De novo Transcriptome Assembly and Associated Studies, with Applications Towards Non-model Organisms

Matthew MacManes
University of New Hampshire
Matthew.MacManes@unh.edu
@MacManes

Who am I?

- Associate Prof at Univ of New Hampshire (USA)
- Lab focus:
  - Ecological, physiological, comparative genomics of non-model (mostly) vertebrates
  - Software development, mostly in assembly methods
  - Always looking for grad students/postdocs

Transcriptome assembly

Transcriptome Assembly Lecture Overview

- General issues in
  - Study design
  - Sample collection and preservation
  - Library Generation
  - Sequencing and Raw data QC
  - Assembly and QC
  - Downstream analysis

Why do I care about today?

- You don't have a reference for your organism
- The existing reference is sh!t
- The specific gene(s) you're interested in are missing

- Because you trust no one, and want to know how to tell if the existing reference is good

Assembled transcriptome may likely form the foundation for months or years of downstream work.

Transcriptome Assembly Lecture Overview

- Explicit Goals
  - Learn about transcriptome assembly
  - To get you thinking about experimental bioinformatics, you are doing an experiment, after all, just like you might do in the wet lab or field
  - The answer to most questions is, “do an experiment”

Transcriptome Assembly Lecture Overview

- Implicit Goals
  - Computational note taking
  - Version control (GitHub)

Transcriptome Assembly Lecture Overview

- General issues in
  - Study design
  - Sample collection and preservation
  - Library Generation
  - Sequencing and Raw data QC
  - Assembly and QC
  - Downstream analysis

Study Design - which platform?

Platform	Project Firefly 2018	MiniSeq	MiSeq	Next Seq 550	HiSeq 2500 RR	Hiseq 2500 V3	HiSeq 2500 V4	HiSeq 4000	HiSeq X	Nova Seq S1 2018	Nova Seq S2	Nova Seq S4	5500 XL	318 HiQ 520	Ion 530	Ion Proton P1	PGM HiQ 540	RS P6-C4	Sequel	R&D end 2018	Smidg ION RnD	Mini ION R9.5	Grid ION X5	Prome thION RnD	Prome thION theor etical	QiaGen Gene Reader	BGI SEQ 500	BGI SEQ 50	#
Reads: (M)	4	25	25	400	600	3000	4000	5000	6000	3300	6600	20000	1400	3-5	15-20	165	60-80	5.5	38.5	--	--	--	--	--	--	400	1600	1600	--
Read length: (paired-end*)	150*	150*	300*	150*	100*	100*	125*	150*	150*	150*	150*	150*	60	200 400	200 400	200	200	15K	12K	32K	--	--	--	--	--	--	100*	50	--
Run time: (d)	0.54	1	2	1.2	1.125	11	6	3.5	3	1.66	1.66	1.66	7	0.37	0.16	--	0.16	4.3	--	--	--	2	2	2	--	--	1	0.4	--
Yield: (Gb)	1	7.5	15	120	120	600	1000	1500	1800	1000	2000	6000	180	1.5	7	10	12	12	5	150	4	8	40	2400	11000	80	200	8	--
Rate: (Gb/d)	1.85	7.5	7.5	100	106.6	55	166	400	600	600	1200	3600	30	5.5	50	--	93.75	2.8	--	--	--	4	20	1200	5500	--	200	20	--
Reagents: ($K)	0.1	1.75	1	5	6.145	23.47	29.9	--	--	--	--	--	10.5	0.6	--	1	1.2	2.4	--	1	--	0.5	1.5	--	--	0.5	--	--	--
per-Gb: ($)	100	233	66	50	51.2	39.1	31.7	20.5	7.08	18	15	5.8	58.33	--	--	100	--	200	80	6.6	--	62.5	37.5	20	4.3	--	--	--	--
hg-30x: ($)	12000	28000	8000	5000	6144	4692	3804	2460	849.6	1800	1564	700	7000	--	--	12000	--	24000	9600	1000	--	7500	4500	2400	500	--	600	--	
Machine: ($)	30K	49.5K	99K	250K	740K	690K	690K	900K	1M	999K	999K	999K	595K	50K	65K	243K	242K	695K	350K	350K	--	--	125K	75K	75K	--	200K	--	--

#Page maintained by http://twitter.com/albertvilella http://tinyurl.com/ngslytics #Editable version: http://tinyurl.com/ngsspecsshared

#curl "https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/export?gid=4&format=csv" | grep -v '^#' | grep

*Not all shown at scale

Study Design - which platform?

Study Design - which platform?

Platform	Throughout	Accuracy	ReadLength	Molecule	Cost
Illumina	High	High	Short	cDNA	Low, but the machine
PacBio	Med	Med	Long	cDNA	High
Nanopore	Low	Med	Longer	RNA/cDNA	Low?

Study Design - which platform?

Platform	Ideal Use, Unlimited Money
Illumina	DE/Quant
PacBio	Assembly
Nanopore	Assembly/nt mod

Study Design - samples

- Think about the question you're trying to answer, as RNAseq (and assembly) is tissue/time/condition dependent.

Technical vs. Biological Replicates

RNA-Seq Technical replicates aren't essential

(Technical variation is well-modeled by the Poisson distribution)

Marioni et al., Genome Research, 2008

However, biological replicates *ARE* essential

total_variance = technical_variance + biological_variance

(Total variance well-modeled by negative binomial distribution)

"… at least six biological replicates should be used, rising to at least 12 when it is important to identify SDE genes for all fold changes."
Schurch et al., RNA, 2016

Study Design

- Assembly: use 1 individual, or 1 per treatment

Transcriptome Assembly Lecture Overview

- General issues in
  - Study design
  - Sample collection and preservation
  - Library Generation
  - Sequencing and Raw data QC
  - Assembly and QC
  - Downstream analysis

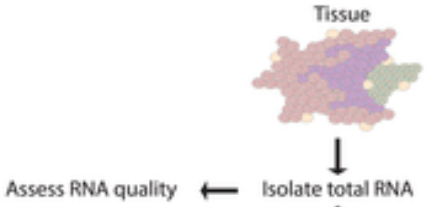
Sample Collection

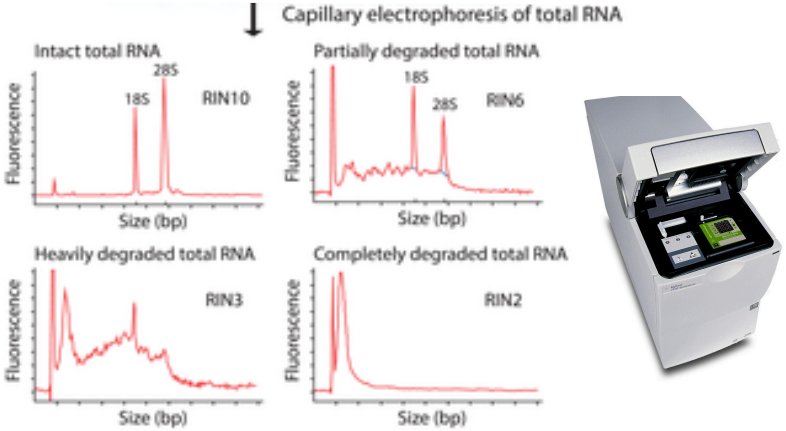
- Transcription changes quickly, may not stop at organismal death…
- Worry about RNA degradation
- The specifics of tissue preservation are taxon specific. I can talk about animal tissues if it's relevant.

Transcriptome Assembly Lecture Overview

- General issues in
  - Study design
  - Sample collection and preservation
  - Library Generation
  - Sequencing and Raw data QC
  - Assembly and QC
  - Downstream analysis

RNA degradation may influence assembly quality

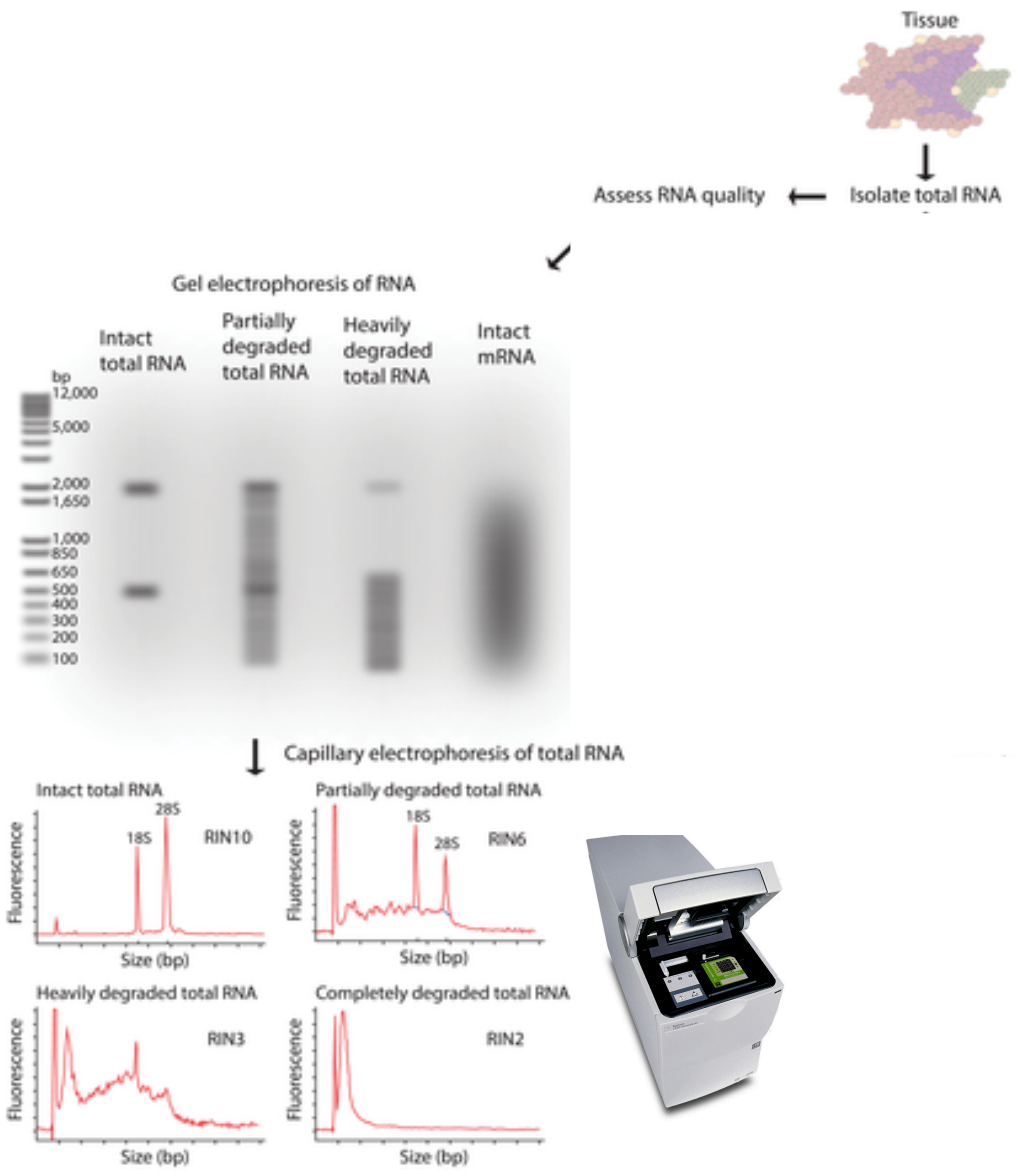




http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

PLOS | COMPUTATIONAL BIOLOGY

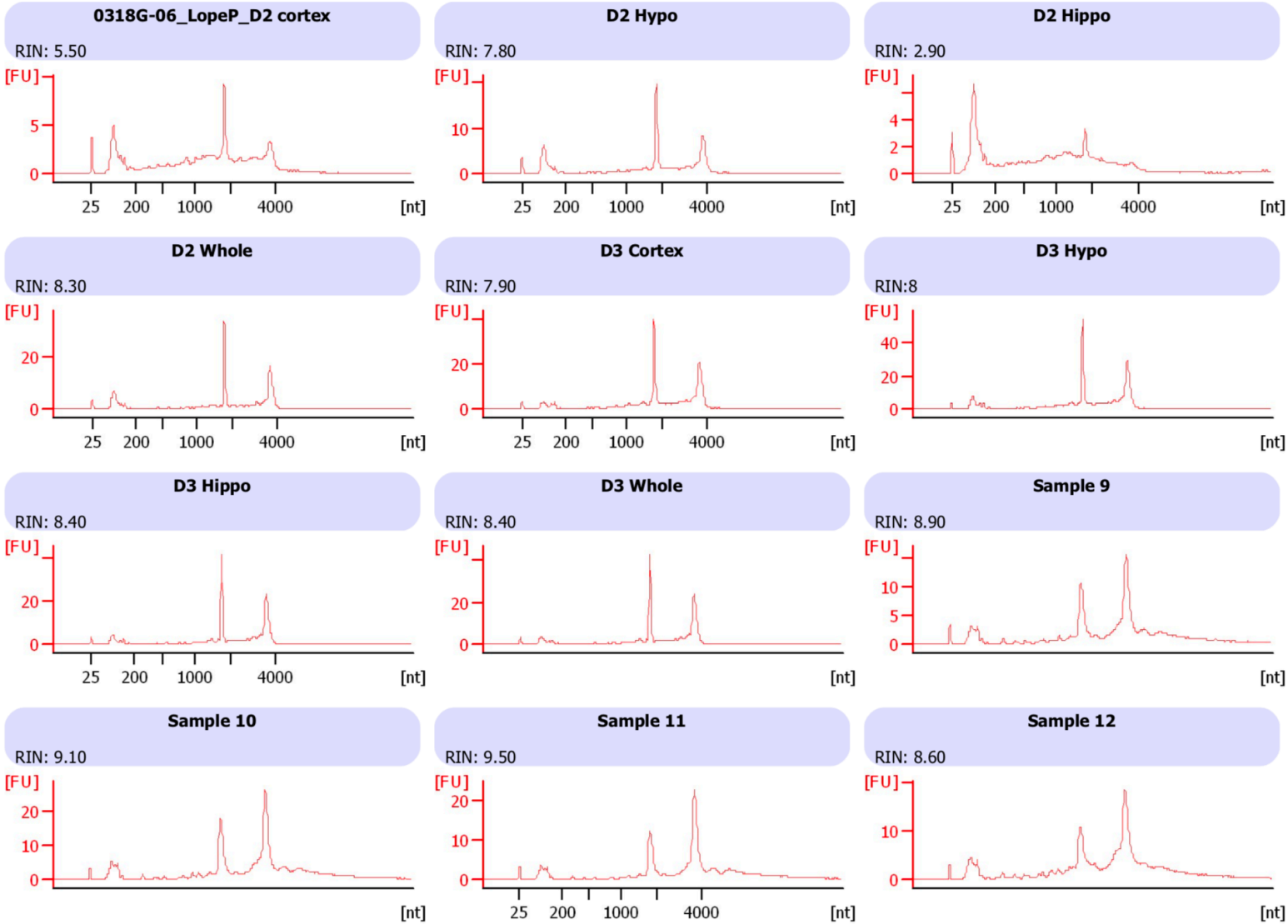
RNA degradation may influence assembly quality



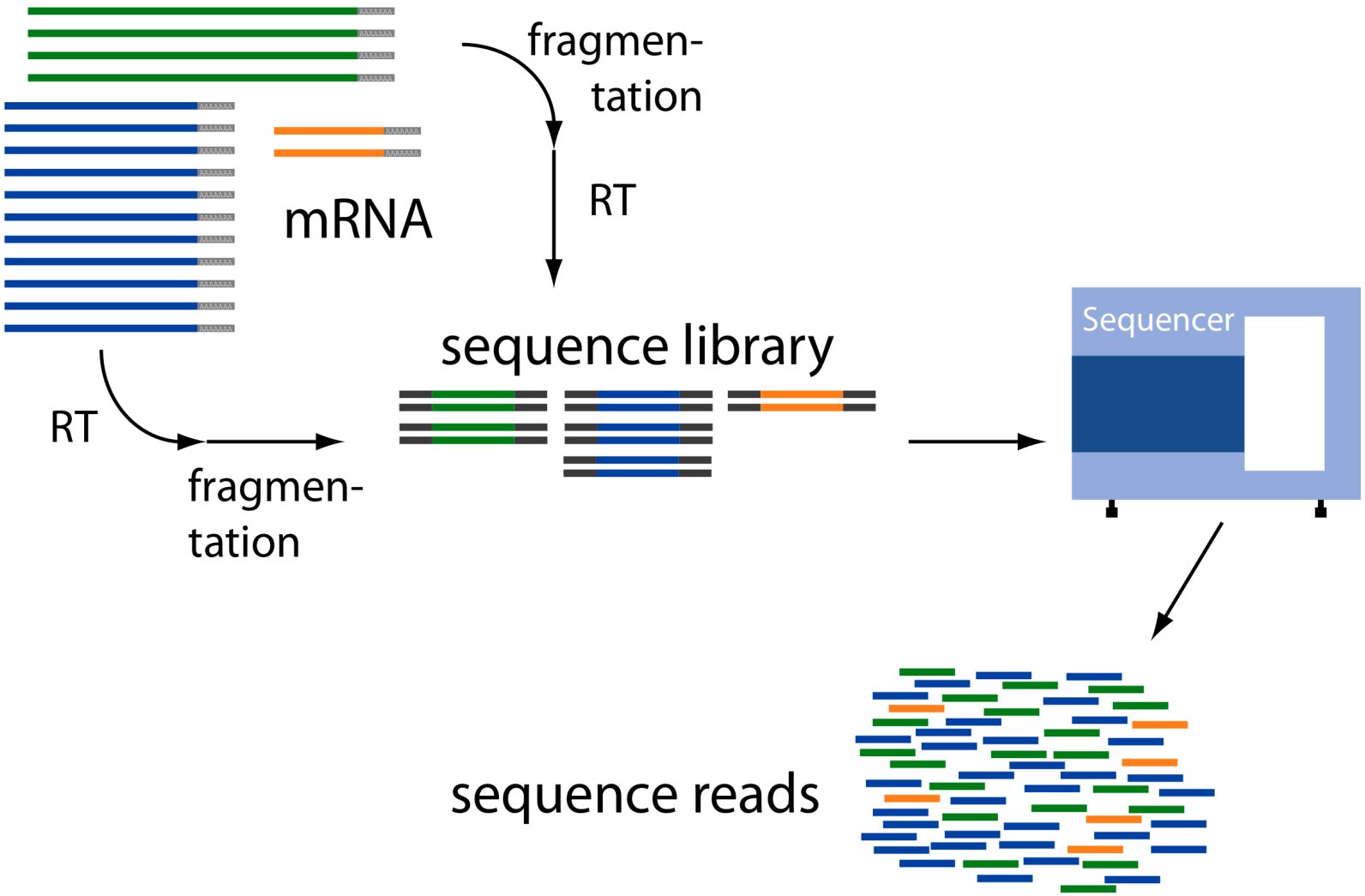
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

PLOS | COMPUTATIONAL BIOLOGY

RNA degradation may influence assembly quality



Overview of RNA-Seq



From: http://www2.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics.html

Transcriptome Assembly Lecture Overview

- General issues in
  - Study design
  - Sample collection and preservation
  - Library Generation
  - Sequencing and Raw data QC
  - Assembly and QC
  - Downstream analysis

Common Data Formats for RNA-Seq

fastA format:

```
>61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
```

fastQ format:

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCCC@@CACCCCCA
```

https://en.wikipedia.org/wiki/FASTQ_format

https://en.wikipedia.org/wiki/Phred_quality_score

Paired-end Sequences (Illumina)

Two FastQ files, read name indicates left (/1) or right (/2) read of paired-end

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

```
@61DFRAAXX100204:1:100:10494:3070/2
CTCAAATGGTTAATTCTCAGGCTGCAAATATTCGTTCAGGATGGAAGAACA
+
C<CCCCCCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBCCCC
```

How many (Illumina) reads??

- For assembly?
  - It depends..
  - 20-40M read pairs is usually sufficient
  - Sequence 1 individual per biological 'unit'

- For quant
  - It depends..
  - Tradeoff between # of replicates and depth per replicate, **choose inc. replicates**.
  - We typically seq to 5-10M pairs

How long (Illumina) reads??

- For assembly?
  - It depends..
  - Longer is better, if you can use length

- For quant
  - It depends..
  - Do you have a good genome?
  - Do you care about splicing?

Strand-specific RNA-Seq is Preferred

Computationally: fewer confounding graph structures in de novo assembly:

Biologically: separate sense vs. antisense transcription

NATURE METHODS | VOL.7 NO.9 | SEPTEMBER 2010 |

BROAD INSTITUTE

Comprehensive comparative analysis of strand-specific RNA sequencing methods

Joshua Z Levin[1,6], Moran Yassour[1–3,6], Xian Adiconis[1], Chad Nusbaum[1], Dawn Anne Thompson[1], Nir Friedman[3,4], Andreas Gnirke[1] & Aviv Regev[1,2,5]

Strand-specific, massively parallel cDNA sequencing (RNA-seq) is a powerful tool for transcript discovery, genome annotation

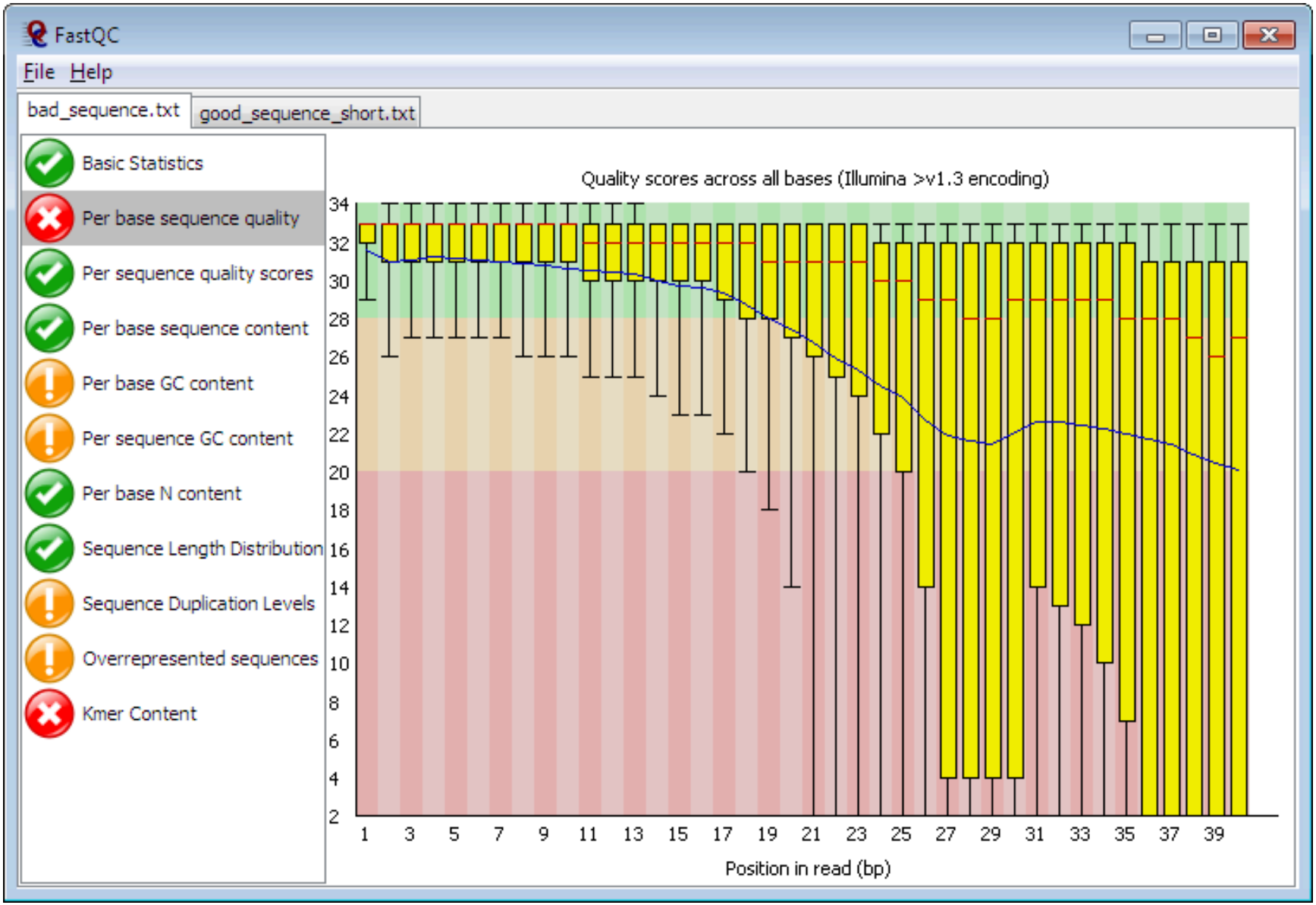
Nevertheless, direct information on the originating strand can substantially enhance the value of an RNA-seq experiment. For

'dUTP second strand marking' identified as the leading protocol

any RNA-seq method. Using the well-annotated *Saccharomyces cerevisiae* transcriptome as a benchmark, we compared seven library-construction protocols, including both published and

resolve the correct expression levels of coding or noncoding overlapping transcripts. These tasks are particularly challenging in small microbial genomes, prokaryotic and eukaryotic, in which

Read Quality Assessment

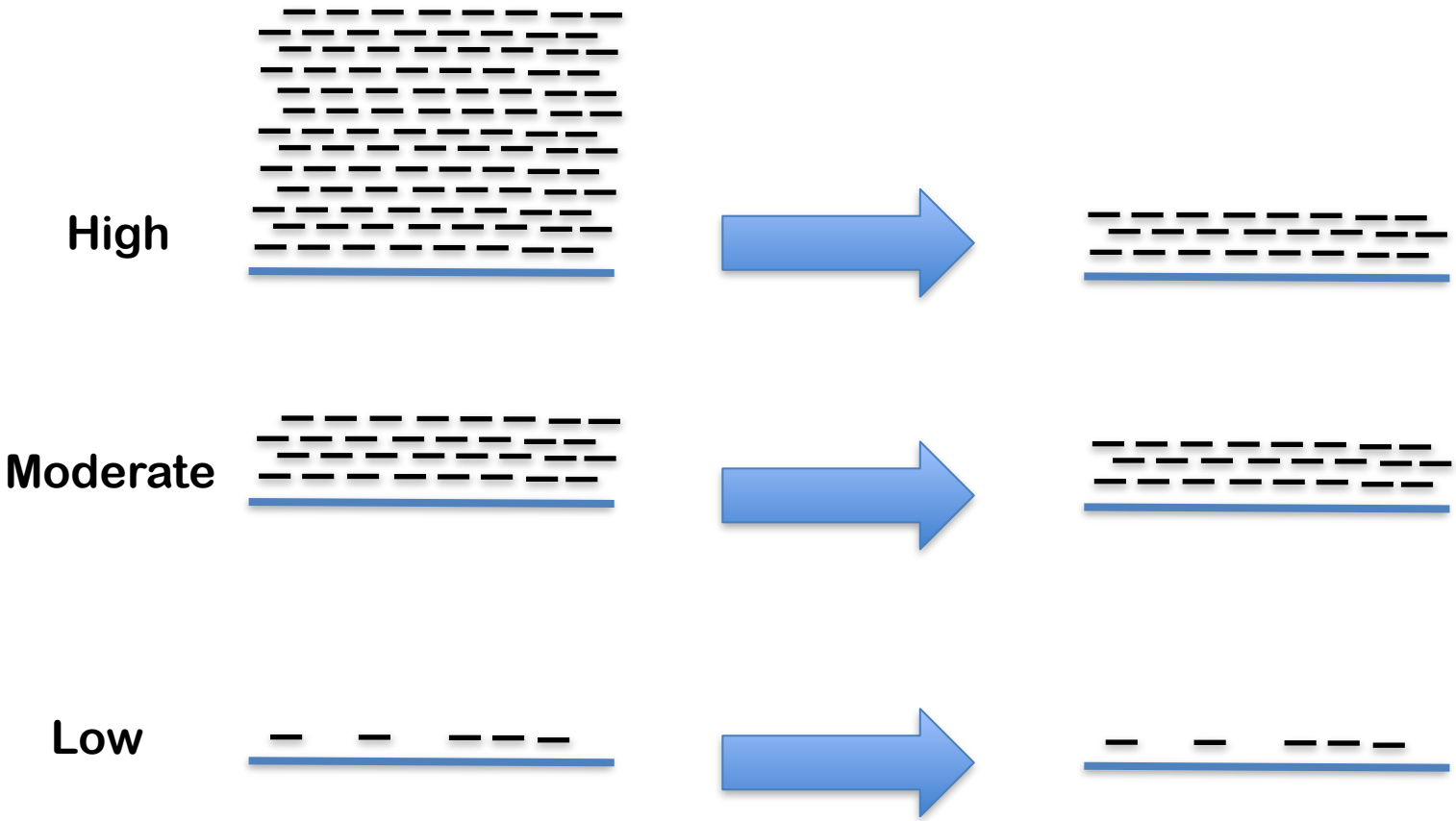


From: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

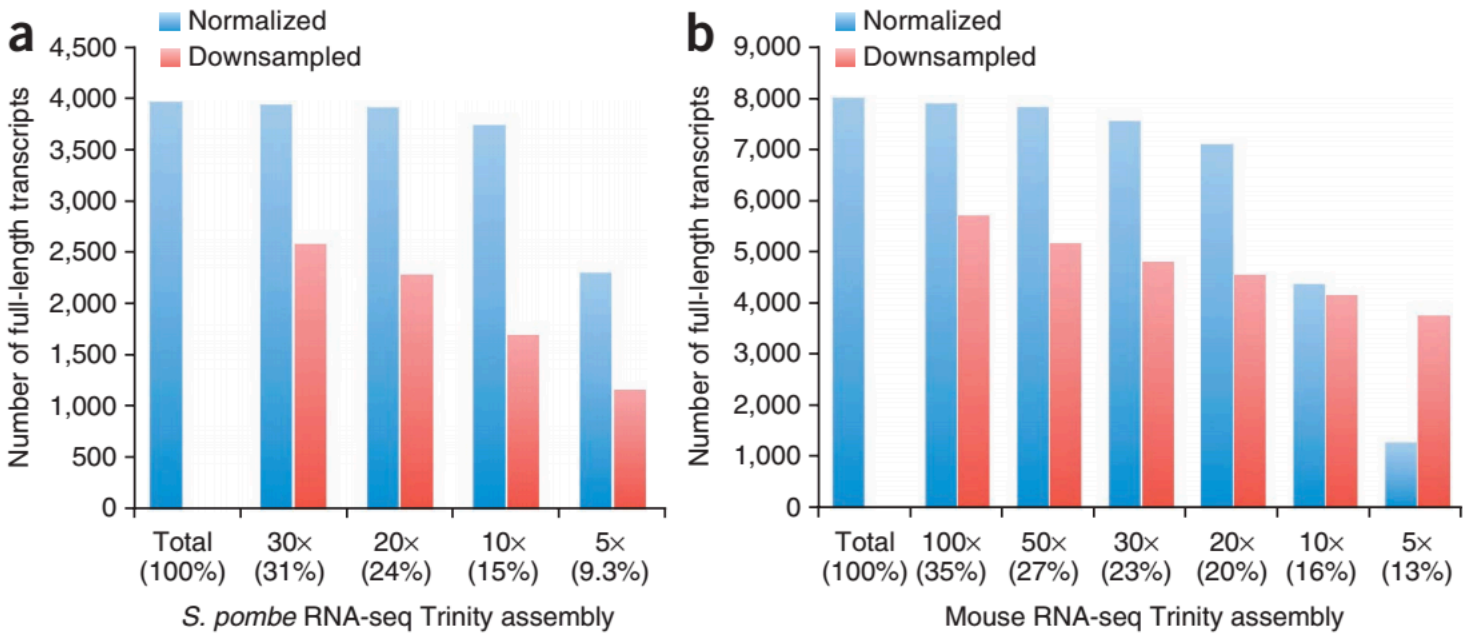
Transcriptome Assembly Lecture Overview

- General issues in
  - Study design
  - Sample collection and preservation
- Library Generation
- Sequencing and Raw data QC
- Pre-assembly read manipulation (ORP STARTS)
- Assembly and QC
- Downstream analysis

In silico normalization of reads



Impact of Normalization on De novo Full-length Transcript Reconstruction



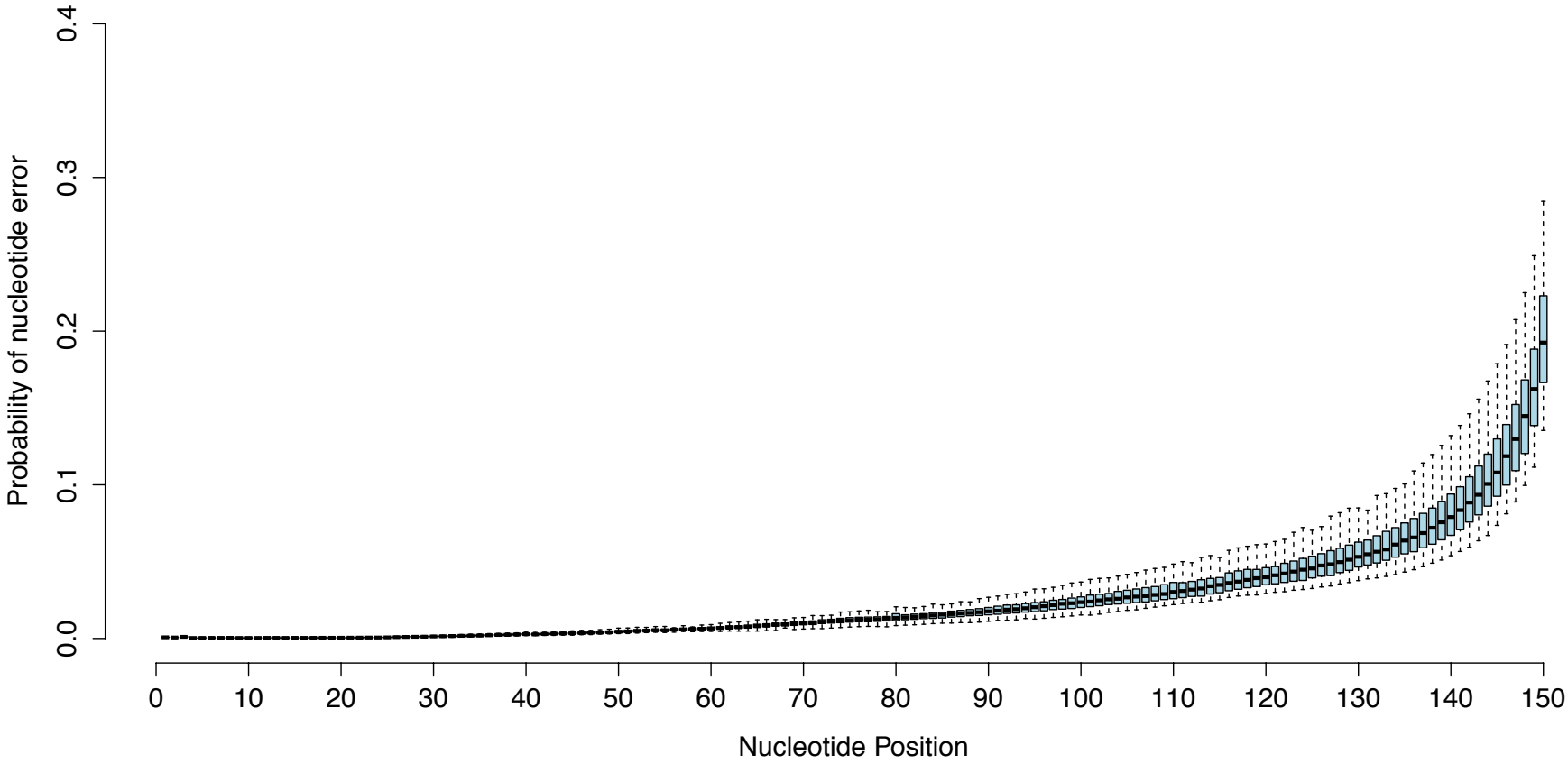
Largely retain full-length reconstruction, but use less RAM and assemble much faster.

Haas et al., 2013

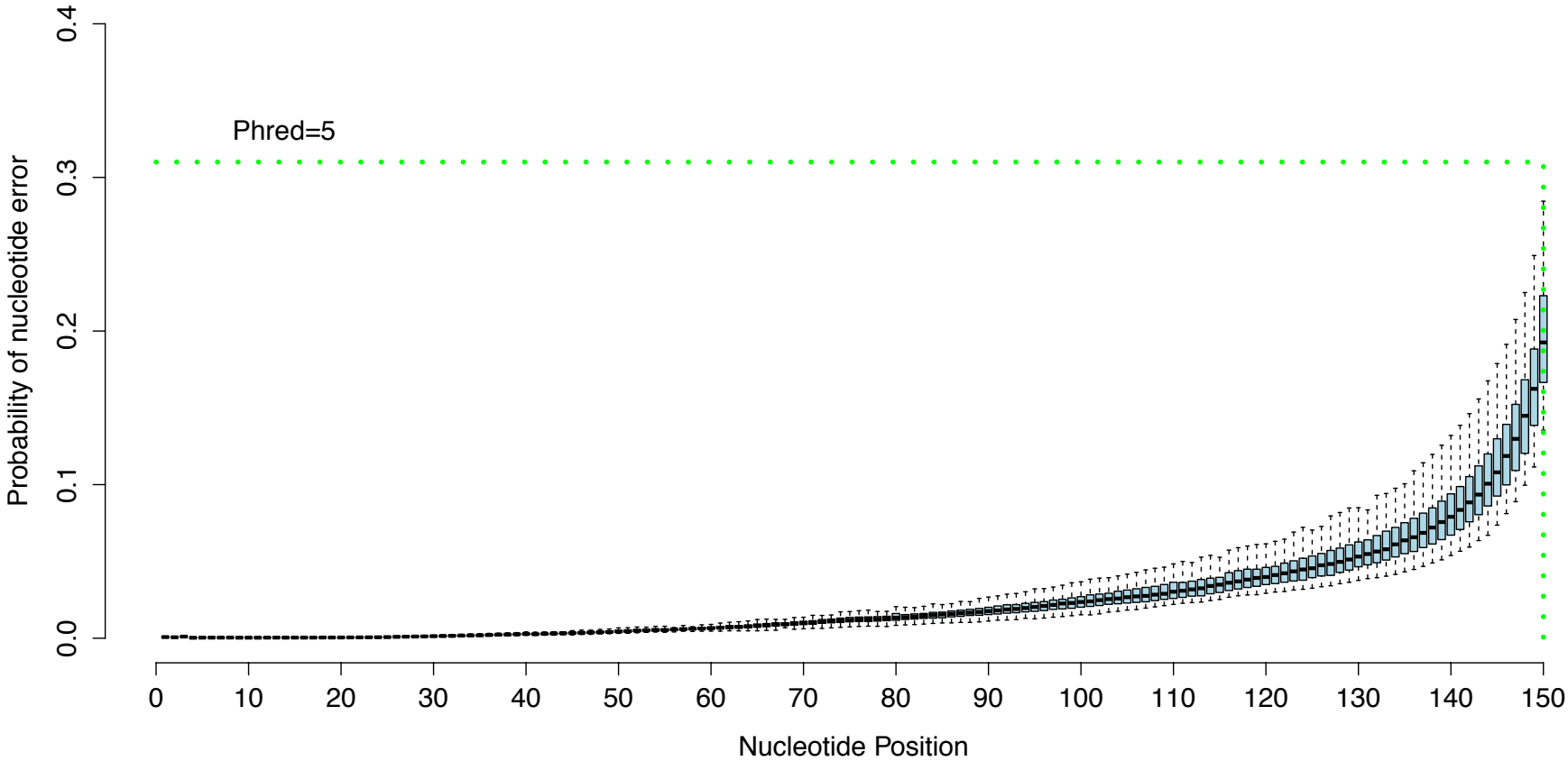
Read Trimming

- What is optimal quality trimming threshold?
- NOTE: Adapter trimming is mandatory.

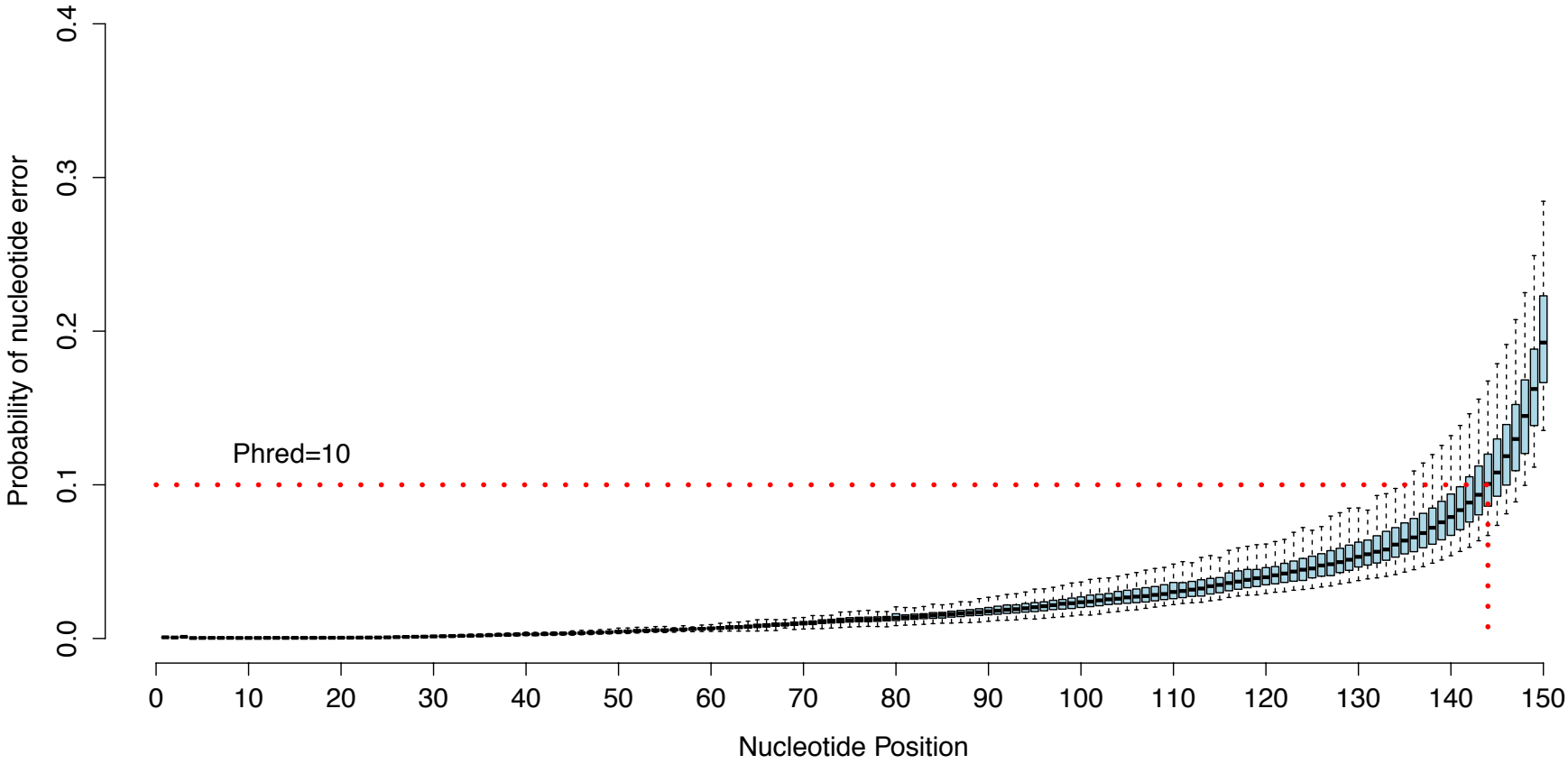
Quality trimming of NGS data



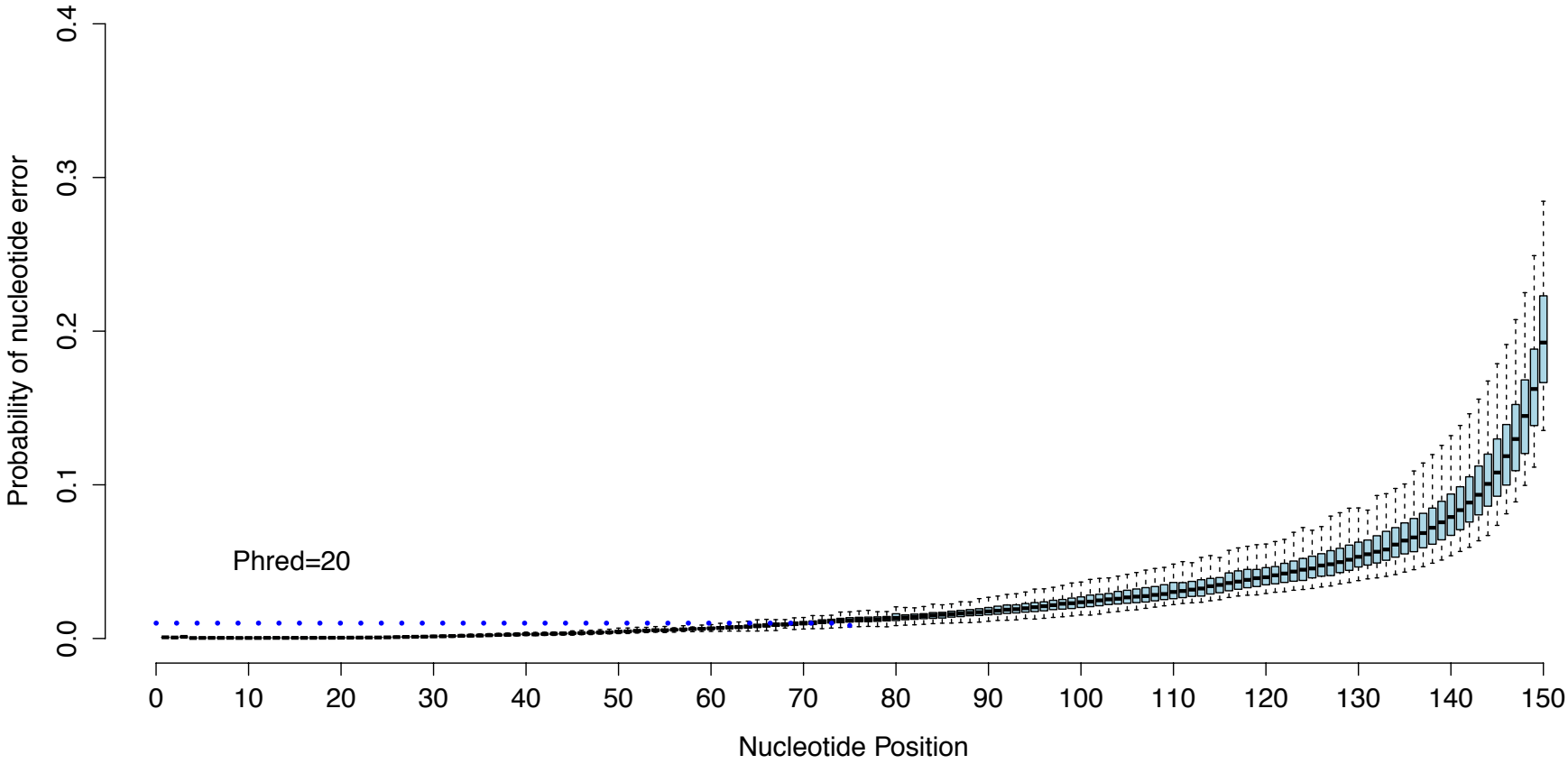
Quality trimming of NGS data



Quality trimming of NGS data



Quality trimming of NGS data

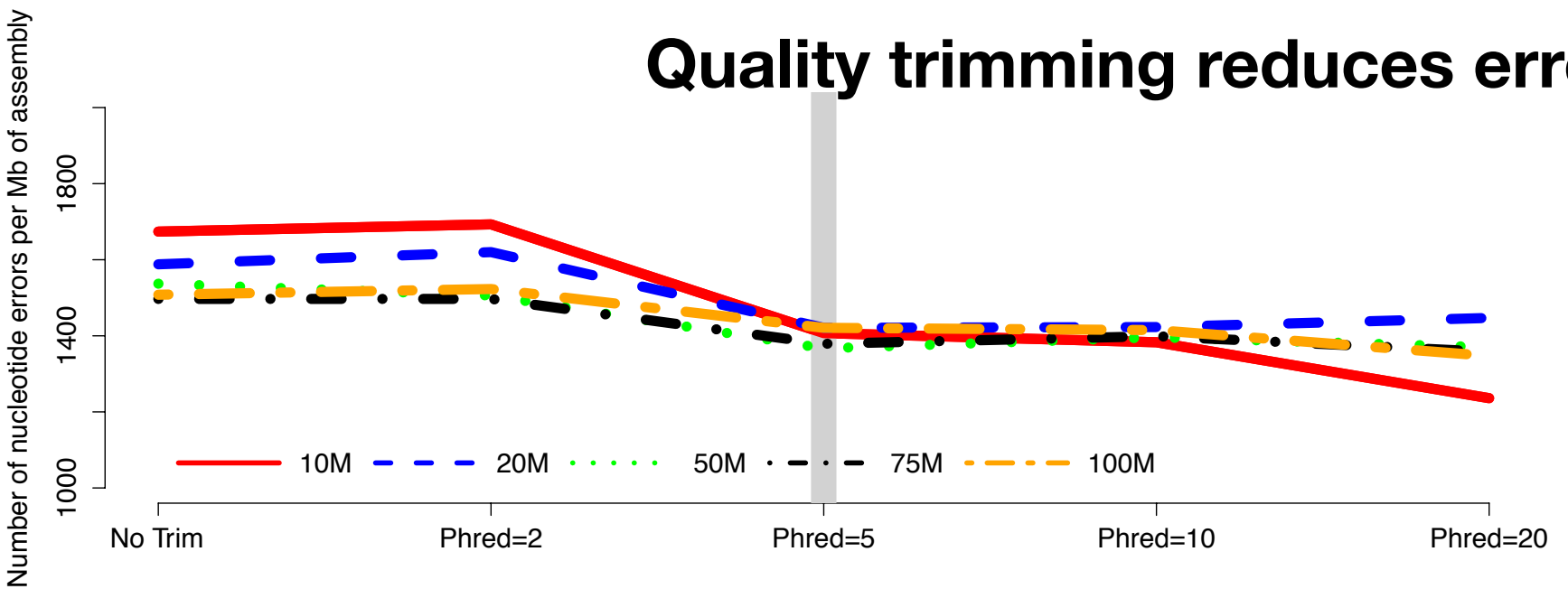


Trimming Experiment

- Illumina datasets that were adapter trimmed.
- Subsampled to 10M, 20M, 50M, 75M, 100M PE reads.
- Trimmed at Phred 0,2,5,10,20

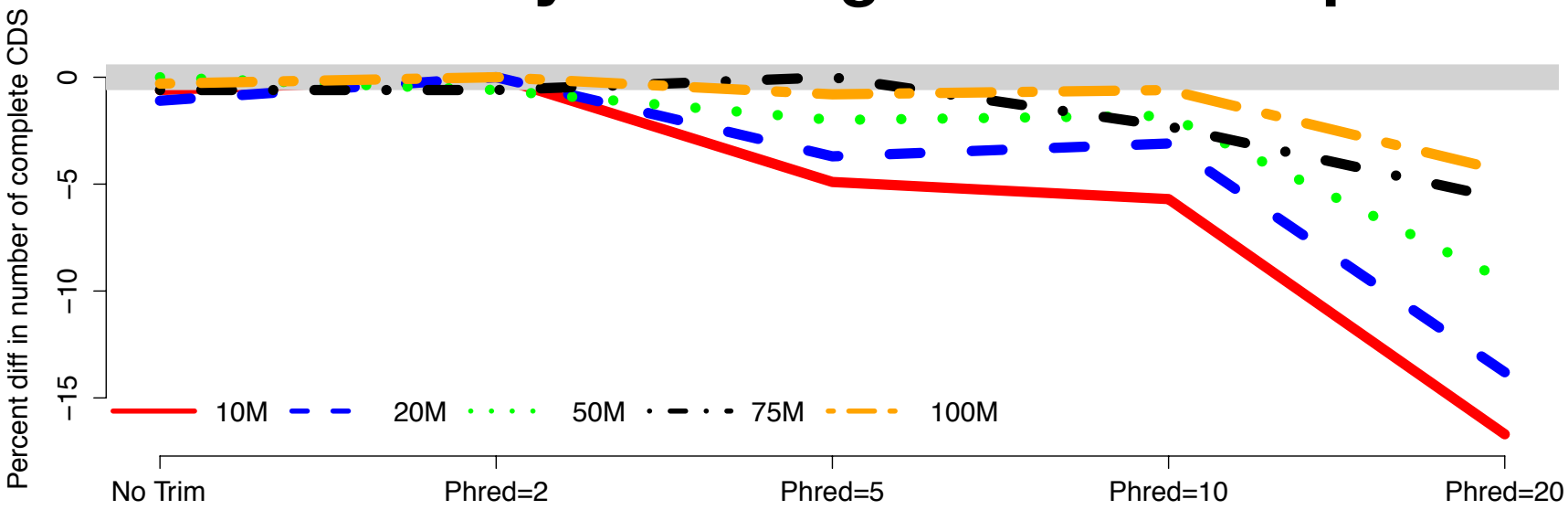
MacManes, Frontiers in Genetics 2014

Quality trimming reduces error

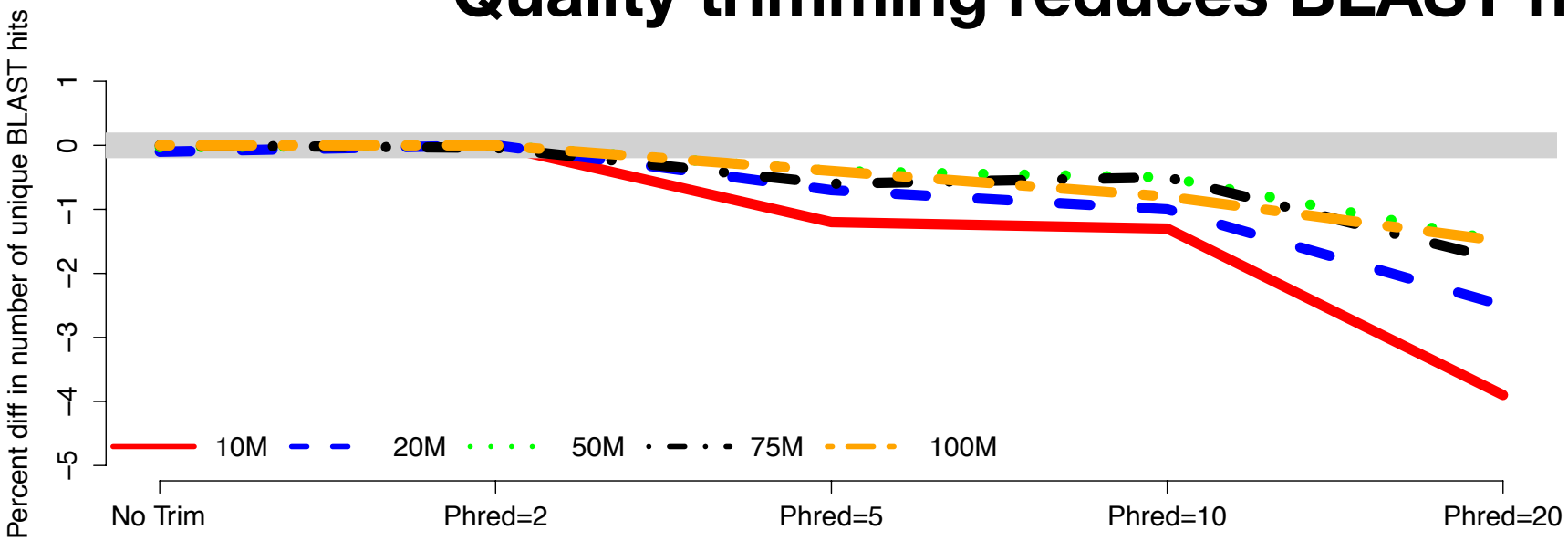


MacManes, Frontiers in Genetics 2014

Quality trimming reduces complete CDS



Quality trimming reduces BLAST hits

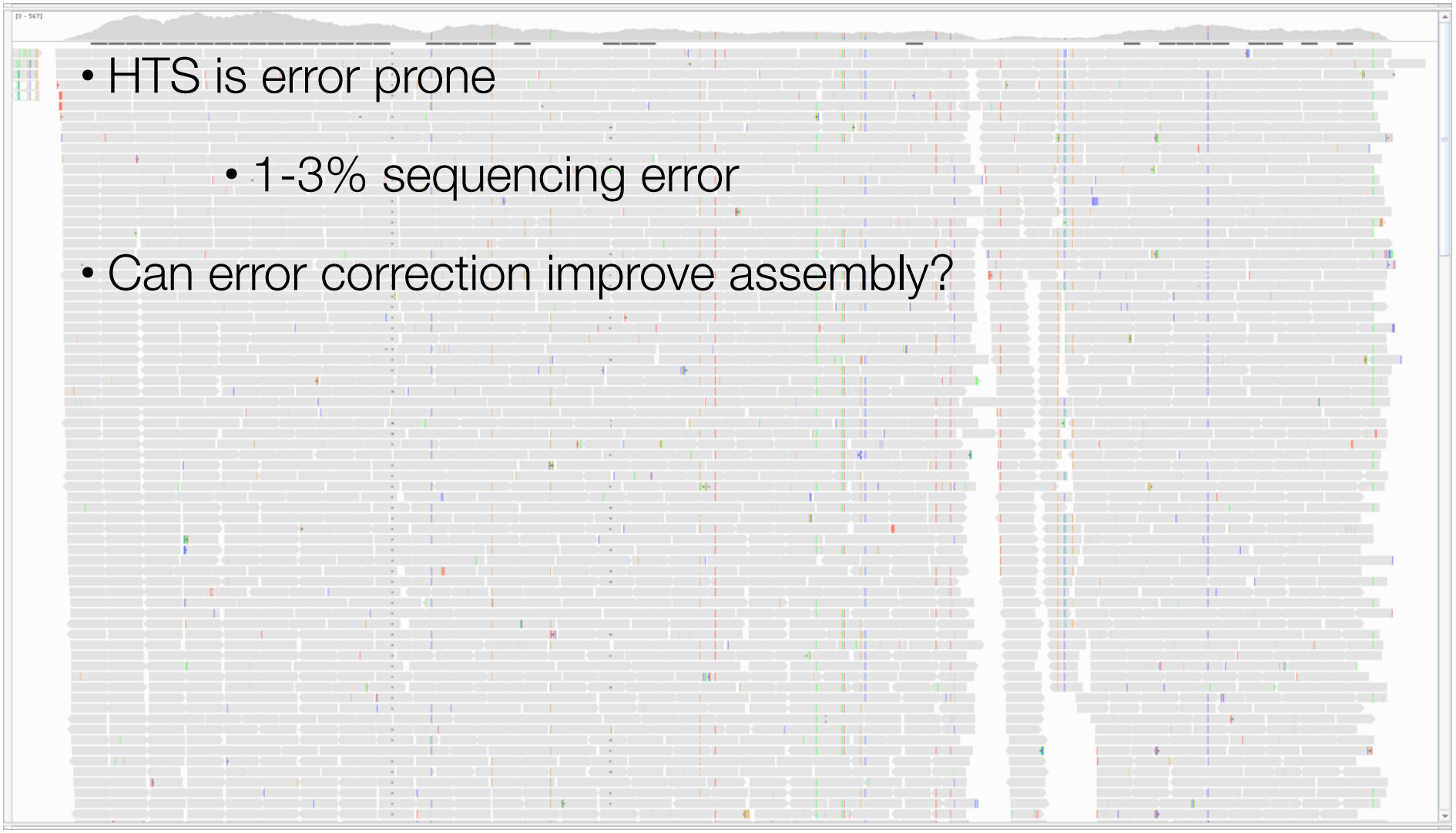


Read Error Correction

Pre-assembly procedures

HTS is error prone

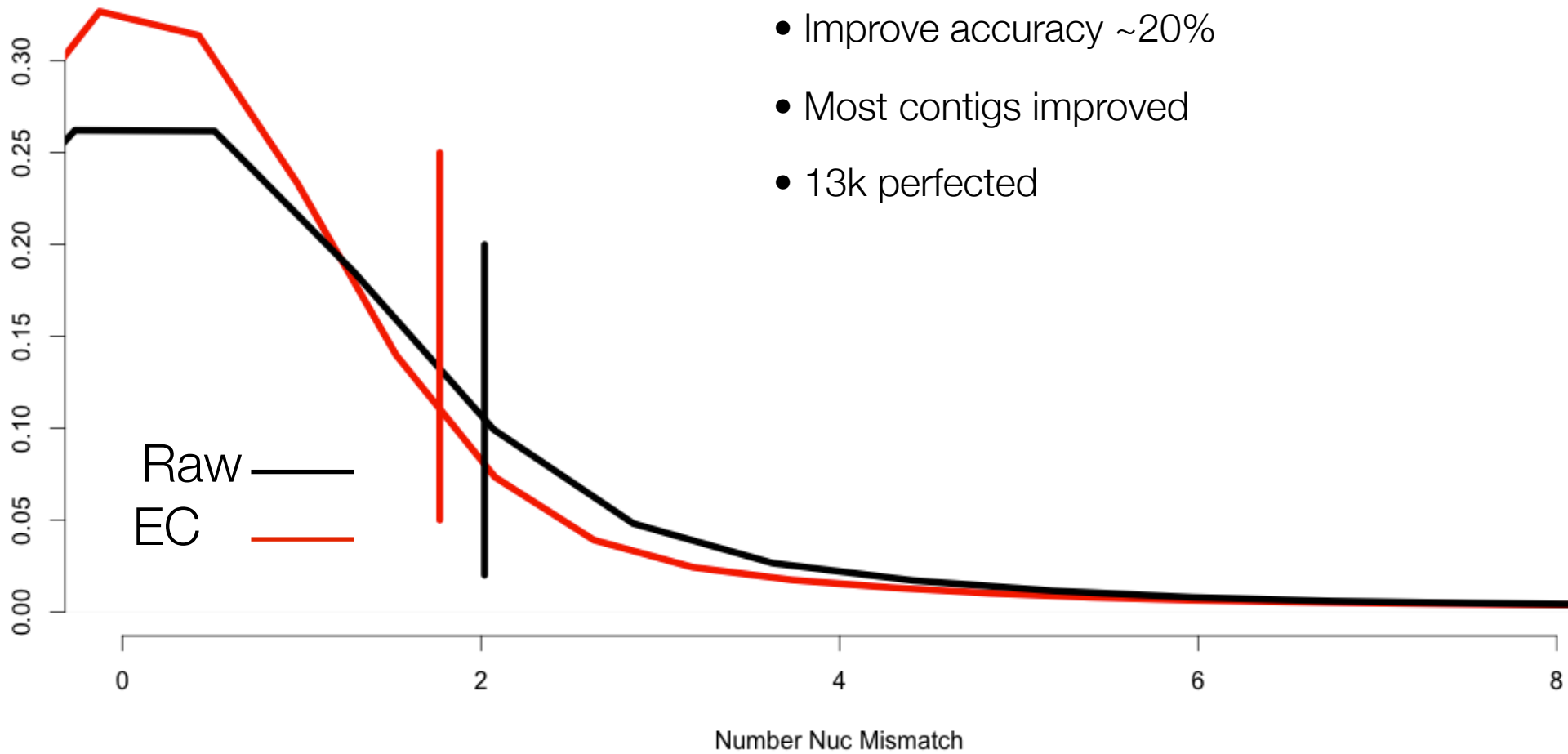
Pre-assembly procedures
Can we correct sequencing error?



- HTS is error prone
    - 1-3% sequencing error
- Can error correction improve assembly?

Pre-assembly procedures
Error correction is important

- 2 billion nucleotide errors
- Improve accuracy ~20%
- Most contigs improved
- 13k perfected

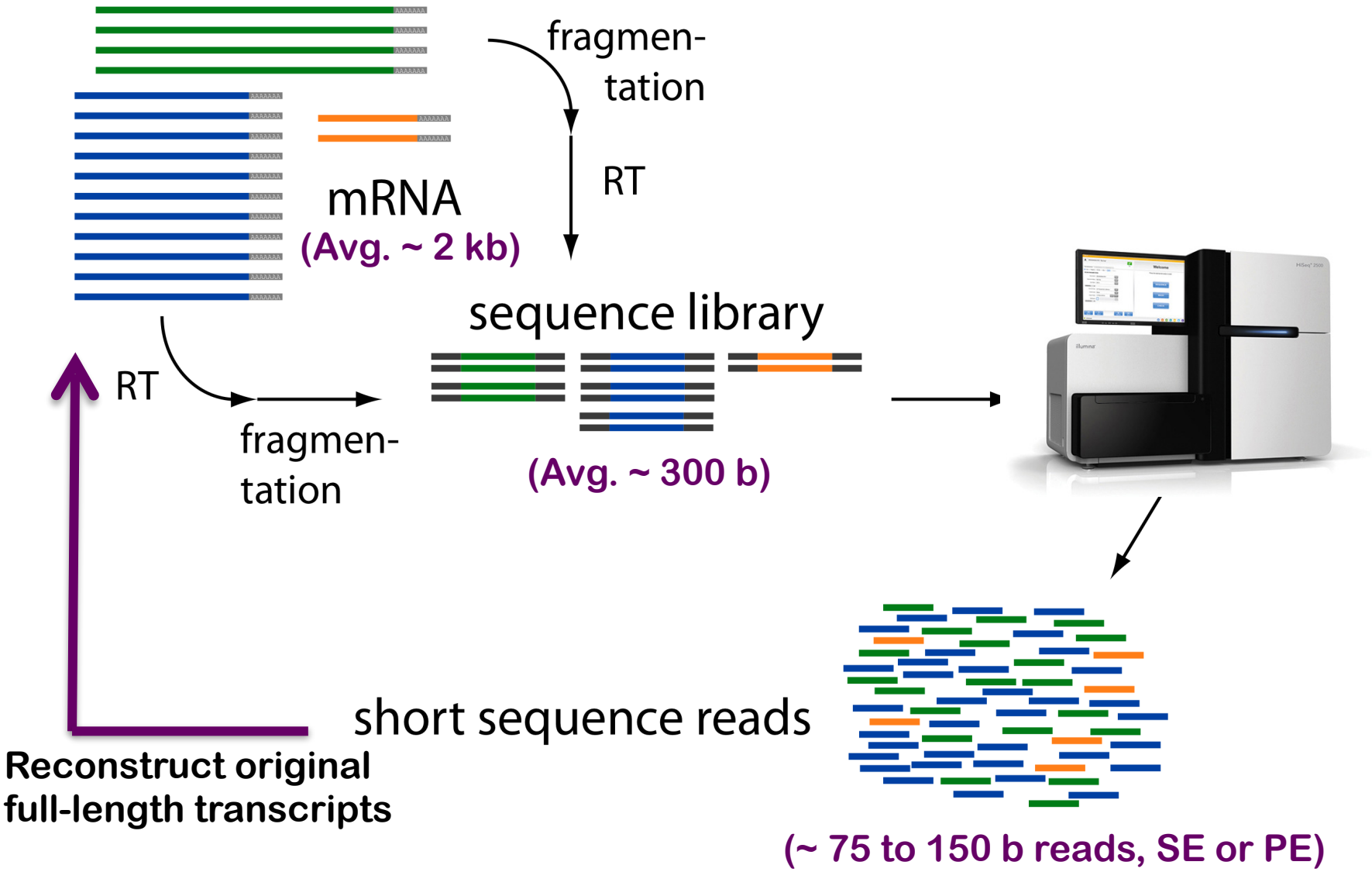


Transcriptomics Lecture Overview

- General issues in
  - Study design
  - Sample collection and preservation
- Library Generation
- Sequencing and Raw data QC
- Assembly and QC
- Downstream analysis

Transcriptome assembly

RNA-Seq Challenge: Transcript Reconstruction



Adapted from: http://www2.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics.html

Transcriptome assembly

- Challenges
  - Highly variable gene expression
  - Many isoforms

Many different *de novo* transcriptome assemblers

ABySS v2

IDBA-tran

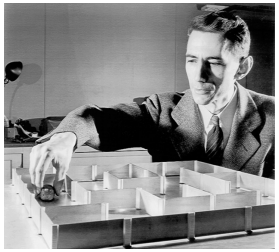
Oases

Trinity

Inchworm

Chrysalis

Butterfly



Binpacker

华大基因 BGI

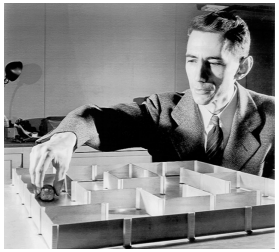
Bridger

Which assembler is “best”?

ABySS v2

IDBA-tran

Oases



Binpacker

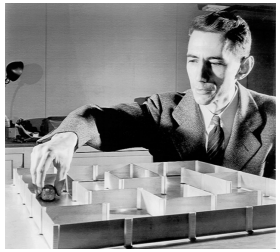
Bridger

Which assembler is “best”?

Type	Accession	Species	Num. Reads	Read Length
Animalia	ERR489297	*Anopheles gambiae*	206M	100bp
Animalia	DRR030368	*Echinococcus multilocularis*	73M	100bp
Animalia	ERR1016675	*Heterorhabditis indica*	51M	100bp
Animalia	SRR2086412	*Mus musculus*	54M	100bp
Animalia	DRR036858	*Mus musculus*	114M	100bp
Animalia	DRR046632	*Oncorhynchus mykiss*	82M	76bp
Animalia	SRR1789336	*Oryctolagus cuniculus*	31M	100bp
Animalia	SRR2016923	*Phyllodoce medipapillata*	86M	100bp
Animalia	ERR1674585	*Schistosoma mansoni*	39M	100bp
Plant	DRR082659	*Aeginetia indica*	69M	90bp
Plant	DRR053698	*Cephalotus follicularis*	126M	90bp
Plant	DRR069093	*Hevea brasiliensis*	103M	100bp
Plant	SRR3499127	*Nicotiana tabacum*	30M	150bp
Plant	DRR031870	*Vigna angularis*	60M	100bp
Protozoa	ERR058009	*Entamoeba histolytica*	68M	100bp

k=55

k=75



k=75

k=25

Assessing assemblers by counting genes

	Genes		
Assembly	Sum	Missing	Unique
Concatenated	14674 ± 3590		
SPAdes55			
SPAdes75			
Shannon			
Trinity			

Each assembler produces a special snowflake

	Genes		
Assembly	Sum	Missing	Unique
Concatenated	14674 ± 3590		
SPAdes55		-1739 ± 758	
SPAdes75		-2711 ± 2047	
Shannon		-4375 ± 3508	
Trinity		-1952 ± 803	

Each assembler produces a special snowflake (there is no One Best Assembler!)

	Genes		
Assembly	Sum	Missing	Unique
Concatenated	14674 ± 3590		
SPAdes55		-1739 ± 758	570 ± 266
SPAdes75		-2711 ± 2047	301 ± 195
Shannon		-4375 ± 3508	302 ± 241
Trinity		-1952 ± 803	520 ± 301

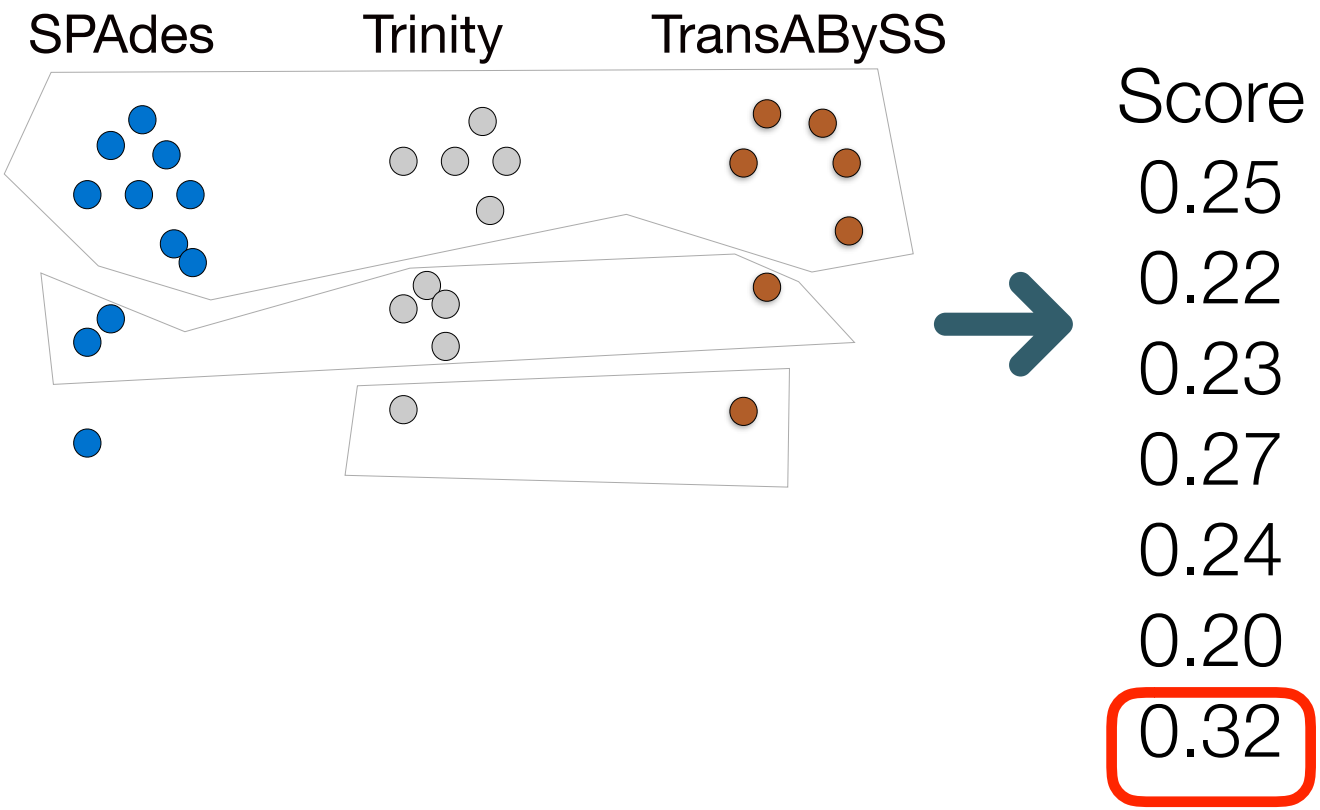
The Oyster River Protocol assembles and merges transcriptomes

- Error Correction (RCorrector) DOI:10.7717/peerj.113
- Trimming (Trimmomatic, Phred=5) DOI:10.3389/fgene.2014.00013
- Assembly (SPAdes (k=55,75), TransABySS, Trinity)

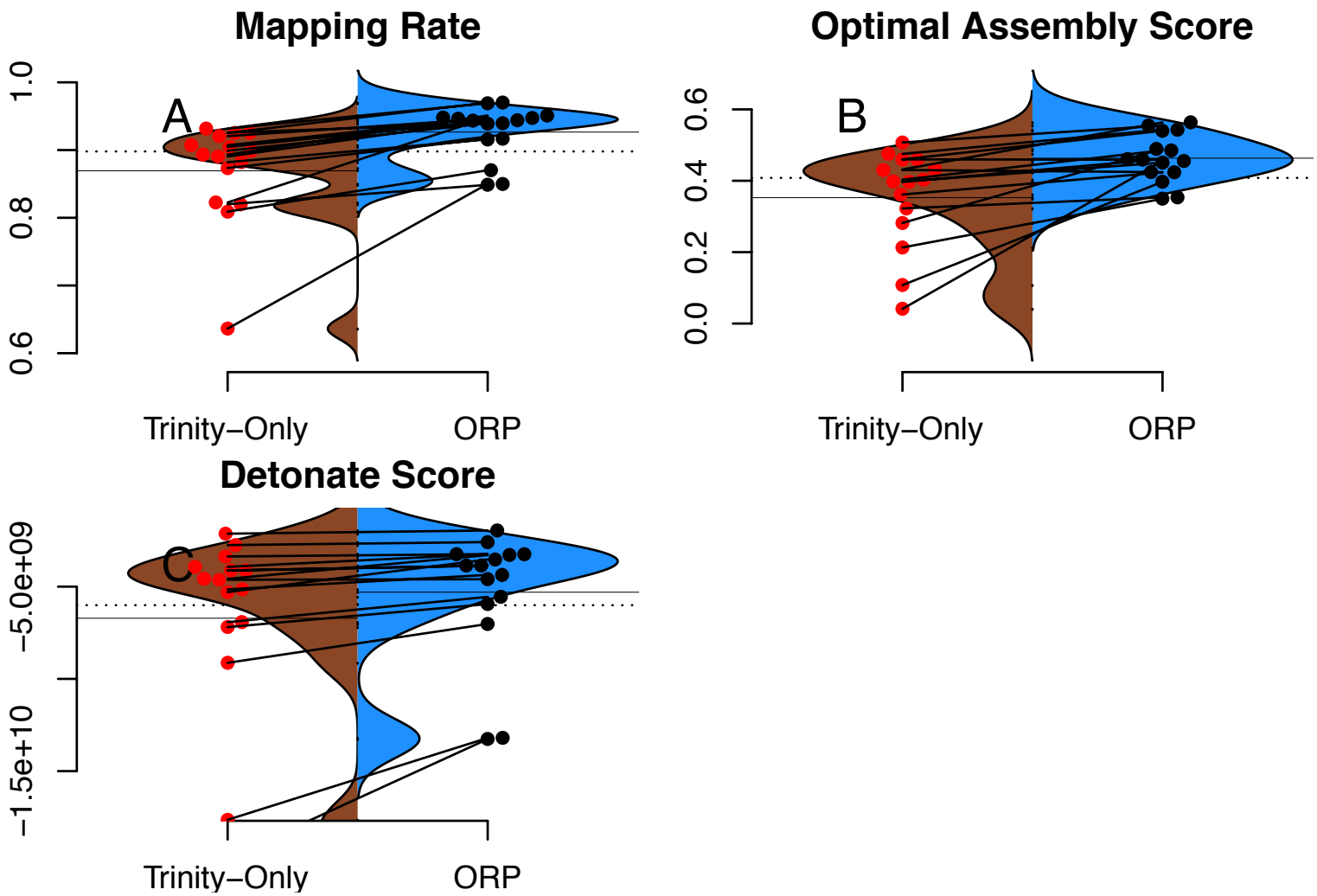
http://oyster-river-protocol.rtfd.io

The Oyster River Protocol assembles and merges transcriptomes

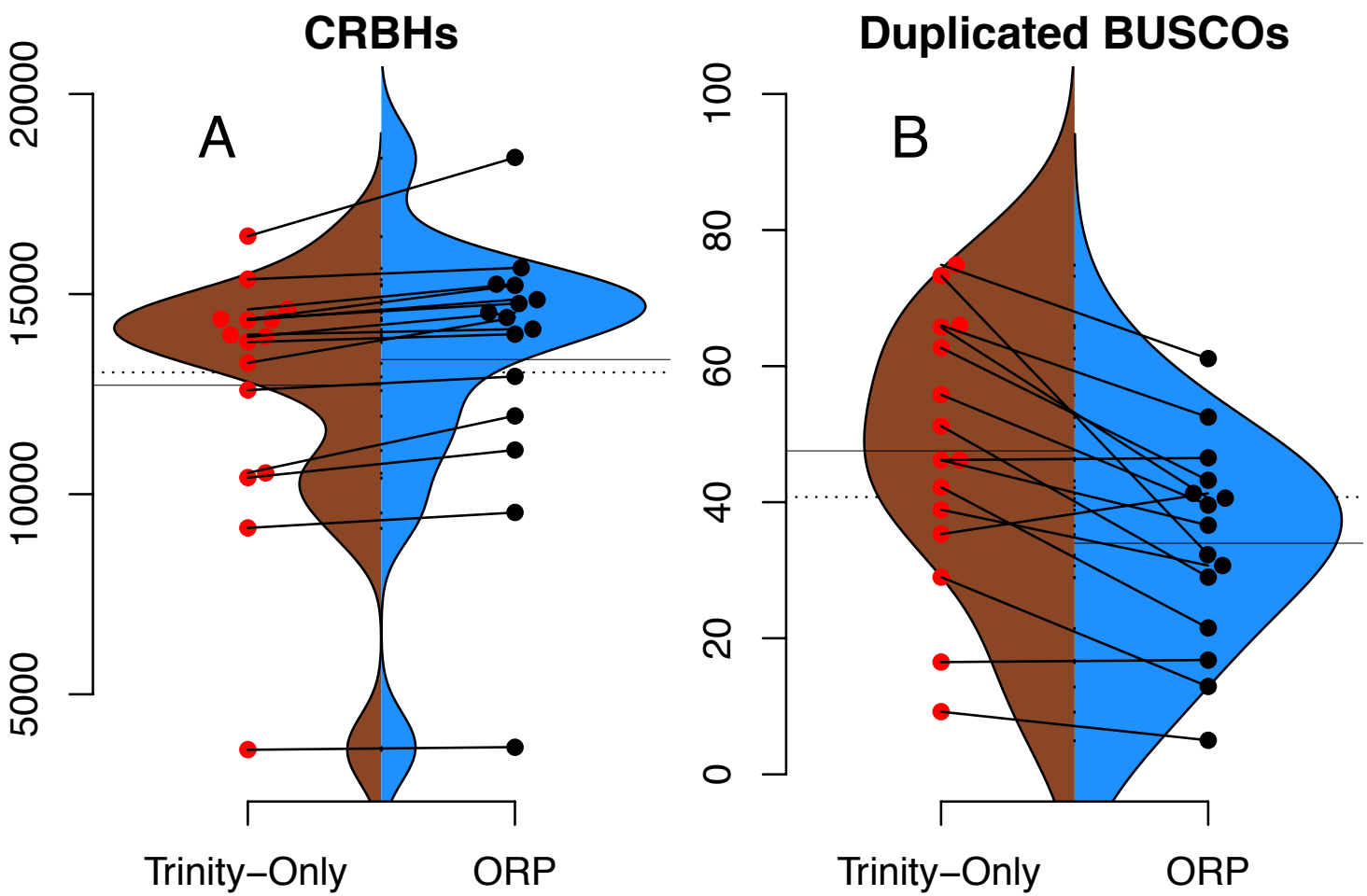
- Orthofuser (defines iso-groups, picks the most representative)



The Oyster River Assembler improves synthetic evaluative metrics



The Oyster River Assembler increases the number of genes recovered, while decreasing duplication



Assembly output: A multi-fasta file

```
>comp0_c0_seq1 len=5528 path=[1:0-3646 10775:3647-3775 3648:3776-5527]
AATTGAATCCCTTTTTGTATTGAAAAAGTTGAAATGAAAGACATATACAGATTGAATGTGGTGATTGCAAAATATAATGCAATTTCGAACAATTAAAATTATGAAAATATACAAAATTGATGGCAGCACACCTAGGTTTG
TGCACTCTCCATCATGTGGAGATACTACAGAGGACTATCCGTCCACAGCAAAGAAGTAACTTGAACCCGATTCCTGCTTTGCAAAGTCTTGACTTAGGATCTCAGTAGTATAAAAAGCAGTGATTGTTTTTTTTCAGTCT
GTAAACAGTAGTCGTGGTTTTTGTTGTTTTAAATATCAATTTACCACACAAAAACAAAACAGAACAAAACCCATATAAACCACAGCAGCAGCAGCACTGGGCCTTGAGCATTCCCCTTAGATGCTAGTGACATACAGGC
GTGGGGAGAGGAGGCGTTCACTCAGTAAGAATATATATATATATACTTTAACTGTACACAATGTATAGTGCATGACAGCTTCCAAAAGAATAGACTGATAAAAGATATTGGCCATTTCATAATTTCACTCTTTTTAC
AGTTATCTCAAAATGTAAGAATTAGATCTGATTGAAATGCTACATTTAGTAAGAAAATCAGCAAGTAACAGAGGAAGTGTAACCCCACCATGACATTATTGTCAACAAGACCAGTGGAGGCCCTACATGTTAGAGCAGG
AGACTGACAGGATAGACCATTAGAGGAAGGAGCTATGCCTGAGGGCCATAGCCAGCCCATGCAGTCCATACGGACCATGGTCAGCTGACAGAGCTTGATTAGAGCATCTGTGACCGTGAACACACAGCACAATTTGAAT
ATCCCAAGAAACATTATTCCAGGAGCCAGGGCAGTGCAAGCAAACGCCCACCCAGTCCCATGGGAAACAGGCCACAACTACAGAACAGTACCACAAGGTGGTGTTCAAGGCAAGTCAAAATCAGGGTCTTCAACCCATCT
TATTCAAAAAGGCCTTTTTTTGGGGATGGAGCACGTGATACTCTGATGCAACCATGATGTGAGGCTCCAGCACATCCTACAAGTAAGAAAGTAGTTTGCACACATTTCCTTCACAAGTAAATTAGTTAACCACTATGCTGA
TTTGTGAATCCCAGACAGTTACGATAAAGAATGCAATGGTGTGCTGCTGGAGCAGTCCATGGGAAAGACCAGTCCTCACCAAGTCATCTTTTCACCTTACAGTTACTCTCAGGAATAAAGTGACAGGGAACAAGAACAGA
AGTGAGAGAGAGGAGATTCAGACACAAAACAGTCACGGAAAGCGCTGTCGGAGCTCGGCATGACATAATCAAGAGCAGTTTTCATCTTCTCGCAGACCAGCCTCTTAAGCTGGAGGCTTAGGGAACAGGCCACCAACCTTAG
GTGAACAACATGAATACACCCTGATGCAGTTCTTAAGTGTCAACAGGACACCAACATCAGGCCCATTATAAAACATACCTTTTTCAACCTAAAACTTAAAAACCTAGGTTAAACCCATATTTAATCCTTCCCATCCCAAAG
CCCCCCTAAAAGCACTAGAGTAGTAAATCTGTTTCCAAGCATATTATGCATATGTGTATCCTGTGTG
```

Evaluating the quality of your transcriptome assembly

- Structure
  - TransRate
  - Detonate
  - Read mapping
- Content
  - BUSCO
  - Blast/Blat?

busco.ezlab.org

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

Zdobnov's Computational Evolutionary Genomics group

CEGG Home | OrthoDB *v9* | BUSCO *v2*

Assessing genome assembly and annotation completeness with **B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

About BUSCO

BUSCO *v2* provides quantitative measures for the assessment of genome assembly, gene set, and transcriptome completeness, based on evolutionarily-informed expectations of gene content from near-universal single-copy orthologs selected from OrthoDB *v9*.

BUSCO assessments are implemented in open-source software, with a large selection of lineage-specific sets of Benchmarking Universal Single-Copy Orthologs. These conserved orthologs are ideal candidates for large-scale phylogenomics studies, and the annotated BUSCO gene models built during genome assessments provide a comprehensive gene predictor training set for use as part of genome annotation pipelines.

Running BUSCO

- Looks for the presence of universal single copy orthologs in your assembly
    - Eukaryote, Vertebrate, Plant, Metazoa, Mammalia, Aves, etc…

- **A highly complete assembly will have most/all of the SCOs**
- **Any number greater that 75% is probably good.**

busco.ezlab.org

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

Zdobnov's Computational Evolutionary Genomics group

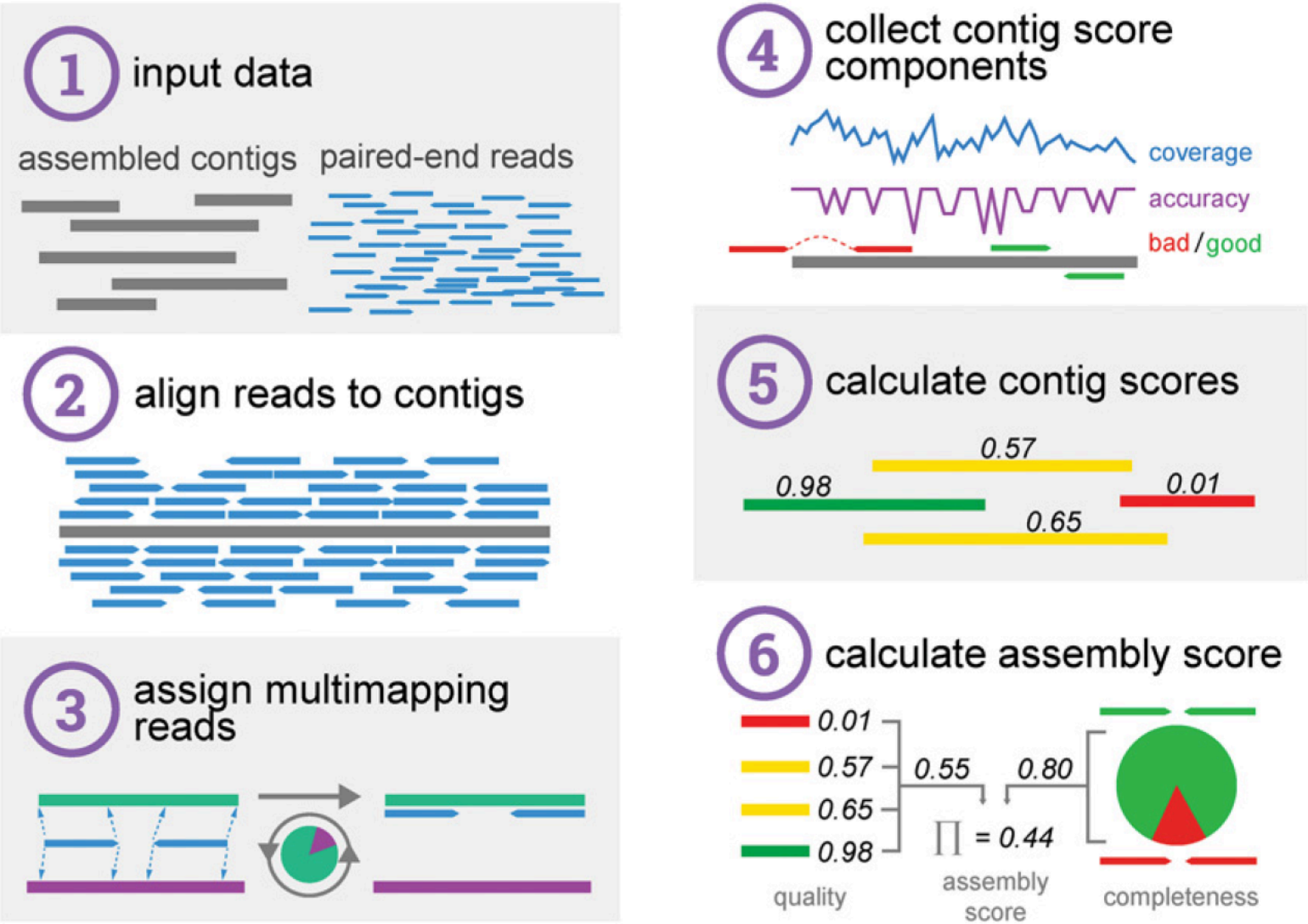
CEGG Home | OrthoDB *v9* | BUSCO *v2*

BUSCO v2

Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs

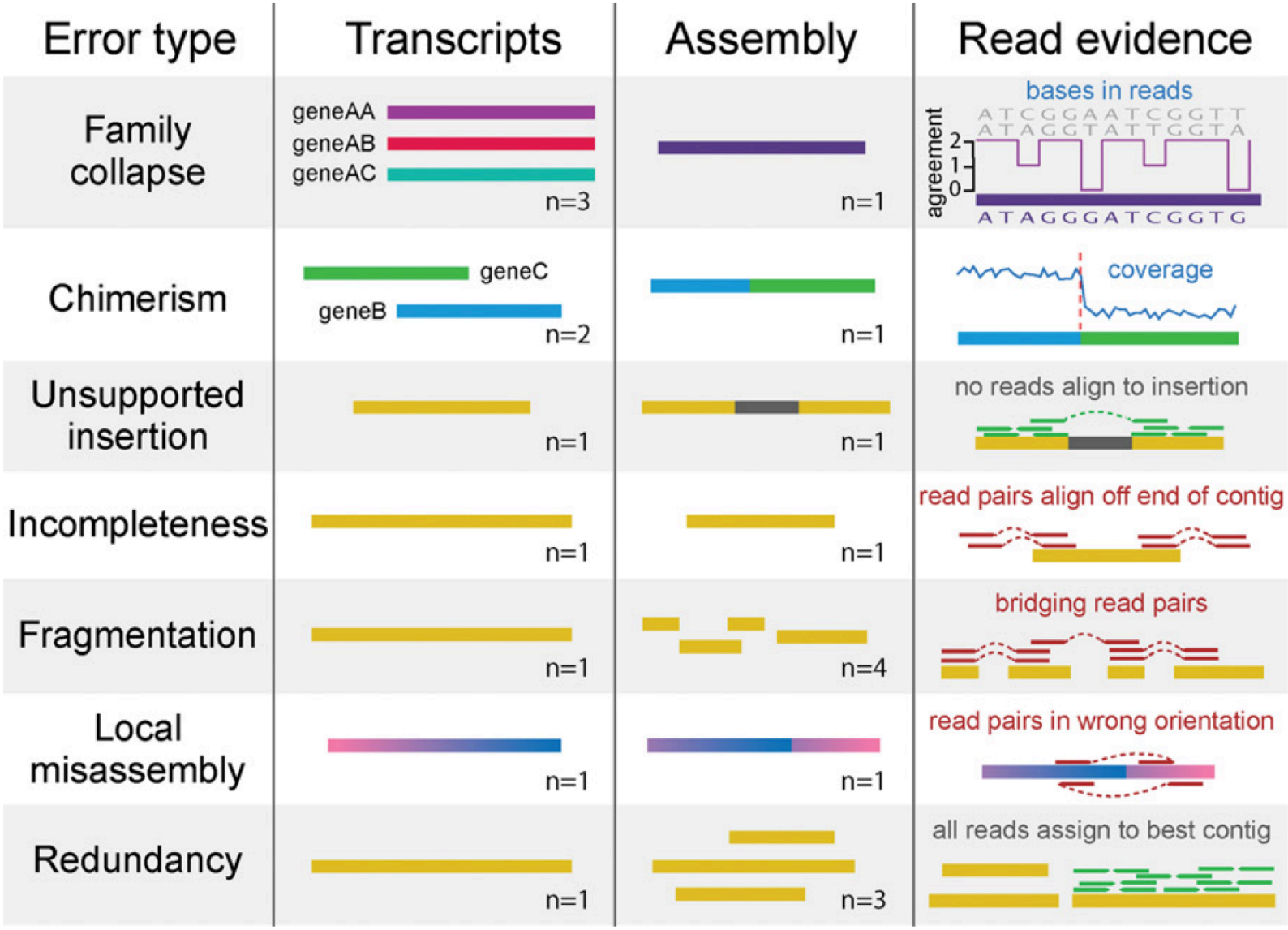
```
#Summarized BUSCO benchmarking for file: Trinity.fasta
#BUSCO was run in mode: trans

Summarized benchmarks in BUSCO notation:
    C:88%[D:53%],F:4.5%,M:7.3%,n:3023

Representing:
    1045    Complete Single-copy BUSCOs
    1617    Complete Duplicated BUSCOs
    139     Fragmented BUSCOs
    222     Missing BUSCOs
    3023    Total BUSCO groups searched
```

TransRate



Unna-Smith et al. Genome Research, 2016

De novo Transcriptome Assembly is Prone to Certain Types of Errors



Smith-Unna et al. Genome Research, 2

Running TransRate

- Analysis of the structure of your assembly, mostly using read mapping
    - Critically, you must use the read dataset used for the assembly

- **Scores greater than 0.22 are acceptable (better than 1/2 of all published transcriptomes at the time of publication.**

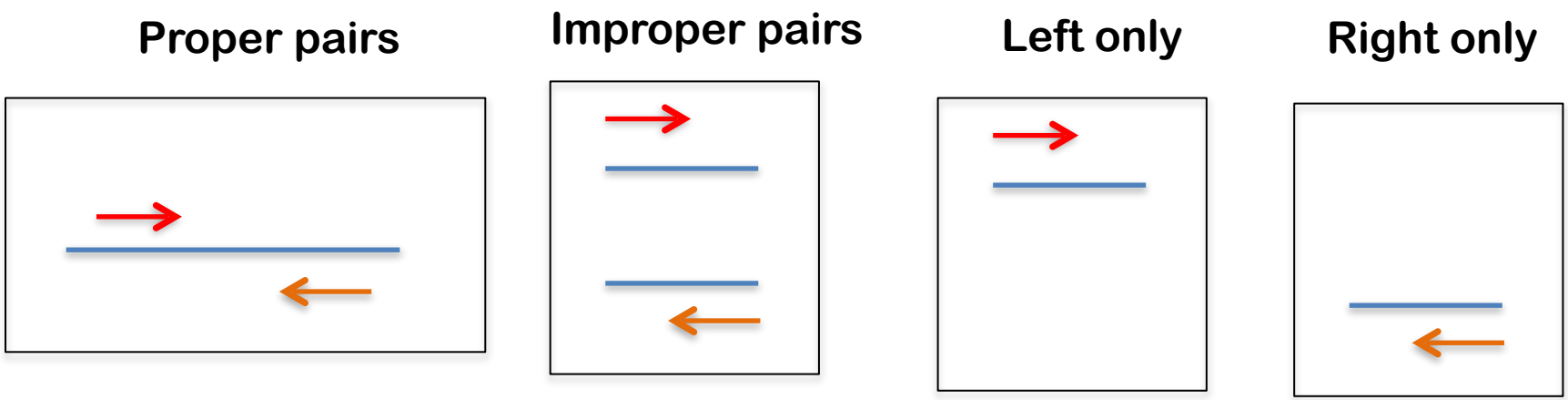
Simple Quantitative and Qualitative Assembly Metrics

Read representation of assembly

Align reads to the assembled transcripts using bwa/bowtie/other.
A typical 'good' assembly has ~80 % reads mapping to the assembly and ~80% are properly paired.

Given read pair: → ←

Possible mapping contexts in the assembly are reported:



Simple Quantitative and Qualitative Assembly Metrics

Assembly content compared to a reference

Imagine you just assembled something closely related to a mouse…

You BLAST your new transcripts to the Mus reference in Ensembl.

What does it mean about your assembly if:

12% match
vs.
95% match

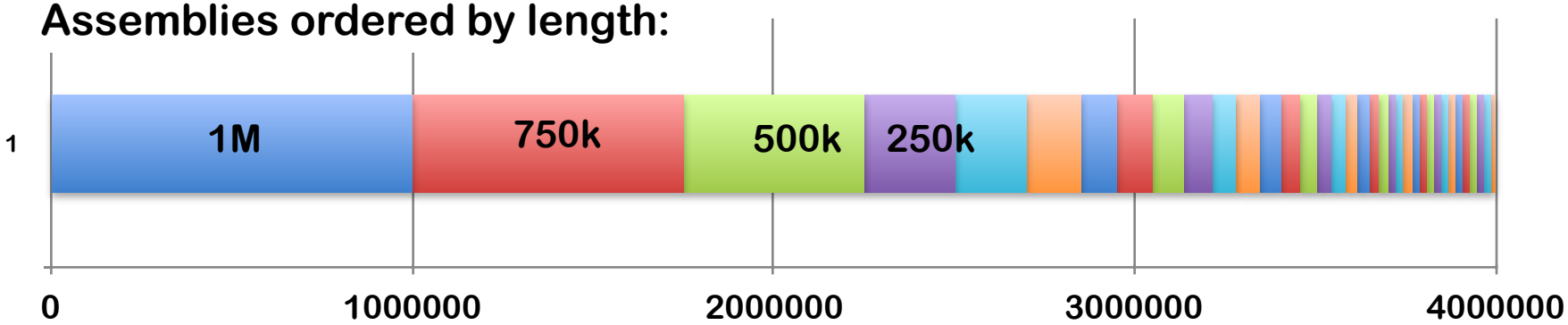
Simple BAD metrics that you should not use

The Contig N50 statistic

"At least half of assembled bases are in contigs that are at least N50 bases in length"

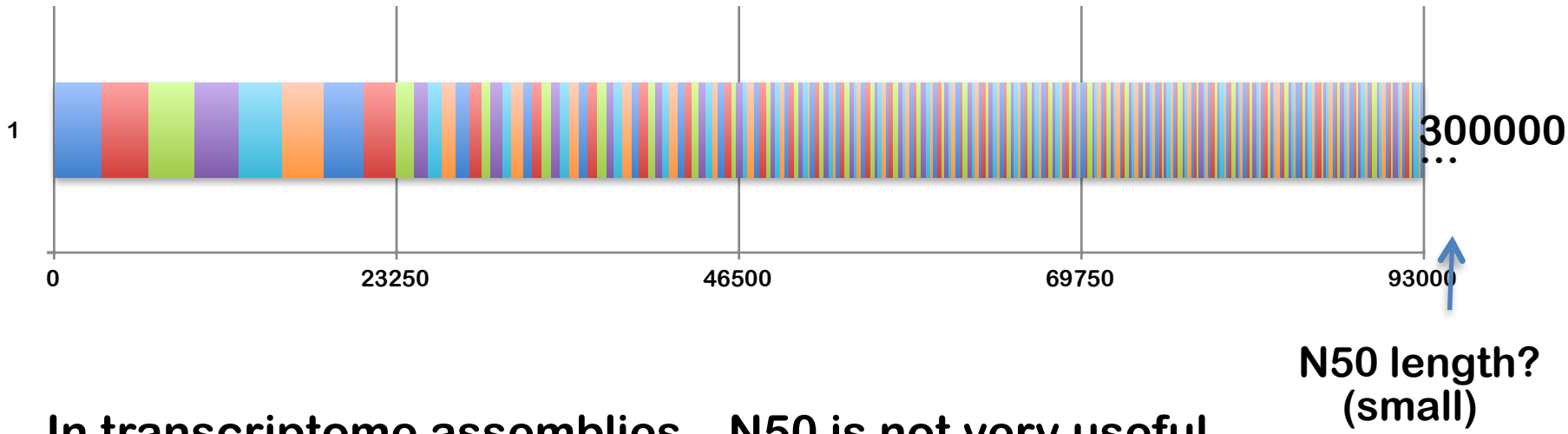
In genome assemblies – used often to judge 'which assembly is better'

Assemblies ordered by length:



N50 contig length = 500k

N50 Calculation for Transcriptome Assemblies??



In transcriptome assemblies – N50 is not very useful.

- Longer != better
- Very sensitive reconstruction for short lowly expressed transcripts drives lower N50

Transcriptome Assembly Lecture Overview

- General issues in
  - Study design
  - Sample collection and preservation
- Library Generation
- Sequencing and Raw data QC
- Assembly and QC
- Downstream analysis

Transcript Functional Annotation

GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGA[illegible]TCGAC
TCTCCC[illegible]TCCCA
AAAGAC[illegible]CCTGG
GGCTTG[illegible]CCTAA
TGACCT[illegible]TGCTG
GAAAAG[illegible]CAGCC
TTGTCA[illegible]TTCCA
GGAAGC[illegible]CGTGG
ATGTGG[illegible]GGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCCAGCTGGCCATAGCCTGGTGCCTGA
GGAATGAGGGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACTTATGGAGA
ACATTGGAGCAATACAGGTCCTTCCAAAATTGTCGTCTTCCATCGTCCACGAGATCGACA
GCATTCTGGGCAATAAACCCTACAGCAAAAAGGACTATAGATCCTAAGTCTACTCCCCAC

Can we gather hints of biological function from sequence?

Methods used to predict function from sequence

- **Sequence homology**

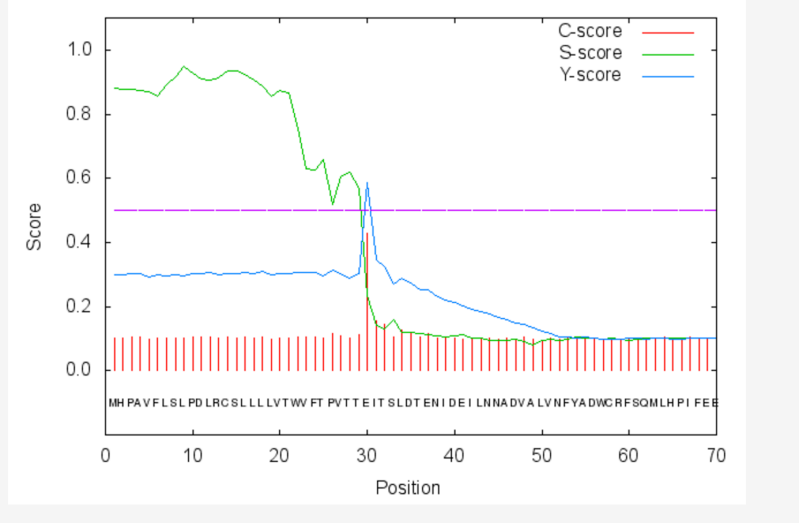
Searching protein database for sequence similarity

```
Query          THVHRPYNEHKSLSGTARYMSINTHLGREQSRRDDLESMGHVFMYFLRGSLPW--QGLKA
               T    P +  K   GT  Y S + HLG    RR DLE +G      L   LPW  Q L A
Database Match TGDFKP-DPKKMHNGTIEYTSRDAHLG-VPTRRADLEILGYNLIEWLGAELPWVTQKLLA
```

- **Sequence composition**

Predict functions of sequence using machine learning methods for pattern recognition.

- Neural Networks
- Hidden Markov Models



Use BLAST to search for sequence similarity to known proteins

Secure https://blast.ncbi.nlm.nih.gov/Blast.cgi

NIH U.S. National Library of Medicine | NCBI National Center for Biotechnology Information | Sign in to NCBI

BLAST®

Home | Recent Results | Saved Strategies | Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

Learn more

NEWS

Magic-BLAST 1.2.0 released

A new version of the BLAST RNA-seq mapping tool is now available.

Mon, 27 Feb 2017 14:00:00 EST

More BLAST news...

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

The Swiss-Prot database is a valuable source of proteins with known functions

www.uniprot.org

UniProt

UniProtKB | Advanced | Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

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 (554,241)
Manually annotated and reviewed.

TrEMBL (84,827,567)
Automatically annotated and not reviewed.

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

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

News

Forthcoming changes
Planned changes for UniProt

UniProt release 2017_04
Death (by insulin) in paradise

UniProt release 2017_03
Viral Short Message Service: peptide texting guides the outcome of infection

UniProt release 2017_02

News archive

(as of April, 2017)

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

BLAST
Find regions of similarity between your sequences

Sequence alignments

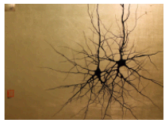
UniProt data

Download latest release
Get the UniProt data

Statistics
View Swiss-Prot and TrEMBL statistics

How to cite us
The UniProt Consortium

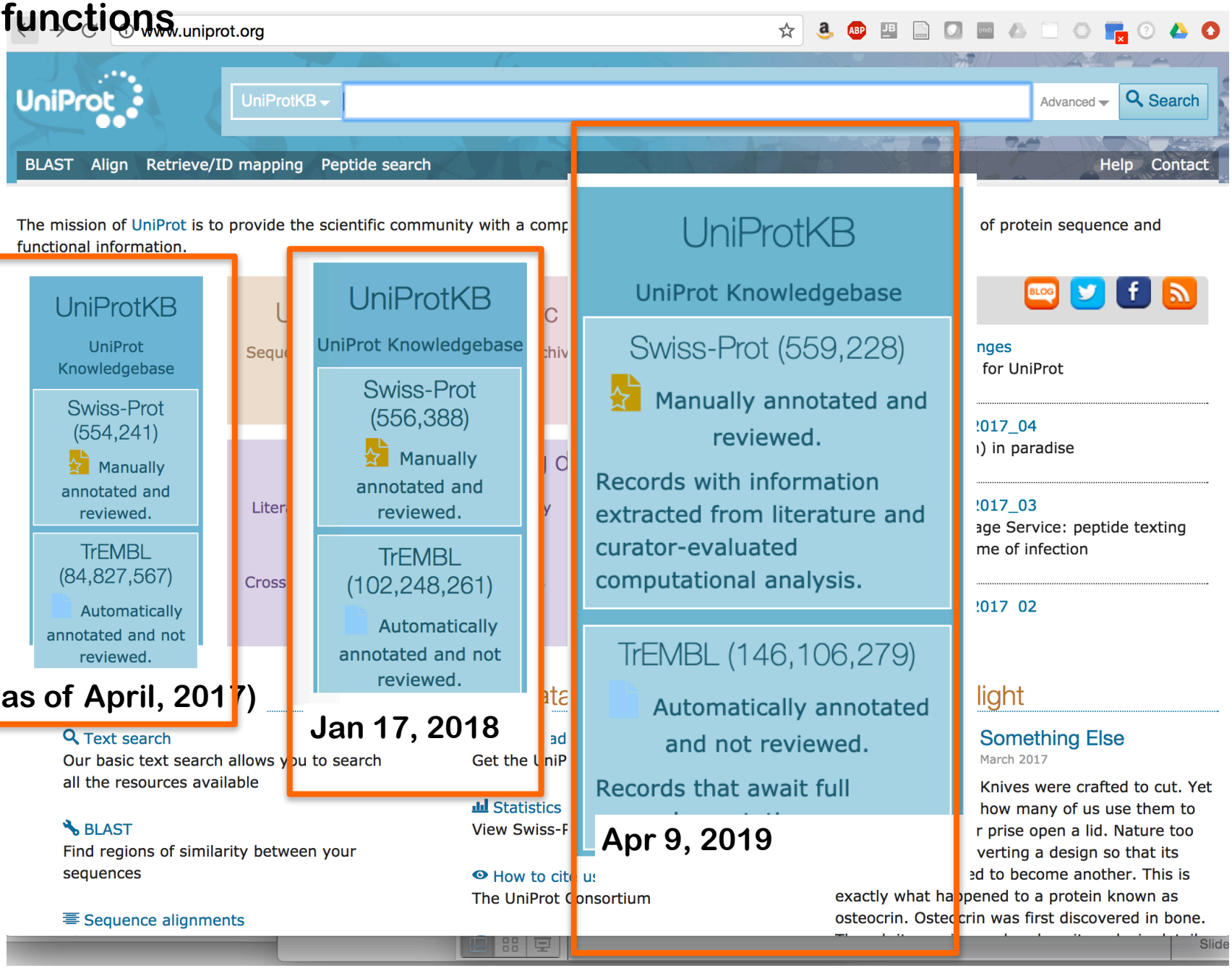
Protein spotlight



Something Else
March 2017

Knives were crafted to cut. Yet how many of us use them to slit an envelope. Or prise open a lid. Nature too has its ways of perverting a design so that its initial role is altered to become another. This is exactly what happened to a protein known as osteocrin. Osteocrin was first discovered in bone.

The Swiss-Prot database is a valuable source of proteins with known functions



Example of a Swiss-Prot Record

www.uniprot.org/uniprot/Q9H479

UniProt | UniProtKB | Advanced | Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

UniProtKB - Q9H479 (FN3K_HUMAN)

Basket

Display

- Entry
- Publications
- Feature viewer
- Feature table

None

- Function
- Names & Taxonomy
- Subcell. location
- Pathol./Biotech
- PTM / Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Cross-references
- Entry information
- Miscellaneous

BLAST | Align | Format | Add to basket | History | Feedback | Help video | Other tutorials and videos

Protein: **Fructosamine-3-kinase**

Gene: **FN3K**

Organism: *Homo sapiens (Human)*

Status: Reviewed - Annotation score: ●●●●○ - Experimental evidence at protein level

Function

May initiate a process leading to the deglycation of fructoselysine and of glycated proteins. May play a role in the phosphorylation of 1-deoxy-1-morpholinofructose (DMF), fructoselysine, fructoseglycine, fructose and glycated lysozyme.

GO - Molecular function

- fructosamine-3-kinase activity — Source: UniProtKB
- kinase activity — Source: Reactome

Complete GO annotation...

GO - Biological process

- epithelial cell differentiation — Source: UniProtKB
- fructosamine metabolic process — Source: GO_Central
- fructoselysine metabolic process — Source: UniProtKB
- post-translational protein modification — Source: Reactome

Complete GO annotation...

Keywords

Molecular | Kinase | Transferase

Gene Ontology (GO): Structured vocabulary for defining molecular functions, biological processes, and cellular components.

No significant sequence similarity… What else?

GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC
TTGTCATCACCACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCCAGCTGGCCATAGCCTGGTGCCTGA
GGAATGAGGGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACTTATGGAGA
ACATTGGAGCAATACAGGTCCTTCCAAAATTGTCGTCTTCCATCGTCCACGAGATCGACA
GCATTCTGGGCAATAAACCCTACAGCAAAAAGGACTATAGATCCTAAGTCTACTCCCCAC

Is there an ORF for a potential Coding Region?

GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC
TTGTCATCACCACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCCAGCTGGCCATAGCCTGGTGCCTGA
GGAATGAGGGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACTTATGGAGA
ACATTGGAGCAATACAGGTCCTTCCAAAATTGTCGTCTTCCATCGTCCACGAGATCGACA
GCATTCTGGGCAATAAACCCTACAGCAAAAAGGACTATAGATCCTAAGTCTACTCCCCAC

Is there an ORF for a potential Coding Region?

GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCAC**ATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT**
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC
TTGTCATCACCACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCCAGCTGGCCATAGCCTGGTGCCTGA
GGAATGAGGGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACTTATGGAGA
ACATTGGAGCAATACAGGTCCTTCCAAAATTGTCGTCTTCCATCGTCCACGAGATCGACA
GCATTCTGGGCAATAAACCCTACAGCAAAAAGGACTATAGATCCTAAGTCTACTCCCCAC

Find all ORFs using ORFfinder

Secure https://www.ncbi.nlm.nih.gov/orffinder/

NCBI Resources How To Sign in to NCBI

ORFfinder PubMed Search

Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for Linux x64.

Examples (click to set values, then click Submit button) :

- NC_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

Enter Query Sequence

Enter accession number, gi, or nucleotide sequence in FASTA format:

```
GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
```

From: **To:**

ORFfinder finds all open reading frames and provides translations

Secure https://www.ncbi.nlm.nih.gov/orffinder/

NCBI Resources How To Sign in to NCBI

ORFfinder PubMed Search

Open Reading Frame Viewer

ORFs can appear in random sequence – so further analysis is required

Sequence

ORFs found: 12 Genetic code: 1 Start codon: 'ATG' only

1: 1..1.8K (1.8Kbp) Find: Tools Tracks

ORFinder_4.25.202734829

ORF4 ORF5 ORF5 ORF6 ORF1 ORF11 ORF3 ORF7 ORF8 ORF9 ORF12 ORF2 ORF10

Ready

Predict coding vs. non-coding ORFs: http://TransDecoder.github.io

Add six-frame translation track

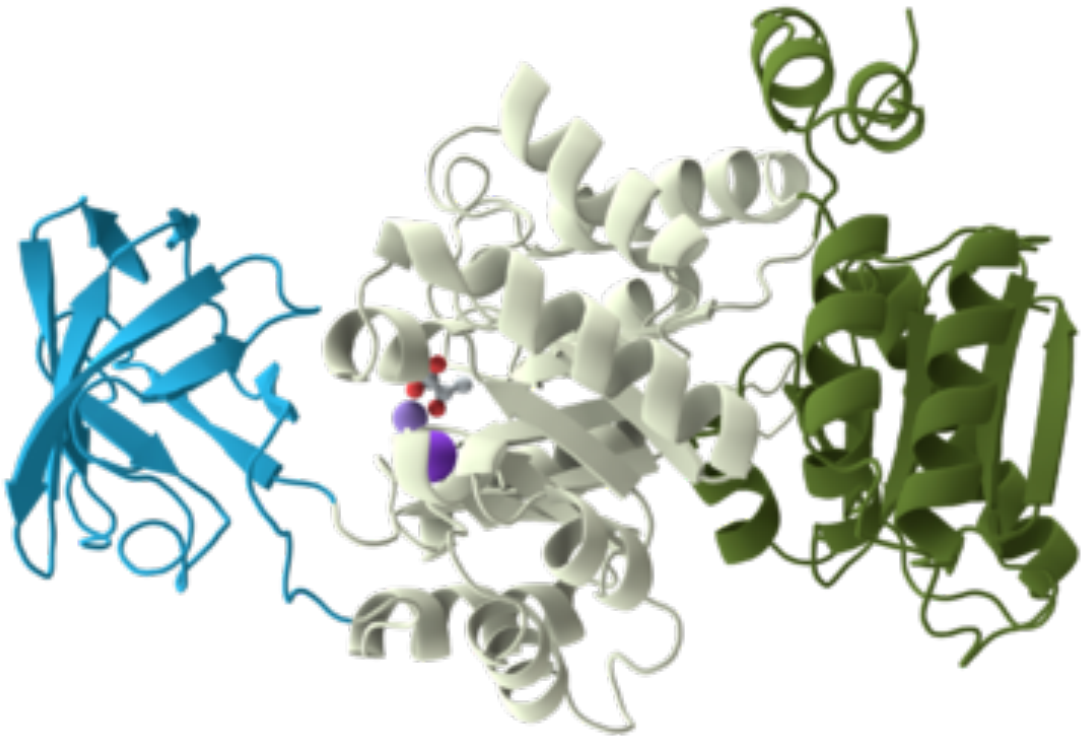
ORF5 (367 aa) Display ORF as... Mark

```
>lcl|ORF5
MYPESTTGSPARLSLRQTGSPGMIYSTRYGSPKRQLQFYR
NLGKSGLRVSCLGLGTWVTFGGQITDEMAEHLMTLAYDNG
INLFDTAEVYAAGKAEVVLGNIIKKKGWRRSSLVITTKIF
WGGKAETERGLSRKHIIEGLKASLERLQLEYVDVVFANRP
DPNTPMEETVRAMTHVINQGMAMYWGTSRWSSMEIMEAYS
VARQFNLIPPICEQAEYHMFQREKVEVQLPELFHKIGVGA
MTWSPLACGIVSGKYDSGIPPYSRASLKGYQWLKDKILSE
EGRRQQAKLKELQAIAERLGCTLPQLAIAWCLRNEGVSSV
LLGASNAEQLMENIGAIQVLPKLSSSIVHEIDSILGNKPY
SKKDYRS
```

Mark subset... Marked: 0 Download marked set as Protein FA

Label	Strand	Frame	Start	Stop	Length (nt |
ORF5	**+**	**3**	**324**	**1427**	**1104 | 36**
ORF3	+	1	1264	1758	495 | 16
ORF7	-	1	492	103	390 | 12
ORF11	-	3	910	590	321 | 10
ORF9	-	3	1384	1130	255 | 8
ORF12	-	3	325	86	240 | 7
ORF8	-	2	848	618	231 | 7

Can we recognize functional domains in putative coding regions?



Hints at substrate binding or catalytic activity

DNA, RNA, calcium, phoshate, etc.

Glycoslase, methylase, kinase, nuclease, lipase, protease, etc.

Search the Pfam library of HMMs to identify potential functional domains

pfam.xfam.org

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

Pfam 31.0 (March 2017, 16712 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. **More...**

QUICK LINKS

SEQUENCE SEARCH

VIEW A PFAM ENTRY

VIEW A CLAN

VIEW A SEQUENCE

VIEW A STRUCTURE

KEYWORD SEARCH

JUMP TO

ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES

Paste your protein sequence here to find matching Pfam entries.

```
METGGRARTGTPQPAAPGVWRARPAGGGGGGASSWLLDGNSWLLCYGFLY
LALYAQVSQSKPCERTGSCFSGRCVNSTCLCDPGWVGDQCQHCQGRFKLT
EPSGYLTDGPINYKYKTKCTWLIEGYPNAVLRLRFNHFATECSWDHMYVY
DGDSIYAPLIAVLSGLIVPEIRGNETVPEVVTTSGYALLHFFSDAAYNLT
GFNIFYSINSCPNNCSGHGKCTTSVSVPSQVYCECDKYWKGEACDIPYCK
ANCGSPDHGYCDLTGEKLCVCNDSWQGPDCSLNVPSTESYWILPNVKPFS
PSVGRASHKAVLHGKFMWVIGGYTFNYSSFQMVLNYNLESSIWNVGTPSR
GPLQRYGHSLALYQENIFMYGGRIETNDGNVTDELWVFNIHSQSWSTKTP
TVLGHGQQYAVEGHSAHIMELDSRDVVMIIIFGYSAIYGYTSSIQEYHIS
SNTWLVPETKGAIVQGGYGHTSVYDEITKSIYVHGGYKALPGNKYGLVDD
LYKYEVNTKTWTILKESGFARYLHSAVLINGAMLIFGGNTHNDTSLSNGA
KCFSADFLAYDIACDEWKILPKPNLHRDVNRFGHSAVVINGSMYIFGGFS
SVLLNDILVYKPPNCKAFRDEELCKNAGPGIKCVWNKNHCESWESGNTNN
ILRAKCPPKTAASDDRCYRYADCASCTANTNGCQWCDDKKCISANSNCSM
SVKNYTKCHVRNEQICNKLTSCKSCSLNLNCQWDQRQQECQALPAHLCGE
GWSHIGDACLRVNSSRENYDNAKLYCYNLSGNLASLTTSKEVEFVLDEIQ
KYTQQKVSPWVGLRKINISYWGWEDMSPFTNTTLQWLPGEPNDSGFCAYL
ERAAVAGLKANPCTSMANGLVCEKPVVSPNQNARPCKKPCSLRTSCSNCT
SNGMECMWCSSTKRCVDSNAYIISFPYGQCLEWQTATCSPQNCSGLRTCG
QCLEQPGCGWCNDPSNTGRGHCIEGSSRGPMKLIGMHHSEMVLDTNLCPK
EKNYEWSFIQCPACQCNGHSTCINNNVCEQCKNLTTGKQCQDCMPGYYGD
PTNGGQCTACTCSGHANICHLHTGKCFCTTKGIKGDQCQLCDSENRYVGN
PLRGTCYYSLLIDYQFTFSLLQEDDRHHTAINFIANPEQSNKNLDISINA
SNNFNLNITWSVGSTAGTISGEETSIVSKNNIKEYRDSFSYEKFNFRSNP
NITFYVYVSNFSWPIKIQIAFSQHNTIMDLVQFFVTFFSCFLSLLLVAAV
VWKIKQTCWASRRREQLLRERQQMASRPFASVDVALEVGAEQTEFLRGPL
EGAPKPIAIEPCAGNRAAVLTVFLCLPRGSSGAPPPGQSGLAIASALIDI
SQQKASDSKDKTSGVRNRKHLSTRQGTCV
```

Go

Example

This search will use and an E-value of 1.0. You can set your own search parameters and perform a range of other searches here.

Example Pfam report illustrating modular domain architecture

pfam.xfam.org/search/sequence

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

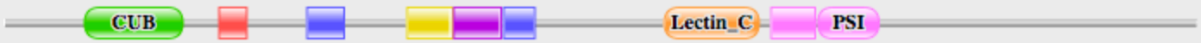
Pfam

keyword search Go

Sequence search results

Show the detailed description of this results page.

We found **9** Pfam-A matches to your search sequence (**all** significant)



Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

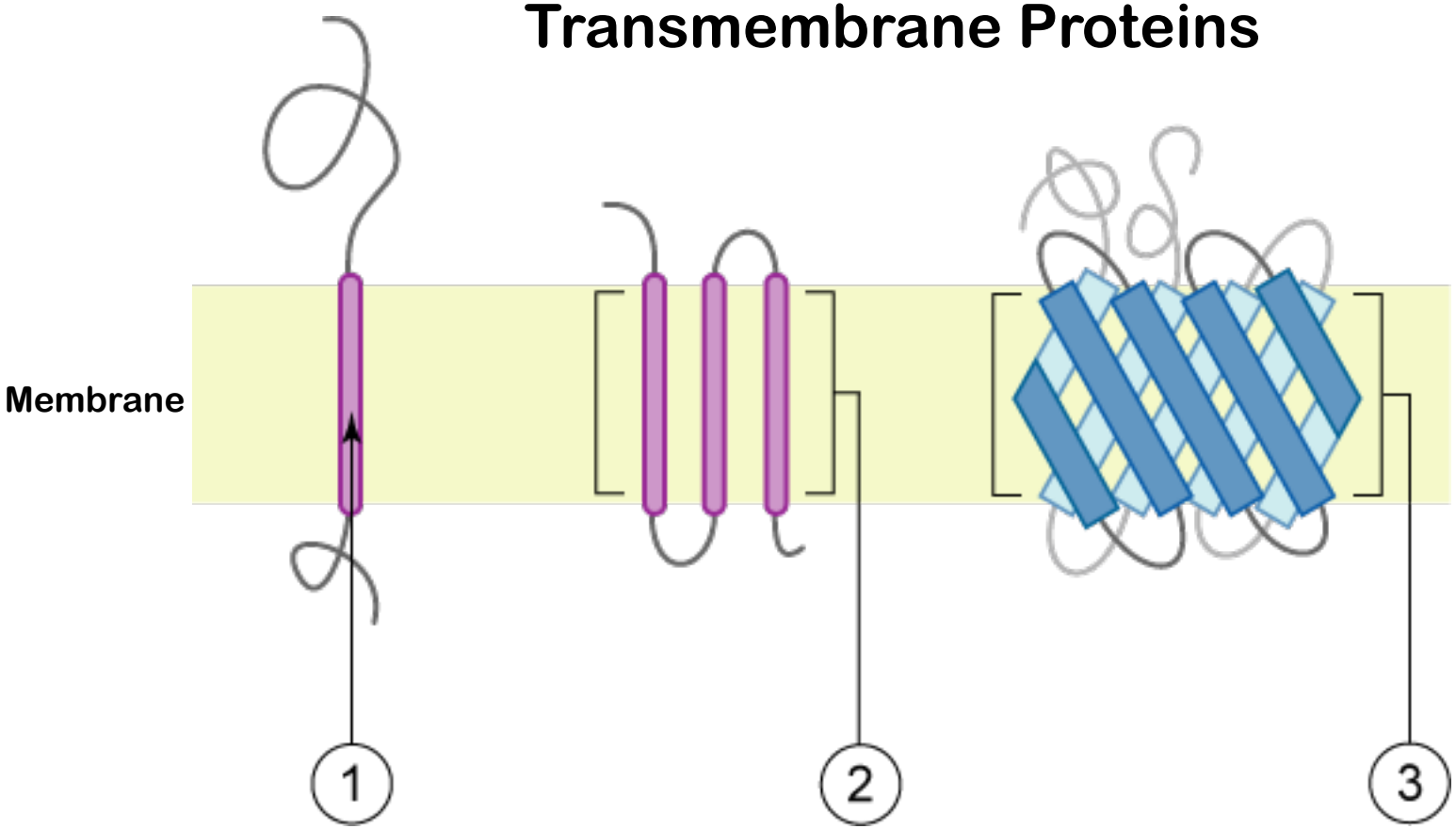
Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
CUB	CUB domain	Domain	CL0164	93	206	93	206	1	110	110	42.2	7.7e-11	n/a	Show
EGF_2	EGF-like domain	Domain	CL0001	249	280	249	280	1	32	32	22.5	0.0001	n/a	Show
Kelch_5	Kelch motif	Repeat	CL0186	351	393	352	392	**2**	**41**	42	33.7	2.2e-08	n/a	Show
Kelch_4	Galactose oxidase, central domain	Repeat	CL0186	466	518	468	514	**3**	**44**	49	20.6	0.0003	n/a	Show
Kelch_1	Kelch motif	Repeat	CL0186	520	574	520	573	1	**45**	46	20.0	0.00033	n/a	Show
Kelch_5	Kelch motif	Repeat	CL0186	579	614	581	613	**5**	**40**	42	25.3	9.7e-06	n/a	Show
Lectin_C	Lectin C-type domain	Domain	CL0056	765	874	766	874	**2**	108	108	70.2	2e-19	n/a	Show
PSI	Plexin repeat	Family	CL0630	889	939	890	938	**2**	**50**	51	27.8	2.5e-06	n/a	Show
PSI	Plexin repeat	Family	CL0630	942	1012	942	1012	1	51	51	50.0	2.9e-13	n/a	Show

Comments or questions on the site? Send a mail to **pfam-help@ebi.ac.uk**.

European Molecular Biology Laboratory

Transmembrane Proteins



Single transmembrane α-helix (bitopic membrane protein)

Polytopic transmembrane α-helical protein

Polytopic transmembrane β-sheet protein

From: https://en.wikipedia.org/wiki/Transmembrane_protein

Using TMHMM to identify putative transmembrane proteins

www.cbs.dtu.dk/services/TMHMM/

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

EVENTS | NEWS | RESEARCH GROUPS | CBS PREDICTION SERVERS | CBS DATA SETS | PUBLICATIONS | EDUCATION

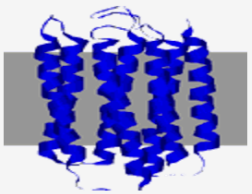
STAFF | CONTACT | ABOUT CBS | INTERNAL | CBS BIOINFORMATICS TOOLS | CBS COURSES | OTHER BIOINFORMATICS LINKS

CBS >> CBS Prediction Servers >> TMHMM

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS ■ TECHNICAL UNIVERSITY OF DENMARK DTU

TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins



Instructions

SUBMISSION

Submission of a local file in FASTA format (HTML 3.0 or higher)

Choose File No file chosen

OR by pasting sequence(s) in FASTA format:

```
MEILCEDNTSLSSIPNSLMQVDGDSGLYRNDFNSRDANSSDASNWTIDGENRTNLSFEG
YLPPTCLSILHLQEKNWSALLTAVVIILTIAGNILVIMAVSLEKKLQNATNYFLMSLAIADMLL
GFLVMPVSMLTILYGYRWPLPSKLCAVWIYLDVLFSTASIMHLCAISLDRYVAIQNPIHHSR
FNSRTKAFLKIIAVWTISVGVSMPIPVFGLQDDSKVFKQGSCLLADDNFVLIGSFVAFFIPLTI
MVITYFLTIKSLQKEATLCVSDLSTRAKLASFSFL
```

Output format:

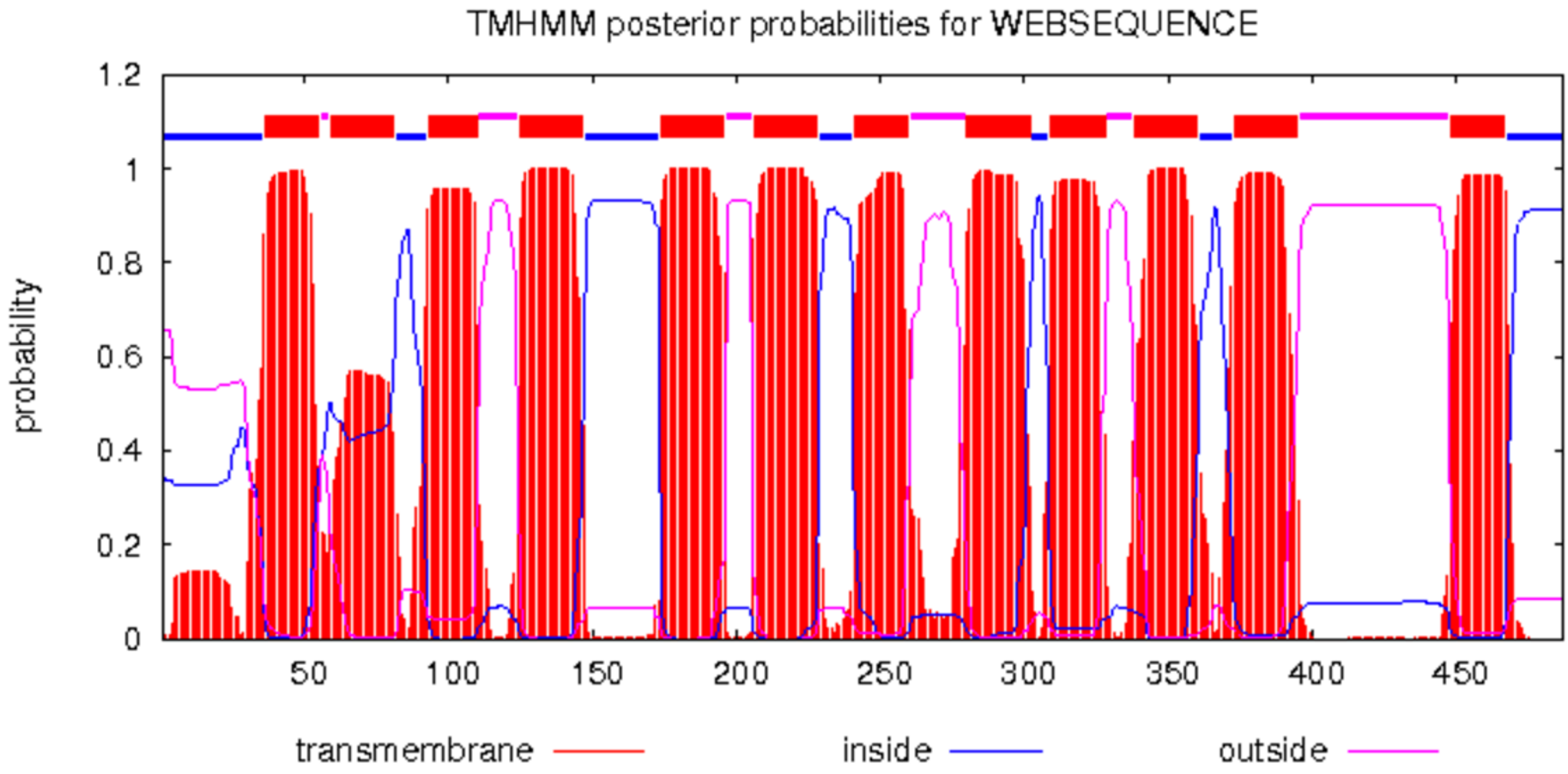
- (•) Extensive, with graphics
- ( ) Extensive, no graphics
- ( ) One line per protein

Other options:

- [ ] Use old model (version 1)

Submit Clear

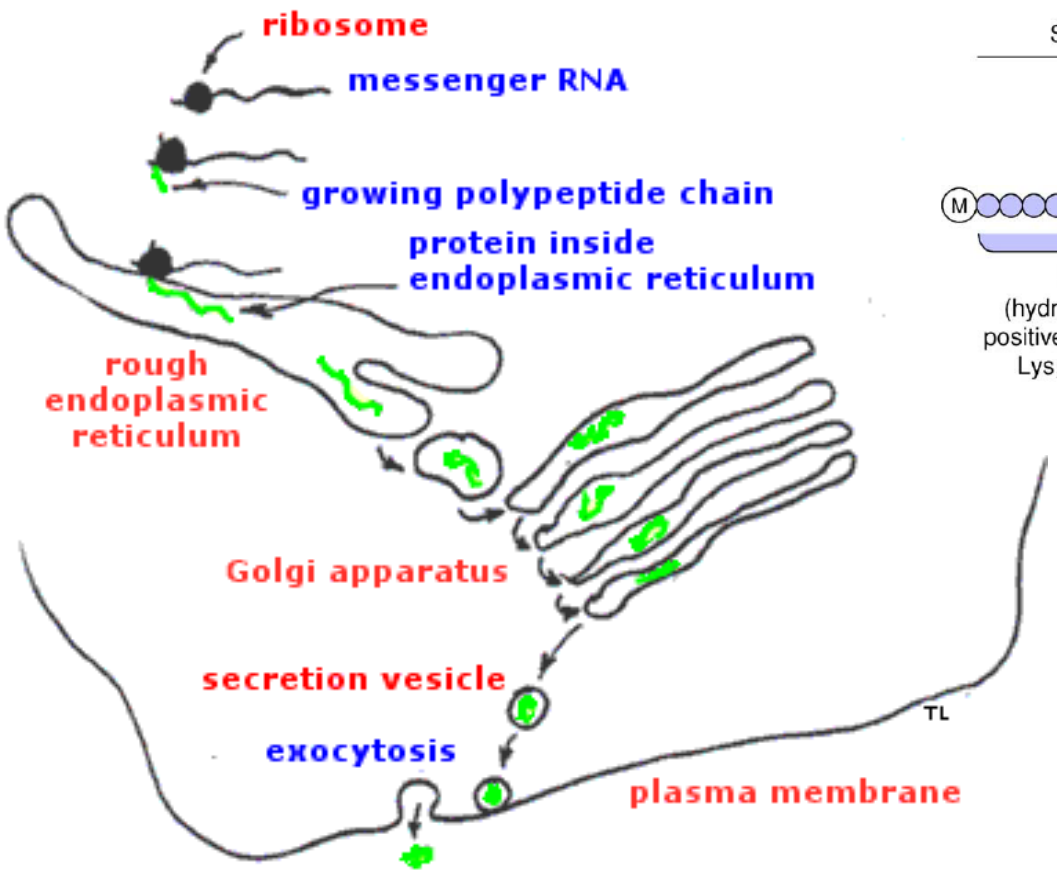
Trans-membrane Domains via TmHMM



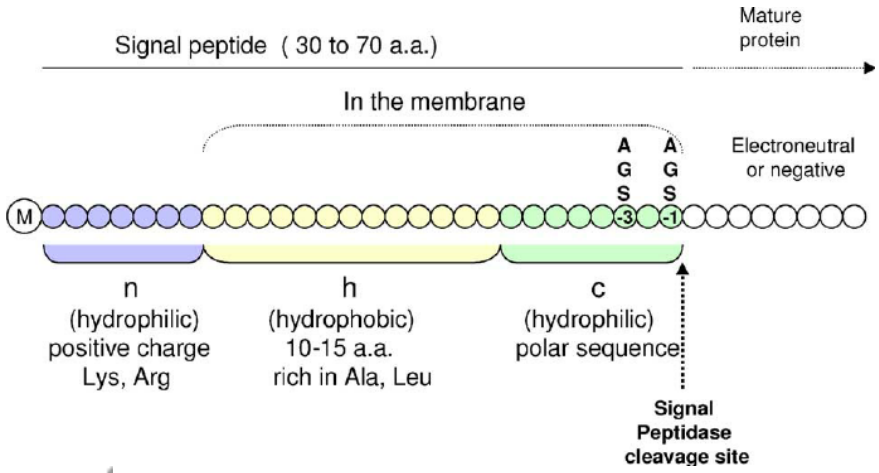
**Topology=i36-55o59-81i93-110o125-147i174-196o206-228i241-260o280-302i309-328o338-360i373-395
448-467i**

http://www.cbs.dtu.dk/services/TMHMM/

Predicting Secreted Proteins



(from: https://courses.washington.edu/conj/cell/secretion.htm)



(from: Vaccine 23(15):1770-8)

SignalP: Prediction of N-terminal signal peptides
(predict secreted proteins)

www.cbs.dtu.dk/services/SignalP/

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

EVENTS | NEWS | RESEARCH GROUPS | CBS PREDICTION SERVERS | CBS DATA SETS | PUBLICATIONS | EDUCATION

STAFF | CONTACT | ABOUT CBS | INTERNAL | CBS BIOINFORMATICS TOOLS | CBS COURSES | OTHER BIOINFORMATICS LINKS

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS ■ TECHNICAL UNIVERSITY OF DENMARK DTU

CBS >> CBS Prediction Servers >> SignalP

SignalP 4.1 Server

SignalP 4.1 server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks.

View the version history of this server. All the previous versions are available on line, for comparison and reference.

NEW: The portable version of SignalP 4.1, previously only available for Mac (Darwin), Linux, and IRIX, is now also available for Windows systems. Academic users: select the "CYGWIN" option at the download page. Cygwin or MobaXterm is required to install SignalP under Windows. For details, read the installation instructions.

FAQ | Article abstracts | Instructions | Output format | Performance | Data

SUBMISSION

Paste a single amino acid sequence or several sequences in FASTA format into the field below:

```
MHPAVFLSLPDLRCSLLLLVTWVFTPVTTEITSLDTENIDEILNNADVALVNFYADWCRFSQMLHPIFEEASDVIKEEFPNENQVVFARVDCDQHSDIAQRYRISKYPTLKLFRNGMMM
KREYRGQRSVKALADYIRQQKSDPIQEIRDLAEITTLDRSKRNIIGYFEQKDSDNYRVFERVANILHDDCAFLSAFGDVSKPERYSGDNIIYKPPGHSAPDMVYLGAMTNFDVTYNWIQ
DKCVPLVREITFENGEELTEEGLPFLILFHMKEDTESLEIFQNEVARQLISEKGTINFLHADCDKFRHPLLHIQKTPADCPVIAIDSFRHMYVFGDFKDVLIPGKLKQFVFDLHSGKLHREF
HHGPDPTDTAPGEQAQDVASSPPESSFQKLAPSEYRYTLLRDRDEL
```

Submit a file in FASTA format directly from your local disk:

Choose File No file chosen

Organism group (explain)
- (●) Eukaryotes
- ( ) Gram-negative bacteria
- ( ) Gram-positive bacteria

D-cutoff values (explain)
- (●) Default (optimized for correlation)
- ( ) Sensitive (reproduce SignalP 3.0's sensitivity)
- ( ) User defined:
  - 0.4 *D-cutoff for SignalP-noTM networks*
  - 0.5 *D-cutoff for SignalP-TM networks*

Graphics output (explain)
- ( ) No graphics
- (●) **PNG** (inline)
- ( ) **PNG** (inline) and **EPS** (as links)

Output format (explain)
- (●) Standard
- ( ) Short (no graphics)
- ( ) Long
- ( ) All - SignalP-noTM and SignalP-TM output (no graphics)

Method (explain)
- (●) Input sequences may include TM regions
- ( ) Input sequences do not include TM regions

Positional limits (explain)
- [ ] Minimal predicted signal peptide length. *Default: 10*
- [ ] N-terminal truncation of input sequence (0 means no truncation). *Default: Truncate sequence to a length of 70 aa*

Example SignalP predicted signal peptide

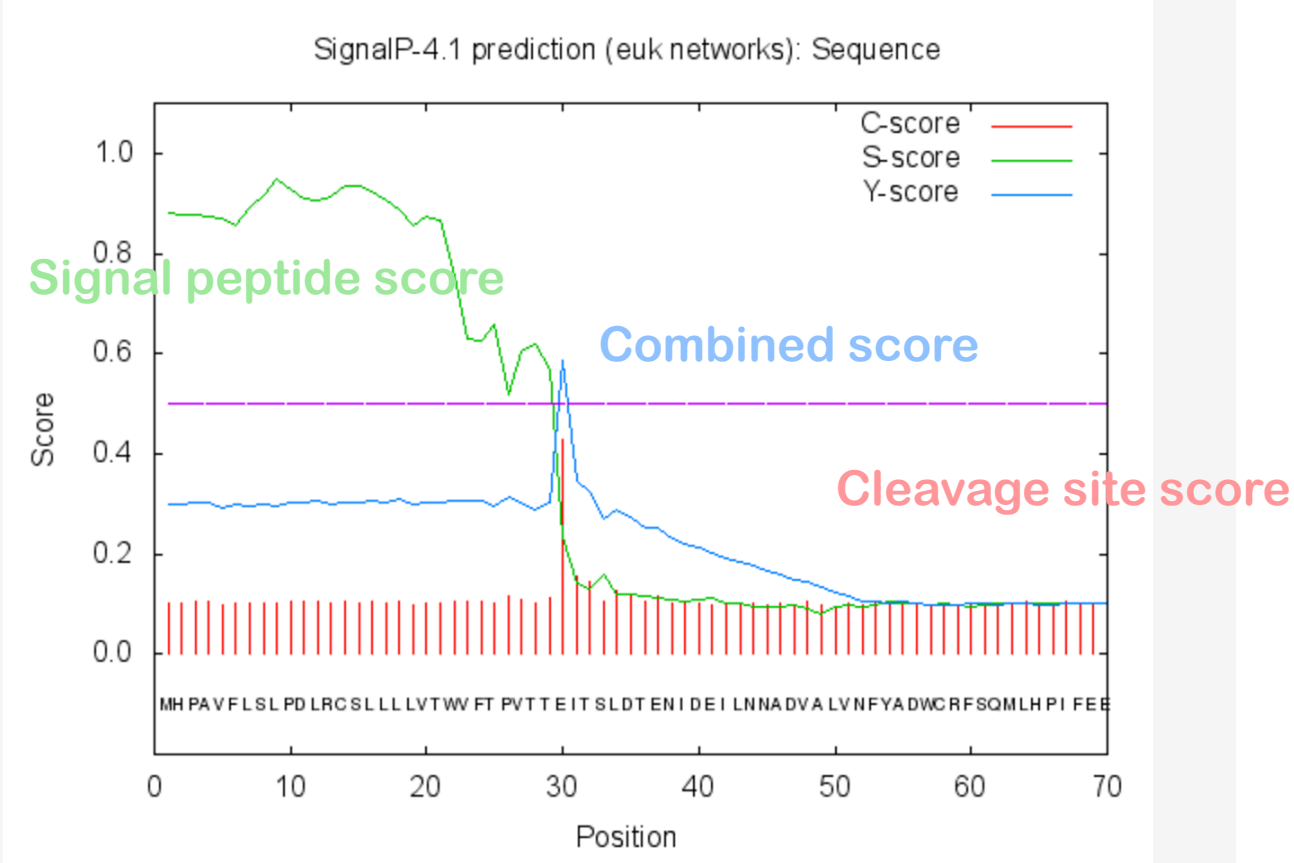
www.cbs.dtu.dk/cgi-bin/webface2.fcgi?jobid=58FFF29C00005F854B357EEA&w...

CENTERFOR BIOLOGICAL SEQUENCE ANALYSIS CBS

SignalP 4.1 Server - prediction results

Technical University of Denmark

```
# SignalP-4.1 euk predictions
>Sequence
```



Acknowledgements

BROAD INSTITUTE

* Brian Haas

University of New Hampshire

Andrew Lang
Jennifer Hill
Nhen Hunter

Other:

C. Titus Brown
Rob Patro
Torsten Seemann