From RNA-seq reads to gene models

Biodiversity Genomics Academy 2023

Thursday 28th September 2023

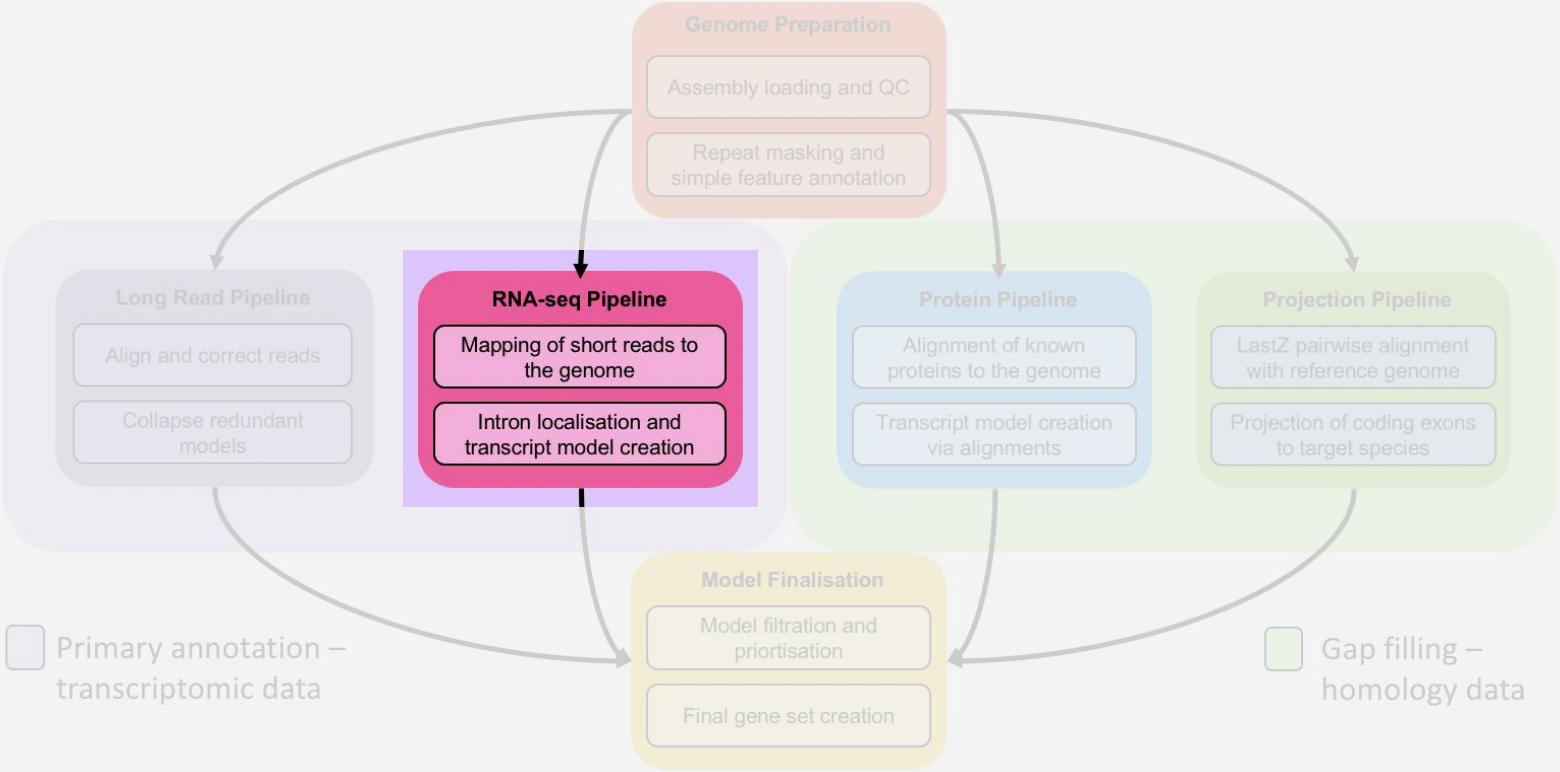
Jose Perez-Silva

Bioinformatician at Genebuild

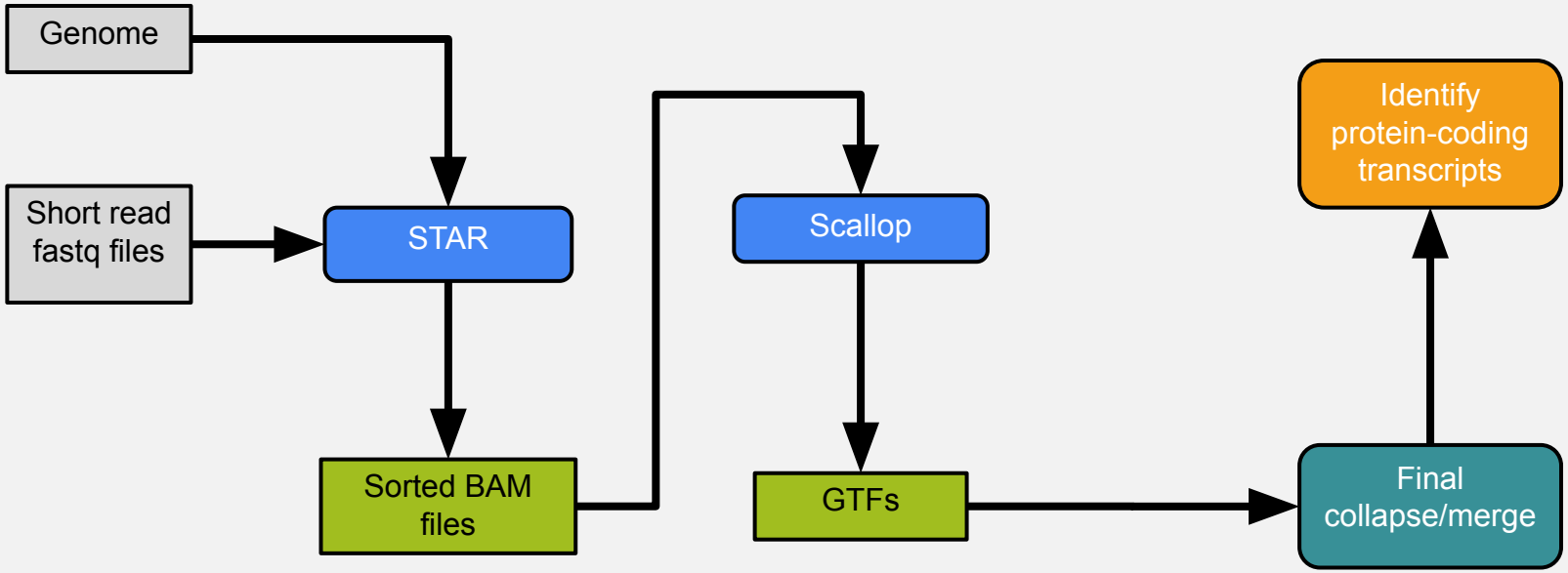
ereboperezsilva@ebi.ac.uk

EMBL-EBI

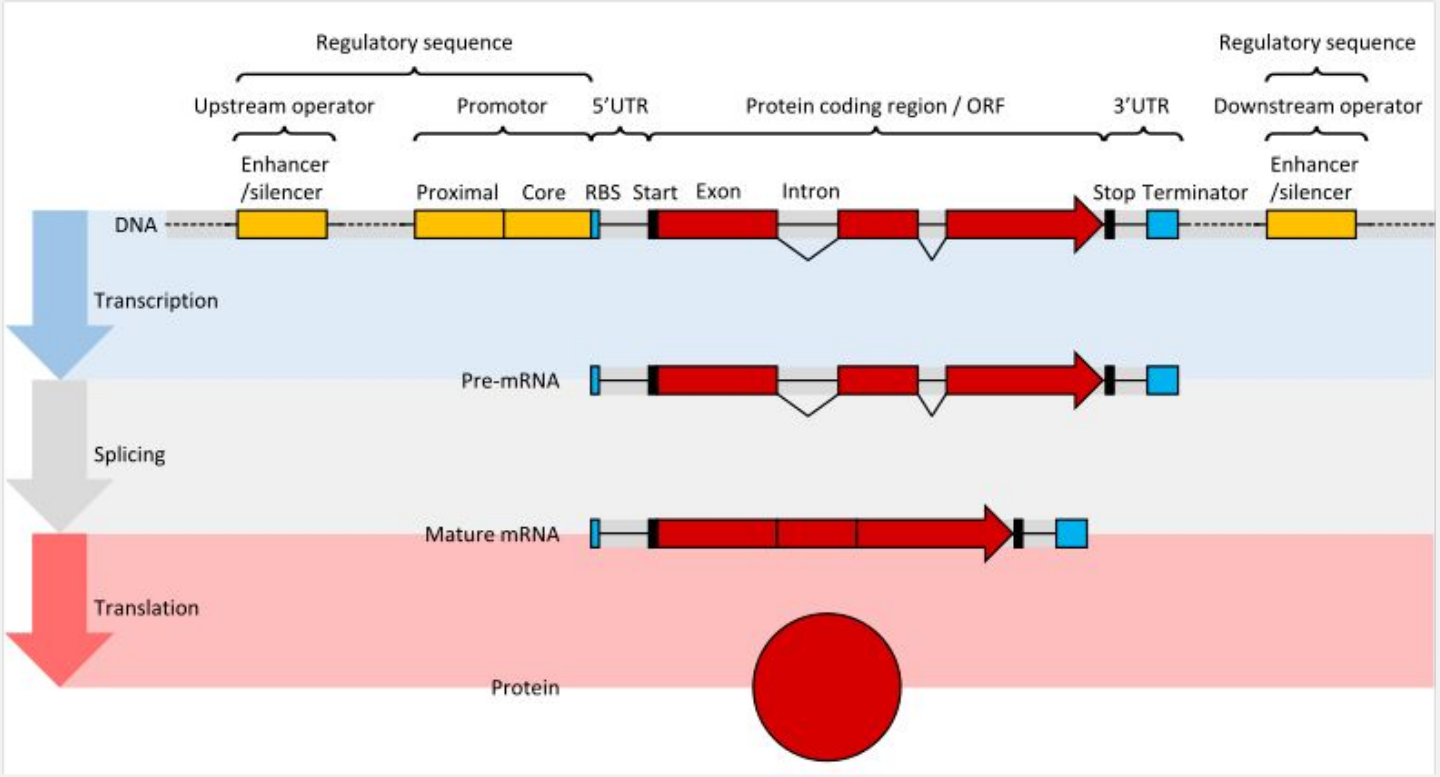
introduction - genome annotation



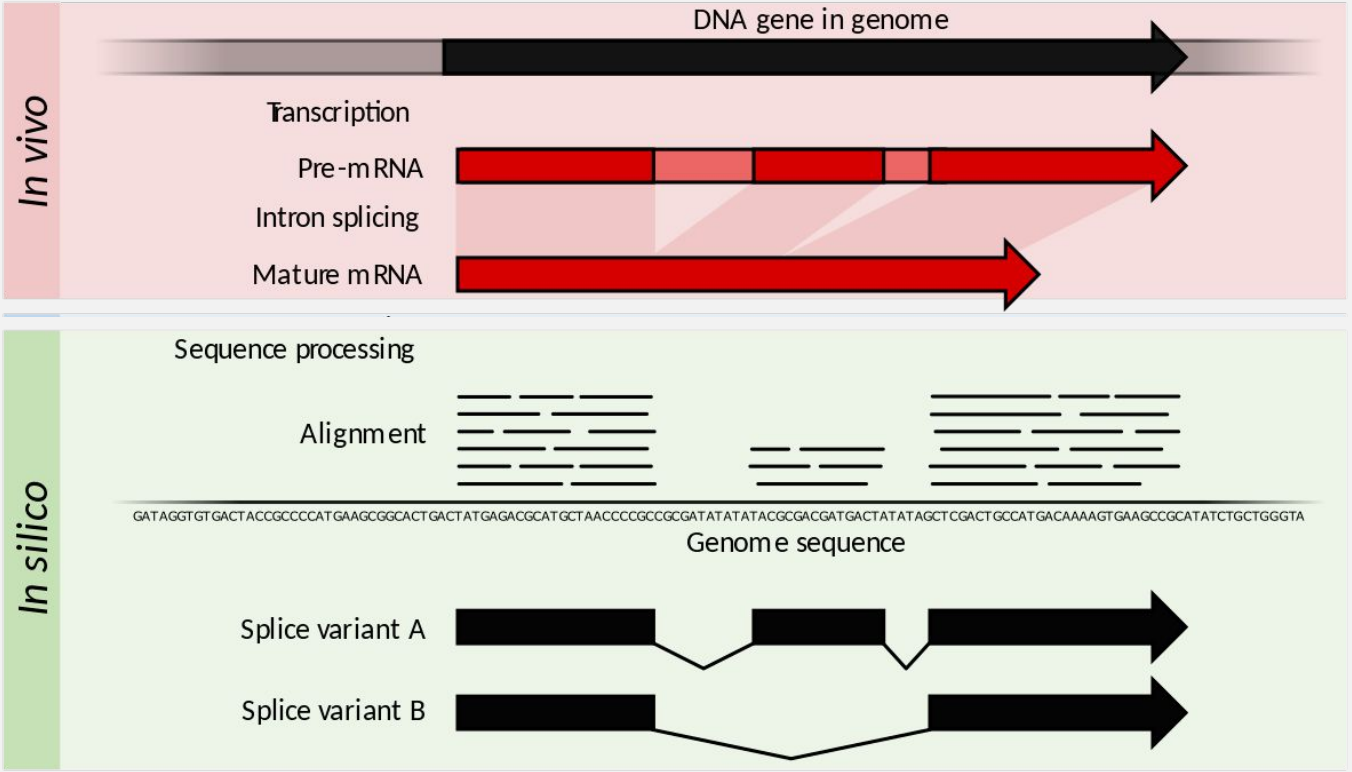
introduction - genome annotation



introduction - transcription and RNA-seq



introduction - transcription and RNA-seq



introduction - data

- Annotating_the_genome_with_rnaseq_data.txt
- Fastq/
  - 2cell_chr12_R1.fastq 2cell_chr12_R2.fastq
  - 6hpf_chr12_R1.fastq 6hpf_chr12_R2.fastq
- Genome/
  - Danio_rerio.GRCz11.dna.chromosome.12.fa

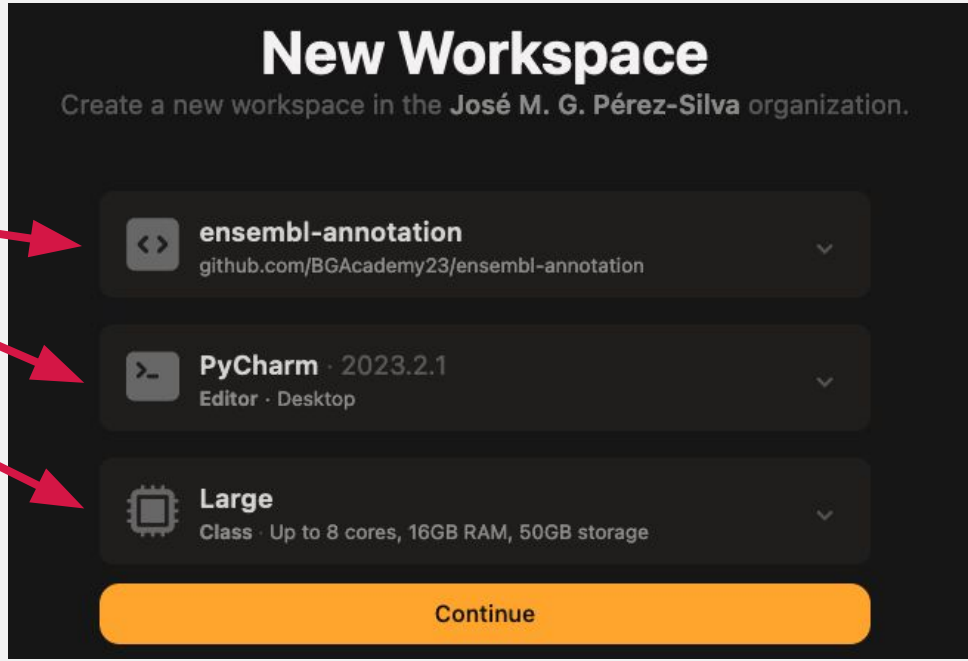
From: Kimmel et al. Stages of embryonic development of the zebrafish Dev. Dyn. 203:253-310, 1995

introduction - tools

- STAR: Spliced Transcripts Alignment to a Reference ©
  - “Ultrafast universal RNA-seq aligner”
- SCALLOP:
  - “Accurate reference-based transcript assembler”
- SAMTOOLS:
  - “Suite of programs for interacting with high-throughput sequencing data”

let’s get starting

- Access: http://gitpod.io/#https://github.com/BGAcademy23/ensembl-annotation
- From the options:
  - Leave this
  - Choose your favourite editor
  - Choose “Large” in the third



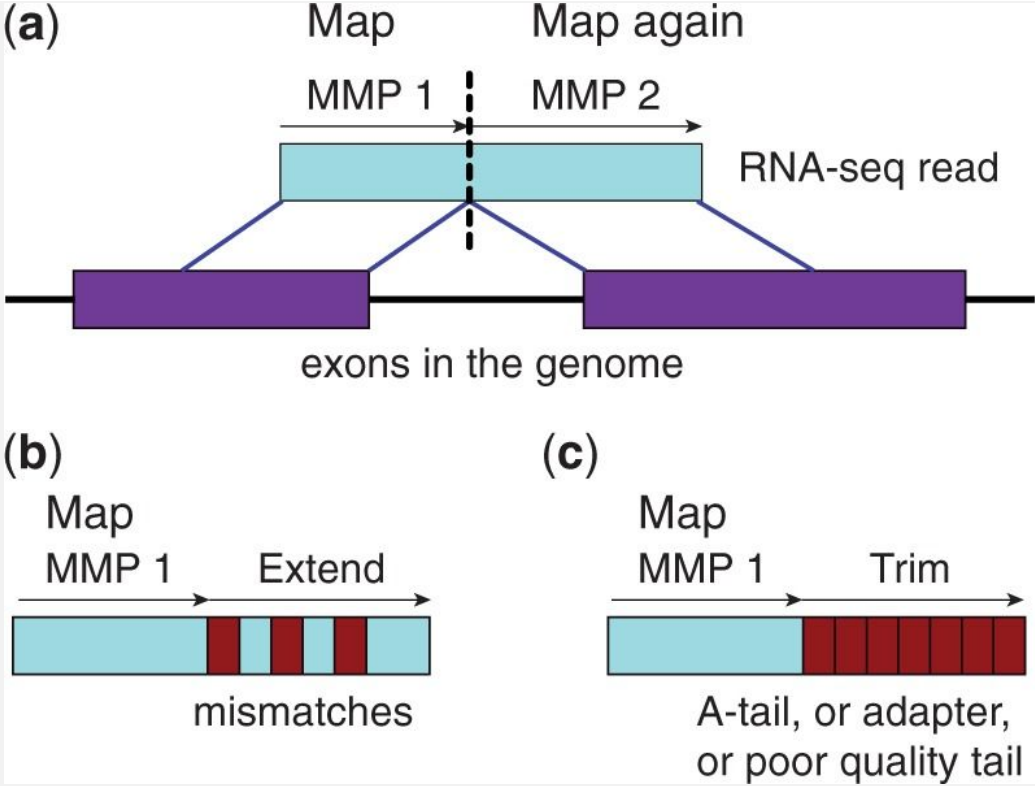
let’s get starting

- Wait while it loads. This may take a second or two.
- Navigate the files in the left, open **`ensembl-annotation`**, then

  `Annotating_the_genome`, and PREVIEW **`instructions.md`** by left-clicking it.
- You should now have a markdown document with instructions.

STAR

- Maps over splices, mismatches and excludes unwanted seqs
- Requires a genome index
- Docs available in github page
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530905/



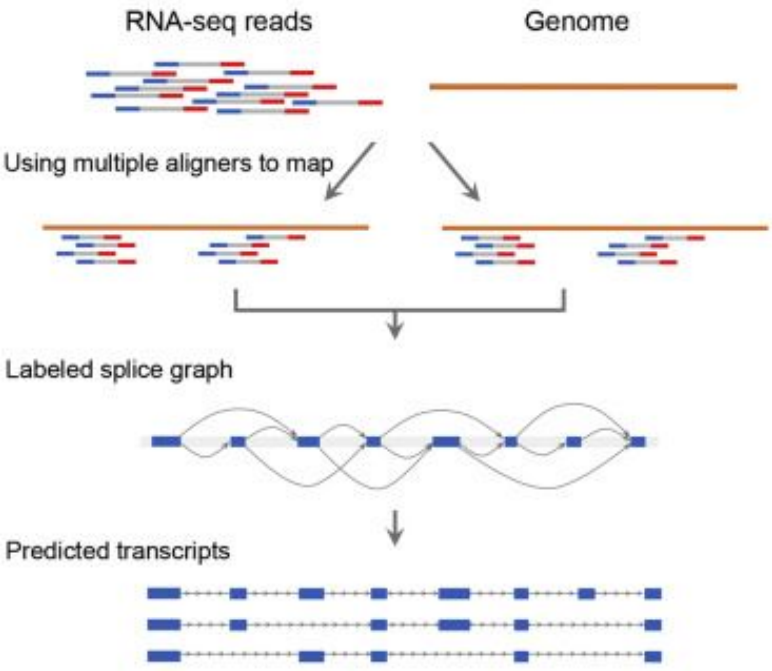
SAMtools

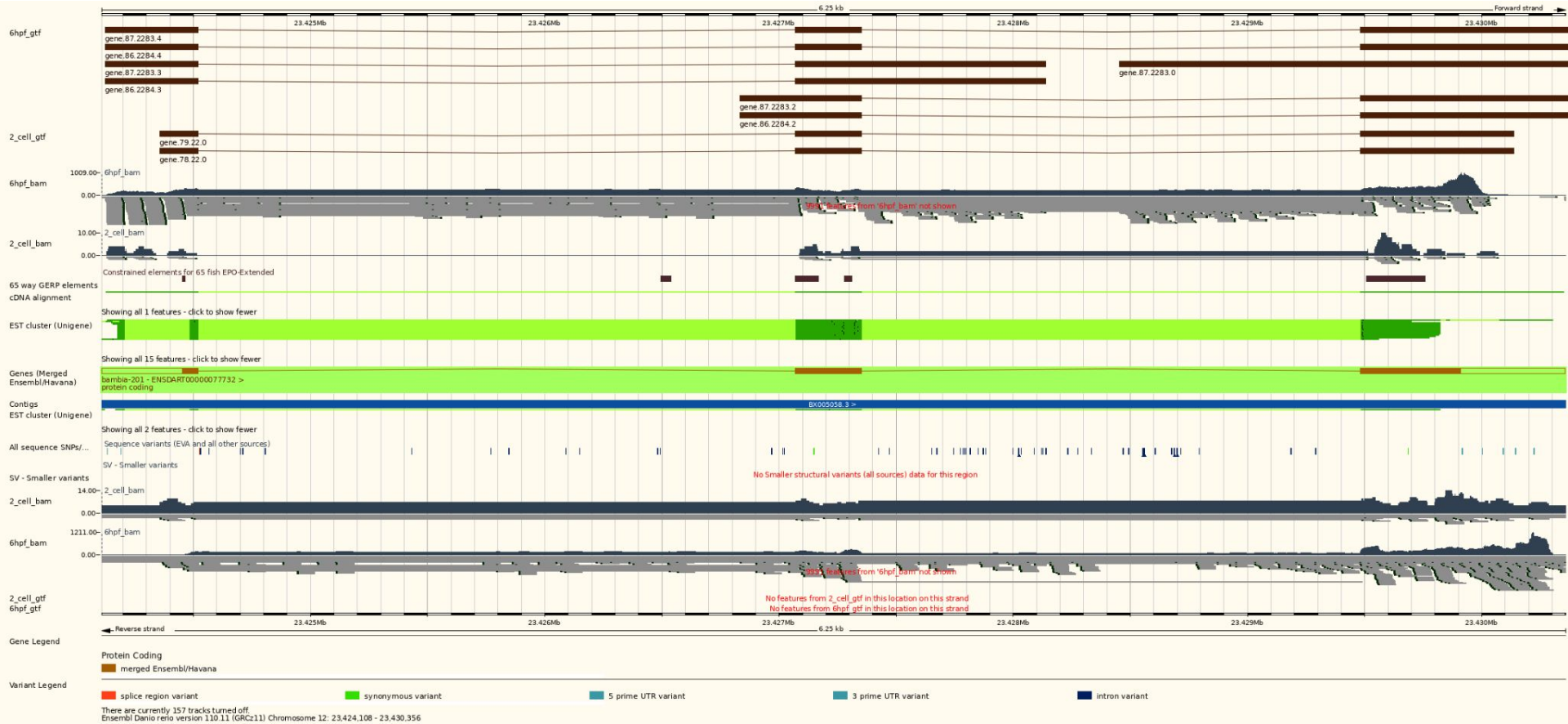
- A suite of software for various tasks:
  - Fastq to BAM/CRAM
  - WGS/WES mapping to variant calls
  - Filtering of VCF files
  - Several, BAM-workflow related
- Docs available in github page
- https://pubmed.ncbi.nlm.nih.gov/33590861/

```
samtools addreplacerg -r 'ID:fish' -r 'LB:1334' -r 'SM:alpha' -o output.bam input.bam
samtools ampliconclip -b bed.file input.bam
samtools ampliconstats primers.bed in.bam
samtools bedcov aln.sorted.bam
samtools calmd in.sorted.bam ref.fasta
samtools cat out.bam in1.bam in2.bam in3.bam
samtools collate -o aln.name_collated.bam aln.sorted.bam
samtools consensus -o out.fasta in.bam
samtools coverage aln.sorted.bam
samtools cram-size -v -o out.size in.cram
samtools depad input.bam
samtools depth aln.sorted.bam
samtools dict -a GRCh38 -s "Homo sapiens" ref.fasta
samtools faidx ref.fasta
samtools fasta input.bam > output.fasta
samtools fastq input.bam > output.fastq
samtools fixmate in.namesorted.sam out.bam
samtools flags PAIRED,UNMAP,MUNMAP
samtools flagstat aln.sorted.bam
samtools fqidx ref.fastq
samtools head in.bam
samtools idxstats aln.sorted.bam
samtools import input.fastq > output.bam
samtools index aln.sorted.bam
samtools markdup in.algnsorted.bam out.bam
samtools merge out.bam in1.bam in2.bam in3.bam
samtools mpileup -C50 -f ref.fasta -r chr3:1,000-2,000 in1.bam in2.bam
samtools phase input.bam
samtools quickcheck in1.bam in2.cram
samtools reference -o ref.fa in.cram
samtools reheader in.header.sam in.bam > out.bam
samtools reset -o /tmp/reset.bam processed.bam
samtools samples input.bam
samtools sort -T /tmp/aln.sorted -o aln.sorted.bam aln.bam
samtools split merged.bam
samtools stats aln.sorted.bam
samtools targetcut input.bam
samtools tview aln.sorted.bam ref.fasta
samtools view -bt ref_list.txt -o aln.bam aln.sam.gz
```

scallop

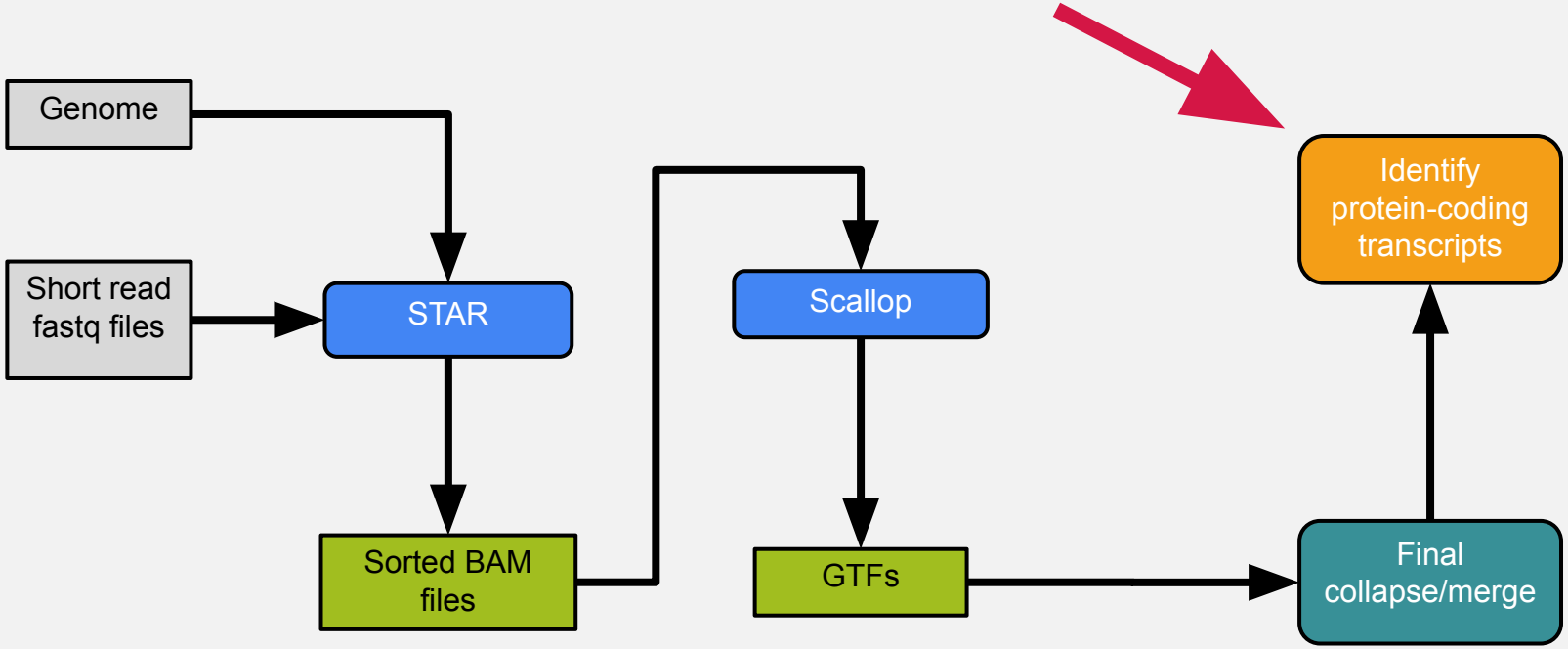
- Splice graph
- Docs available in github page
- https://www.nature.com/articles/nbt.4020




6.25 kb
Forward strand
6hpf_gtf
gene:87.2283.4
gene:86.2284.4
gene:87.2283.3
gene:86.2284.3
gene.87.2283.0
gene.87.2283.2
gene.86.2284.2
2_cell_gtf
gene.79.22.0
gene.78.22.0
6hpf_bam
1009.00
0.00
6hpf_bam
2933 features from '6hpf_bam' not shown
2_cell_bam
10.00
0.00
2_cell_bam
65 way GERP elements
Constrained elements for 65 fish EPO-Extended
cDNA alignment
EST cluster (Unigene)
Showing all 1 features - click to show fewer
Genes (Merged Ensembl/Havana)
Showing all 15 features - click to show fewer
bambia-201 - ENSDART00000077732 >
protein coding
Contigs
BX005058.3 >
EST cluster (Unigene)
Showing all 2 features - click to show fewer
All sequence SNPs/...
Sequence variants (EVA and all other sources)
SV - Smaller variants
SV - Smaller variants
No Smaller structural variants (all sources) data for this region
14.00
2_cell_bam
0.00
2_cell_bam
1211.00
6hpf_bam
0.00
6hpf_bam
2933 features from '6hpf_bam' not shown
2_cell_gtf
No features from 2_cell_gtf in this location on this strand
6hpf_gtf
No features from 6hpf_gtf in this location on this strand
23.425Mb
23.426Mb
23.427Mb
23.428Mb
23.429Mb
23.430Mb
Reverse strand
6.25 kb
Gene Legend
Protein Coding
merged Ensembl/Havana
Variant Legend
splice region variant
synonymous variant
5 prime UTR variant
3 prime UTR variant
intron variant
There are currently 157 tracks turned off.
Ensembl Danio rerio version 110.11 (GRCz11) Chromosome 12: 23,424,108 - 23,430,356

EMBL-EBI

what comes next?

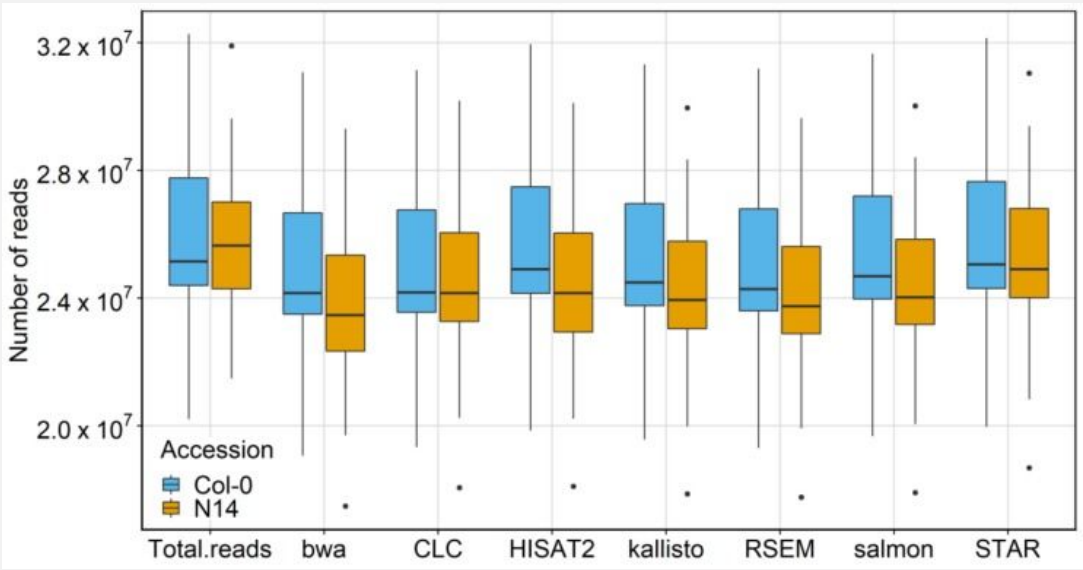


what comes next?

- There are tons of transcripts (m, snc, lnc, r, …)
- We must differentiate among them
- Translation, gene model generation, final geneset

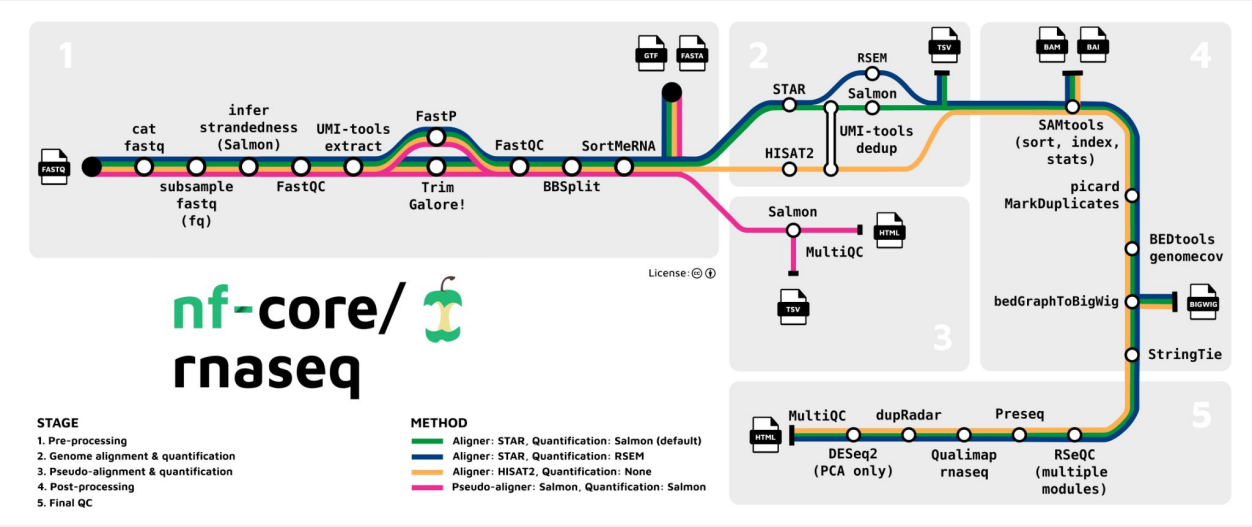
overview

- Simplified show of 1 process among many in our pipeline
- Alternatives:
  - HISAT2, salmon
  - StringTie2



overview

- A different alternative:
  - NextFlow/nf-core
- Different functionalities and uses
- Modules and pipelines



The Eukaryotic Annotation Team

Fergal Martin
Eukaryotic Annotation Team Leader

The Genebuild Team

Leanne Haggerty
Ensembl Genome Annotation Project Lead

Swati Sinha
Senior Bioinformatician

Francesca Floriana Tricomi
Bioinformatician

Jose Maria Gonzalez Perez-Silva
Bioinformatician

Vianey Paola Barrera Enriquez
Bioinformatician

The Comparative Genomics Team

Thiago Augusto Lopes Genez
Software Engineer

Thomas Walsh
Senior Bioinformatician

Botond Sipos
Bioinformatics Developer

Ivana Pilizota
Bioinformatics Developer

Simarpreet Kaur Bhurji
Bioinformatician

EMBL-EBI

Acknowledgements

Funding

EMBL

Open Targets

National Human Genome Research Institute (NHGRI)

National Institute of Allergy and Infectious Diseases (NIAID)

Funded by the European Union

wellcome

UKRI UK Research and Innovation

TRANSFORMING GENOMIC MEDICINE INITIATIVE

elixir