Tutorial 6 – RNA-seq Alignment with STAR

MICB405 – BIOINFORMATICS – 2021W-T1

22 OCTOBER 2021

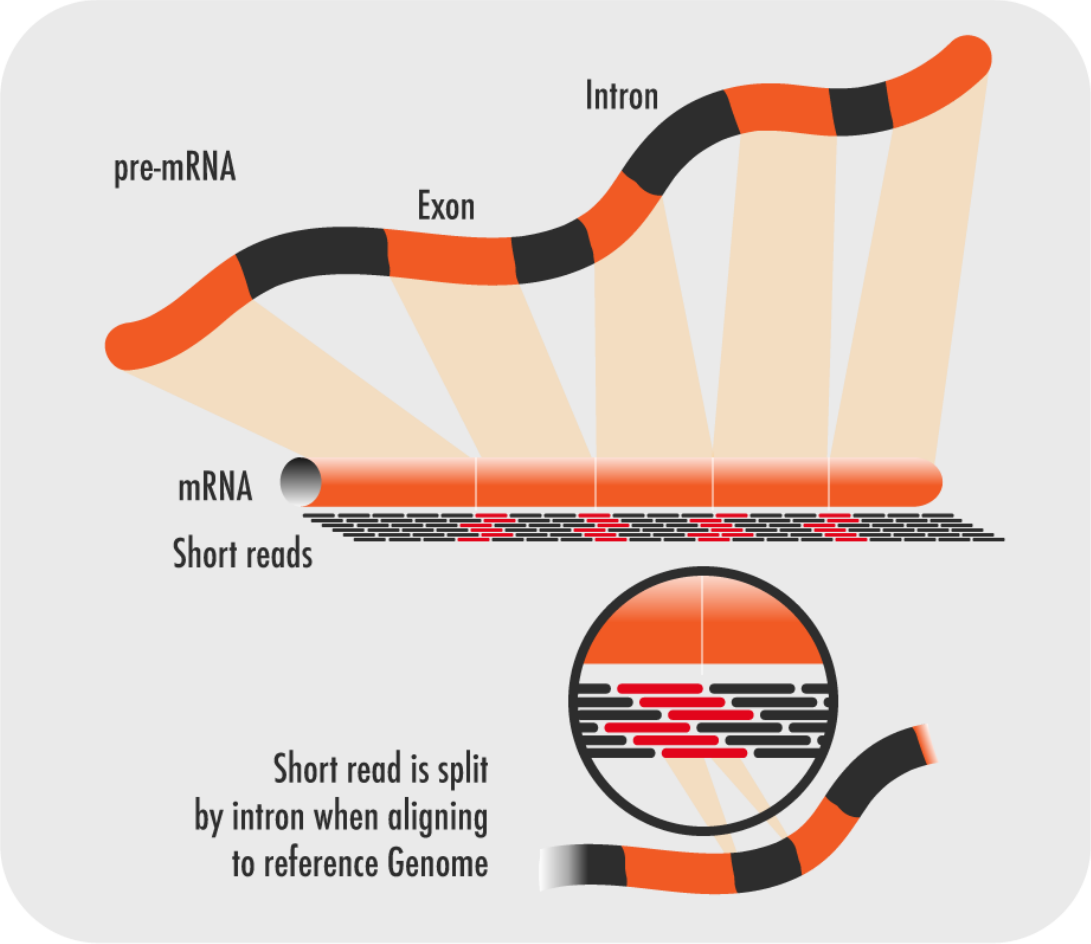
AXEL HAUDUC

RNA-seq review

Examines the **quantity** and **sequences** of RNA from a sample

- *Transcriptome*

Different tissues and conditions, even within the same individual, will yield different results



What is **STAR**

Alignment software for RNA-seq experiments

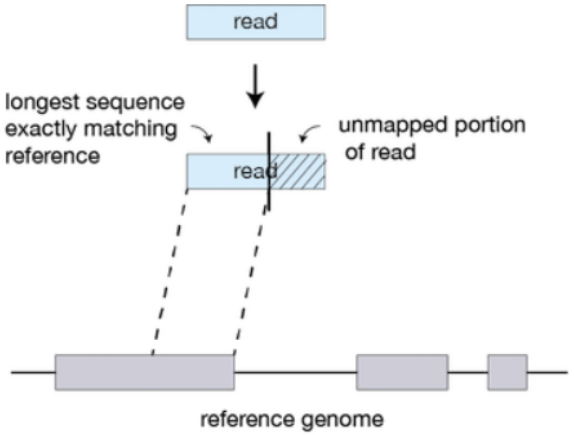
- Intron/exon-aware

Inputs

- FASTA
- FASTQ
- GTF

Seed searching

For every read that STAR aligns, STAR will search for the longest sequence that exactly matches one or more locations on the reference genome. These longest matching sequences are called the Maximal Mappable Prefixes (MMPs):



The different parts of the read that are mapped separately are called 'seeds'. So the first MMP that is mapped to the genome is called *seed1*.

STAR will then search again for only the unmapped portion of the read to find the next longest sequence that exactly matches the reference genome, or the next MMP, which will be *seed2*.

Gene Transfer Format (GTF)

Consists of one line per feature, each containing 9 columns of data, plus optional track definition lines. The following documentation is based on the Version 2 specifications.

Fields must be tab-separated. Also, all but the final field in each feature line must contain a value; "empty" columns should be denoted with a '.'

- **seqname** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. Important note: the seqname 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.
- **source** - name of the program that generated this feature, or the data source (database or project name)
- **feature** - feature type name, e.g. Gene, Variation, Similarity
- **start** - Start position* of the feature, with sequence numbering starting at 1.
- **end** - End position* of the feature, with sequence numbering starting at 1.
- **score** - A floating point value.
- **strand** - defined as + (forward) or - (reverse).
- **frame** - 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..
- **attribute** - A semicolon-separated list of tag-value pairs, providing additional information about each feature.

Tutorial Background

Open Access | Published: 12 December 2017

An RNA-Seq atlas of gene expression in mouse and rat normal tissues

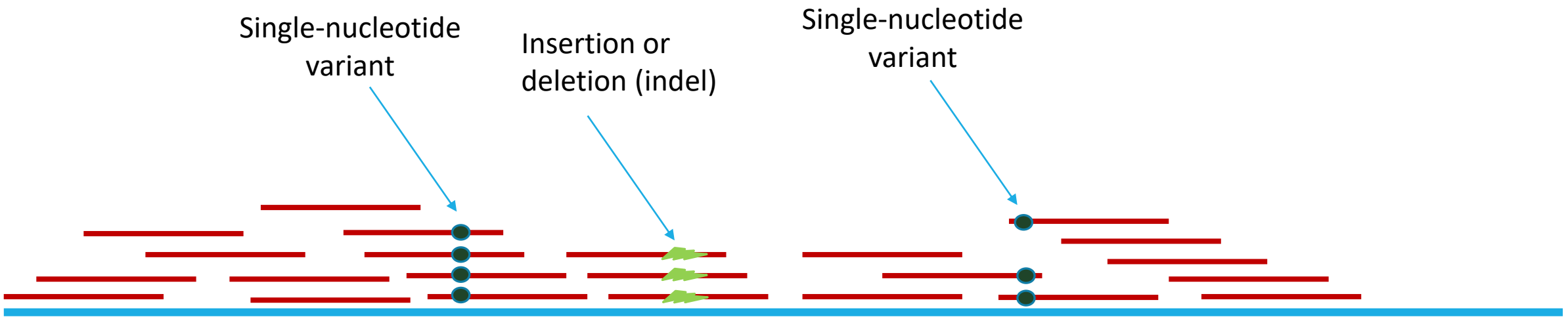
Julia F. Söllner, German Leparc, Tobias Hildebrandt, Holger Klein, Leo Thomas, Elia Stupka & Eric Simon

Scientific Data **4**, Article number: 170185 (2017) | Cite this article

- Freely-accessible data
- Panel of normal tissue RNA-seq assays done on multiple organs for 3 different male mouse individuals
- Study of inter- and intra-species variation in gene expression

Variant Calling

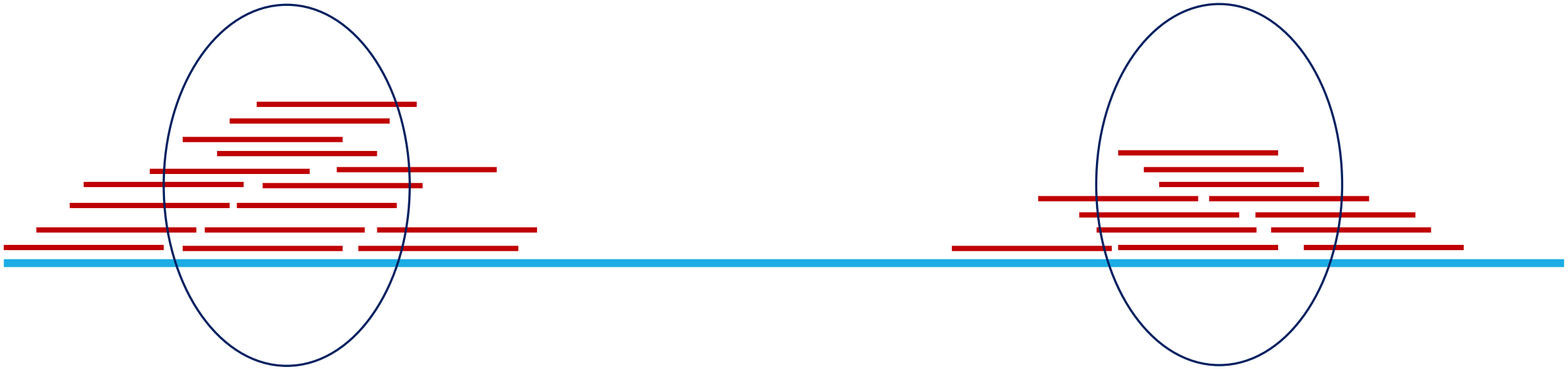
Variant Calling



Let’s find them!

Peak Calling

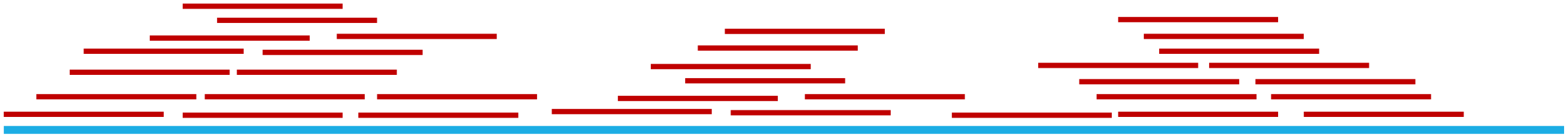
Peak Calling



These are peaks – their locations are linked to the epigenomic mark that is being studied

RNA-seq Analysis

RNA-seq reads



Genome,
hypothetically
ignoring introns

RNA-seq Analysis