Transcriptome Assembly

BIOINF 3005 /7160: Transcriptomics Applications

Dr Terry Bertozzi
South Australian Museum
University of Adelaide

June 1, 2020

Outline

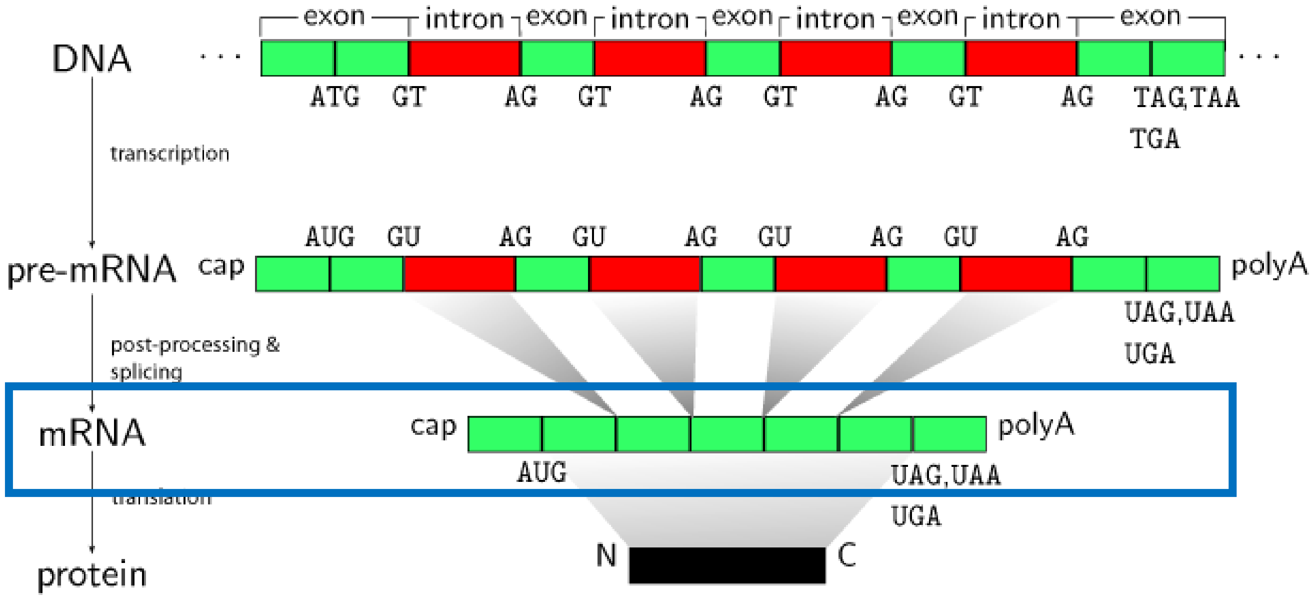
- Transcriptome recap
- Genome vs Transcriptome assembly
- Transcriptome assembly
    - Short read assembly
    - Long read methods
    - Guided assembly
- Transcriptome evaluation

Transcriptome

- **The set of all RNA transcripts, including coding and non-coding, in an individual or a population of cells.**

- mRNA
- lncRNA
- tRNA
- rRNA
- Small RNAs (e.g. miRNA, siRNA)

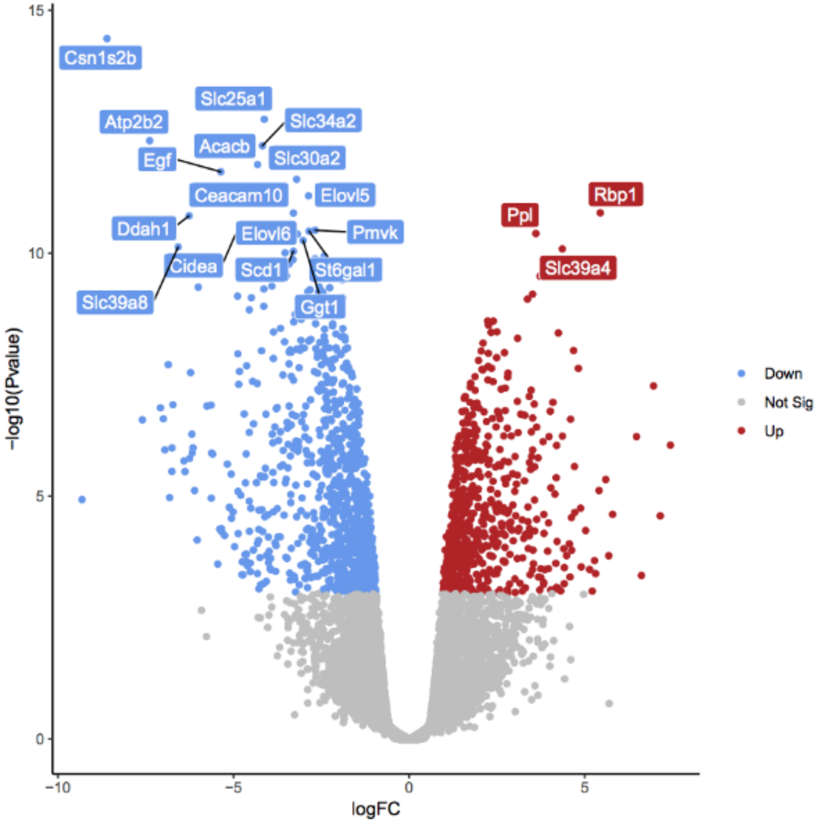
- Typically avoid rRNA
  - >80% total RNA
  - ~5% mRNA



http://svmcompbio.tuebingen.mpg.de

Transcriptome

- Differential expression
- Single cell clustering



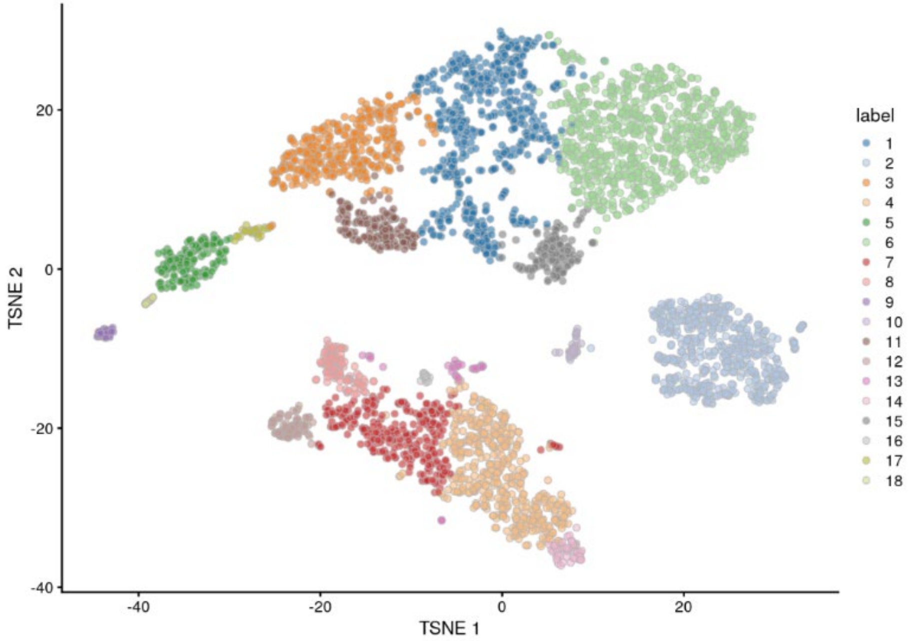


Image from https://galaxyproject.github.io

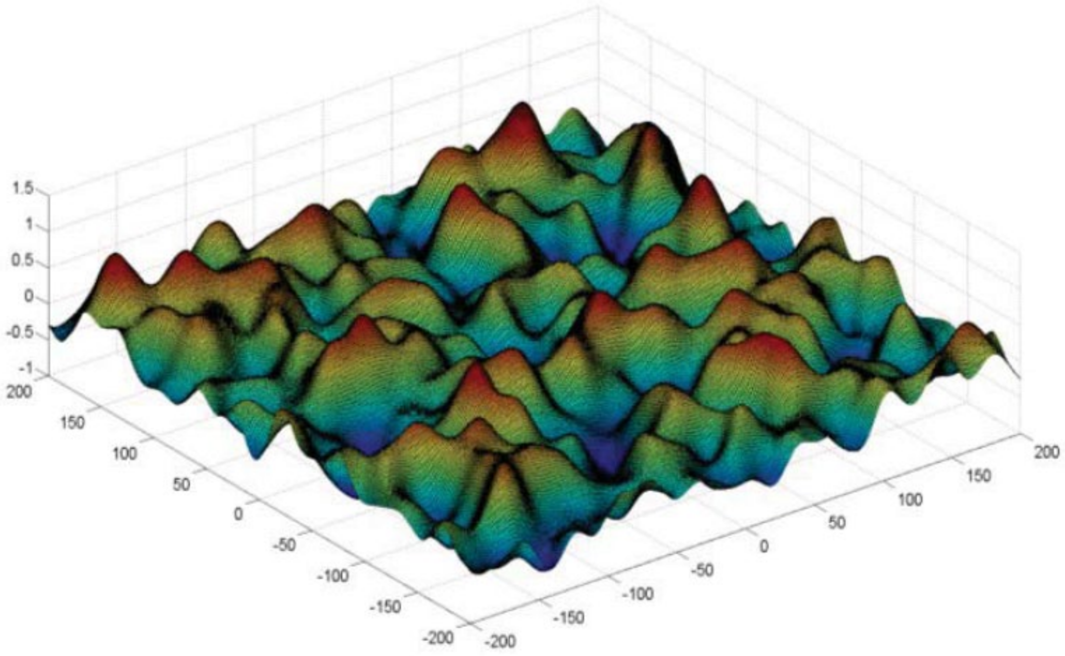
Genome vs Transcriptome assembly

Genome

- Uniform coverage

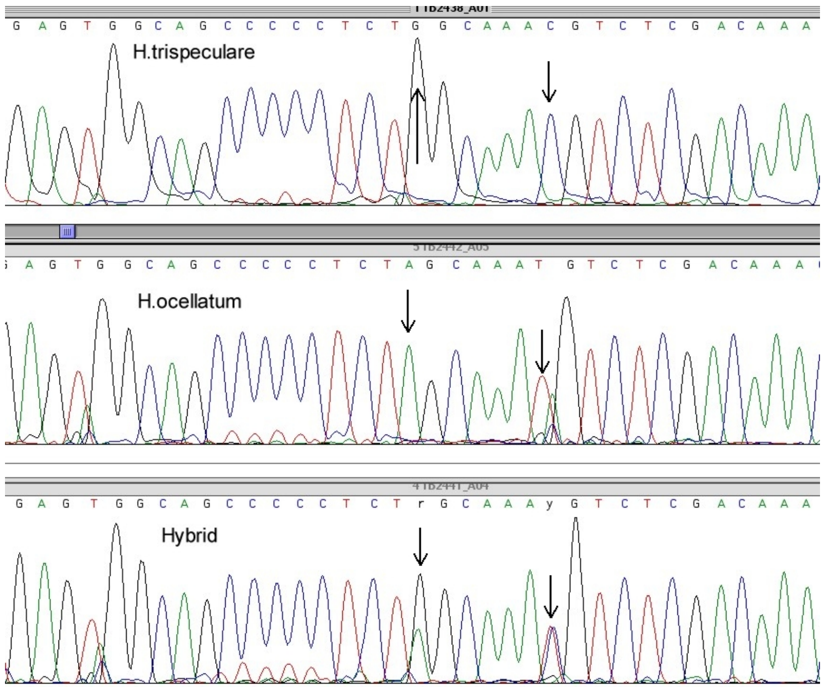
Transcriptome

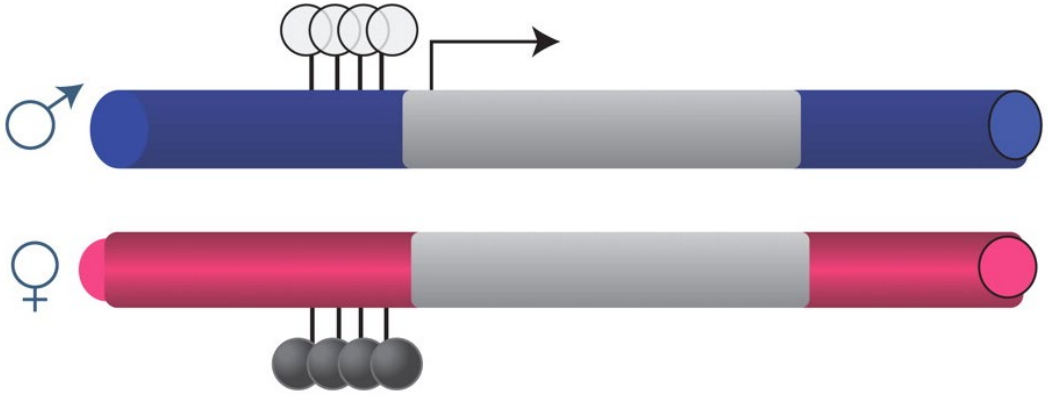
- Exponentially distributed coverage



Uneven coverage

- Cell/tissue compartmentalisation/function
  - Rare transcripts
- Allele specific expression
  - X-inactivation (Xist: X-inactive specific transcript)
  - Full / partial imprinting





Bartolomei 2011 Cold Spring Harbor Perspectives in Biology

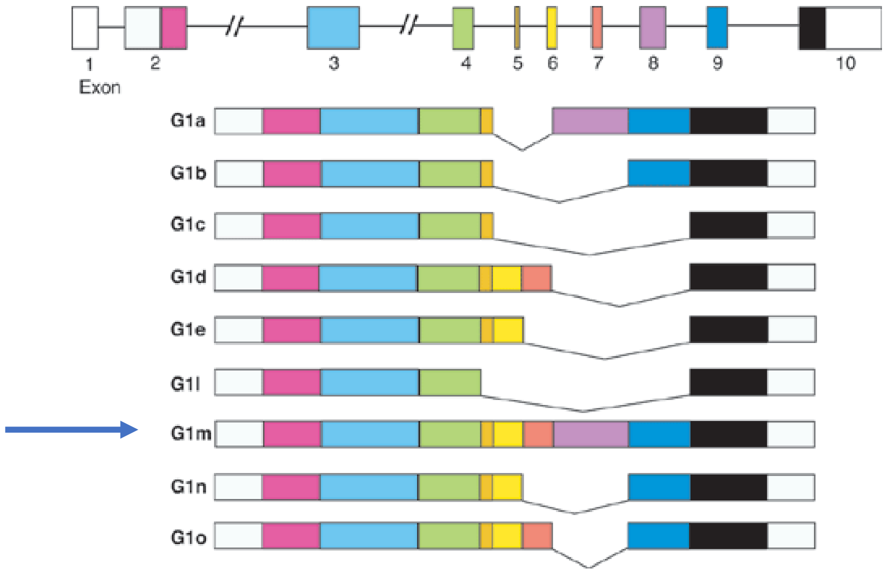
Genome vs Transcriptome assembly

Genome

- Uniform coverage
- Two contigs / locus

Transcriptome

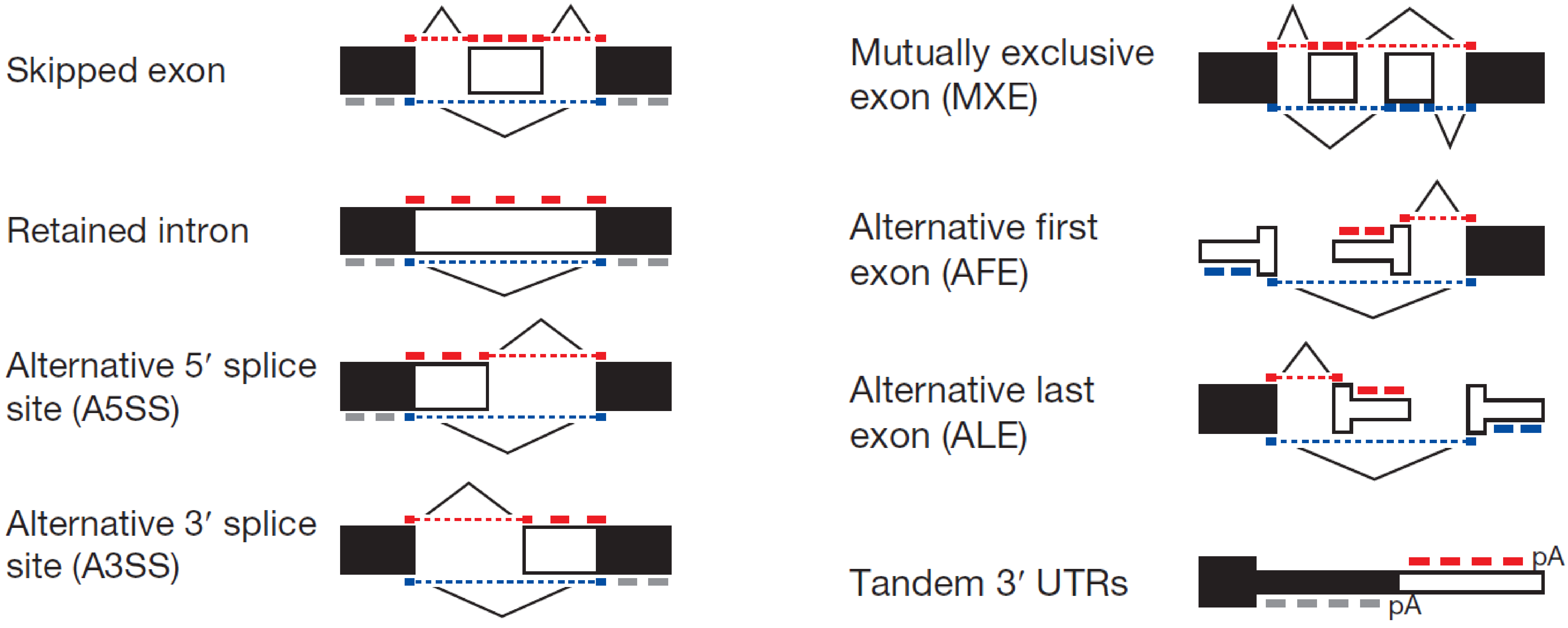
- Exponentially distributed coverage
- Multiple contigs / locus
  - isoforms



Canonical transcript

http://bsi.riken.jp/bsi-news/bsinews31/files/research0101-big.gif

Alternate Splicing



Skipped exon

Retained intron

Alternative 5′ splice site (A5SS)

Alternative 3′ splice site (A3SS)

Mutually exclusive exon (MXE)

Alternative first exon (AFE)

Alternative last exon (ALE)

Tandem 3′ UTRs

Wang et al 2008 Nature

Genome vs Transcriptome assembly

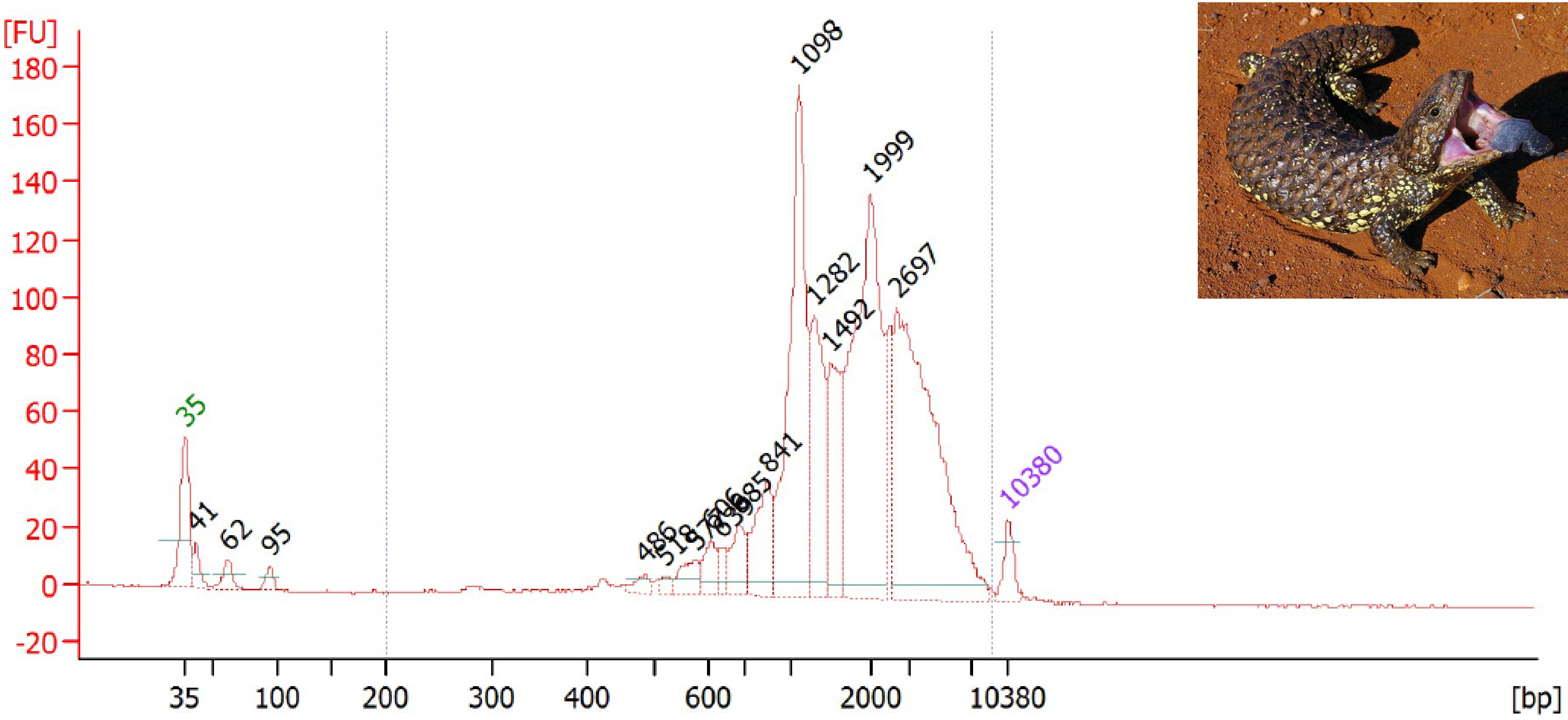
Genome

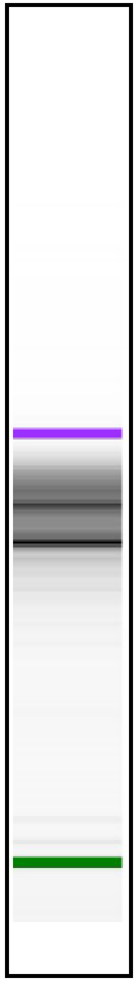
- Uniform coverage
- Single contig / locus
- Assemble small number of large Mb-length contigs

Transcriptome

- Exponentially distributed coverage
- Multiple contigs / locus
- Assemble thousands of Kb-length transcripts

Transcript distribution





Genome vs Transcriptome assembly

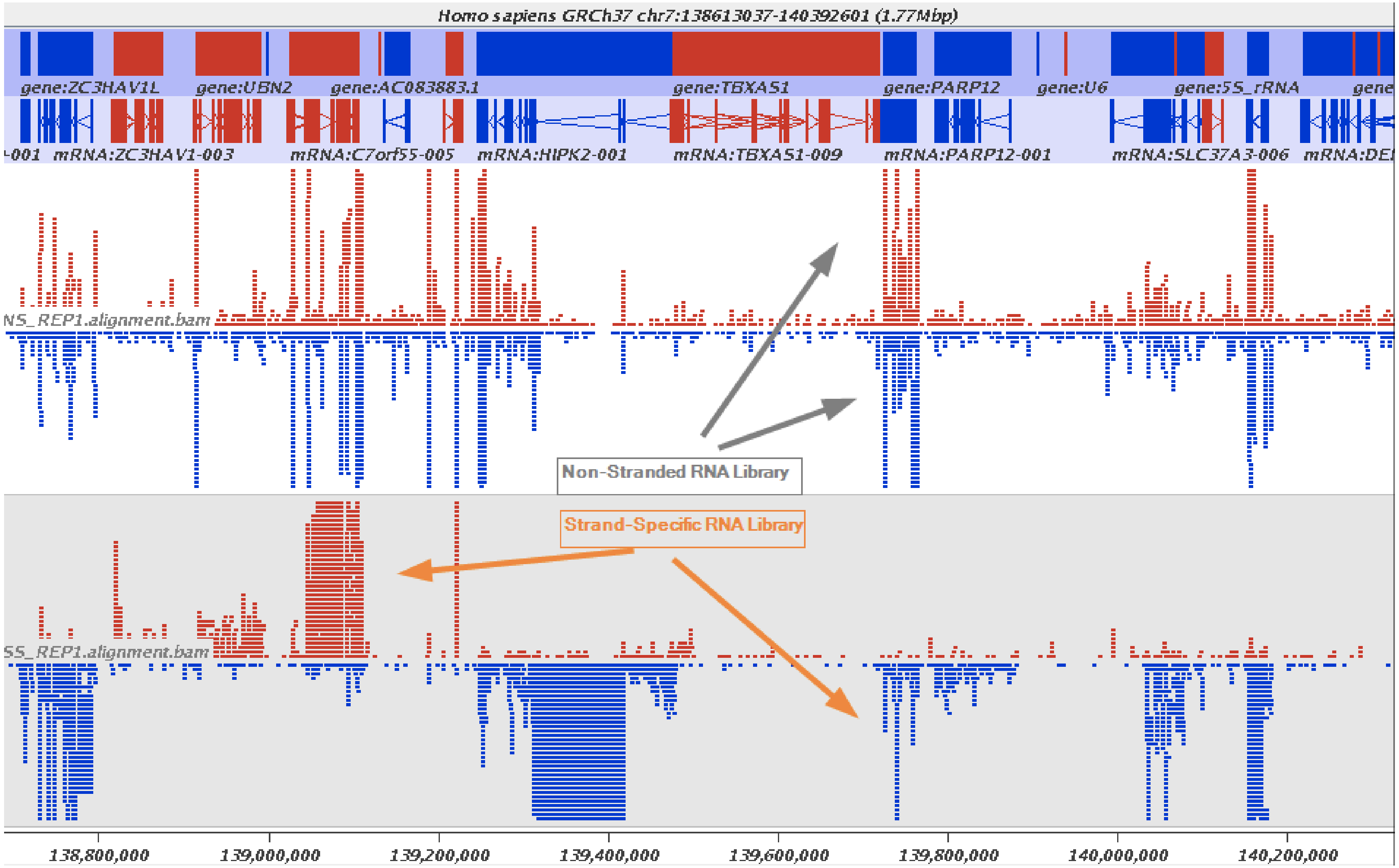
Genome

- Uniform coverage
- Single contig / locus
- Assemble small number of large Mb-length contigs
- Double stranded

Transcriptome

- Exponentially distributed coverage
- Multiple contigs / locus
- Assemble thousands of Kb-length transcripts
- Strand specific

Strand specific expression

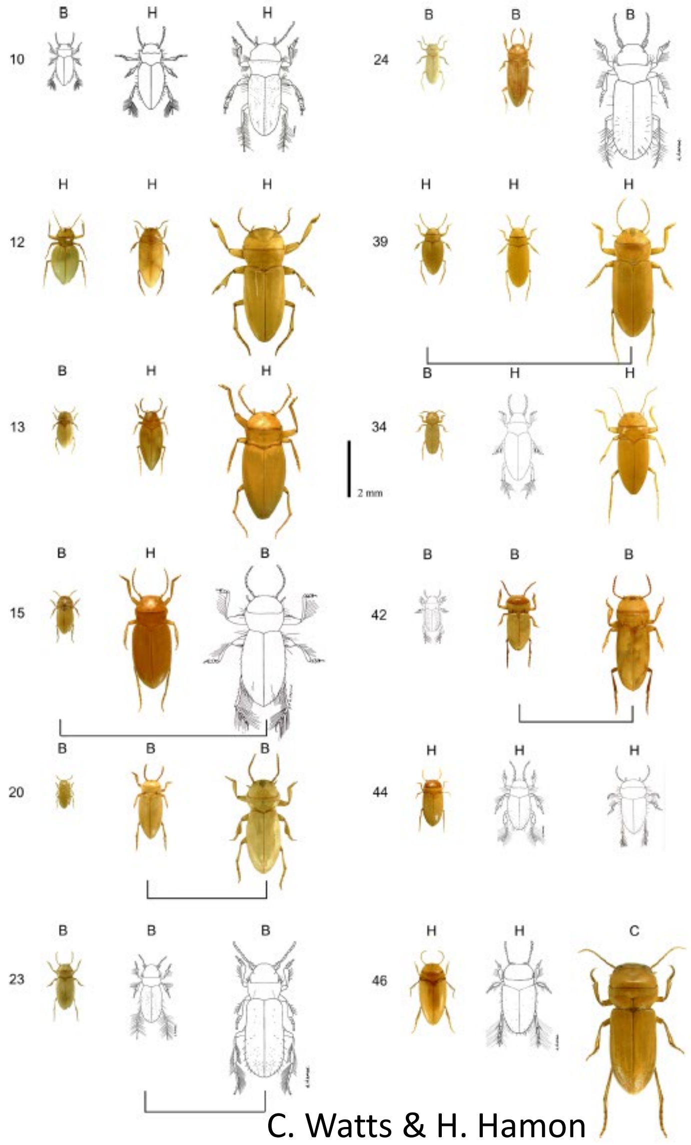


Non-Stranded RNA Library

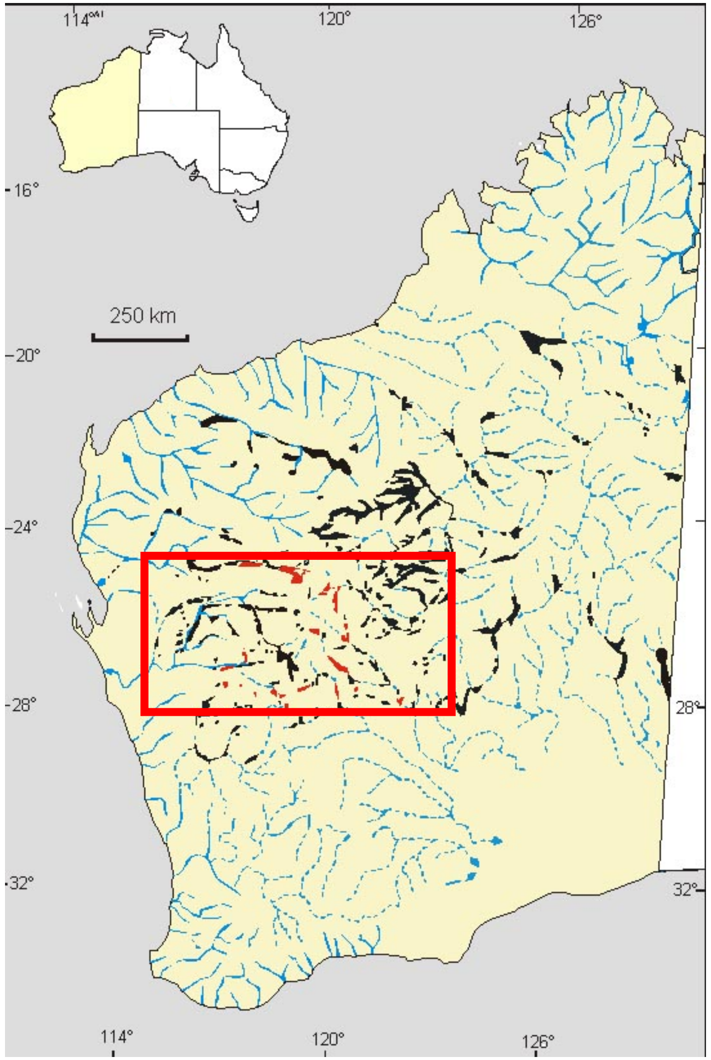
Strand-Specific RNA Library

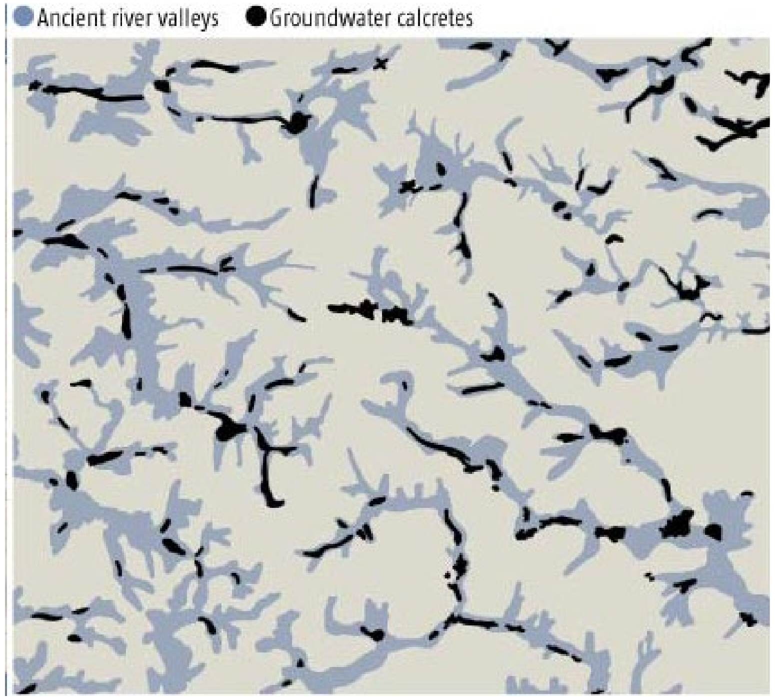
Assembly tools

Phototransduction in blind beetles

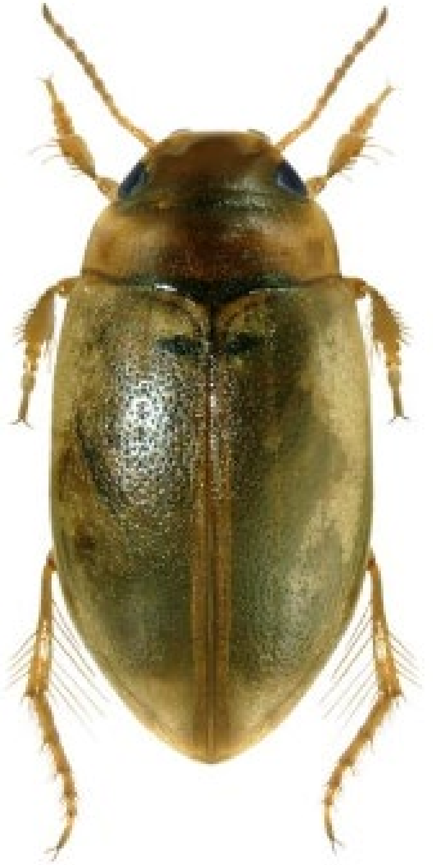


C. Watts & H. Hamon

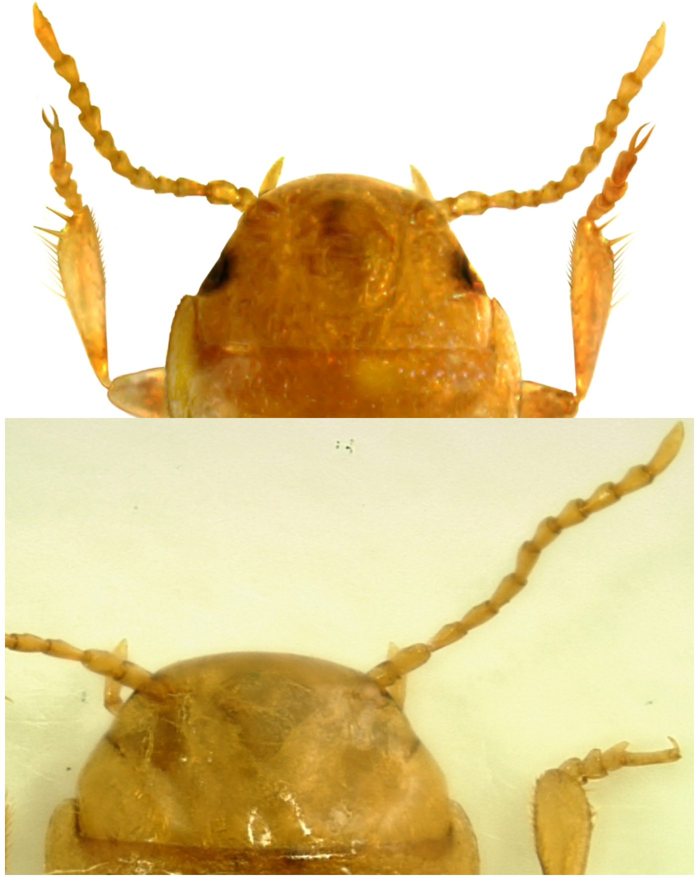




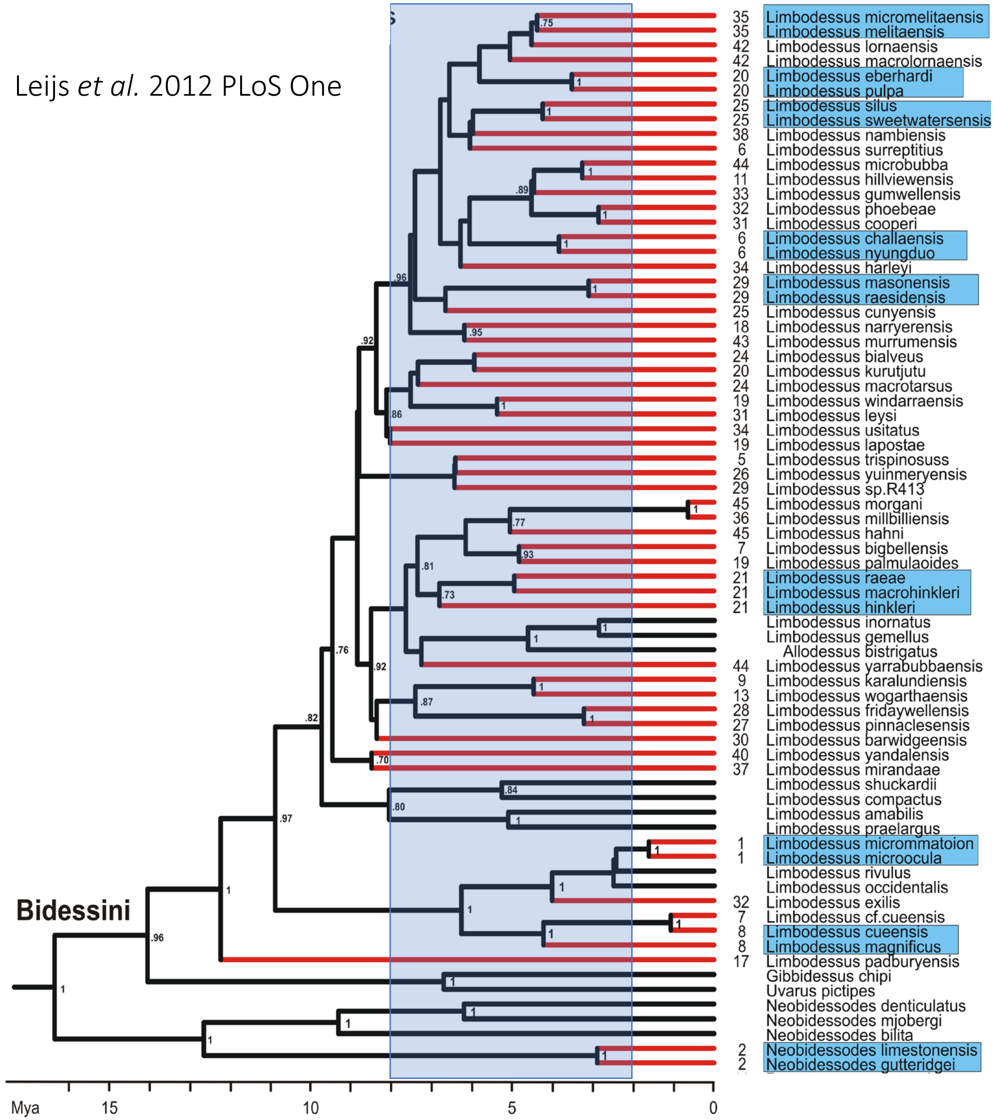
Phototransduction in blind beetles



Surface

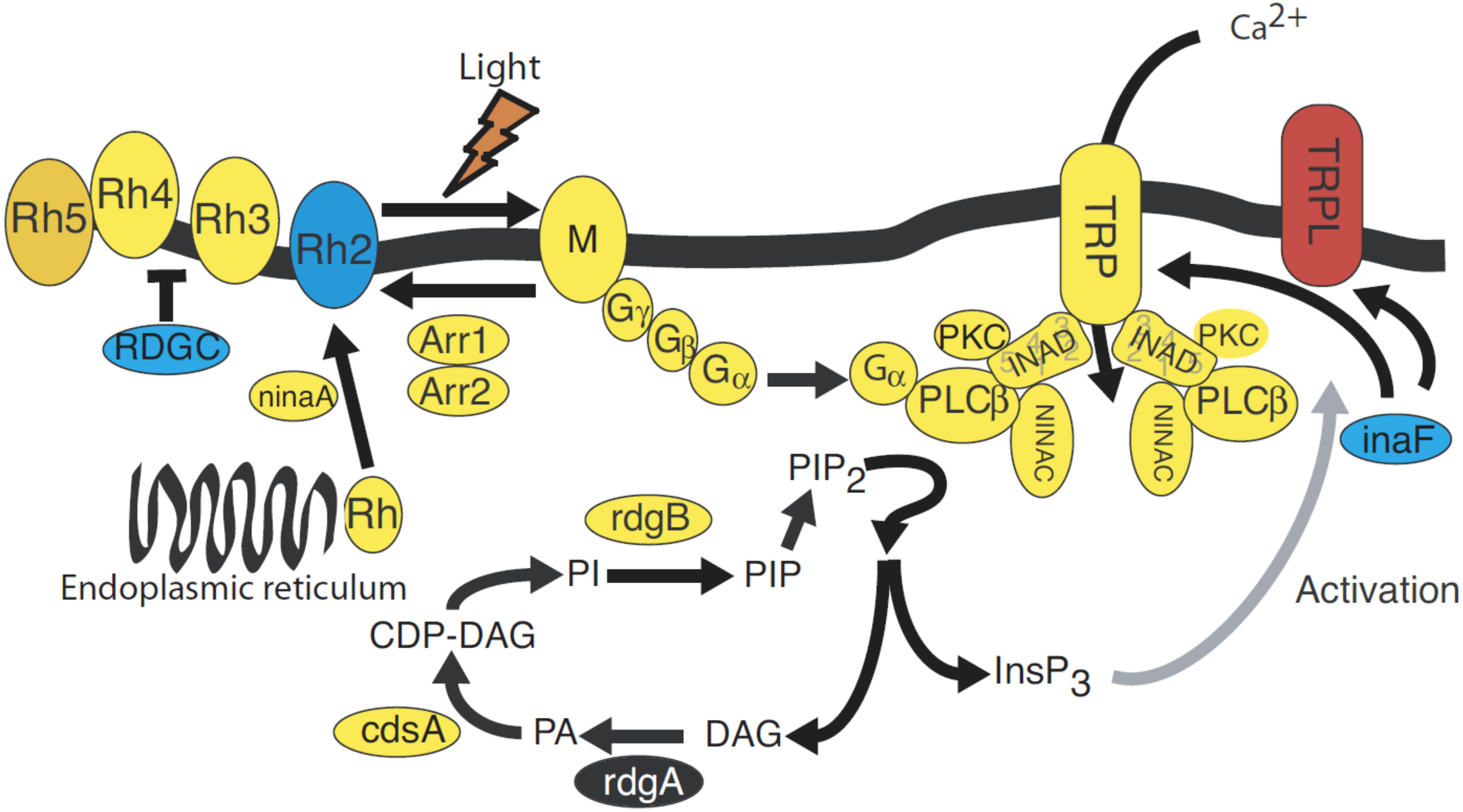


Subterranean


Bidessini
35 Limbodessus micromelitaensis
35 Limbodessus melitaensis
42 Limbodessus lornaensis
42 Limbodessus macrolornaensis
20 Limbodessus eberhardi
20 Limbodessus pulpa
25 Limbodessus silus
25 Limbodessus sweetwatersensis
38 Limbodessus nambiensis
6 Limbodessus surreptitius
44 Limbodessus microbubba
11 Limbodessus hillviewensis
33 Limbodessus gumwellensis
32 Limbodessus phoebeae
31 Limbodessus cooperi
6 Limbodessus challaensis
6 Limbodessus nyungduo
34 Limbodessus harleyi
29 Limbodessus masonensis
29 Limbodessus raesidensis
25 Limbodessus cunyensis
18 Limbodessus narryerensis
43 Limbodessus murrumensis
24 Limbodessus bialveus
20 Limbodessus kurutjutu
24 Limbodessus macrotarsus
19 Limbodessus windarraensis
31 Limbodessus leysi
34 Limbodessus usitatus
19 Limbodessus lapostae
5 Limbodessus trispinosuss
26 Limbodessus yuinmeryensis
29 Limbodessus sp.R413
45 Limbodessus morgani
36 Limbodessus millbilliensis
45 Limbodessus hahni
7 Limbodessus bigbellensis
19 Limbodessus palmulaoides
21 Limbodessus raeae
21 Limbodessus macrohinkleri
21 Limbodessus hinkleri
Limbodessus inornatus
Limbodessus gemellus
Allodessus bistrigatus
44 Limbodessus yarrabubbaensis
9 Limbodessus karalundiensis
13 Limbodessus wogarthaensis
28 Limbodessus fridaywellensis
27 Limbodessus pinnaclesensis
30 Limbodessus barwidgeensis
40 Limbodessus yandalensis
37 Limbodessus mirandaae
Limbodessus shuckardii
Limbodessus compactus
Limbodessus amabilis
Limbodessus praelargus
1 Limbodessus micrommatoion
1 Limbodessus microocula
Limbodessus rivulus
Limbodessus occidentalis
32 Limbodessus exilis
7 Limbodessus cf.cueensis
8 Limbodessus cueensis
8 Limbodessus magnificus
17 Limbodessus padburyensis
Gibbidessus chipi
Uvarus pictipes
Neobidessodes denticulatus
Neobidessodes mjobergi
Neobidessodes bilita
2 Neobidessodes limestonensis
2 Neobidessodes gutteridgei
Mya
15
10
5
0

Leijs *et al.* 2012 PLoS One

Phototransduction in blind beetles



Landry et al 2007 PNAS

Short Read Assembly

- Trinity Assembler & post assembly tools

- External required software (trinity):
  - Jellyfish (k-mer counting), samtools, Bowtie2, kallisto, salmon

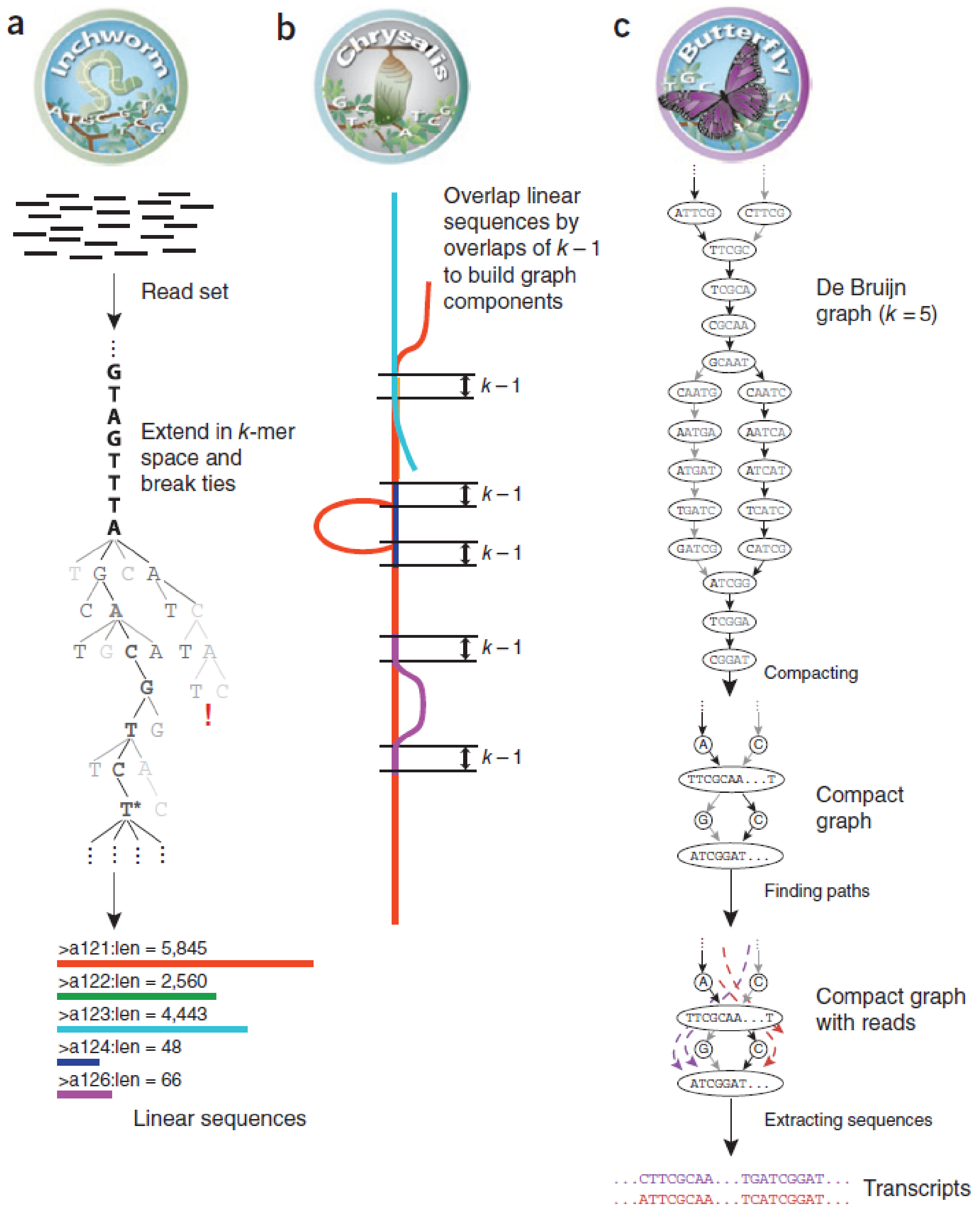
- External required software (post-assembly):
  - R with Bioconductor
  - BLAST, Picard, GATK4, Hisat2, STAR
  - RSEM, express
  - Transdecoder

Trinity

sequence **ATGGAAGTCGCGGAATC**

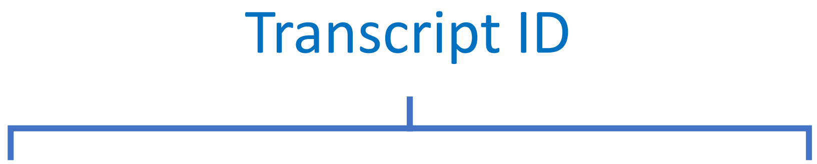
7mers

```
ATGGAAG
 TGGAAGT
  GGAAGTC
   GAAGTCG
    AAGTCGC
     AGTCGCG
      GTCGCGG
       TCGCGGA
        CGCGGAA
         GCGGAAT
          CGGAATC
```



Trinity

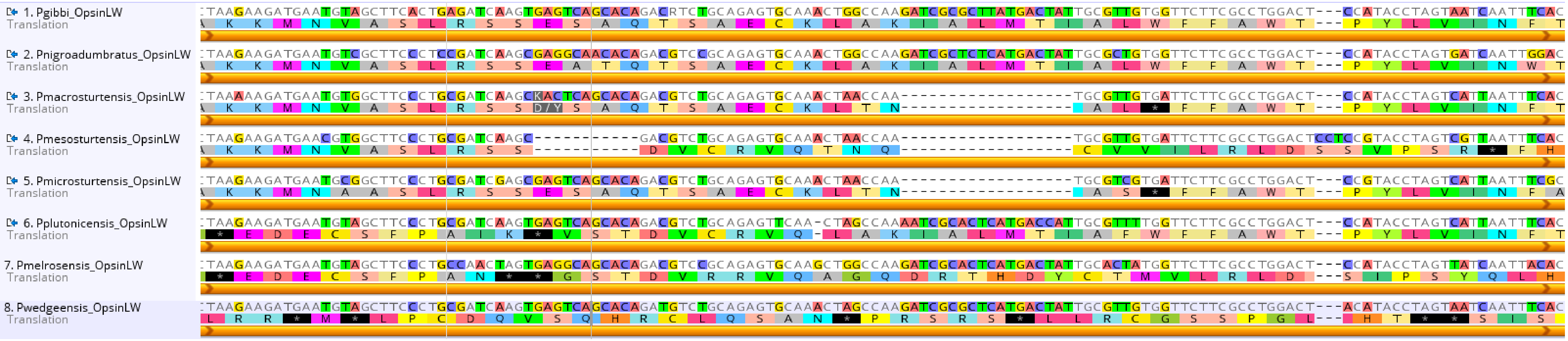
- Output is a bunch of files and folders
- `trinity.fasta` contains the assembled transcripts



Transcript ID

```
>TRINITY_DN1000|c115_g5_i1 len=247 path=[31015:0-148 23018:149-246]
AATCTTTTTTGGTATTGGCAGTACTGTGCTCTGGGTAGTGATTAGGGCAAAAGAAGACAC
ACAATAAAGAACCAGGTGTTAGACGTCAGCAAGTCAAGGCCTTGGTTCTCAGCAGACAGA
AGACAGCCCTTCTCAATCCTCATCCCTTCCCTGAACAGACATGTCTTCTGCAAGCTTCTC
CAAGTCAGTTGTTCACAGGAACATCATCAGAATAAATTTGAAATTATGATTAGTATCTGA
TAAAGC
```

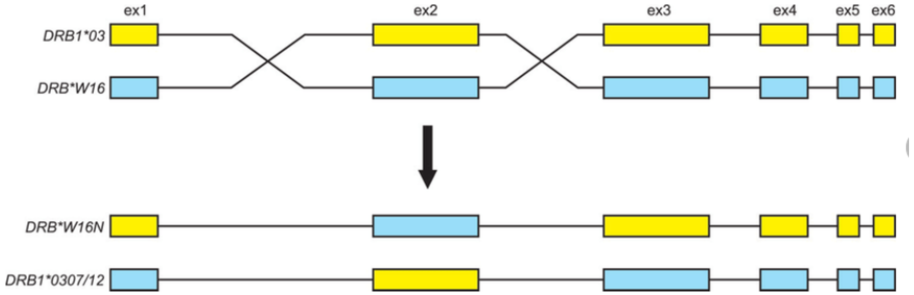
UV Opsin



```
1. Pgibbi_OpsinLW
   CTAAGAAGATGAATGTAGCTTCACTGAGATCAAGTGAGTCAGCACAGACRTCTGCAGAGTGCAAACTGGCCAAGATCGCGCTTATGACTATTGCGTTGTGGTTCTTCGCCTGGACT---CCATACCTAGTAATCAATTTCAC
Translation
   A K K M N V A S L R S S E S A Q T S A E C K L A K I A L M T I A L W F F A W T --- P Y L V I N F T
2. Pnigroadumbratus_OpsinLW
   CTAAGAAGATGAATGTCGCTTCCCTCCGATCAAGCGAGGCAACACAGACGTCCGCAGAGTGCAAACTGGCCAAGATCGCTCTCATGACTATTGCGCTGTGGTTCTTCGCCTGGACT---CCATACCTAGTGATCAATTGGAC
Translation
   A K K M N V A S L R S S E A T Q T S A E C K L A K I A L M T I A L W F F A W T --- P Y L V I N W T
3. Pmacrosturtensis_OpsinLW
   CTAAAAAGATGAATGTGGCTTCCCTGCGATCAAGCKACTCAGCACAGACGTCTGCAGAGTGCAAACTAACCAA------------------TGCGTTGTGATTCTTCGCCTGGACT---CCATACCTAGTCATTAATTTCAC
Translation
   A K K M N V A S L R S S D/Y S A Q T S A E C K L T N------------------ A L * F F A W T --- P Y L V I N F T
4. Pmesosturtensis_OpsinLW
   CTAAGAAGATGAACGTGGCTTCCCTGCGATCAAGC-----------GACGTCTGCAGAGTGCAAACTAACCAA------------------TGCGTTGTGATTCTTCGCCTGGACTCCTCCGTACCTAGTCGTTAATTTCAC
Translation
   A K K M N V A S L R S S ----------- D V C R V Q T N Q ------------------ C V V I L R L D S S V P S R * F H
5. Pmicrosturtensis_OpsinLW
   CTAAGAAGATGAATGCGGCTTCCCTGCGATCGAGCGAGTCAGCACAGACGTCTGCAGAGTGCAAACTAACCAA------------------TGCGTCGTGATTCTTCGCCTGGACT---CCGTACCTAGTCATTAATTTCGC
Translation
   A K K M N A A S L R S S E S A Q T S A E C K L T N------------------ A S * F F A W T --- P Y L V I N F A
6. Pplutonicensis_OpsinLW
   CTAAGAAGATGAATGTAGCTTCCCTGCGATCAAGTGAGTCAGCACAGACGTCTGCAGAGTTCAA-CTAGCCAAAATCGCACTCATGACCATTGCGTTTTGGTTCTTCGCCTGGACT---CCATACCTAGTCATTAATTTCAC
Translation
   * E D E C S F P A I K * V S T D V C R V Q - L A K I A L M T I A F W F F A W T --- P Y L V I N F T
7. Pmelrosensis_OpsinLW
   CTAAGAAGATGAATGTAGCTTCCCTGCCAACTAGTGAGGCAGCACAGACGTCCGCAGAGTGCAAGCTGGCCAAGATCGCACTCATGACTATTGCACTATGGTTCTTCGCCTGGACT---CCATACCTAGTTATCAATTACAC
Translation
   * E D E C S F P A N * * G S T D V R R V Q A G Q D R T H D Y C T M V L R L D --- S I P S Y Q L H
8. Pwedgeensis_OpsinLW
   CTAAGAAGATGAATGTAGCTTCCCTGCGATCAAGTGAGTCAGCACAGATGTCTGCAGAGTGCAAACTAGCCAAGATCGCGCTCATGACTATTGCGTTGTGGTTCTTCGCCTGGACT---ACATACCTAGTAATCAATTTCAC
Translation
   L R R * M * L P C D Q V S Q H R C L Q S A N * P R S R S * L L R C G S S P G L --- H T * * S I S
```

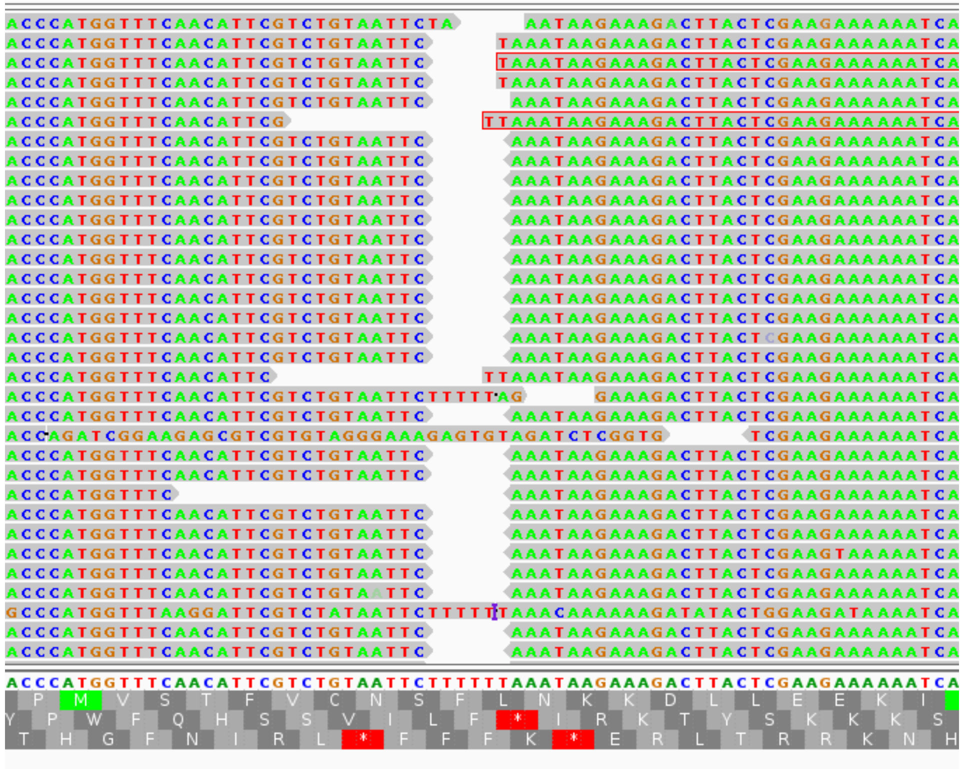
Incorrectly assembled transcripts

- Multicopy gene families



Doxiadis et al 2006 PNAS

- Coverage
  - Incomplete assembly
  - Extreme expression
- Alternative splicing


ACCCATGGTTTCAACATTCGTCTGTAATTCTTTTTTAAATAAGAAAGACTTACTCGAAGAAAAAATCA

Long read methods

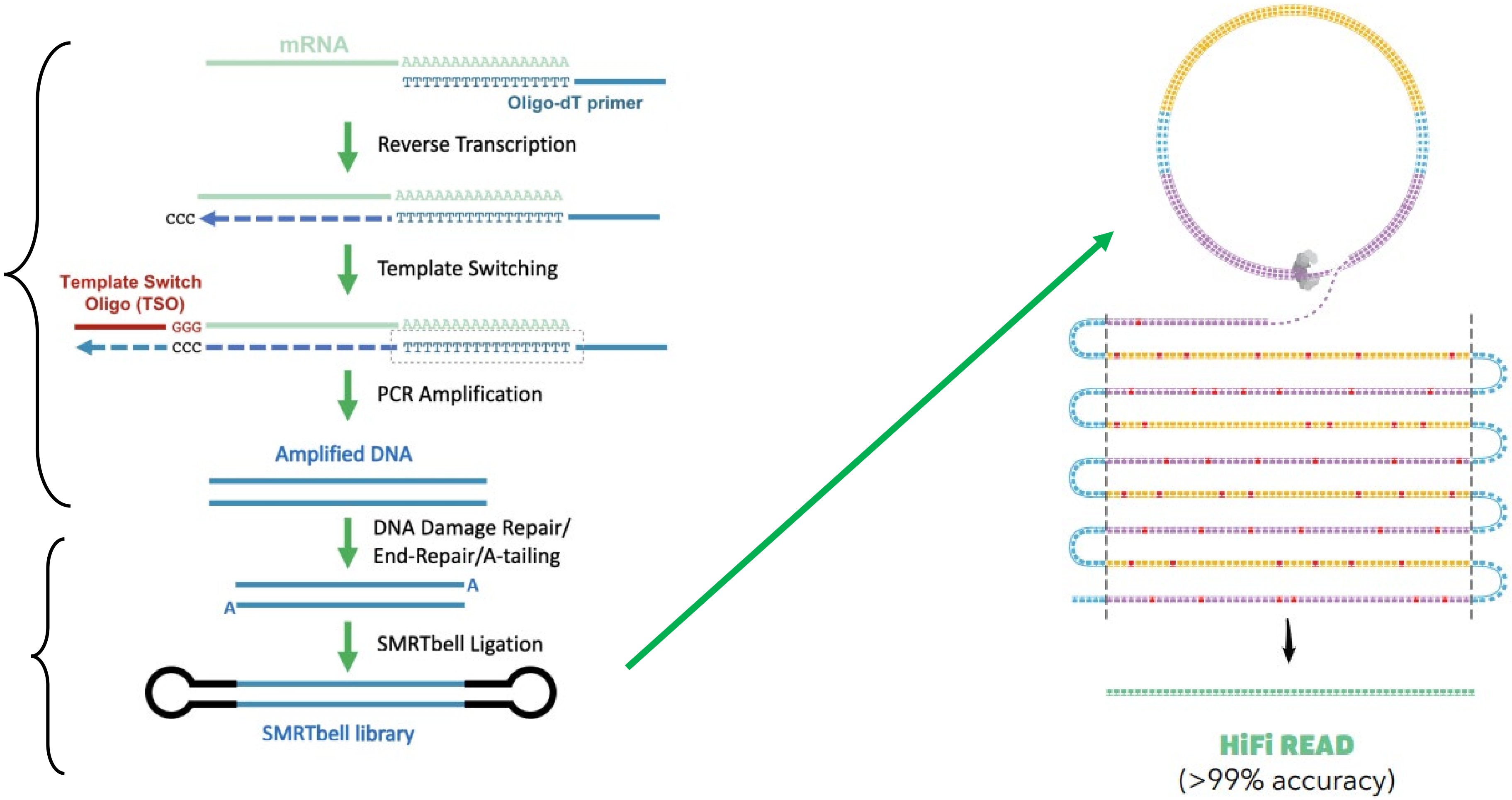
- Alternative splicing
  - 90% genes[1]; 30% ncRNA[2]
- Long read sequencing
  - No assembly required!
- Nanopore RNA sequencing
  - ✓Native RNA sequencing
  - X Error prone reads
- Iso-Seq (Pacific Biosciences)
  - X Native RNA sequencing
  - ✓ Highly accurate reads

[1]Wang et al 2008 Nature; [2]Cabili et al 2011 Genes & Development

Iso-Seq

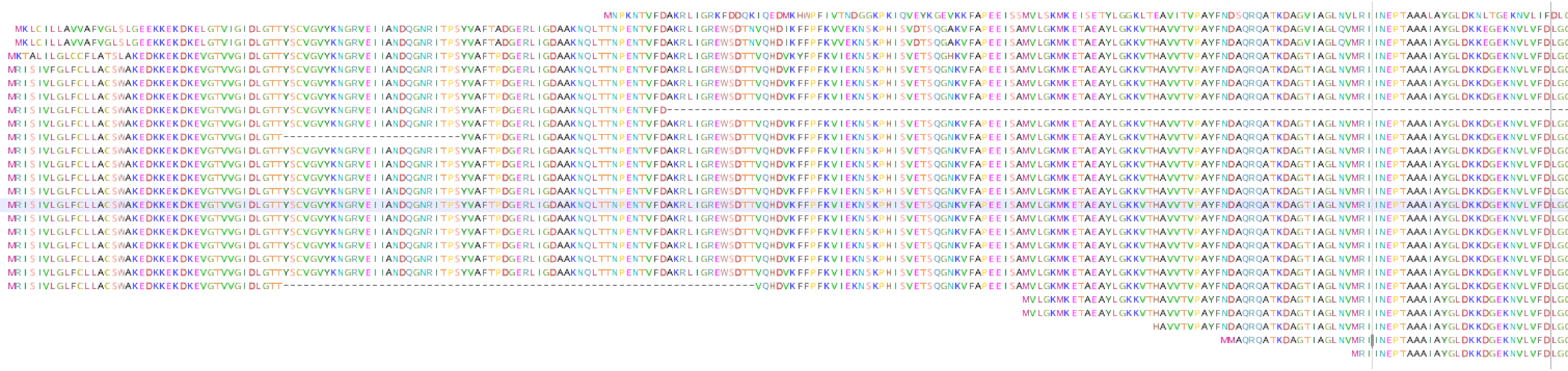
cDNA synthesis

Library construction



SMRT sequencing

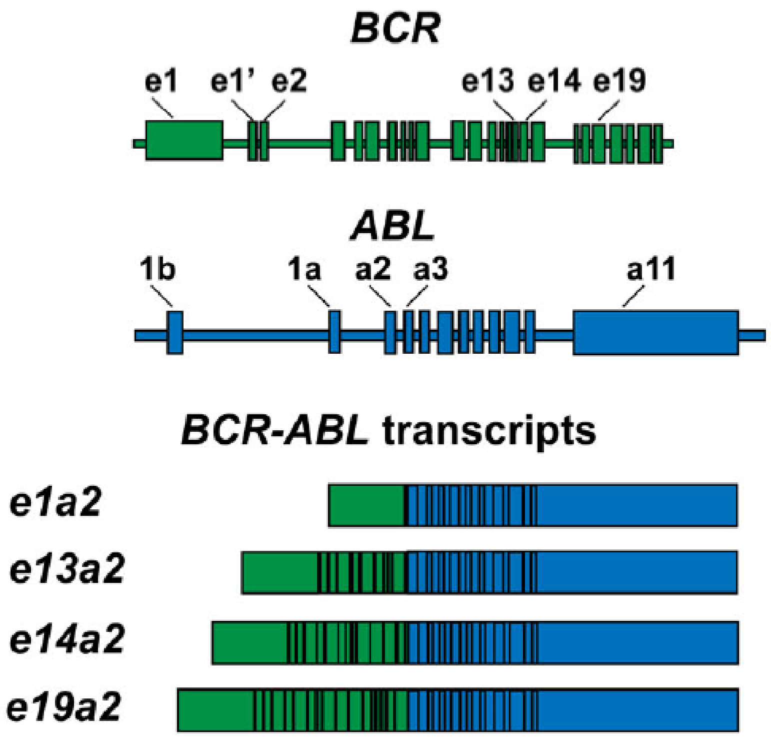
Heat Shock Proteins (HSP70)



```
                                                                                       MNPKNTVFDAKRLIGRKFDDQKIQEDMKHWPFIVTNDGGKPKIQVEYKGEVKKFAPEEISSMVLSKMKEISETYLGGKLTEAVITVPAYFNDSQRQATKDAGVIAGLNVLRIINEPTAAALAYGLDKNLTGEKNVLIFDLG
 MKLCILLAVVAFVGLSLGEEKKEKDKELGTVIGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTADGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTNVQHDIKFFPFKVVEKNSKPHISVDTSQGAKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGVIAGLQVMRIINEPTAAAIAYGLDKKEGEKNVLVFDLG
 MKLCILLAVVAFVGLSLGEEKKEKDKELGTVIGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTADGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTNVQHDIKFFPFKVVEKNSKPHISVDTSQGAKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGVIAGLQVMRIINEPTAAAIAYGLDKKEGEKNVLVFDLG
MKTALILGLCCFLATSLAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKYFPFKVIEKNSKPHISVETSQGHKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVFGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFD-----------------------------------------------------------------------------------------------------------------------------------
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTT--------------------------YVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTTYSCVGVYKNGRVEIIANDQGNRITPSYVAFTPDGERLIGDAAKNQLTTNPENTVFDAKRLIGREWSDTTVQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
MRISIVLGLFCLLACSWAKEDKKEKDKEVGTVVGIDLGTT---------------------------------------------------------------------VQHDVKFFPFKVIEKNSKPHISVETSQGNKVFAPEEISAMVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
                                                                                                                                                   MVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
                                                                                                                                                   MVLGKMKETAEAYLGKKVTHAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
                                                                                                                                                                       HAVVTVPAYFNDAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
                                                                                                                                                                                 MMAQRQATKDAGTIAGLNVMRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
                                                                                                                                                                                                    MRIINEPTAAAIAYGLDKKDGEKNVLVFDLG
```

Genome Guided Assembly

- Fusion transcripts



Neckles et al 2109 WIREs RNA

- RNA editing
  - A → I; C → U

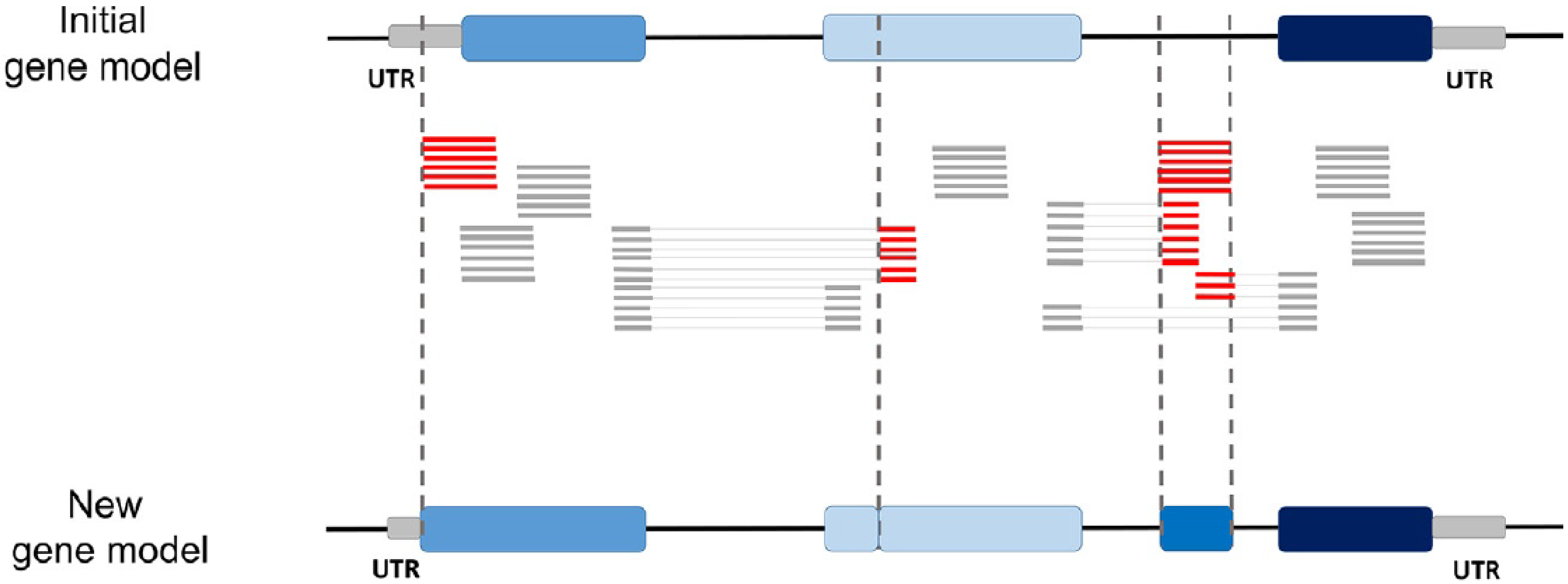
DNA: CACTGGACG

mRNA: GIGACCUGC

Protein: G T C

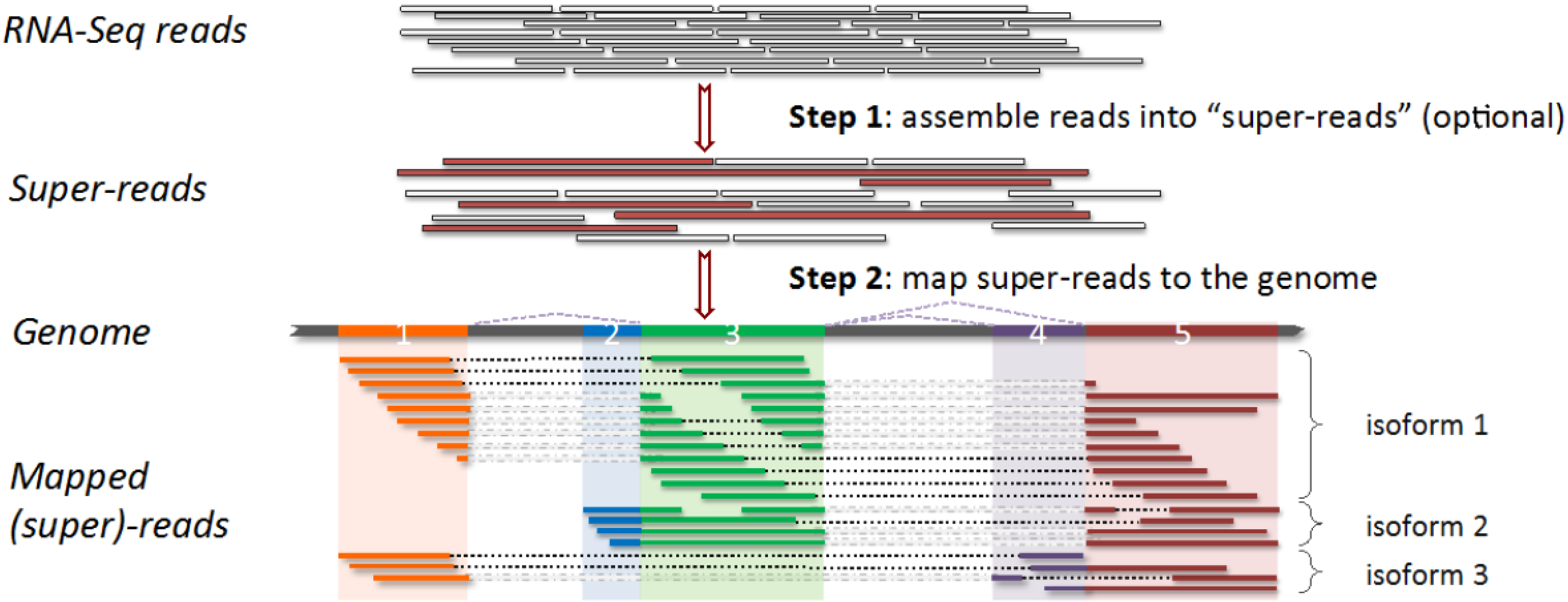
Genome Guided Assembly

- Gene models



Mcafee 2016 Methods in enzymology

StringTie



RNA-Seq reads

Step 1: assemble reads into "super-reads" (optional)

Super-reads

Step 2: map super-reads to the genome

Genome

Mapped (super)-reads

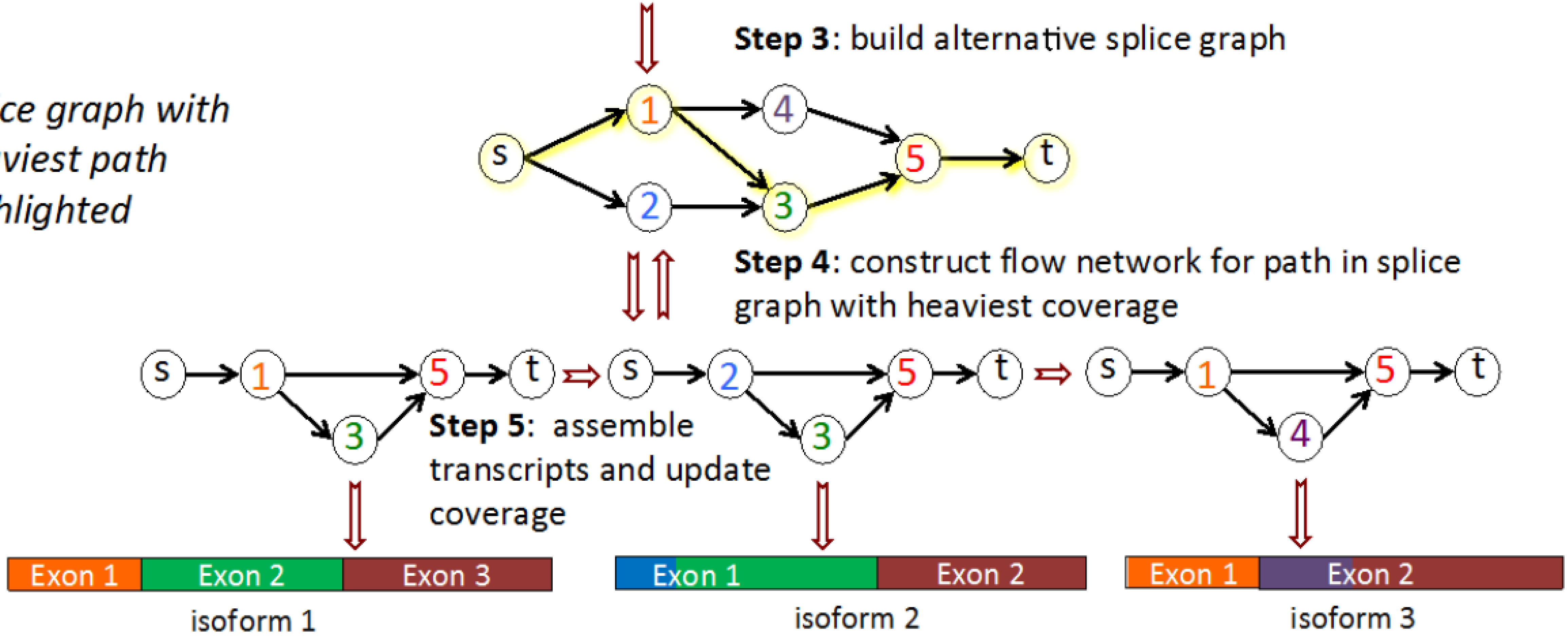
isoform 1

isoform 2

isoform 3

StringTie

Splice graph with heaviest path highlighted



Step 3: build alternative splice graph

Step 4: construct flow network for path in splice graph with heaviest coverage

Step 5: assemble transcripts and update coverage

Pertea et al 2015 Nature Biotechnology

StringTie

- General Transfer Format (GTF)

```
seqname source      feature     start   end     score   strand  frame attributes
chrX    StringTie   transcript  281394  303355  1000    +       .     gene_id "ERR188044.1"; transcript_id "ERR188044.1.1"; reference_id "NM
chrX    StringTie   exon        281394  281684  1000    +       .     gene_id "ERR188044.1"; transcript_id "ERR188044.1.1"; exon_number "1";
...
```

https://asia.ensembl.org/info/website/upload/gff.html

- Can use current annotation (GTF)
- Also calculates coverage for expression analysis
  - Ballgown, DESeq2, edgeR

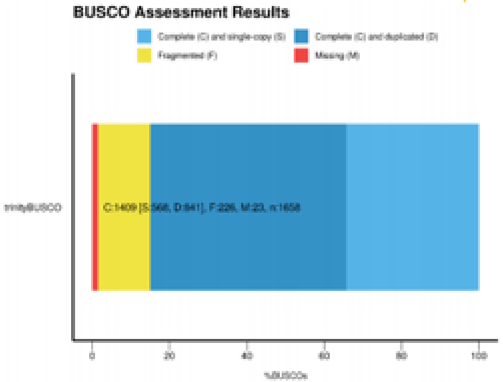
Transcriptome Evaluation

- Benchmarking of Universal Single Copy Orthologs (BUSCO)

C: 85% [S:34%,D:51%],
F: 14%,
M: 1%,
n: 1658

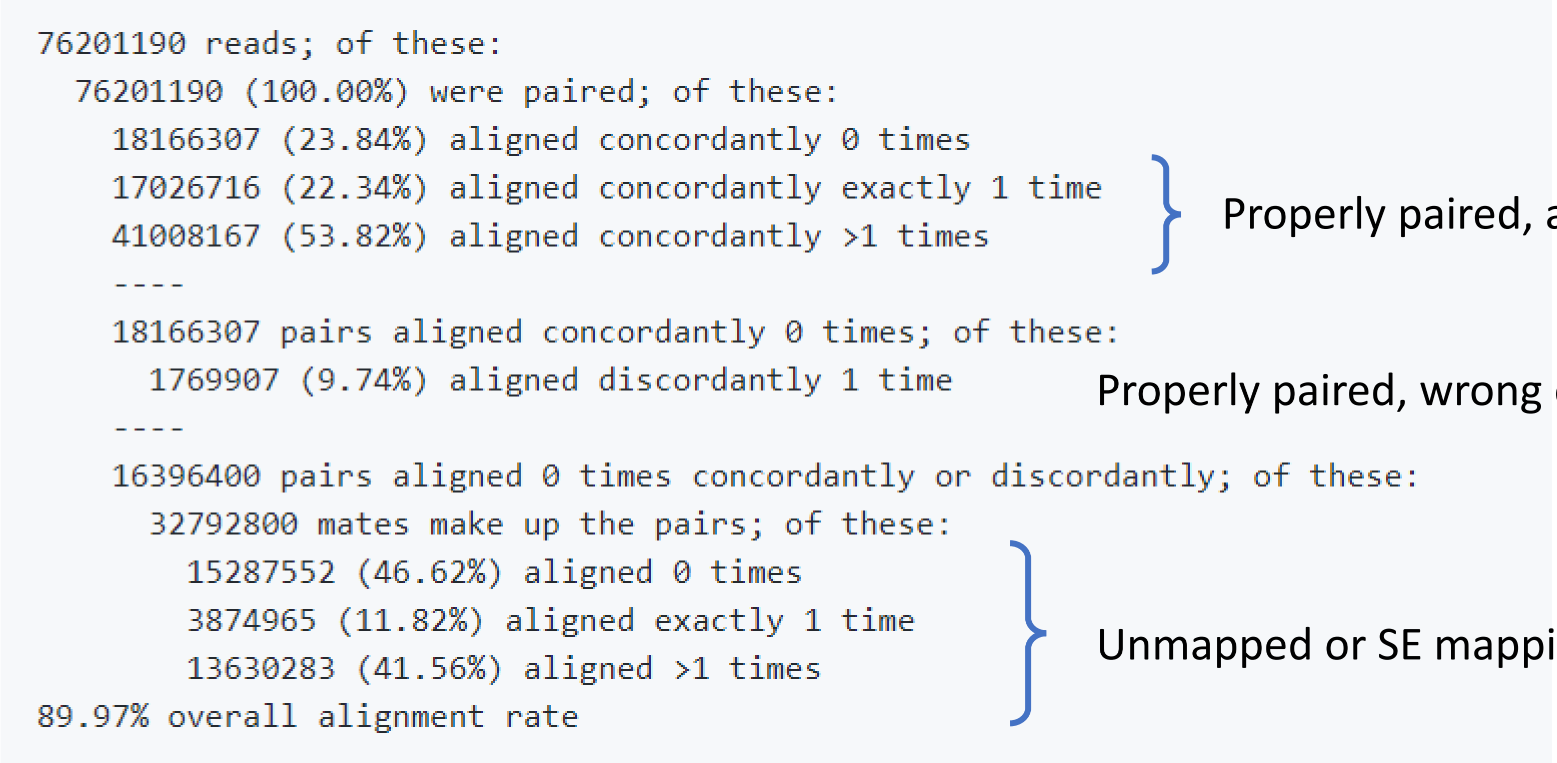
C: Complete
- S: single copy;
- D: duplicated;

F: Fragmented;
M: Missing;
n: Total groups;



Transcriptome Evaluation

- BLAST to proteome
- Quantify read support



```
76201190 reads; of these:
  76201190 (100.00%) were paired; of these:
    18166307 (23.84%) aligned concordantly 0 times
    17026716 (22.34%) aligned concordantly exactly 1 time
    41008167 (53.82%) aligned concordantly >1 times
    ----
    18166307 pairs aligned concordantly 0 times; of these:
      1769907 (9.74%) aligned discordantly 1 time
    ----
    16396400 pairs aligned 0 times concordantly or discordantly; of these:
      32792800 mates make up the pairs; of these:
        15287552 (46.62%) aligned 0 times
        3874965 (11.82%) aligned exactly 1 time
        13630283 (41.56%) aligned >1 times
89.97% overall alignment rate
```

Properly paired, aligning as expected

Properly paired, wrong orientation or distance

Unmapped or SE mapping reads

Summary

- Transcriptome assembly is hard!
  - Diverse population of RNA
  - Non-uniform coverage
  - Tissue specificity
  - Multigene families
  - Alternative splicing