Transcriptome and isoform reconstruction with short reads

Tangled up in reads

Enabler for Life Sciences

Karolinska Institutet

KTH VETENSKAP OCH KONST

ROYAL INSTITUTE OF TECHNOLOGY

Stockholm University

UPPSALA UNIVERSITET

SciLifeLab

Topics of this lecture

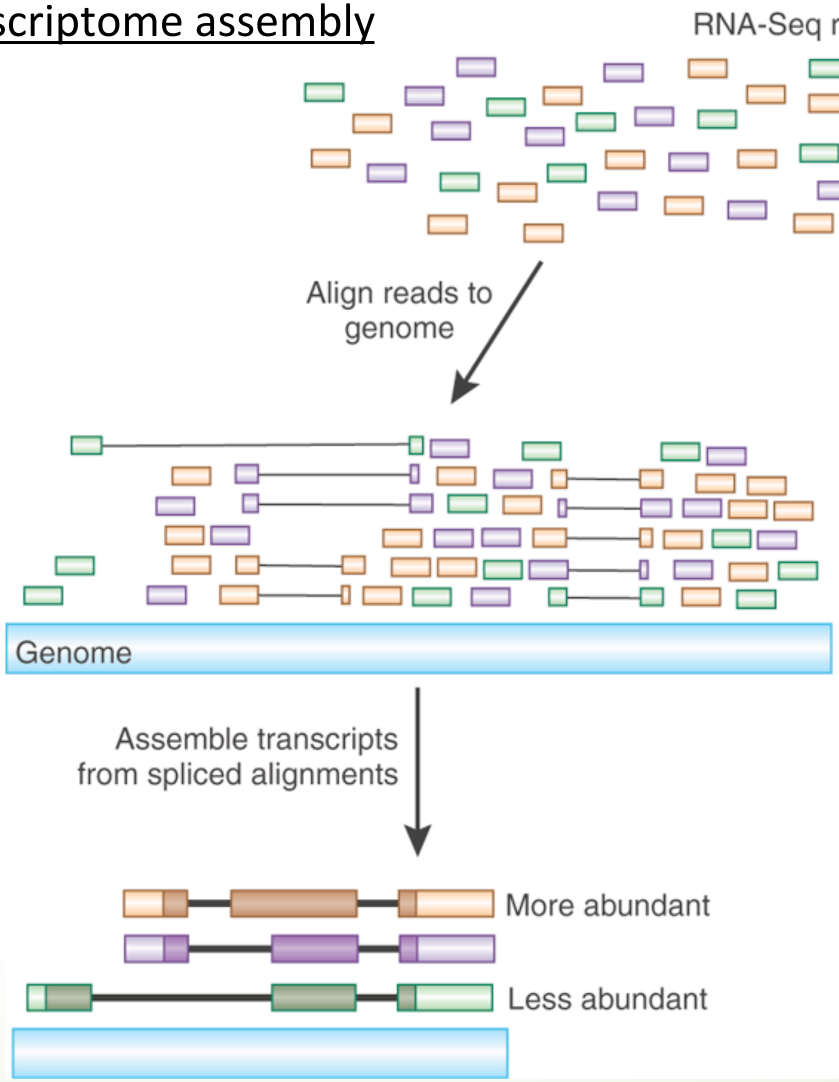
- Mapping-based reconstruction methods
  - Case study: The domestic dog
- De-novo reconstruction method
  - Trinity

Transcriptome assembly

RNA-Seq reads

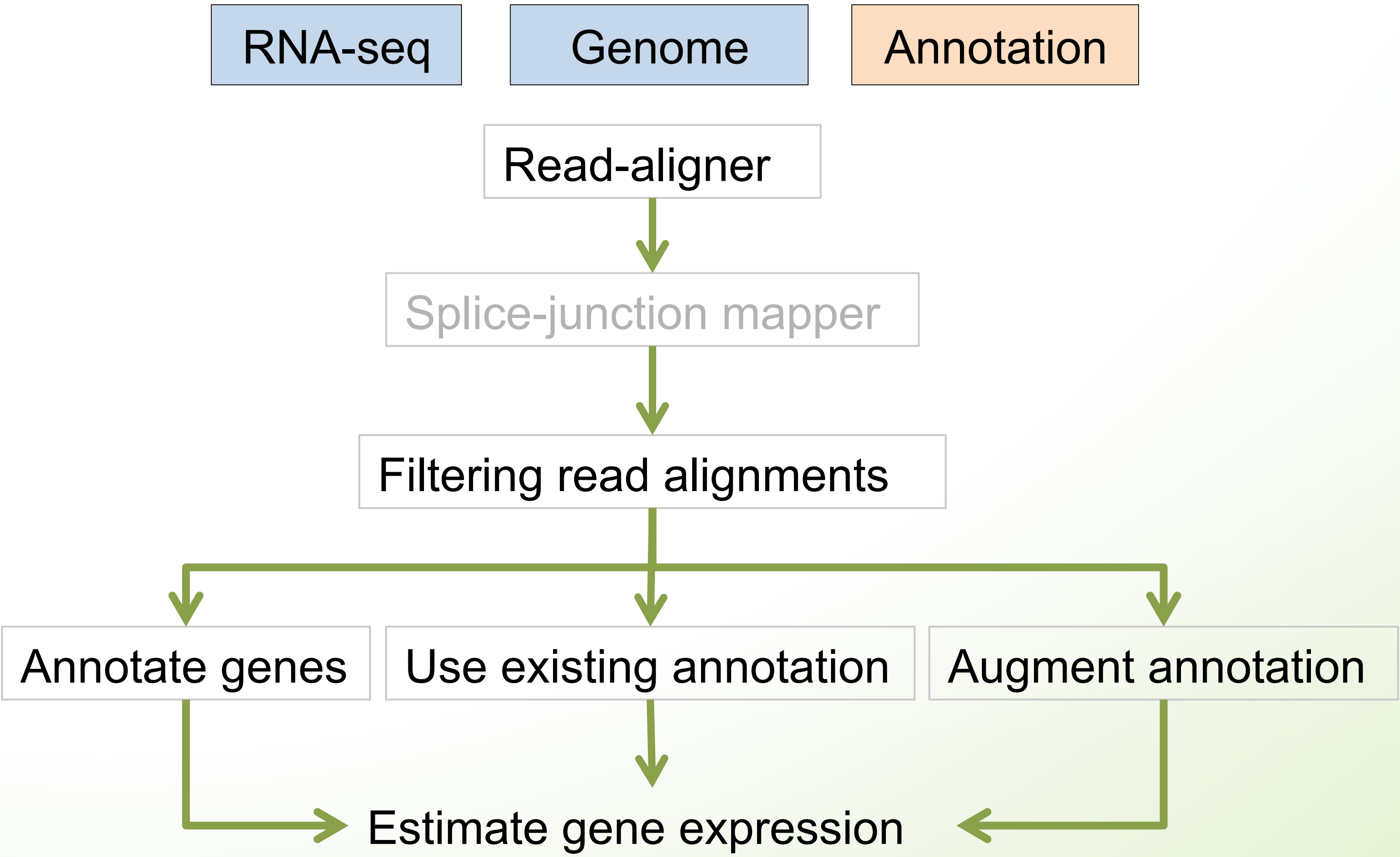
Haas and Zody, Nature Biotechnology 28, 421–423 (2010)

Transcriptome assembly



Haas and Zody, Nature Biotechnology 28, 421–423 (2010)

Mapping-based transcriptome reconstruction



RNA-seq

Genome

Annotation

Read-aligner

Splice-junction mapper

Filtering read alignments

Annotate genes

Use existing annotation

Augment annotation

Estimate gene expression

Case study: The transcriptome of the domestic dog

Case study: The transcriptome of the domestic dog

Has shared an environment with humans for > 10.000 years

> Exposed to many of the same environ. influences

Affected by many of the same diseases as man

> Cancer

> Heart disease

Extensive breeding and selection

> Many dog breeds are prone to certain diseases

> Long haplotypes ideal for association studies

Question: what genes are located in my region of interest?

Requires a high quality genome...and detailed annotation!

Case study: The transcriptome of the domestic dog

Recently, the Broad institute released an updated build, canFam3.1

- 85 Mb of additional sequence integrated
- 99.8% of euchromatic portion of genome covered, high quality
- Recovered 100s of GC-rich promoter regions

Now approaches level of quality/completion of mouse or human

> the annotation...not so much.

Case study: The transcriptome of the domestic dog

strong discrepancy between well-annotated human genome and dog. Why?

> largely homology-based

> almost no isoform information

> only few dog-specific gene annotations

Majority of loci likely incomplete, many dog-specific genes probably missing

Case study: The transcriptome of the domestic dog

10 tissues at great depth (> 20 million reads)

blood, brain, heart, kidney, liver, lung, muscle, ovary, skin, testes

Stranded paired-end libraries

Poly-A selected: default approach, recovers mostly protein-coding genes

DSN prep: Targets all RNAs, but normalizes library to avoid strong biases

An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. Hoeppner MP et al. PLoS One 2014 Mar 13;9(3):e91172

Mapping-based transcriptome reconstruction

Align reads with Tophat/Bowtie

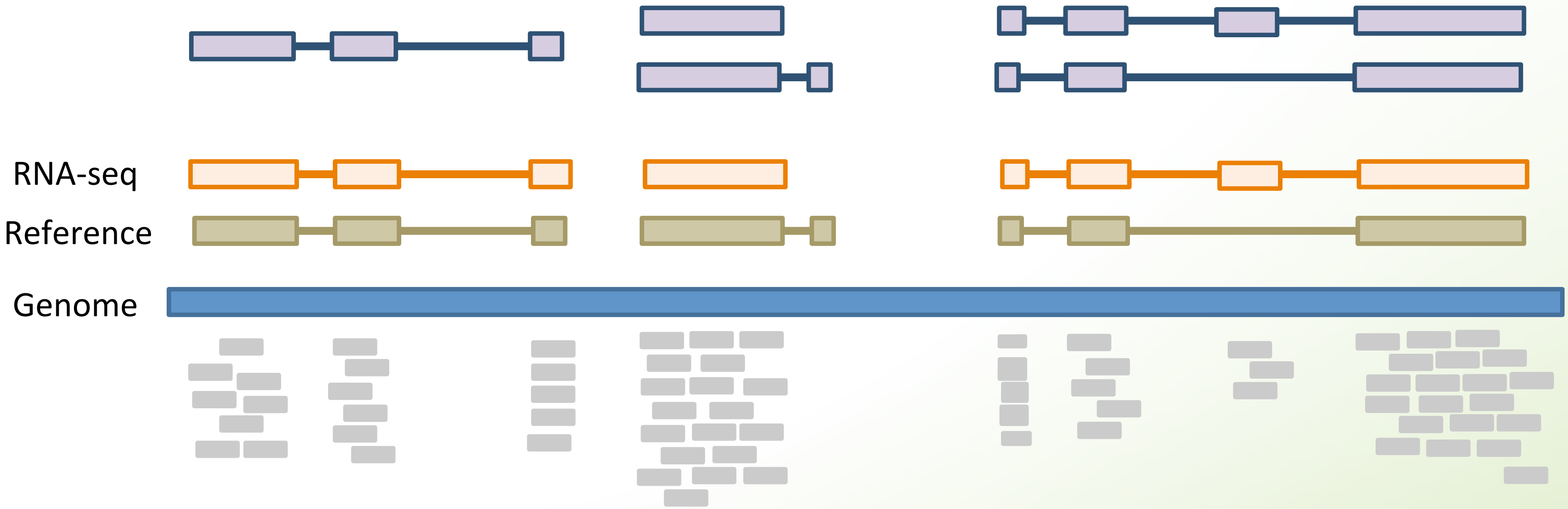
Reconstruct transcripts with Cufflinks

Reconcile de-novo annotation with reference

Annotate novel transcripts

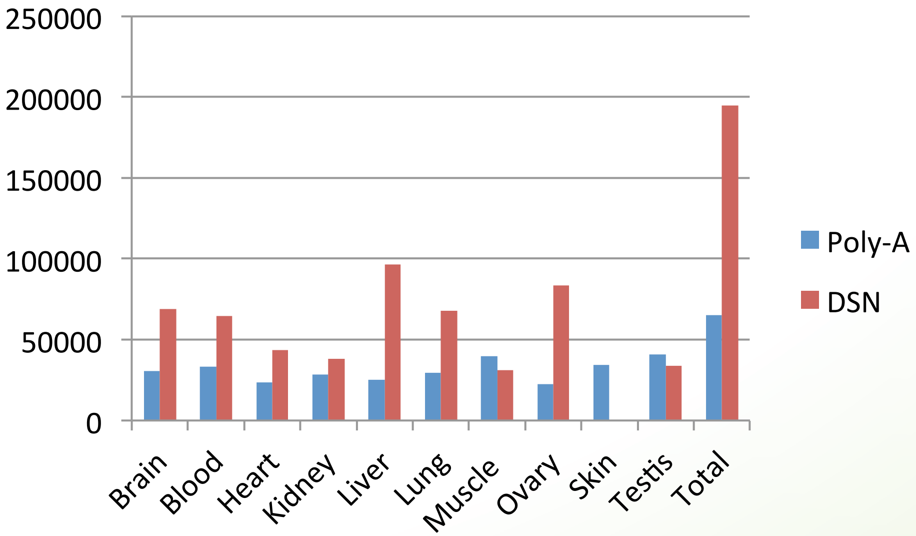
Quantify

Mapping-based transcriptome reconstruction



Case study: The transcriptome of the domestic dog

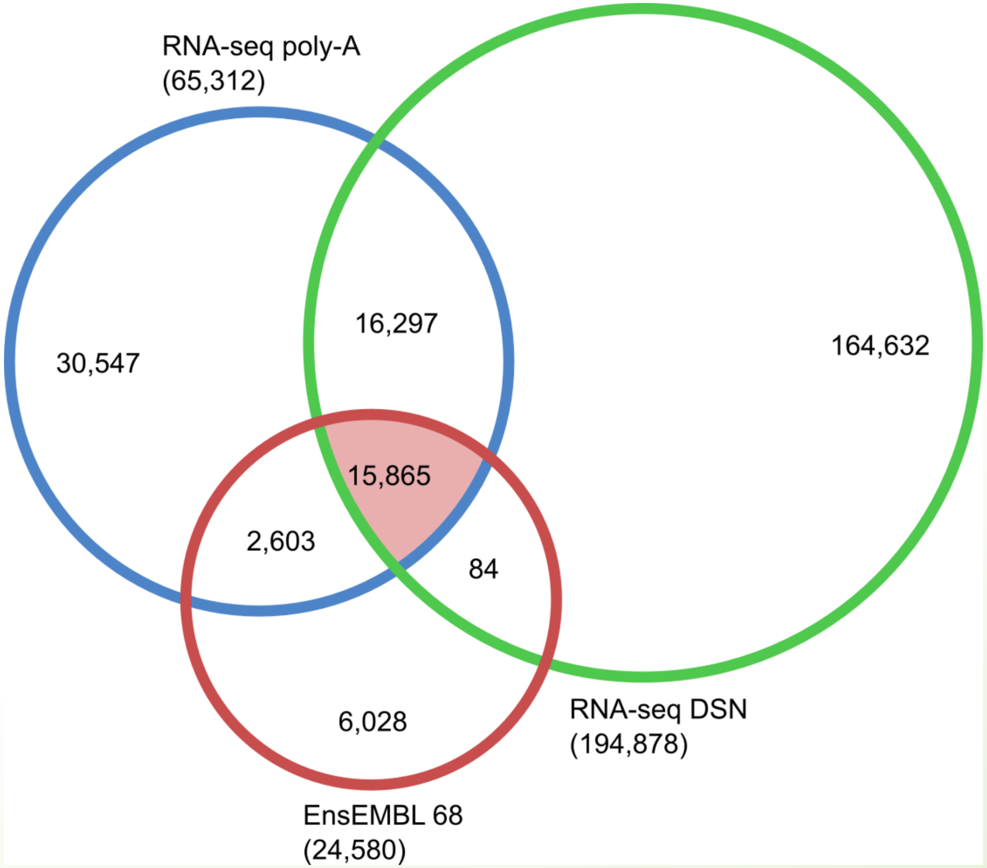
Transcript reconstruction using cufflinks for both libraries



DSN recovers more transcripts than polyA

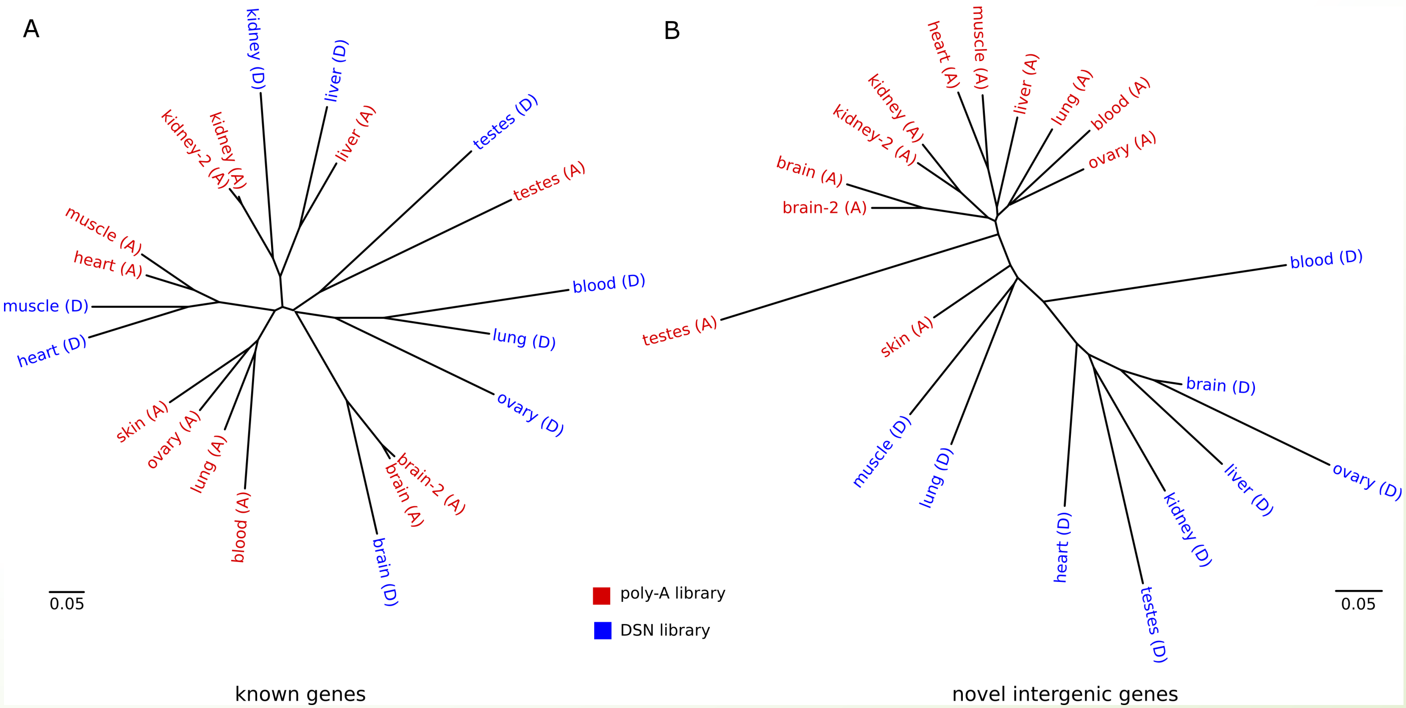
Transcriptional diversity is highest in testes

Case study: The transcriptome of the domestic dog



Case study: The transcriptome of the domestic dog

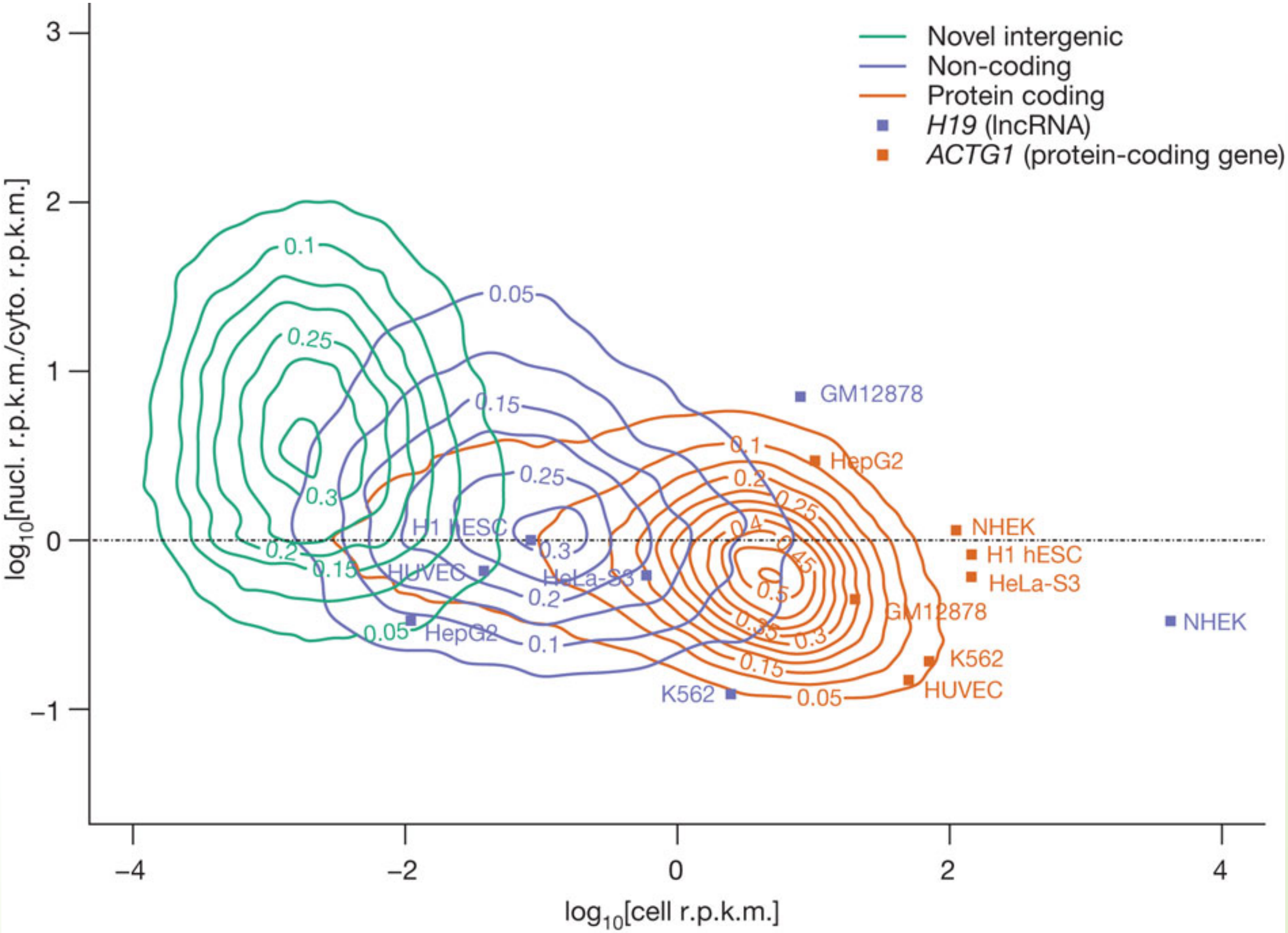
Transcript reconstruction using cufflinks for both libraries



🟥 poly-A library

🟦 DSN library

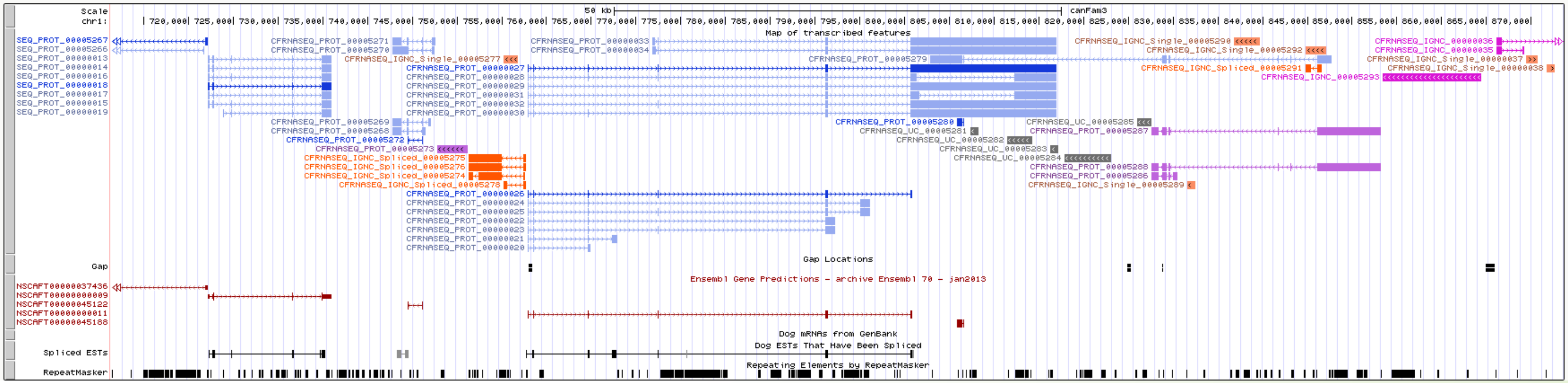
RNA flavors



Landscape of transcription in human cells, S Djebali *et al.* ***Nature 2012***

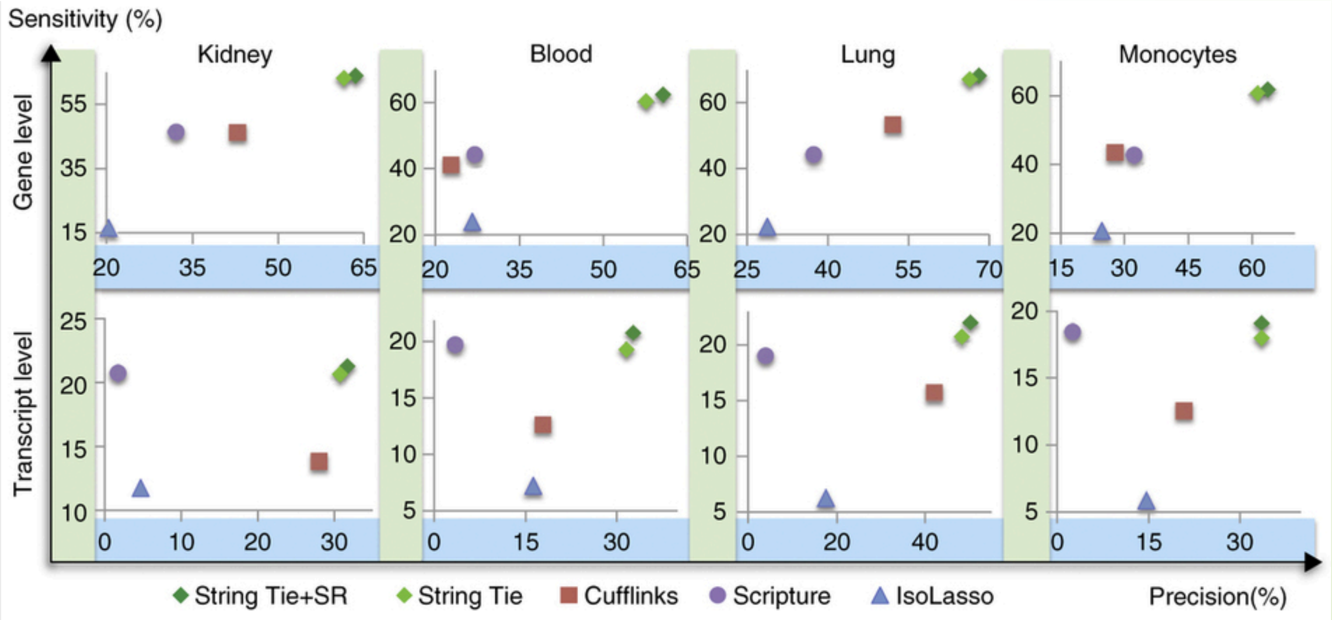
Case study: The transcriptome of the domestic dog

Augmented annotation and transcript classification

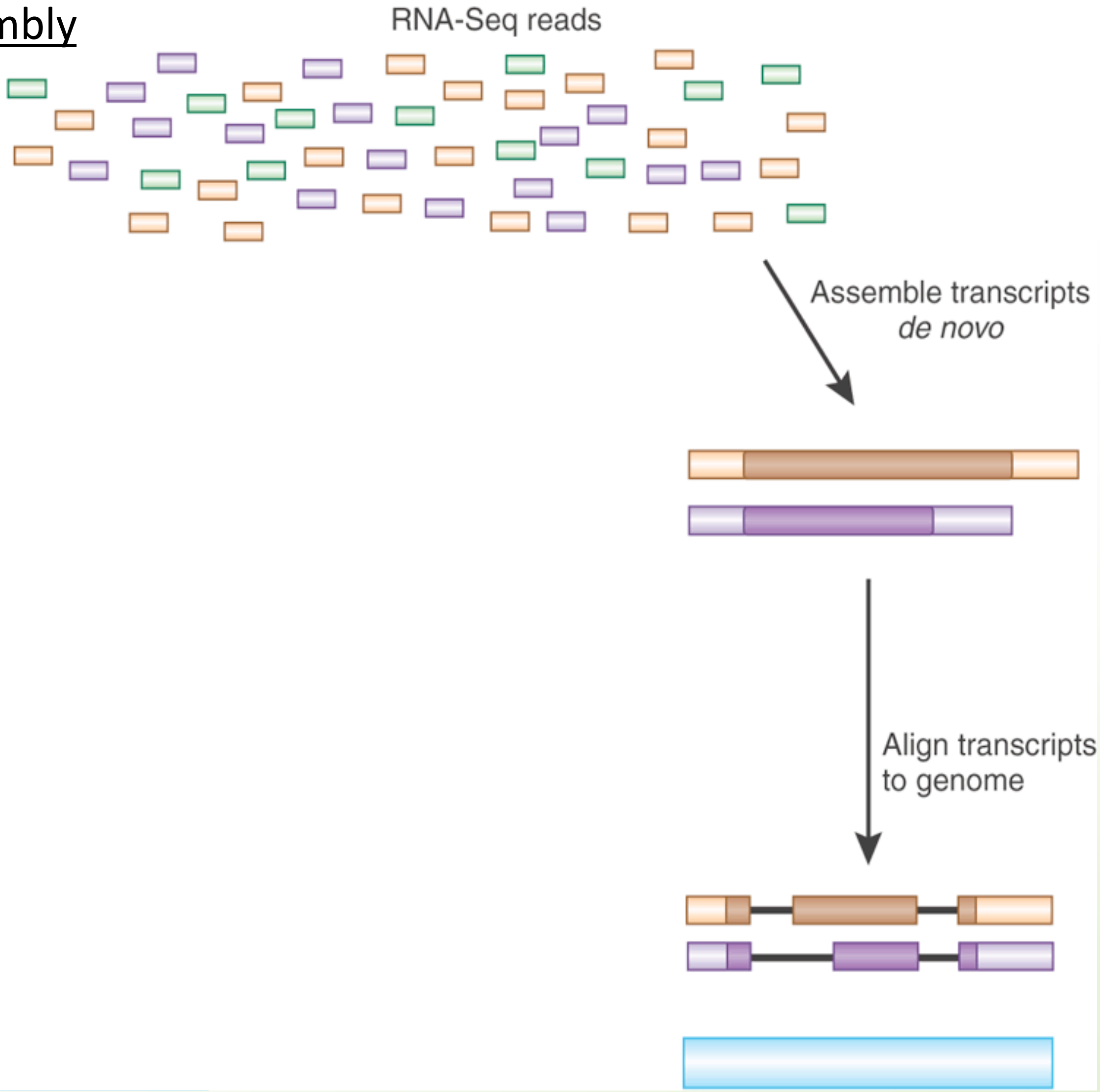


Several softwares

- Cufflinks
- Scripture
- Ballgown
- StringTie



Transcriptome assembly



Haas and Zody, Nature Biotechnology 28, 421–423 (2010)

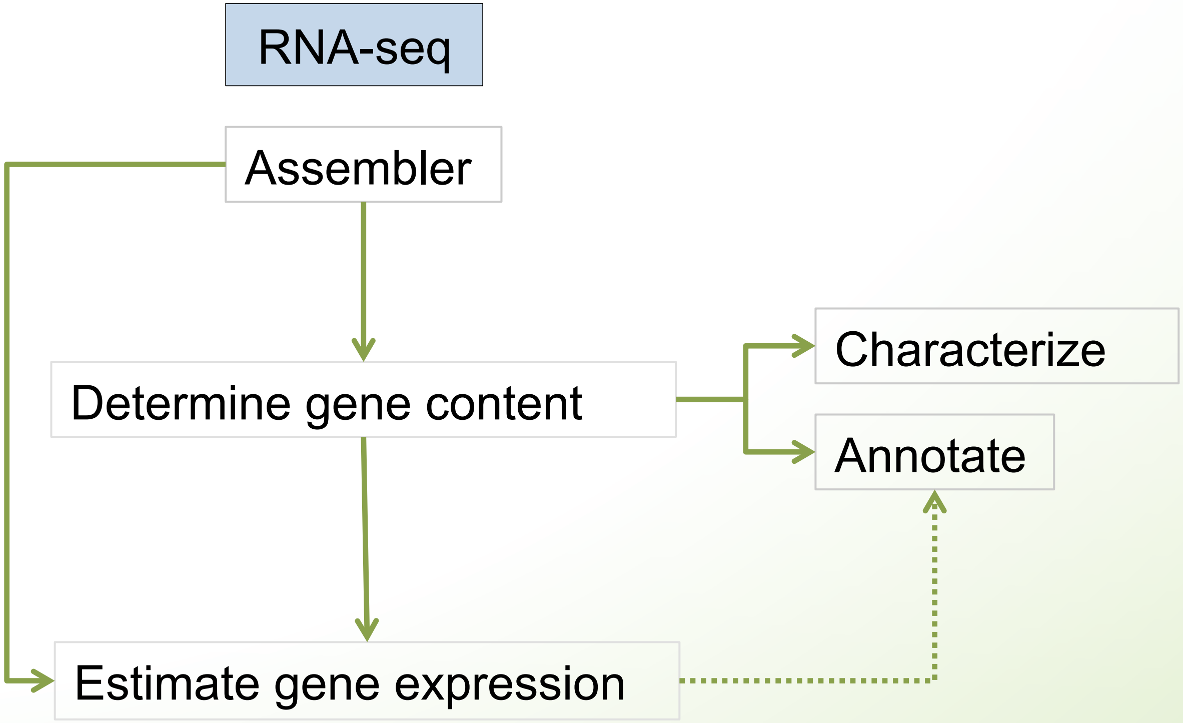
De-novo transcriptome assembly

For the majority of species, there are no comprehensive genome sequences…

Transcriptomics can inform a broad range of questions without reference

→ De-novo transcriptome assembly from extracted RNA

De-novo transcriptome reconstruction

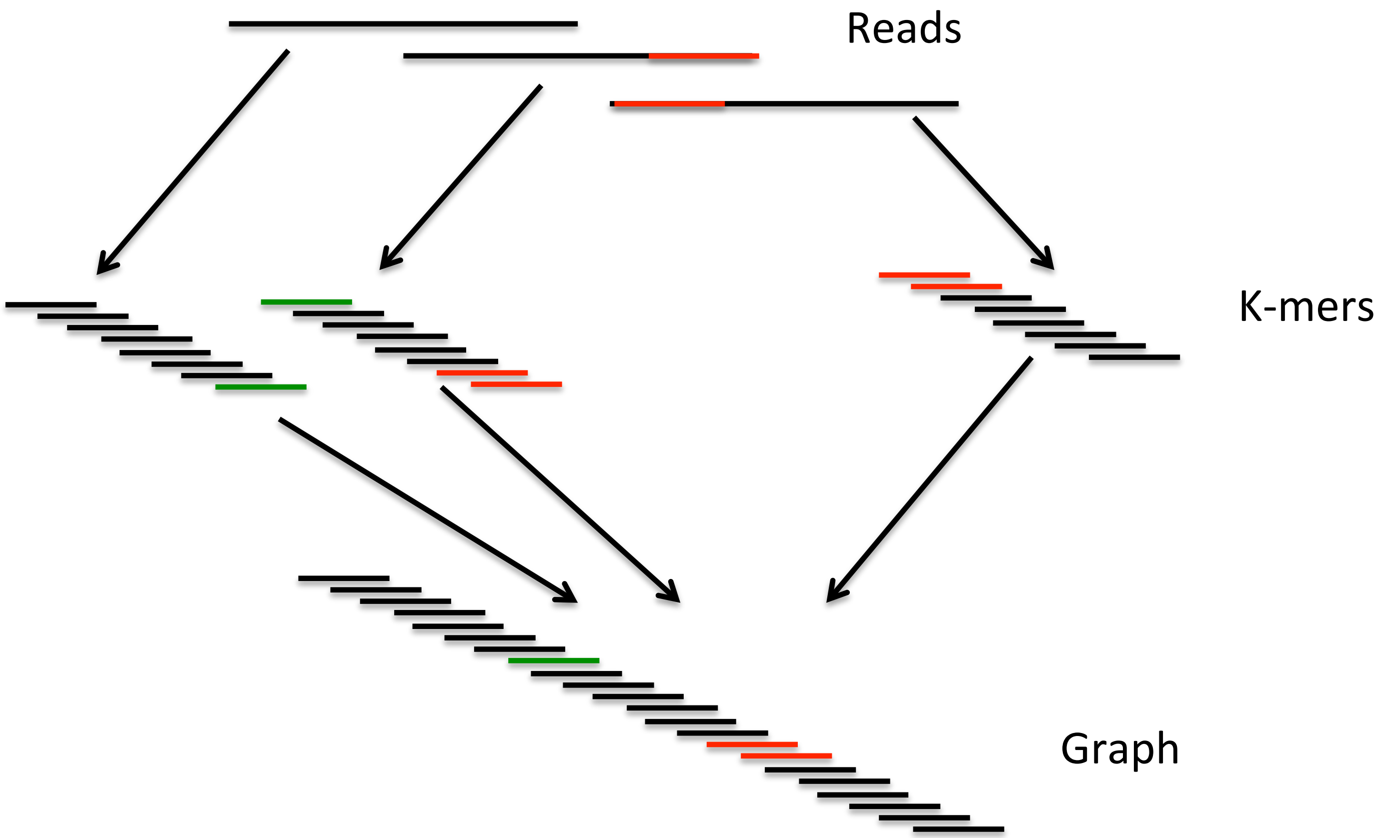


De-novo transcriptome assembly

Manfred Grabherr
Brian Haas
Moran Yassour
Kerstin Lindblad-Toh
Aviv Regev
Nir Friedman
David Eccles
Alexie Papanicolaou
Michael Ott
…

The k-mer

- K consecutive nucleotides



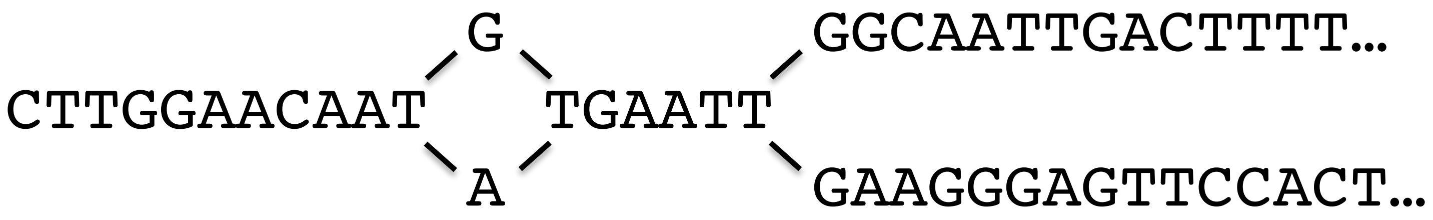
The de Bruijn Graph

- Graph of overlapping sequences
- Intended for cryptology
- Fixed length element: *k*

```
CTTGGAA
 TTGGAAC
  TGGAACA
   GGAACAA
    GAACAAT
```

The de Bruijn Graph

- Graph has “nodes” and “edges”



G

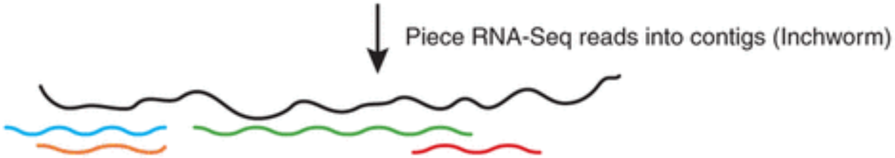
CTTGGAACAAT

A

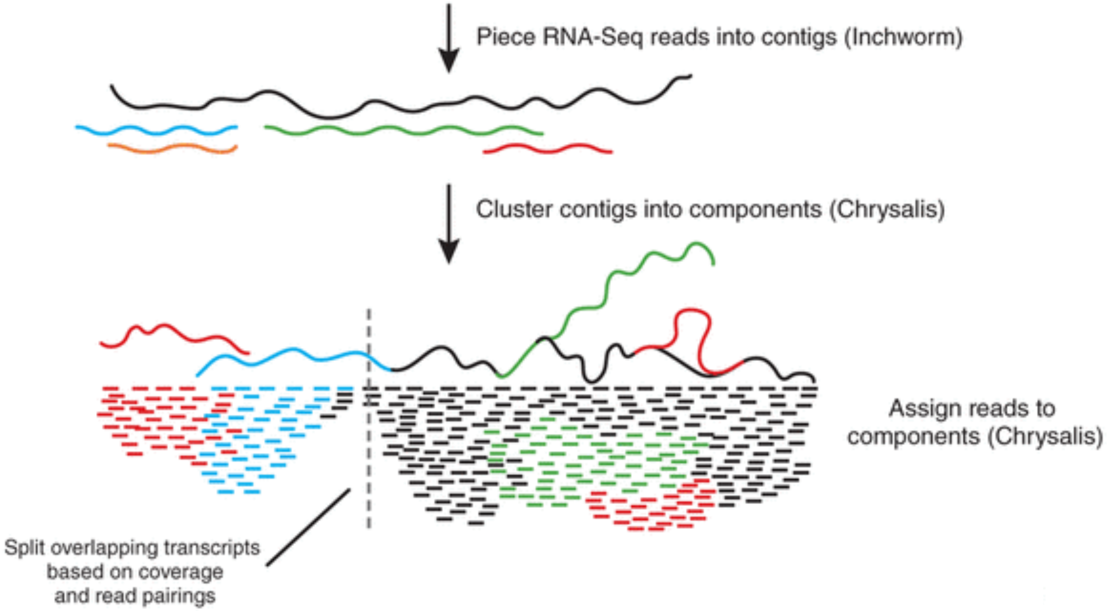
TGAATT

GGCAATTGACTTTT…

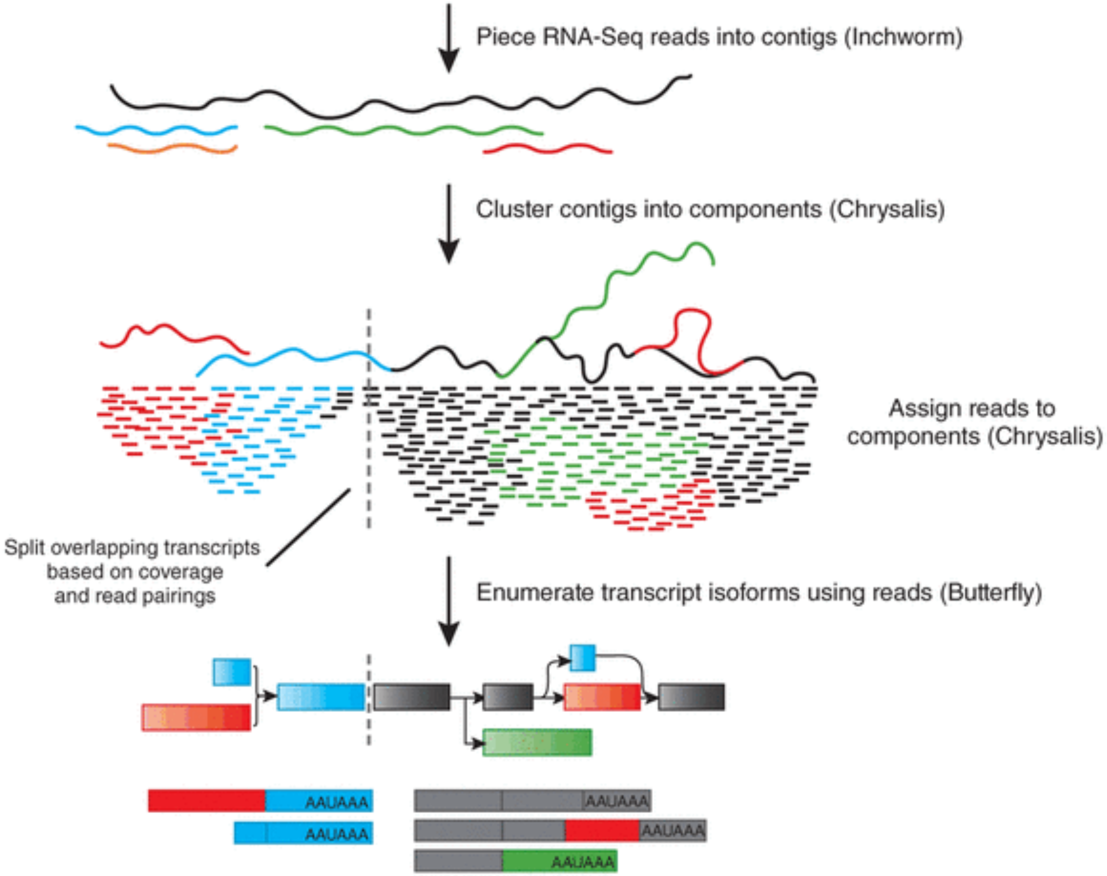
GAAGGGAGTTCCACT…



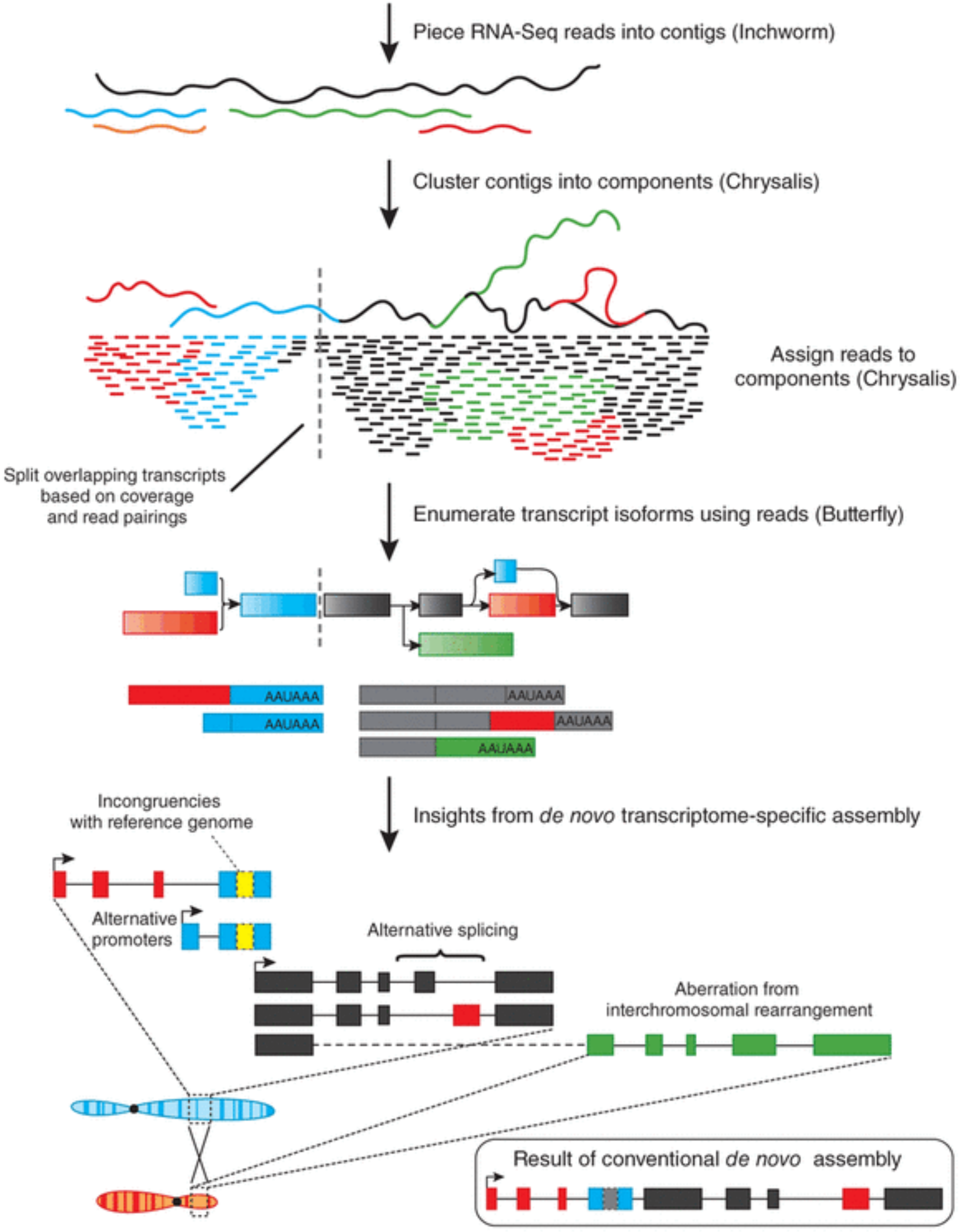
Iyer MK, Chinnaiyan AM (2011)
Nature Biotechnology **29**, 599–600


Piece RNA-Seq reads into contigs (Inchworm)
Cluster contigs into components (Chrysalis)
Assign reads to components (Chrysalis)
Split overlapping transcripts based on coverage and read pairings

Iyer MK, Chinnaiyan AM (2011)
Nature Biotechnology **29**, 599–600


Piece RNA-Seq reads into contigs (Inchworm)
Cluster contigs into components (Chrysalis)
Assign reads to components (Chrysalis)
Split overlapping transcripts based on coverage and read pairings
Enumerate transcript isoforms using reads (Butterfly)
AAUAAA
AAUAAA
AAUAAA
AAUAAA
AAUAAA

Iyer MK, Chinnaiyan AM (2011)
Nature Biotechnology **29**, 599–600


Piece RNA-Seq reads into contigs (Inchworm)
Cluster contigs into components (Chrysalis)
Assign reads to components (Chrysalis)
Split overlapping transcripts based on coverage and read pairings
Enumerate transcript isoforms using reads (Butterfly)
AAUAAA
AAUAAA
AAUAAA
AAUAAA
AAUAAA
Insights from de novo transcriptome-specific assembly
Incongruencies with reference genome
Alternative promoters
Alternative splicing
Aberration from interchromosomal rearrangement
Result of conventional de novo assembly

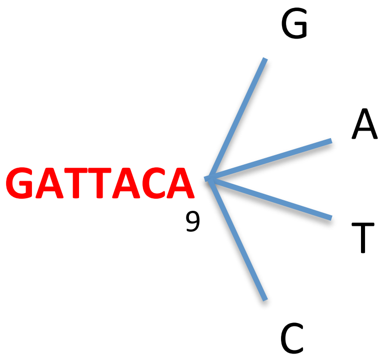
Iyer MK, Chinnaiyan AM (2011) *Nature Biotechnology* **29**, 599–600

Inchworm Algorithm

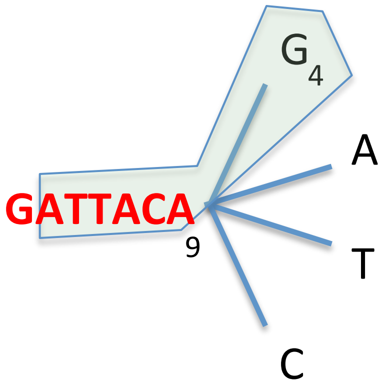
Decompose all reads into overlapping Kmers (25-mers)

Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.

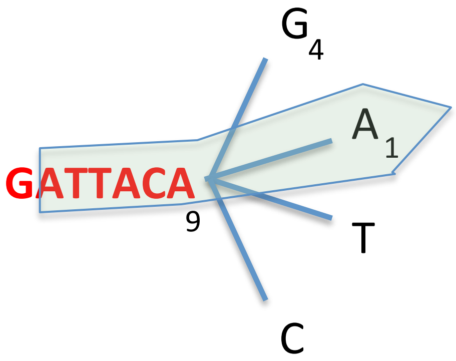
Extend kmer at 3' end, guided by coverage.



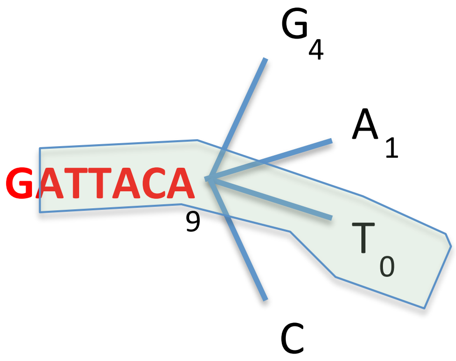
Inchworm Algorithm



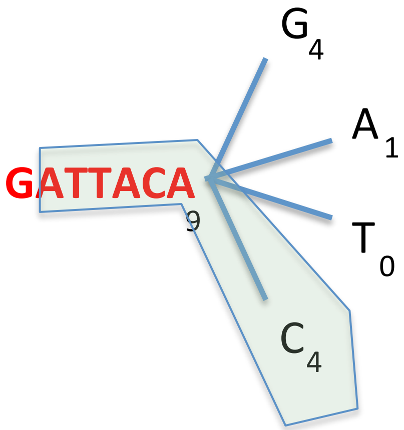
Inchworm Algorithm



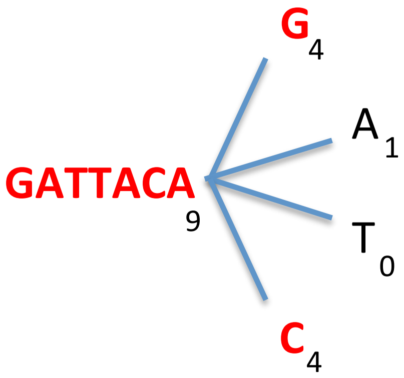
Inchworm Algorithm



Inchworm Algorithm



Inchworm Algorithm



G_4

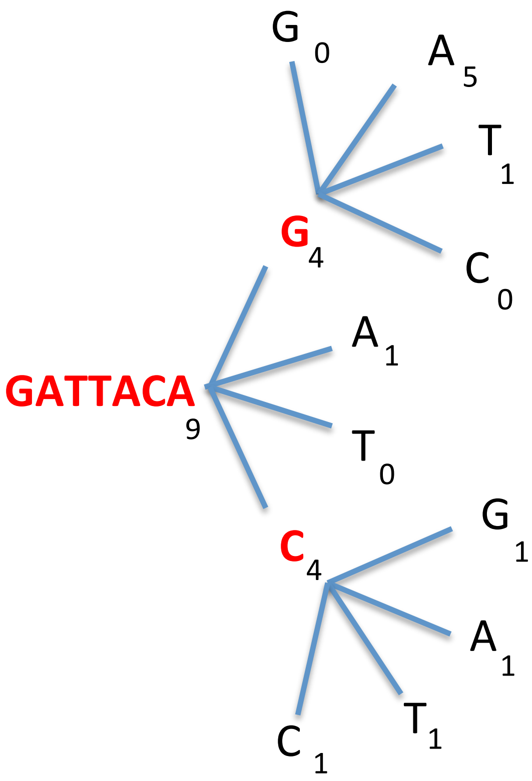
A_1

GATTACA_9

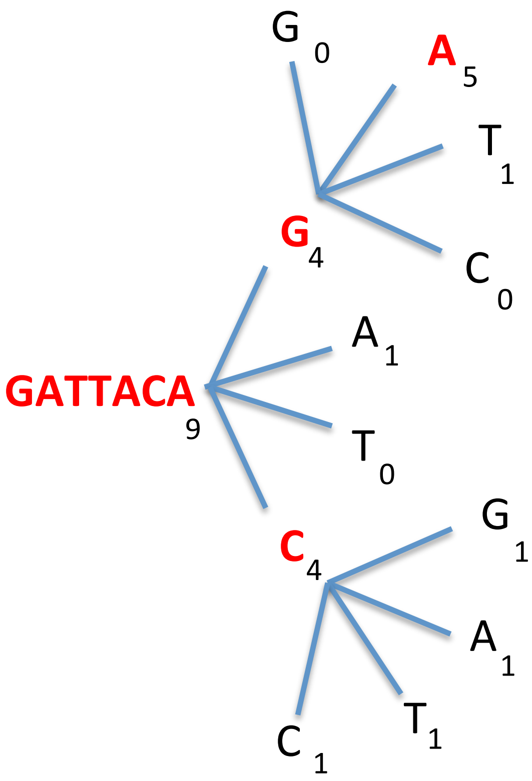
T_0

C_4

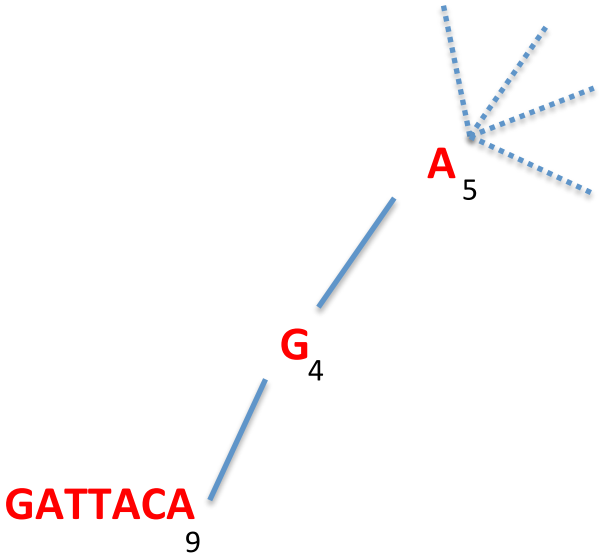
Inchworm Algorithm



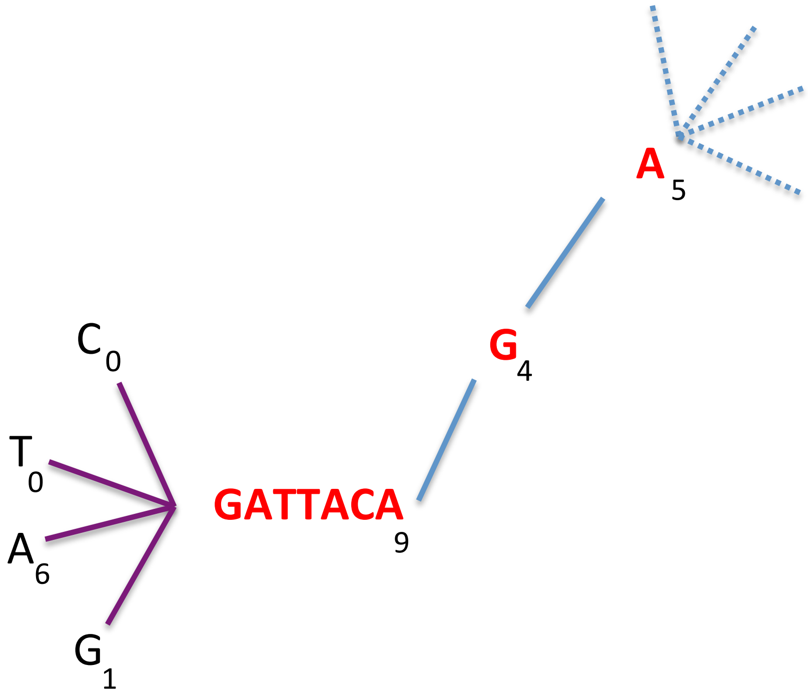
Inchworm Algorithm



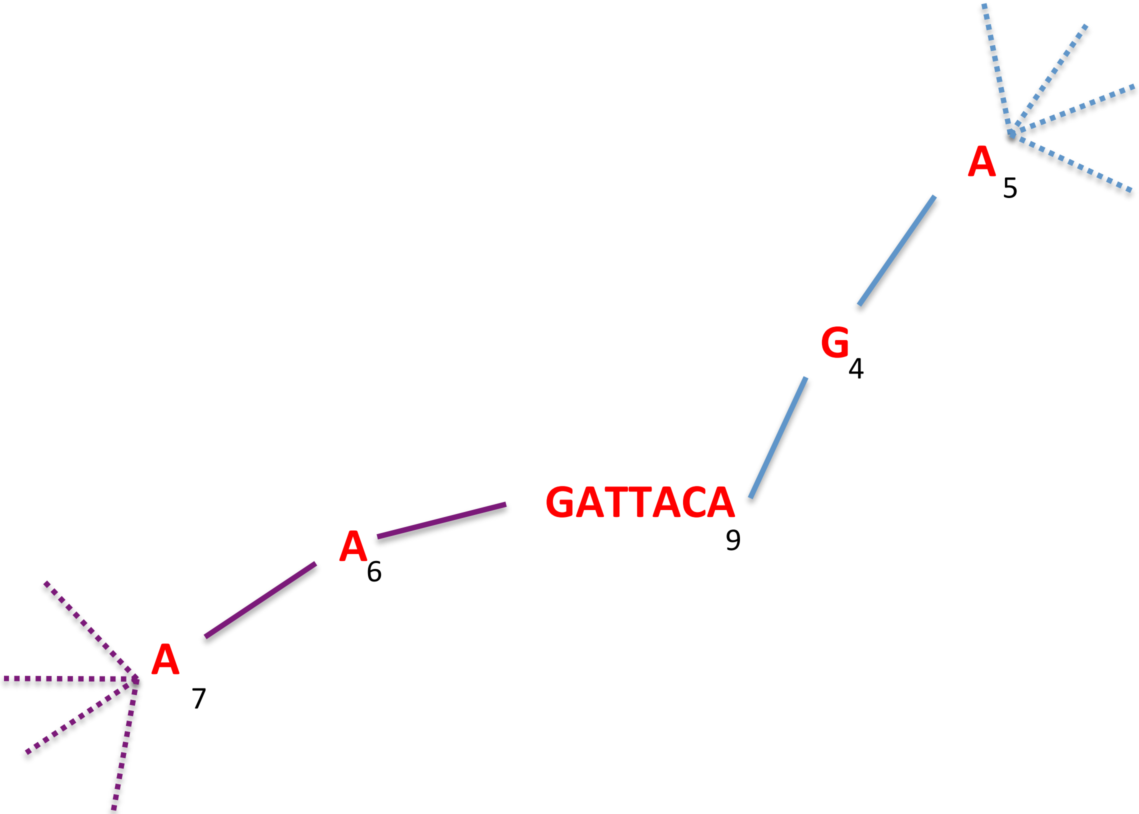
Inchworm Algorithm



Inchworm Algorithm



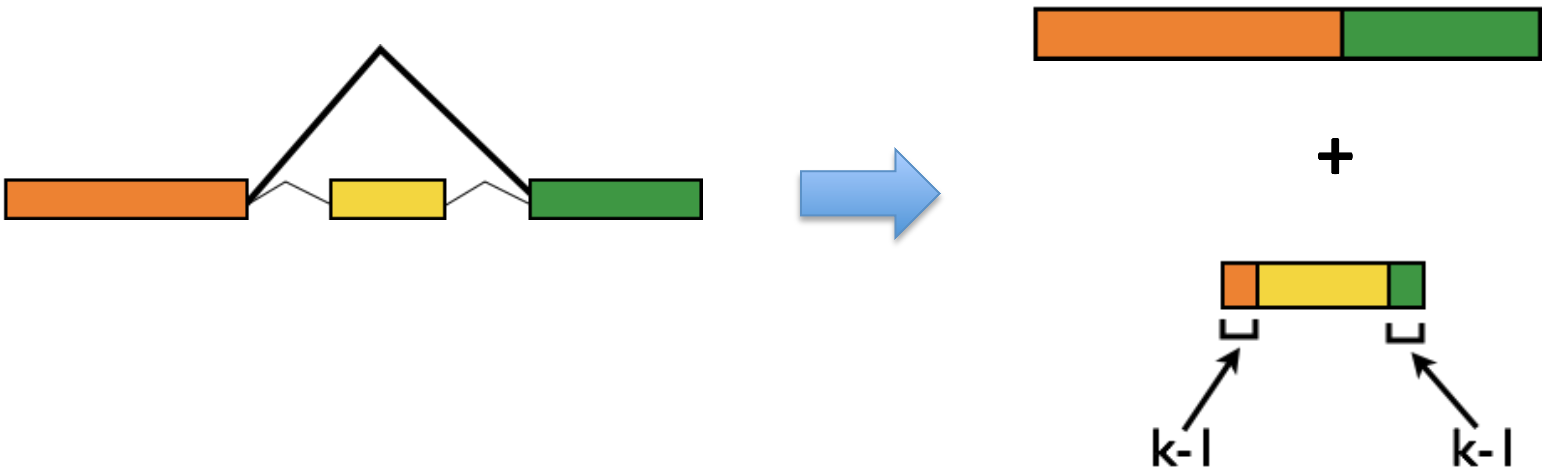
Inchworm Algorithm



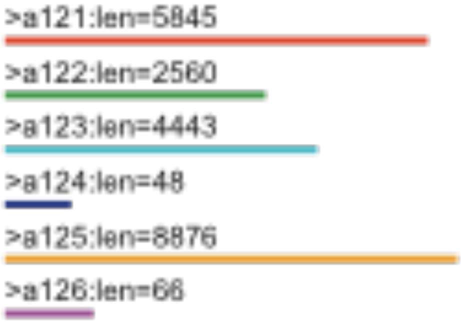
Report contig: ….AAGATTACAGA….

Remove assembled kmers from catalog, then repeat the entire process.

Inchworm Contigs from Alt-Spliced Transcripts => Minimal lossless representation of data

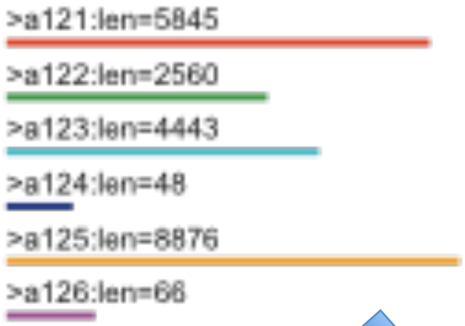


Chrysalis

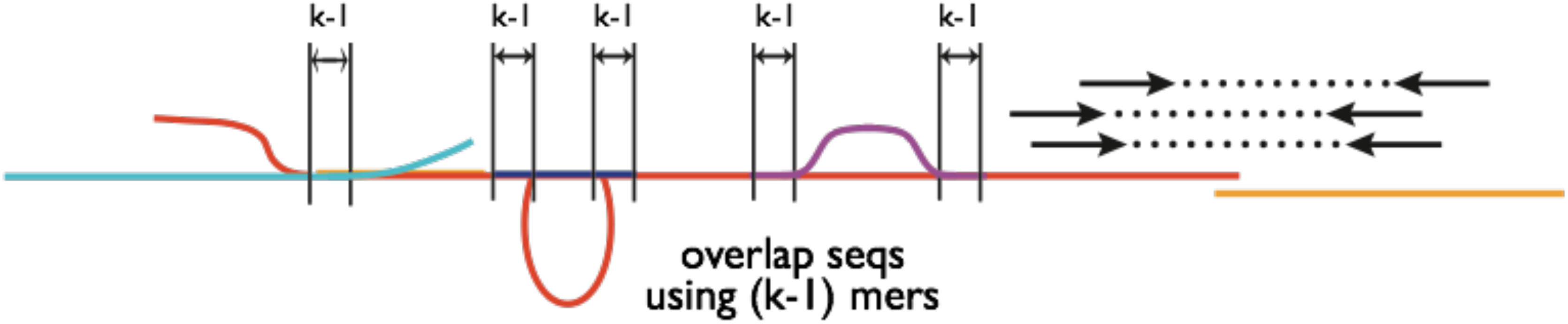


Integrate isoforms
via k-1 overlaps

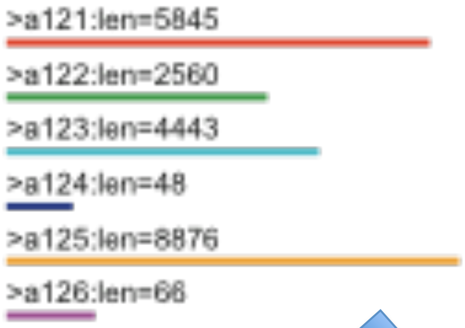
Chrysalis



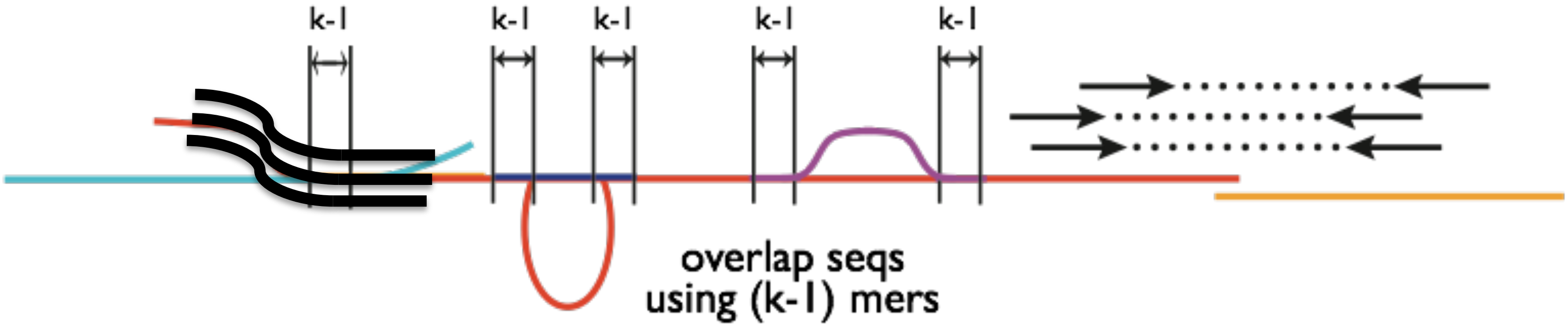
Integrate isoforms via k-1 overlaps



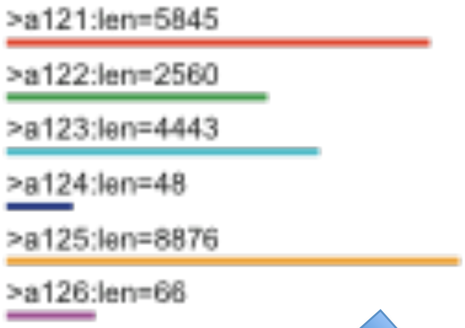
Chrysalis



Integrate isoforms
via k-1 overlaps
Verify via “welds”

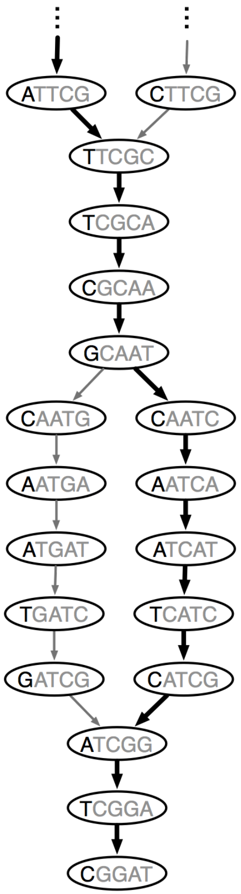


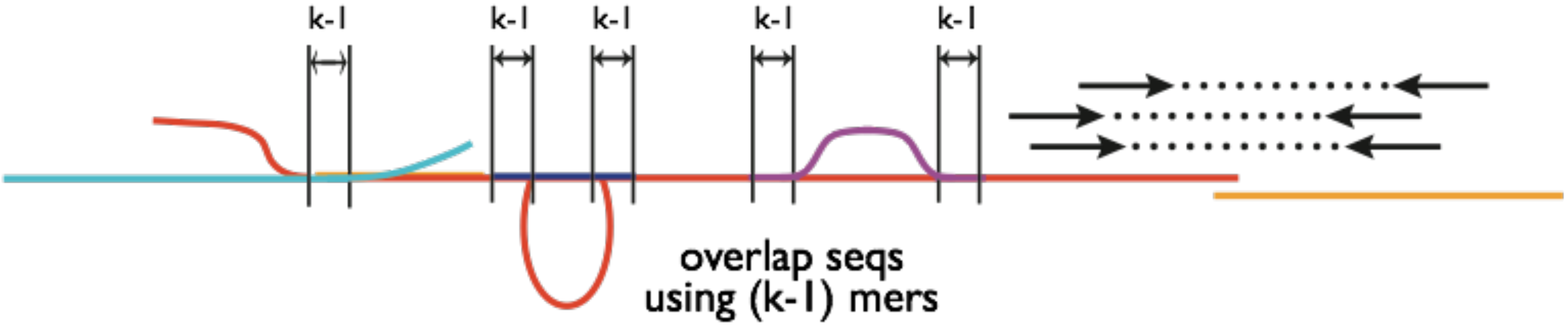
Chrysalis



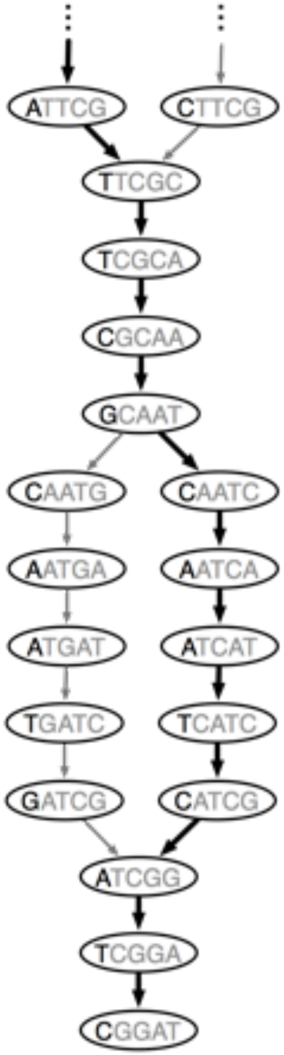
Integrate isoforms
via k-1 overlaps
Verify via “welds”

Build de Bruijn Graphs
(ideally, one per gene)



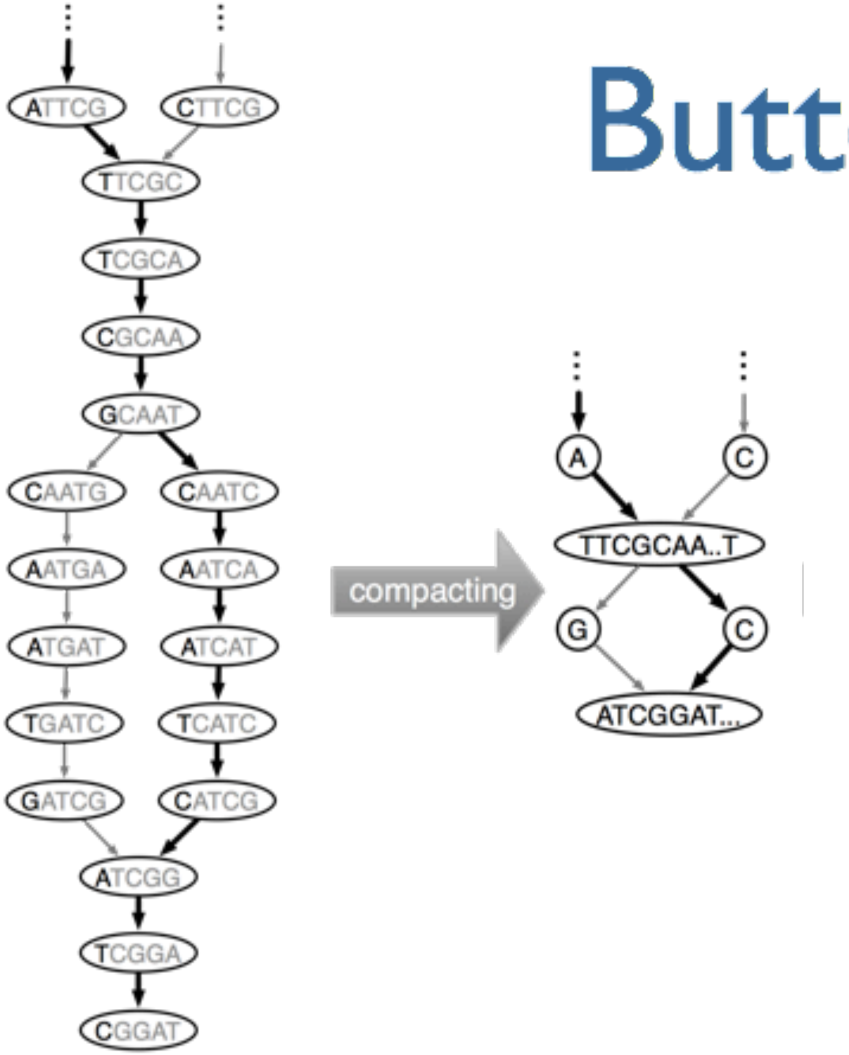


Butterfly



de Bruijn graph

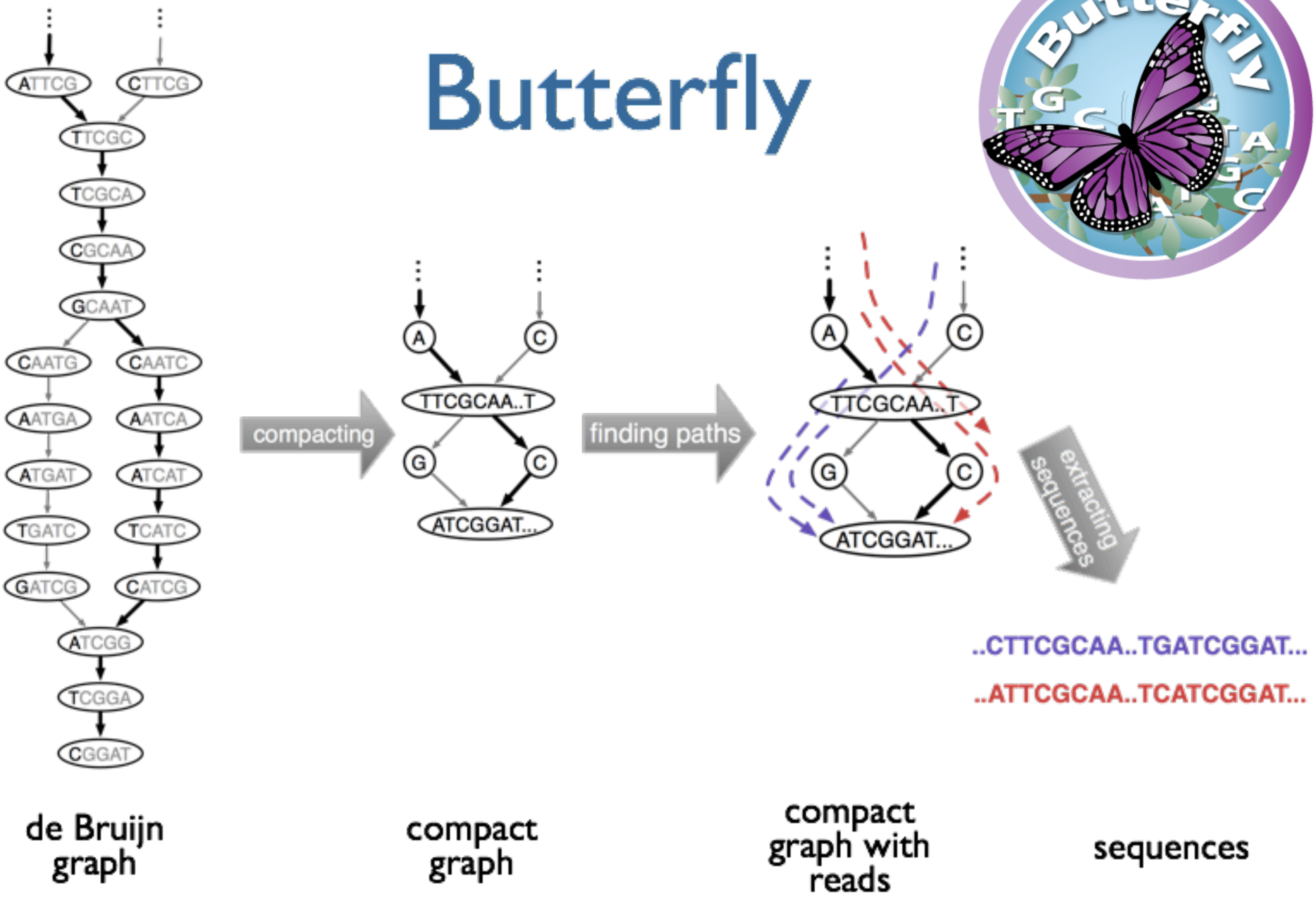
Butterfly



de Bruijn graph

compact graph

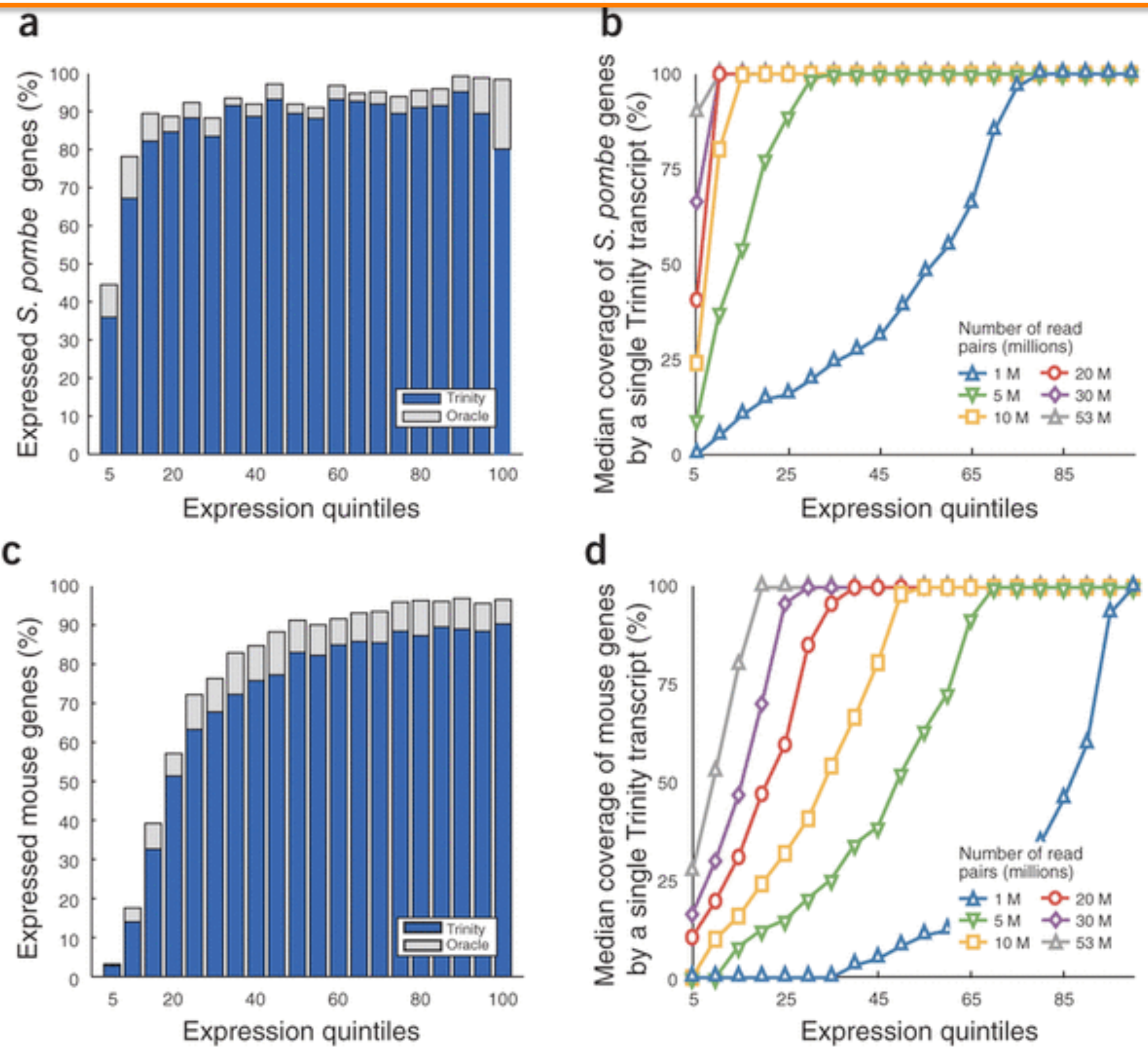
Butterfly



..CTTCGCAA..TGATCGGAT...

..ATTCGCAA..TCATCGGAT...

Completeness and coverage as function of read counts



Grabherr et al. Nature Biotechnology 29, 644–652 (2011)