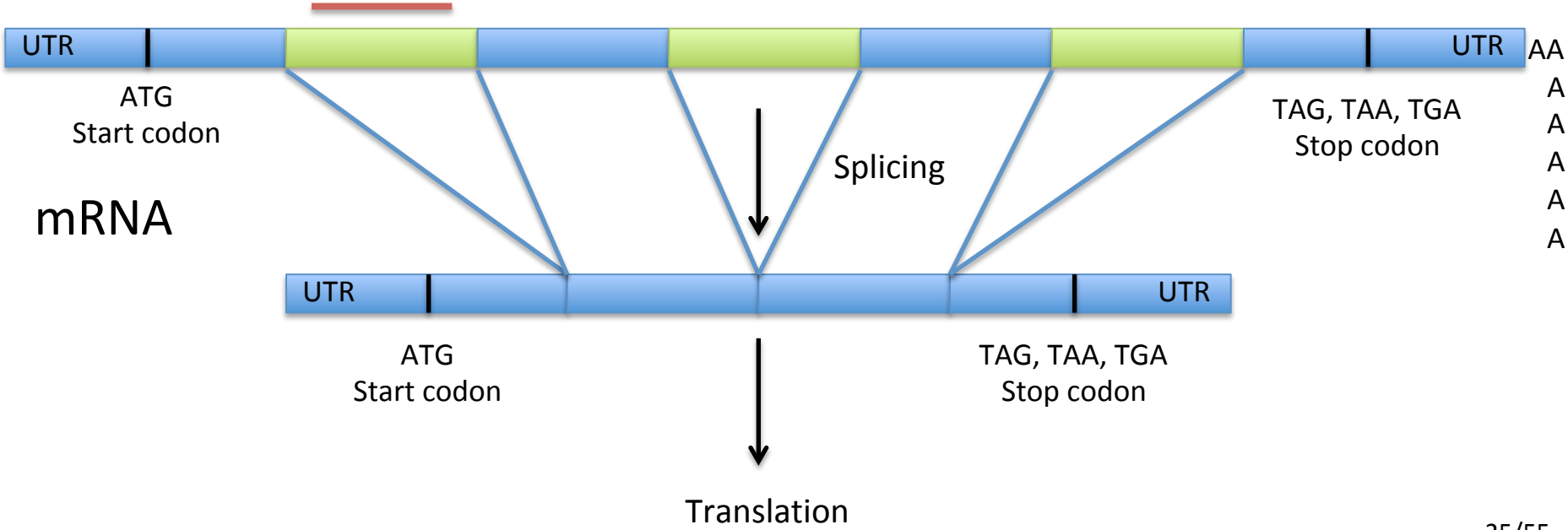


Pre-mRNA

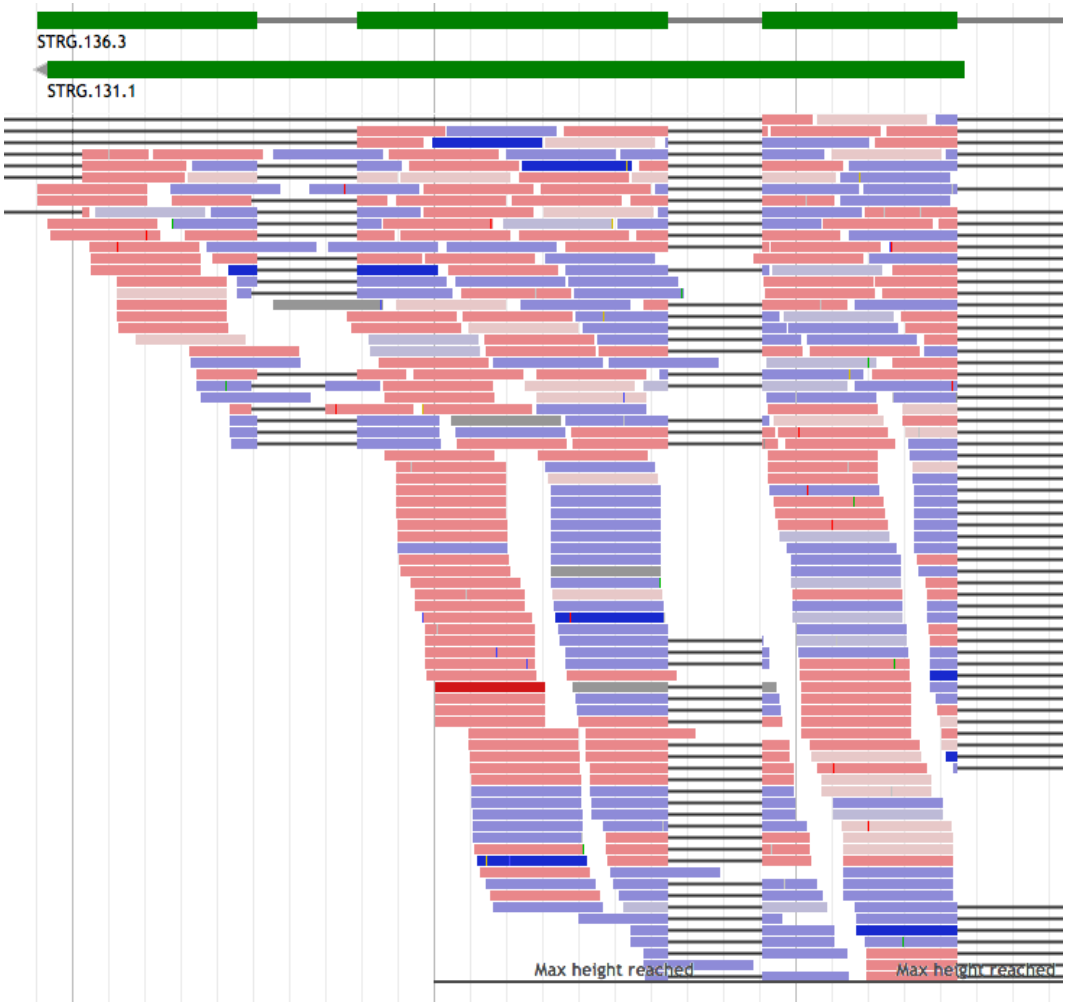
DNA



Pre-mRNA



RNA-seq – pre-mRNA noise



Types of data used: RNA-seq

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

- Genome guided assembly

=> e.g., Stringtie: mapped reads -> transcripts

- *De novo*

=> e.g., Trinity: assembles transcripts without a genome