2ND EDITION

COURSE ON

COMPUTATIONAL BIOSCIENCES USING HPC SYSTEMS

6, 7, 8 FEBRUARY 2024

@ NOVA SCHOOL OF
SCIENCE AND TECHNOLOGY

UCIBIO · i4HB Institute for Health and Bioeconomy · INCD · LAQV requimte · LIP · EURO · PRR

MODULE 1 - OMICS

Transcriptome assembly

Pedro M. Costa
(pmcosta@fct.unl.pt)

RNA-Seq. Basics

- RNA-Seq is quantitative.
- Can yield more than 100K validated transcripts.
- It is usually non-targeted (i.e. “transcriptome-wide”)
- Depending on sequencing depth and length, may not yield full-length mRNAs.

For instance: 10-20 M reads, 150 bp single-end reads is the basic for expression analysis IF the transcriptome is reasonably annotated (unlikely in marine organisms). 100 M reads, 150-300 bp paired end is great for quantification AND characterisation of mRNAs (but it is also very expensive and challenging computationally).

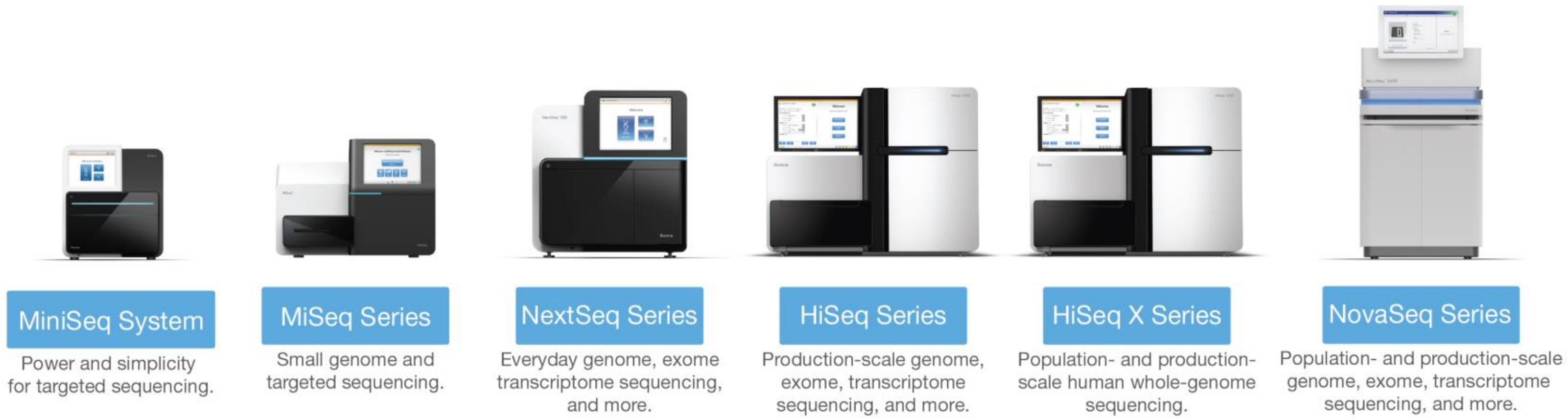
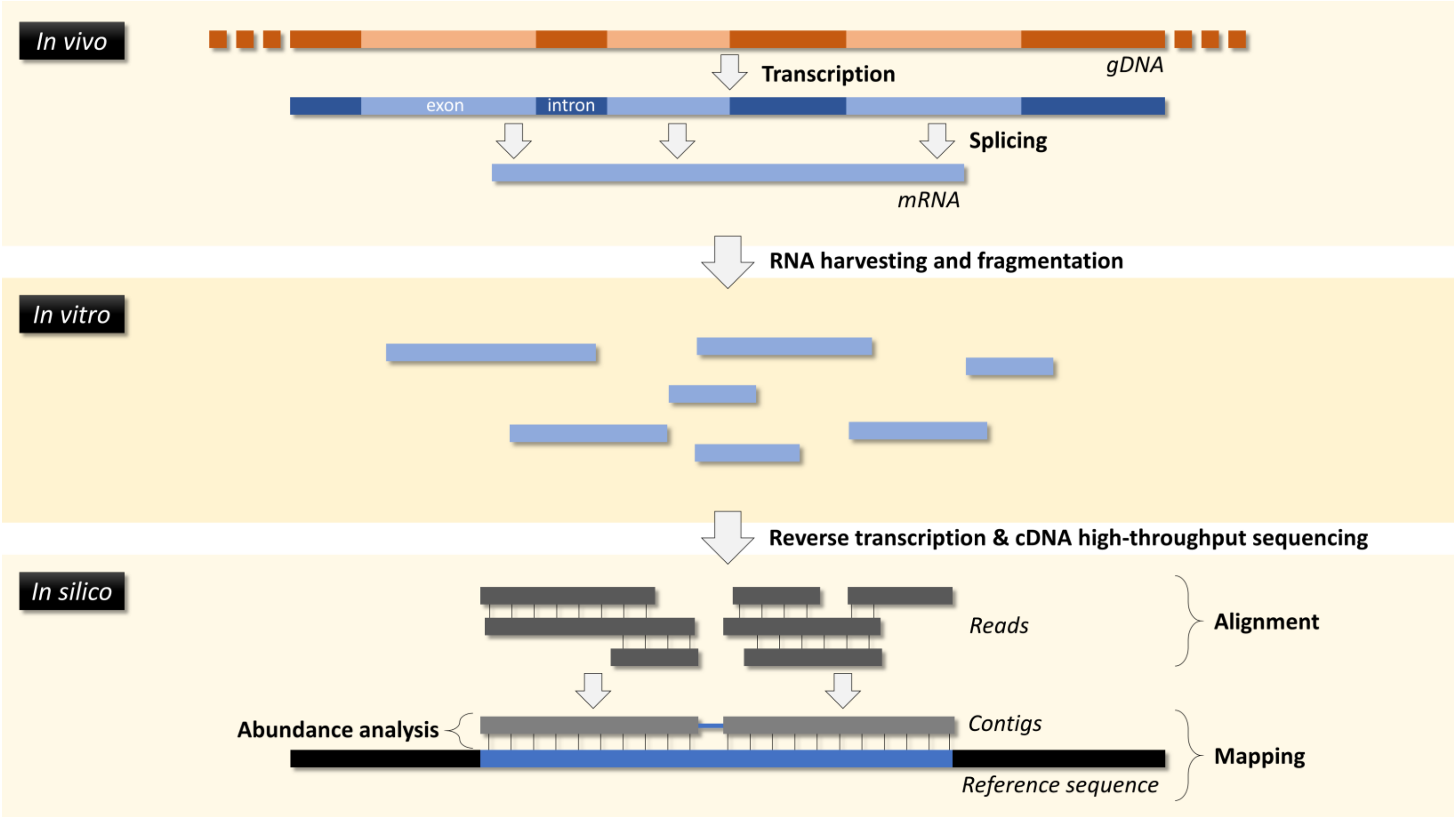

MiniSeq System
Power and simplicity for targeted sequencing.
MiSeq Series
Small genome and targeted sequencing.
NextSeq Series
Everyday genome, exome transcriptome sequencing, and more.
HiSeq Series
Production-scale genome, exome, transcriptome sequencing, and more.
HiSeq X Series
Population- and production-scale human whole-genome sequencing.
NovaSeq Series
Population- and production-scale genome, exome, transcriptome sequencing, and more.

Figure 6: Sequencing Systems for Virtually Every Scale—Illumina offers innovative NGS platforms that deliver exceptional data quality and accuracy over a wide scale, from small benchtop sequencers to production-scale sequencing systems.

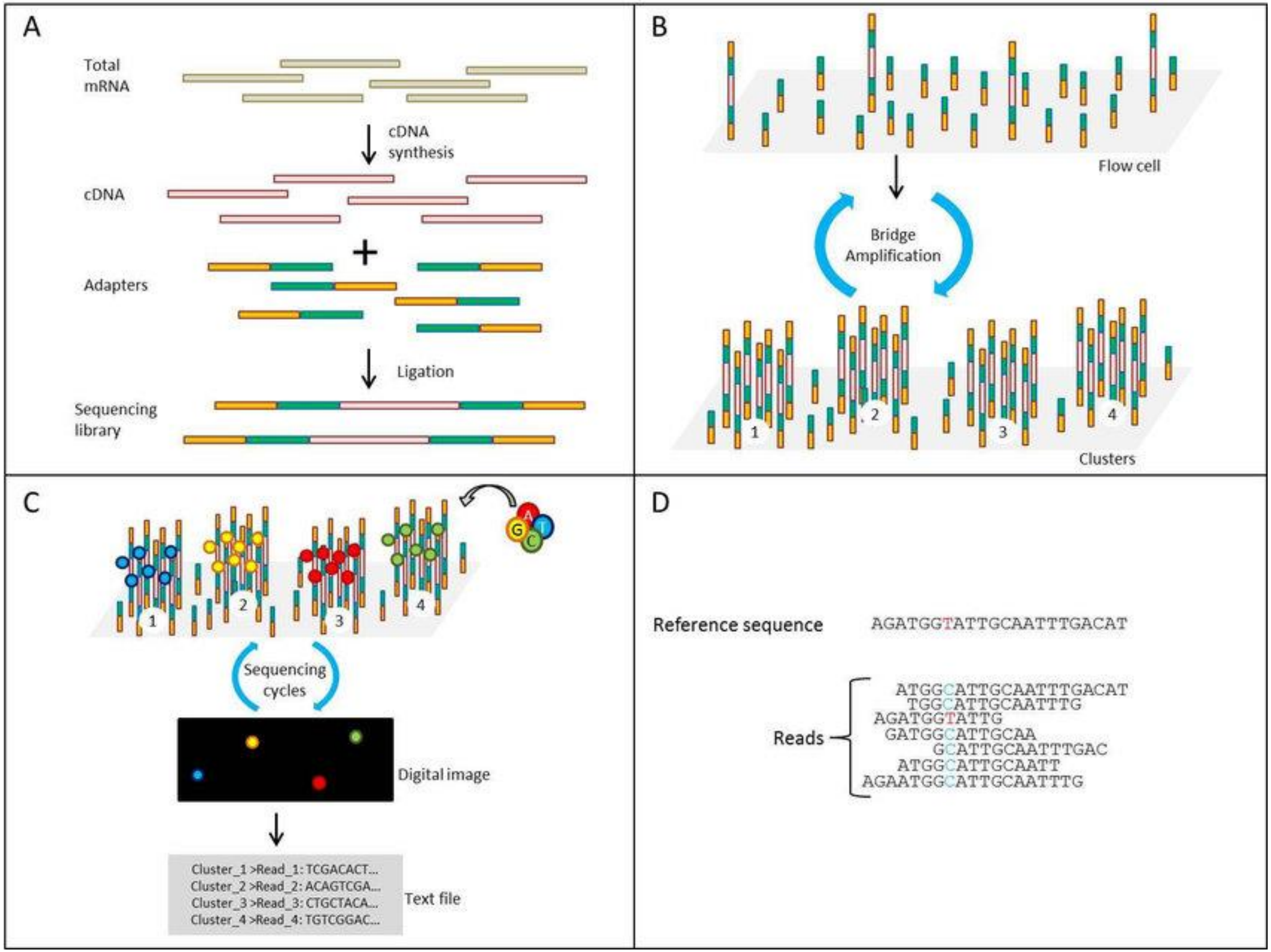
RNA-Seq. *Basics*



Martins et al. (2019). Int. J. Human Environ. Health 16, 4718. (doi: 10.3390/ijerph16234718)

This is called Next-Generation Sequencing (NGS). A similar process applies to genomes.

RNA-Seq. *Basics*



```
Cluster_1 >Read_1: TCGACACT...
Cluster_2 >Read_2: ACAGTCGA...
Cluster_3 >Read_3: CTGCTACA...
Cluster_4 >Read_4: TGTCGGAC...
```

Reference sequence

```
 AGATGGTATTGCAATTTGACAT
```

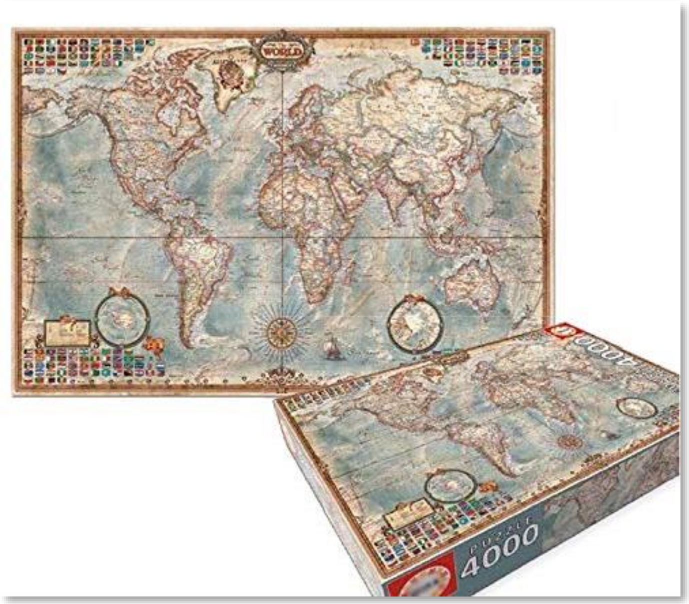
Reads

```
   ATGGCATTGCAATTTGACAT
    TGGCATTGCAATTTG
 AGATGGTATTG
  GATGGCATTGCAA
      GCATTGCAATTTGAC
   ATGGCATTGCAATT
AGAATGGCATTGCAATTTG
```

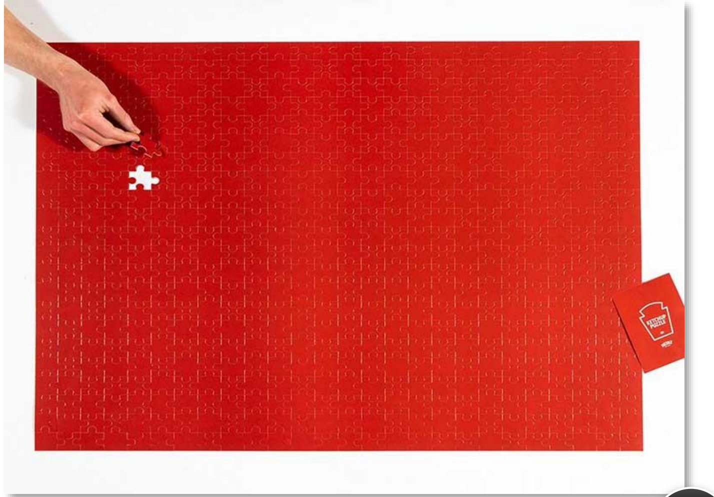
P. Juarez©

Transcriptome mapping and assembly. *Model vs. non-model organisms*

- Model organisms such as humans, rats, mice, zebrafish and a few other benefit from a high degree of genomic resources, including available transcriptomes/genomes against which RNA-Seq raw data can be **mapped**.



- Non-model/novel organisms have limited or null information on gene, peptide or mRNA sequences. In these cases, the transcriptome needs to be ***de novo* assembled**. Pretty much like a 10K+ pieces without a reference photo…



Transcriptome mapping and assembly. *K-mer*

Genome: ACGATGCTCAGACCCCCCCC

Short reads: ACGATGCTCAGA CTCAGACCC AGACCCC CCCCCCC

k-mers:

ACGAT
CGATG
GATGC
ATGCT
TGCTC
GCTCA
CTCAG
TCAGA

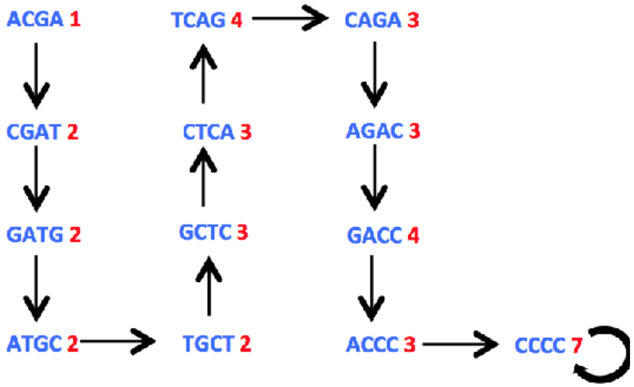
CTCAG
TCAGA
CAGAC
AGACC
GACCC

AGACC
GACCC
ACCCC

CCCCC
CCCCC
CCCCC

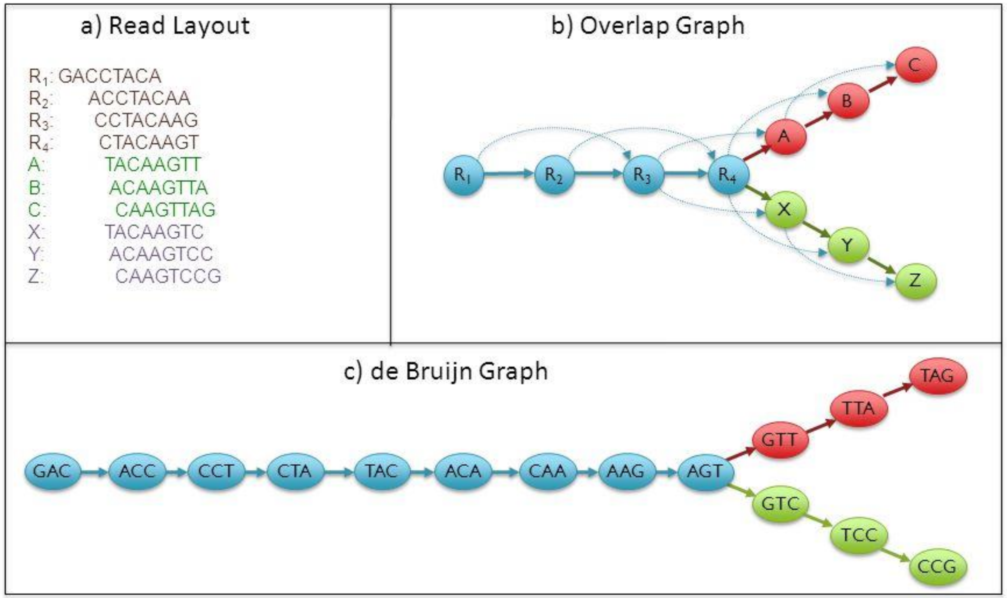
Commonly, k = 25 is the target

De Bruijn graph:



Assembled Contigs: ACGATGCTCAGACCCC

Two Paradigms for Assembly



R_1:	GACCTACA
R_2:	ACCTACAA
R_3:	CCTACAAG
R_4:	CTACAAGT
A:	TACAAGTT
B:	ACAAGTTA
C:	CAAGTTAG
X:	TACAAGTC
Y:	ACAAGTCC
Z:	CAAGTCCG

Assembly of Large Genomes using Second Generation Sequencing
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research* 20, 1165-73.

Trinity. *workflow*

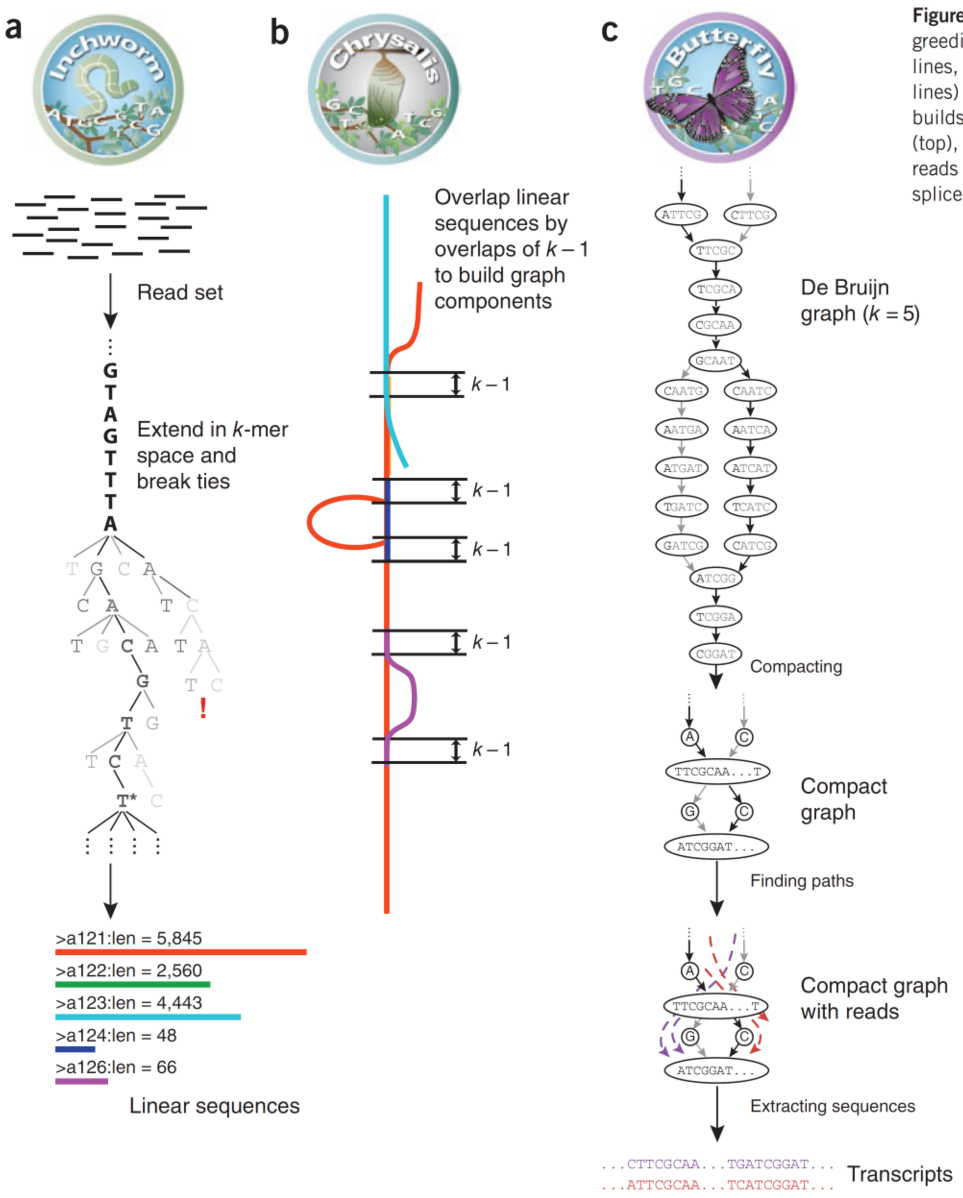


Figure 1 Overview of Trinity. (**a**) Inchworm assembles the read data set (short black lines, top) by greedily searching for paths in a *k*-mer graph (middle), resulting in a collection of linear contigs (color lines, bottom), with each *k*-mer present only once in the contigs. (**b**) Chrysalis pools contigs (colored lines) if they share at least one *k* – 1-mer and if reads span the junction between contigs, and then it builds individual de Bruijn graphs from each pool. (**c**) Butterfly takes each de Bruijn graph from Chrysalis (top), and trims spurious edges and compacts linear paths (middle). It then reconciles the graph with reads (dashed colored arrows, bottom) and pairs (not shown), and outputs one linear sequence for each splice form and/or paralogous transcript represented in the graph (bottom, colored sequences).

Grabherr et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644–652. https://doi.org/10.1038/nbt.1883

Trinity. *workflow*

Let’s get practical and assemble a transcriptome from raw data!

Transcriptome assembly and annotation. *“Lab” Practice!*

Training set

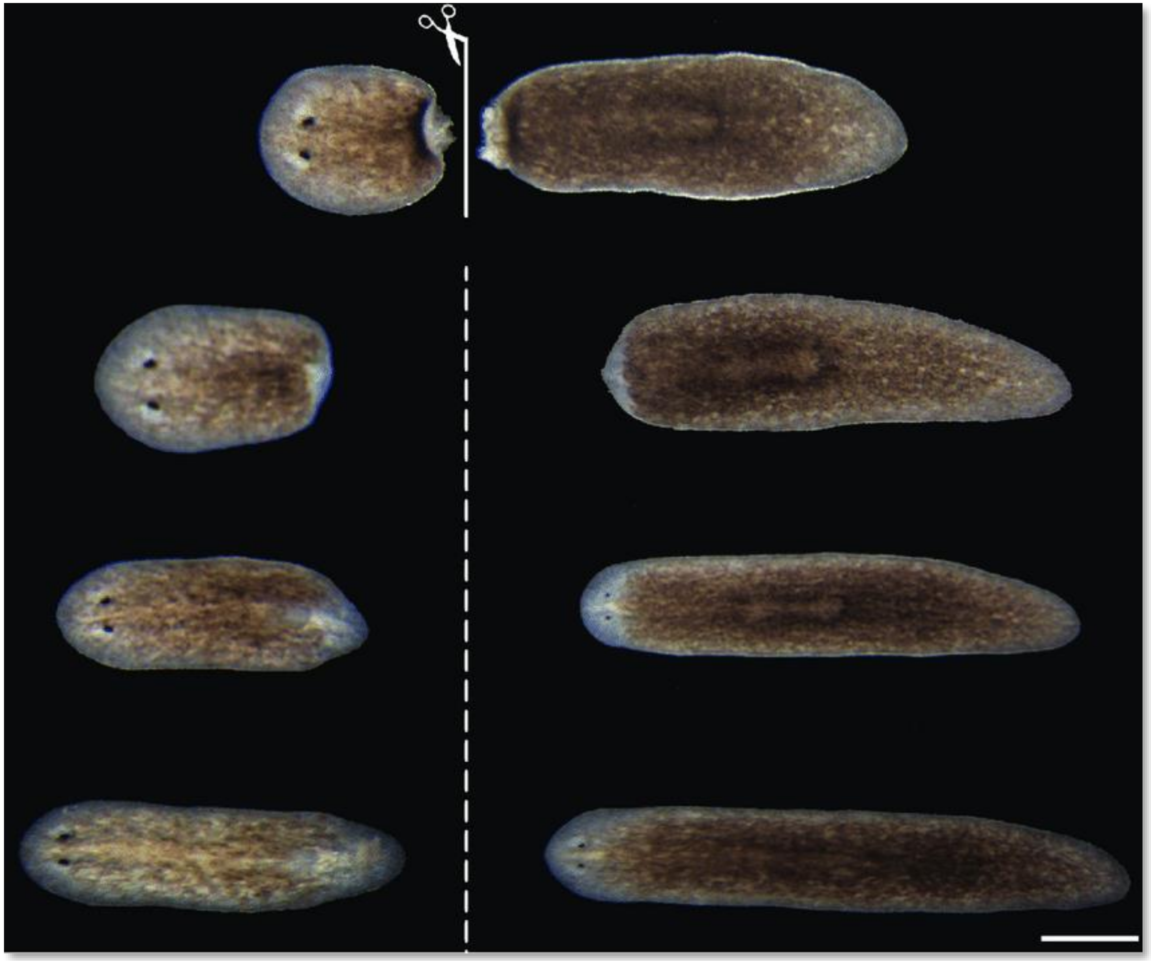
Raw Fastq data set

?M reads (<10)

36 bp paired-end sequencing Illumina platform (but we will work only with L)

GEO GSM767958

Today, we will put aside *quantification* and focus on *identification*



Cebrià et al. (2016). Regeneration and Growth as Modes of Adult Development: the Platyhelminthes as a Case Study. Doi: 10.1007/978-3-7091-1871-9_4

Transcriptome assembly and annotation. *Objectives*

- De novo transcriptome assembly from *SmedIllumina_R1.fastq.gz* using Trinity

- Basic quality assessment using *TrinityStats* (*Nx* analyses)

- Predict ORFs using *Transdecoder*

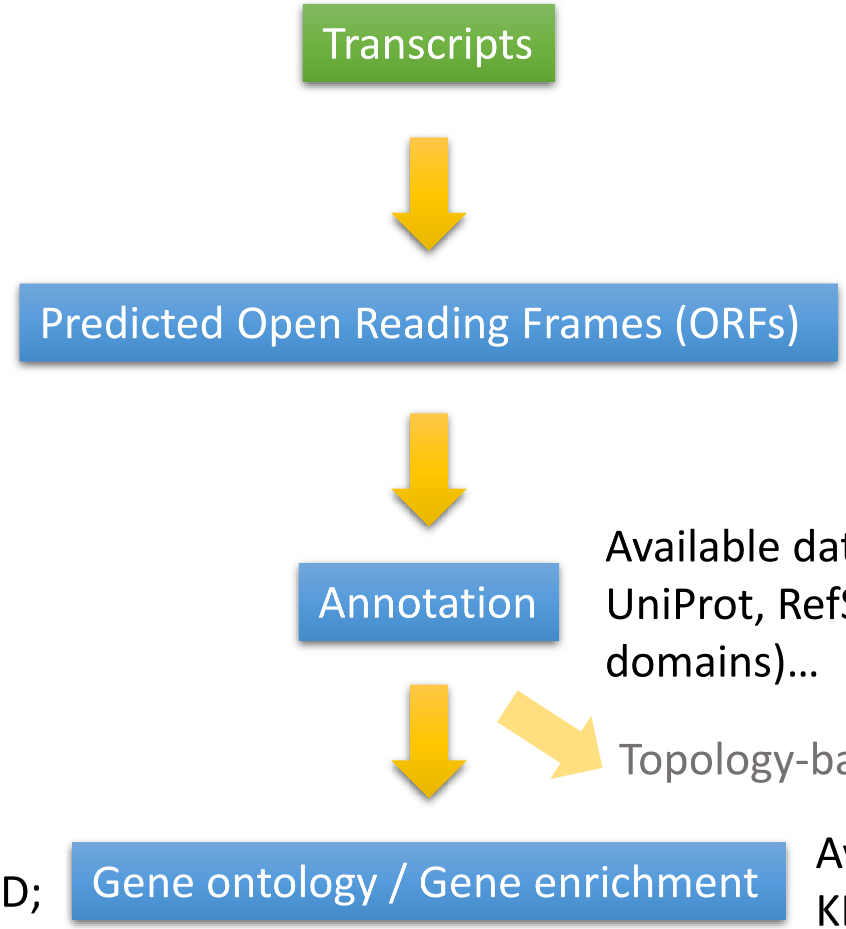
- Annotate the resulting ORFs using *Pblast* (homology-matching):
  -Against UniProt
  -Against Uniprot (Human proteome only)

Trinity. *Constraints to be considered* in novo *assembly*

Fortunately there are tools to clean and normalise data before assembly and to check its quality afterwards…

- Inter-specific RNA contamination
- Adapters and low-quality reads must be removed
- Sequences too short
- Over-represented genes (i.e. high vs low expression genes)
- Heterogenous representation of the transcriptome
- Gene-dense genomes resulting in a large number of transcriptional variants (for instance, resulting from chromosome duplication, cryptic genes…)

Transcriptome annotation. *Standard procedure*



Transcripts

E.g., Transdecoder

Predicted Open Reading Frames (ORFs)

E.g., Blast; HMMER

Annotation

Available databases include UniProt, RefSeq, Pfam (conserved domains)…

Topology-based pathway enrichment

E.g., Blast2GO; DAVID; several R packages

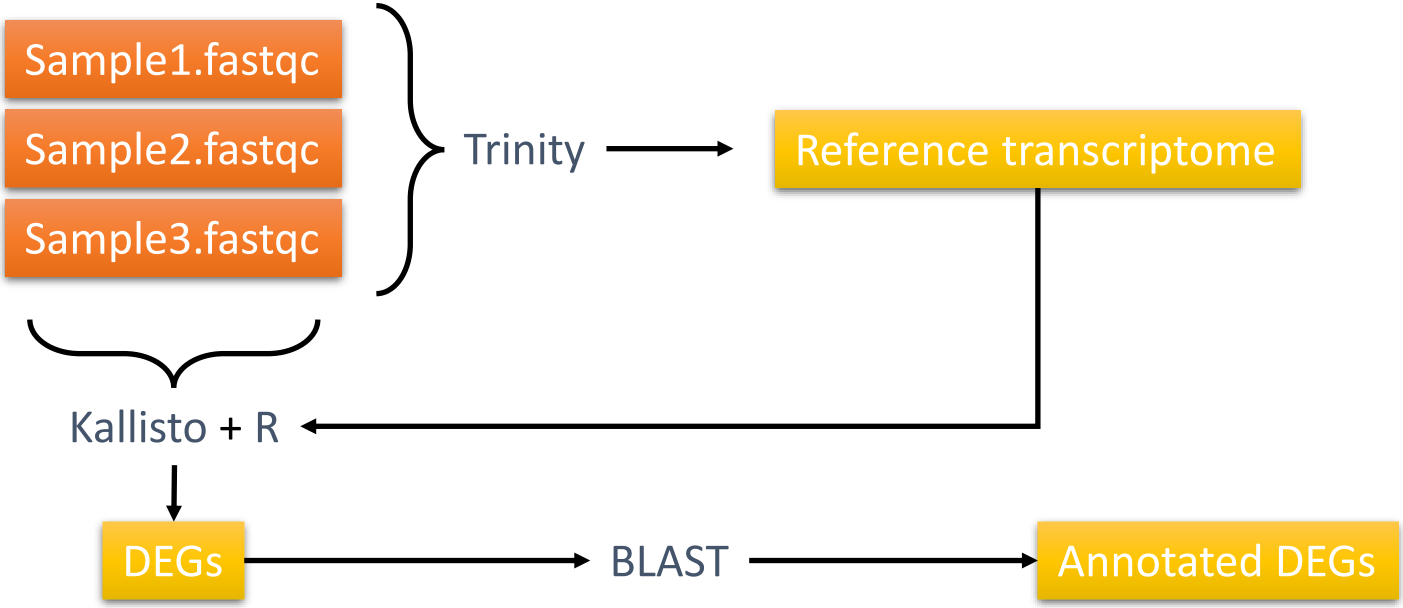
Gene ontology / Gene enrichment

Available databases include KEGG, Reactome…

Transcriptome annotation. Standard procedure

Q: Can we quantify expression when dealing with *de novo* assembly? Yes!

First we use a programme like Trinity to assembly the transcriptome of our tart organism/tissue/organ (this can/should be done using several samples). We will then use this transcriptome as reference for mapping using, e.g. Kallisto.



Sample1.fastqc

Sample2.fastqc

Sample3.fastqc

Trinity

Reference transcriptome

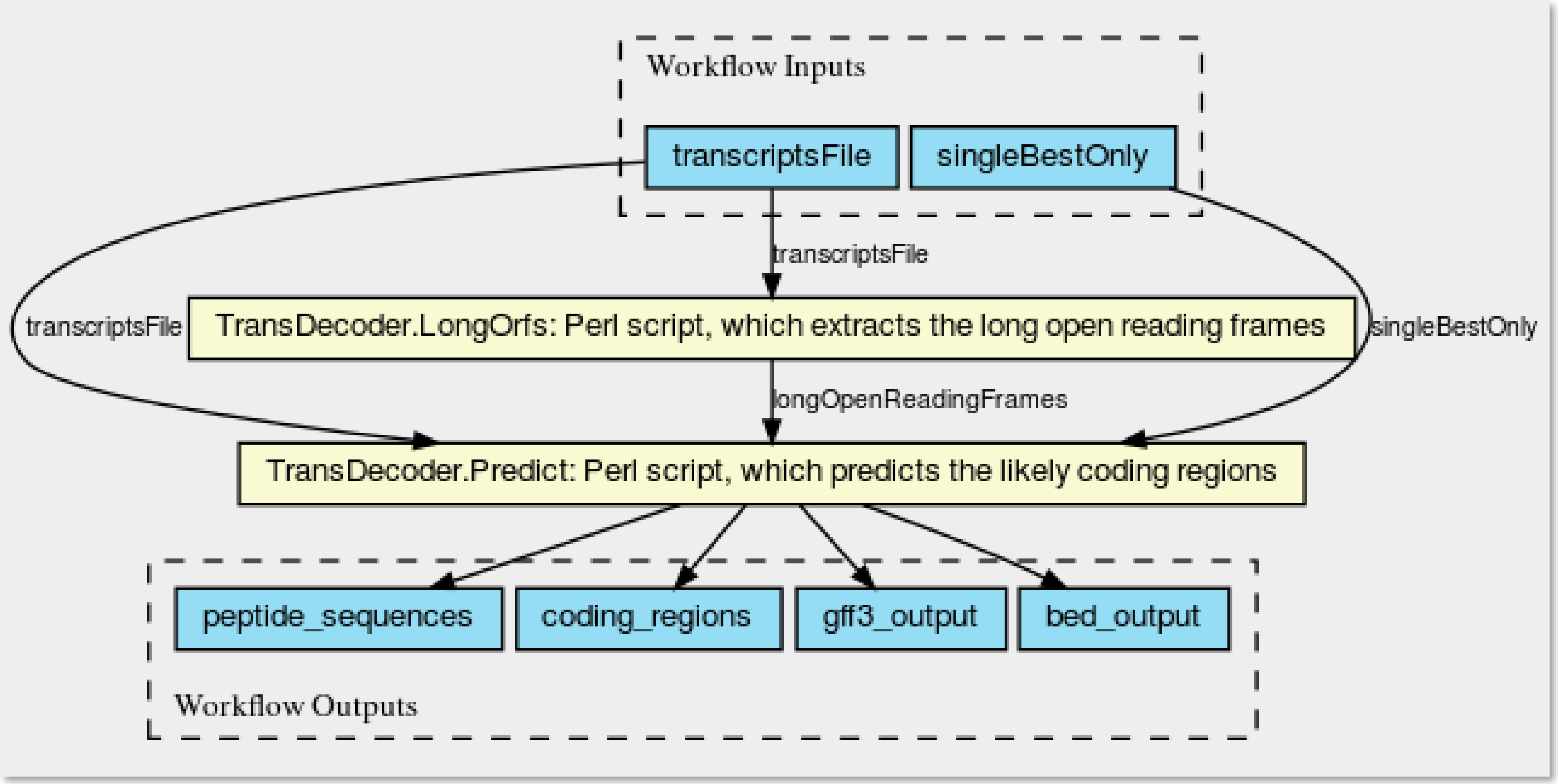
Kallisto + R

DEGs

BLAST

Annotated DEGs

Pre-annotation tools. *Predicting ORFs*

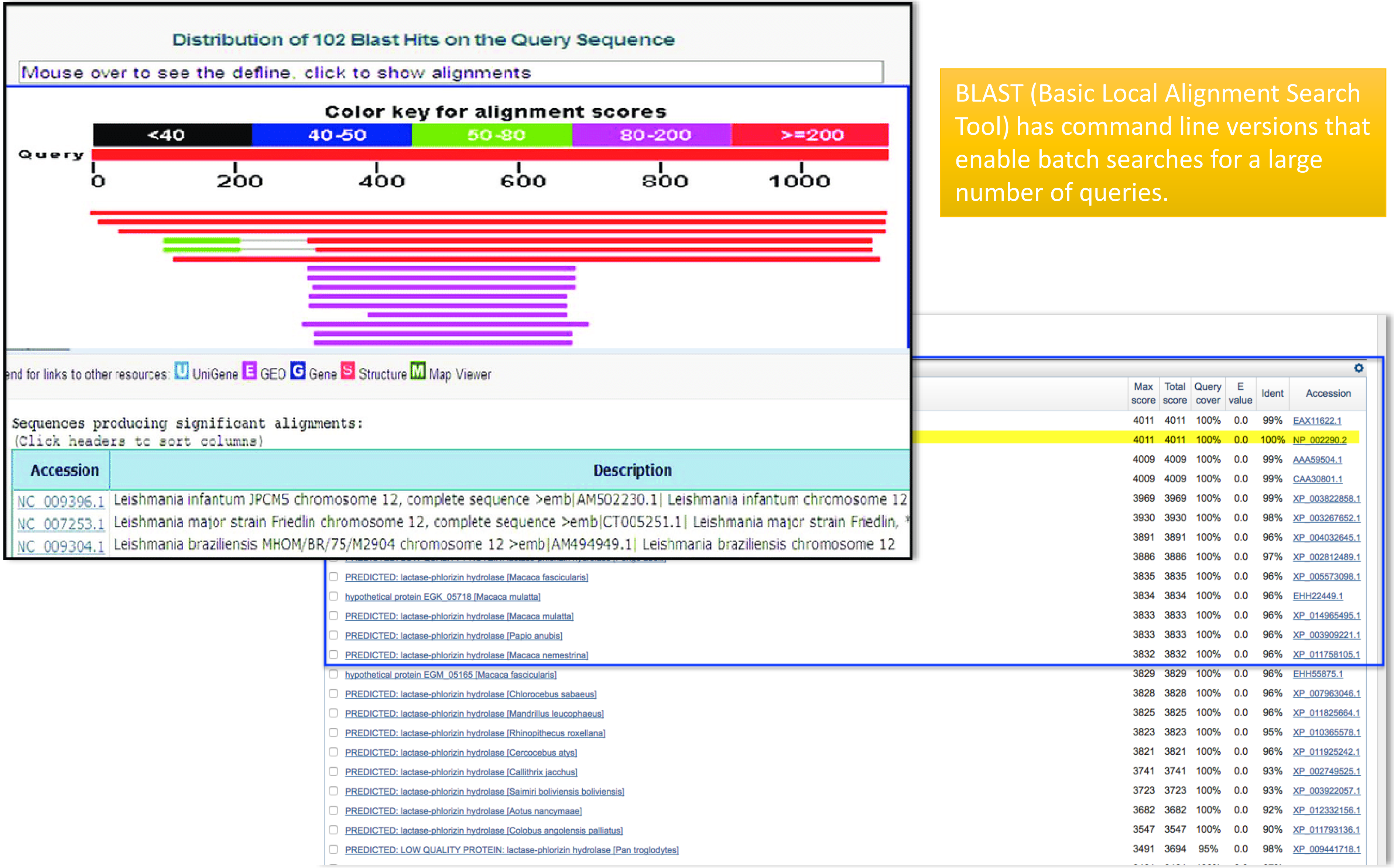


Transdecoder workflow

Homology-matching can be done with cDNA or AA sequences. However, the later can filter sequences by isolating coding from non-coding and reducing variability.

Annotation tools. *Protein BLAST*

BLAST (Basic Local Alignment Search Tool) has command line versions that enable batch searches for a large number of queries.

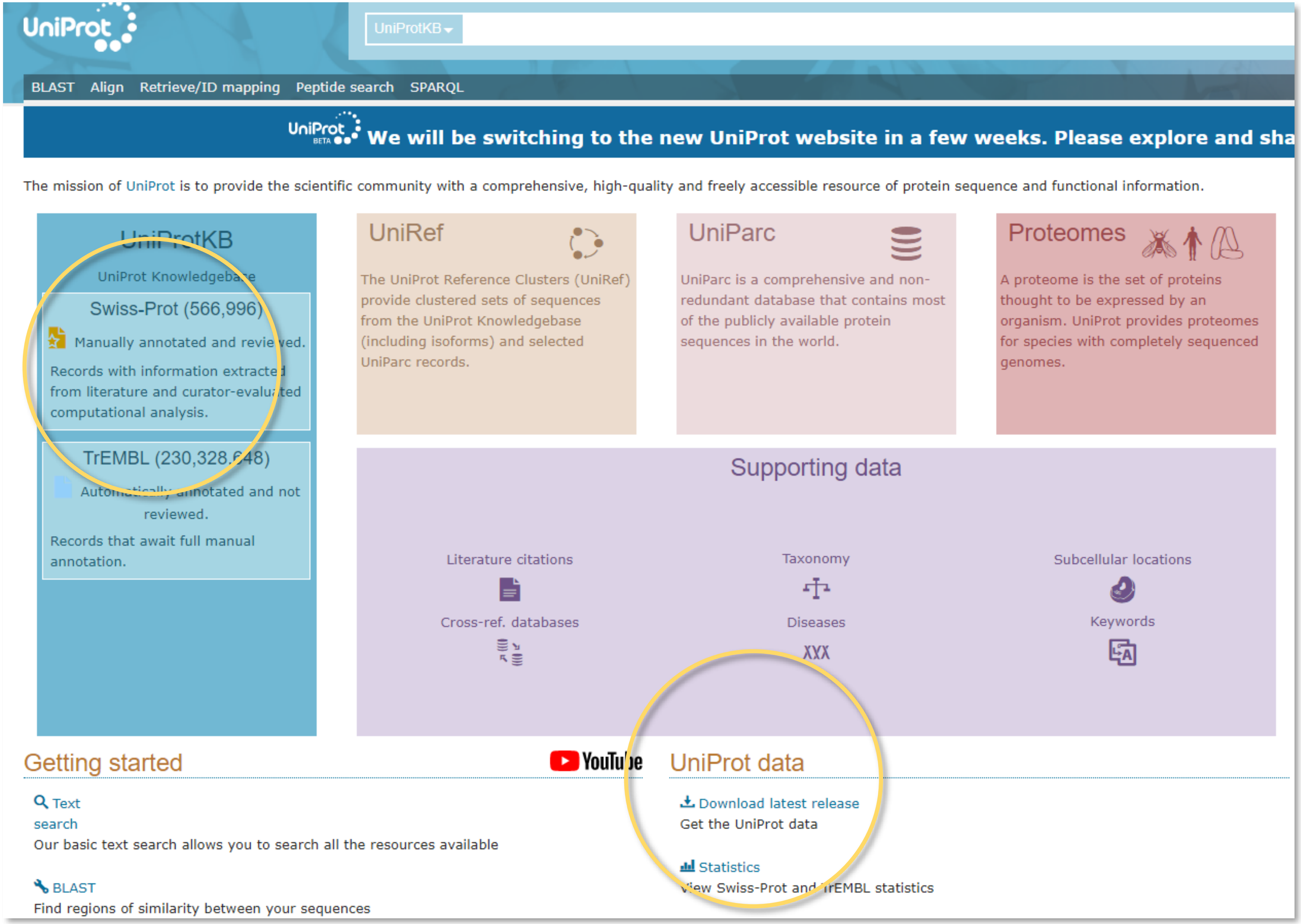

Distribution of 102 Blast Hits on the Query Sequence
Mouse over to see the define, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
200
400
600
800
1000
end for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description
NC_009396.1	Leishmania infantum JPCM5 chromosome 12, complete sequence >emb|AM502230.1| Leishmania infantum chromosome 12
NC_007253.1	Leishmania major strain Friedlin chromosome 12, complete sequence >emb|CT005251.1| Leishmania major strain Friedlin, *
NC_009304.1	Leishmania braziliensis MHOM/BR/75/M2904 chromosome 12 >emb|AM494949.1| Leishmania braziliensis chromosome 12

	Max score	Total score	Query cover	E value	Ident	Accession
	4011	4011	100%	0.0	99%	EAX11622.1
	4011	4011	100%	0.0	100%	NP_002290.2
	4009	4009	100%	0.0	99%	AAA59504.1
	4009	4009	100%	0.0	99%	CAA30801.1
	3969	3969	100%	0.0	99%	XP_003822858.1
	3930	3930	100%	0.0	98%	XP_003267652.1
	3891	3891	100%	0.0	96%	XP_004032645.1
	3886	3886	100%	0.0	97%	XP_002812489.1
□ PREDICTED: lactase-phlorizin hydrolase [Macaca fascicularis]	3835	3835	100%	0.0	96%	XP_005573098.1
□ hypothetical protein EGK_05718 [Macaca mulatta]	3834	3834	100%	0.0	96%	EHH22449.1
□ PREDICTED: lactase-phlorizin hydrolase [Macaca mulatta]	3833	3833	100%	0.0	96%	XP_014965495.1
□ PREDICTED: lactase-phlorizin hydrolase [Papio anubis]	3833	3833	100%	0.0	96%	XP_003909221.1
□ PREDICTED: lactase-phlorizin hydrolase [Macaca nemestrina]	3832	3832	100%	0.0	96%	XP_011758105.1
□ hypothetical protein EGM_05165 [Macaca fascicularis]	3829	3829	100%	0.0	96%	EHH55875.1
□ PREDICTED: lactase-phlorizin hydrolase [Chlorocebus sabaeus]	3828	3828	100%	0.0	96%	XP_007963046.1
□ PREDICTED: lactase-phlorizin hydrolase [Mandrillus leucophaeus]	3825	3825	100%	0.0	96%	XP_011825664.1
□ PREDICTED: lactase-phlorizin hydrolase [Rhinopithecus roxellana]	3823	3823	100%	0.0	95%	XP_010365578.1
□ PREDICTED: lactase-phlorizin hydrolase [Cercocebus atys]	3821	3821	100%	0.0	96%	XP_011925242.1
□ PREDICTED: lactase-phlorizin hydrolase [Callithrix jacchus]	3741	3741	100%	0.0	93%	XP_002749525.1
□ PREDICTED: lactase-phlorizin hydrolase [Saimiri boliviensis boliviensis]	3723	3723	100%	0.0	93%	XP_003922057.1
□ PREDICTED: lactase-phlorizin hydrolase [Aotus nancymaae]	3682	3682	100%	0.0	92%	XP_012332156.1
□ PREDICTED: lactase-phlorizin hydrolase [Colobus angolensis palliatus]	3547	3547	100%	0.0	90%	XP_011793136.1
□ PREDICTED: LOW QUALITY PROTEIN: lactase-phlorizin hydrolase [Pan troglodytes]	3491	3694	95%	0.0	98%	XP_009441718.1

Databases for ORF annotation. *Uniprot*



We will be switching to the new UniProt website in a few weeks. Please explore and sha

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB

UniProt Knowledgebase

Swiss-Prot (566,996)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (230,328,648)

Automatically annotated and not reviewed.

Records that await full manual annotation.

UniRef

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

UniParc

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

Proteomes

A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

Supporting data

Literature citations
Taxonomy
Subcellular locations
Cross-ref. databases
Diseases
Keywords

Getting started

Text search

Our basic text search allows you to search all the resources available

BLAST

Find regions of similarity between your sequences

UniProt data

Download latest release

Get the UniProt data

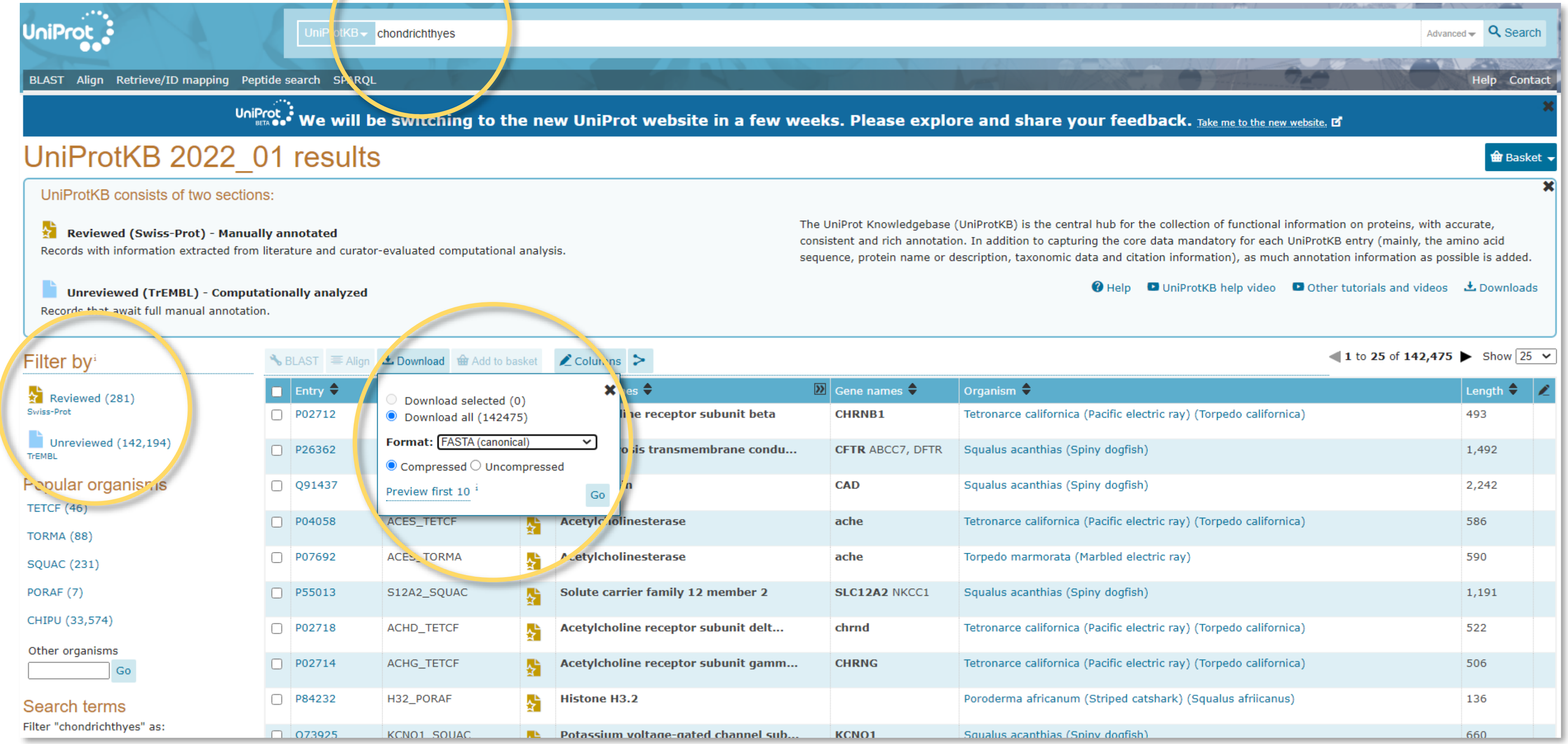
Statistics

View Swiss-Prot and TrEMBL statistics

https://www.uniprot.org/

Databases for ORF annotation. *Uniprot*

You can easily customise and download a database from UniProt's website



chondrichthyes

We will be switching to the new UniProt website in a few weeks. Please explore and share your feedback. Take me to the new website.

UniProtKB 2022_01 results

UniProtKB consists of two sections:

Reviewed (Swiss-Prot) - Manually annotated
Records with information extracted from literature and curator-evaluated computational analysis.

Unreviewed (TrEMBL) - Computationally analyzed
Records that await full manual annotation.

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.

Filter by

Reviewed (281)
Swiss-Prot

Unreviewed (142,194)
TrEMBL

Popular organisms

TETCF (46)
TORMA (88)
SQUAC (231)
PORAF (7)
CHIPU (33,574)

Other organisms
Go

Search terms

Filter "chondrichthyes" as:

1 to 25 of 142,475 ▶ Show 25

○ Download selected (0)
● Download all (142475)

Format: FASTA (canonical)

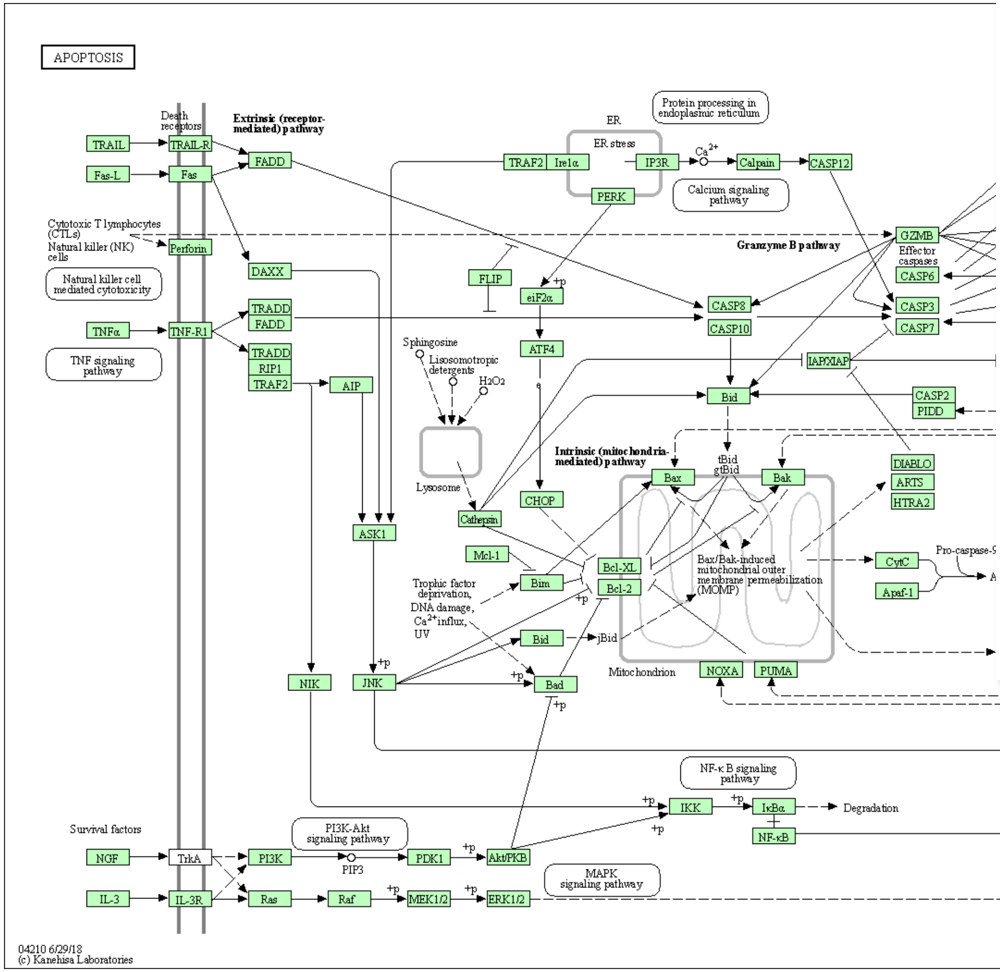
● Compressed ○ Uncompressed

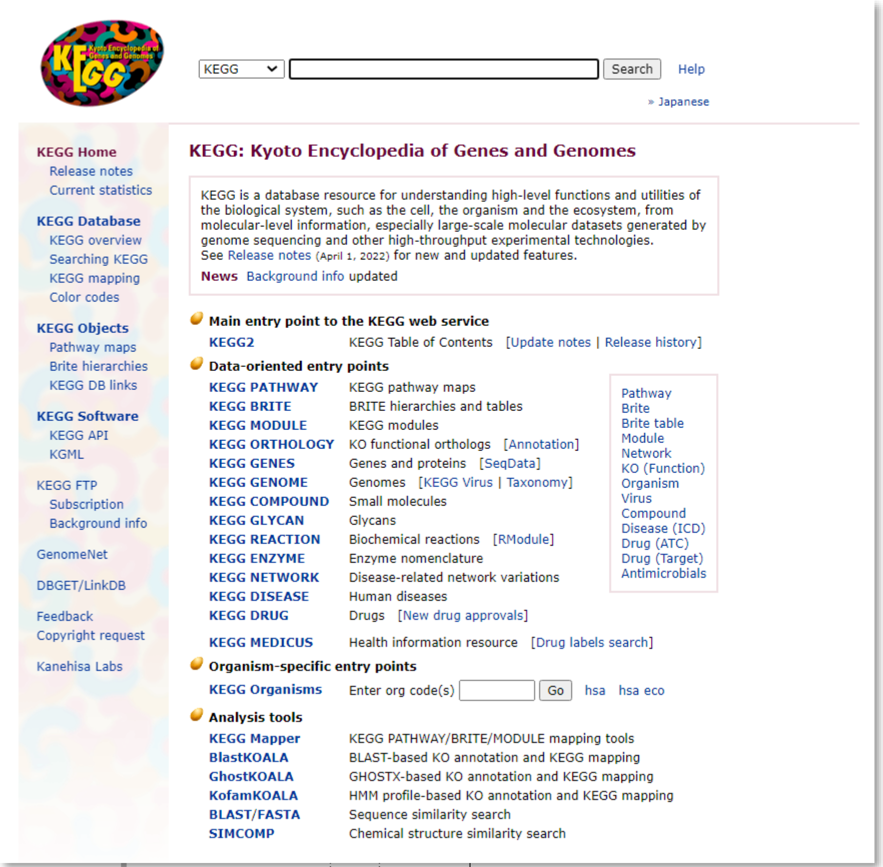
Preview first 10

Entry		es	Gene names	Organism	Length
P02712		ine receptor subunit beta	CHRNB1	Tetronarce californica (Pacific electric ray) (Torpedo californica)	493
P26362		o is transmembrane condu...	CFTR ABCC7, DFTR	Squalus acanthias (Spiny dogfish)	1,492
Q91437		n	CAD	Squalus acanthias (Spiny dogfish)	2,242
P04058	ACES_TETCF	Acetylcholinesterase	ache	Tetronarce californica (Pacific electric ray) (Torpedo californica)	586
P07692	ACES_TORMA	Acetylcholinesterase	ache	Torpedo marmorata (Marbled electric ray)	590
P55013	S12A2_SQUAC	Solute carrier family 12 member 2	SLC12A2 NKCC1	Squalus acanthias (Spiny dogfish)	1,191
P02718	ACHD_TETCF	Acetylcholine receptor subunit delt...	chrnd	Tetronarce californica (Pacific electric ray) (Torpedo californica)	522
P02714	ACHG_TETCF	Acetylcholine receptor subunit gamm...	CHRNG	Tetronarce californica (Pacific electric ray) (Torpedo californica)	506
P84232	H32_PORAF	Histone H3.2		Poroderma africanum (Striped catshark) (Squalus afriicanus)	136
Q73925	KCNQ1_SQUAC	Potassium voltage-gated channel sub...	KCNQ1	Squalus acanthias (Spiny dogfish)	660

Post-annotation tools. *Gene enrichment*

https://www.genome.jp/kegg/





KEGG ▾ Search Help

» Japanese

KEGG Home
Release notes
Current statistics

KEGG Database
KEGG overview
Searching KEGG
KEGG mapping
Color codes

KEGG Objects
Pathway maps
Brite hierarchies
KEGG DB links

KEGG Software
KEGG API
KGML

KEGG FTP
Subscription
Background info

GenomeNet

DBGET/LinkDB

Feedback
Copyright request

Kanehisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.
See Release notes (April 1, 2022) for new and updated features.

News Background info updated

Main entry point to the KEGG web service

KEGG2	KEGG Table of Contents [Update notes | Release history]

Data-oriented entry points

KEGG PATHWAY	KEGG pathway maps
KEGG BRITE	BRITE hierarchies and tables
KEGG MODULE	KEGG modules
KEGG ORTHOLOGY	KO functional orthologs [Annotation]
KEGG GENES	Genes and proteins [SeqData]
KEGG GENOME	Genomes [KEGG Virus | Taxonomy]
KEGG COMPOUND	Small molecules
KEGG GLYCAN	Glycans
KEGG REACTION	Biochemical reactions [RModule]
KEGG ENZYME	Enzyme nomenclature
KEGG NETWORK	Disease-related network variations
KEGG DISEASE	Human diseases
KEGG DRUG	Drugs [New drug approvals]
KEGG MEDICUS	Health information resource [Drug labels search]

Pathway
Brite
Brite table
Module
Network
KO (Function)
Organism
Virus
Compound
Disease (ICD)
Drug (ATC)
Drug (Target)
Antimicrobials

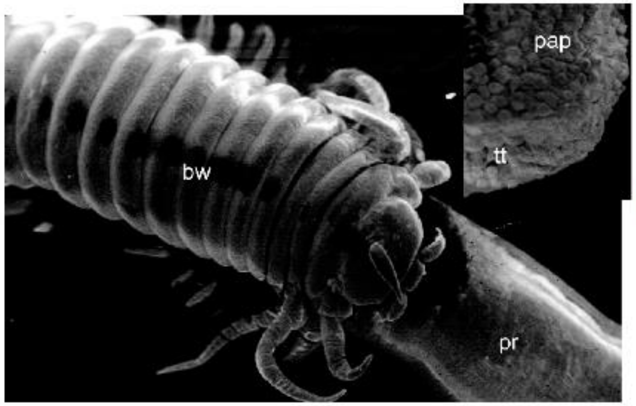
Organism-specific entry points

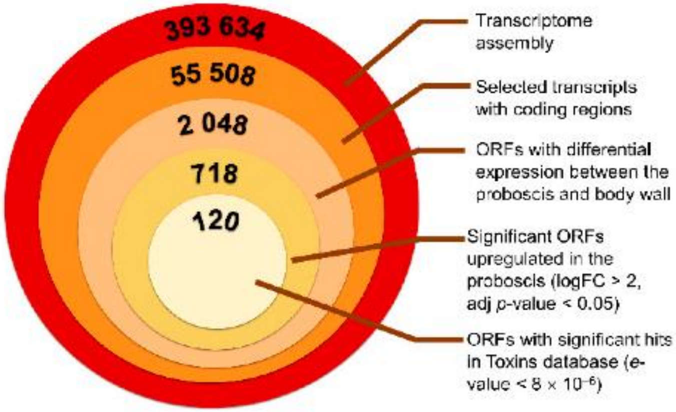
KEGG Organisms	Enter org code(s) Go hsa hsa eco

Analysis tools

KEGG Mapper	KEGG PATHWAY/BRITE/MODULE mapping tools
BlastKOALA	BLAST-based KO annotation and KEGG mapping
GhostKOALA	GHOSTX-based KO annotation and KEGG mapping
KofamKOALA	HMM profile-based KO annotation and KEGG mapping
BLAST/FASTA	Sequence similarity search
SIMCOMP	Chemical structure similarity search

De novo assembly and annotation. *Example*





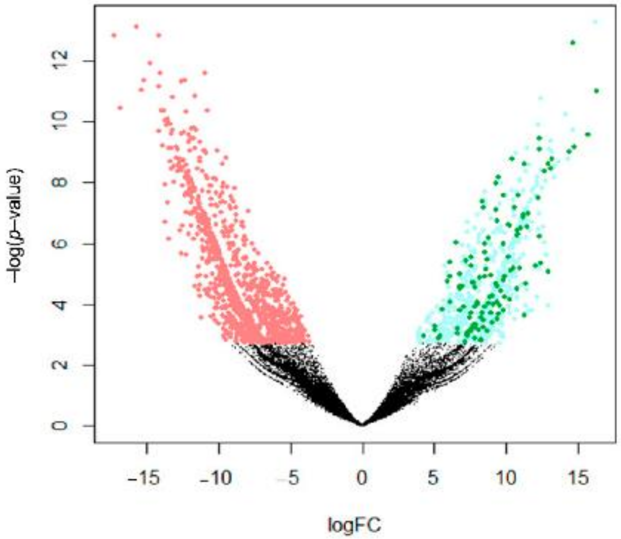
Transcriptome assembly

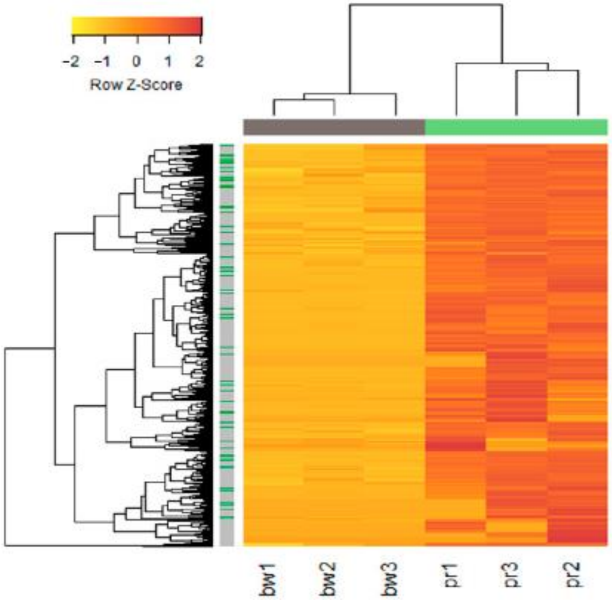
Selected transcripts with coding regions

ORFs with differential expression between the proboscis and body wall

Significant ORFs upregulated in the proboscis (logFC > 2, adj *p*-value < 0.05)

ORFs with significant hits in Toxins database (e-value < 8 × 10^{-6})





Rodrigo et al. (2021). Toxins 13, 97