Transcriptome Assembly, Functional Annotation (and related thoughts)

Monica Britton, Ph.D.

Sr. Bioinformatics Analyst

June 2018

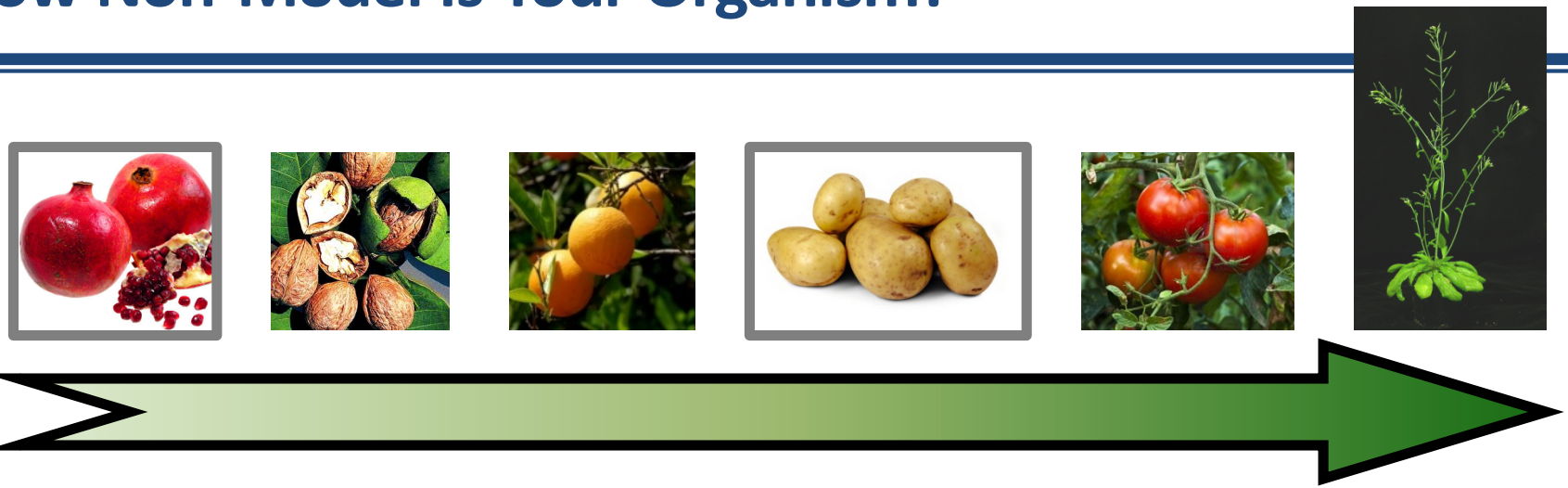
In this week's exercises, you're working with a typical mRNA-Seq project.

The reads are aligned to a genome with existing gene and functional annotation.

We've had the "luxury" of using a model organism.

But what if your favorite organism has poor or little or no gene sequence information?

How Non-Model is Your Organism?



- **Novel** – little/no previous sequencing (may need assembly)
- **Non-Model**
  - some sequence available (draft genome or transcriptome assembly)
  - Thousands of scaffolds, maybe tens of chromosomes
  - Some annotation (*ab initio*, EST-based, etc.)
- **Model** – genome fully sequenced and annotated
  - Multiple genomes available for comparison
  - Well-annotated transcriptome based on experimental evidence
  - Genetic maps with markers available
  - Basic research can be conducted to verify annotations (mutants available)

Questions ...

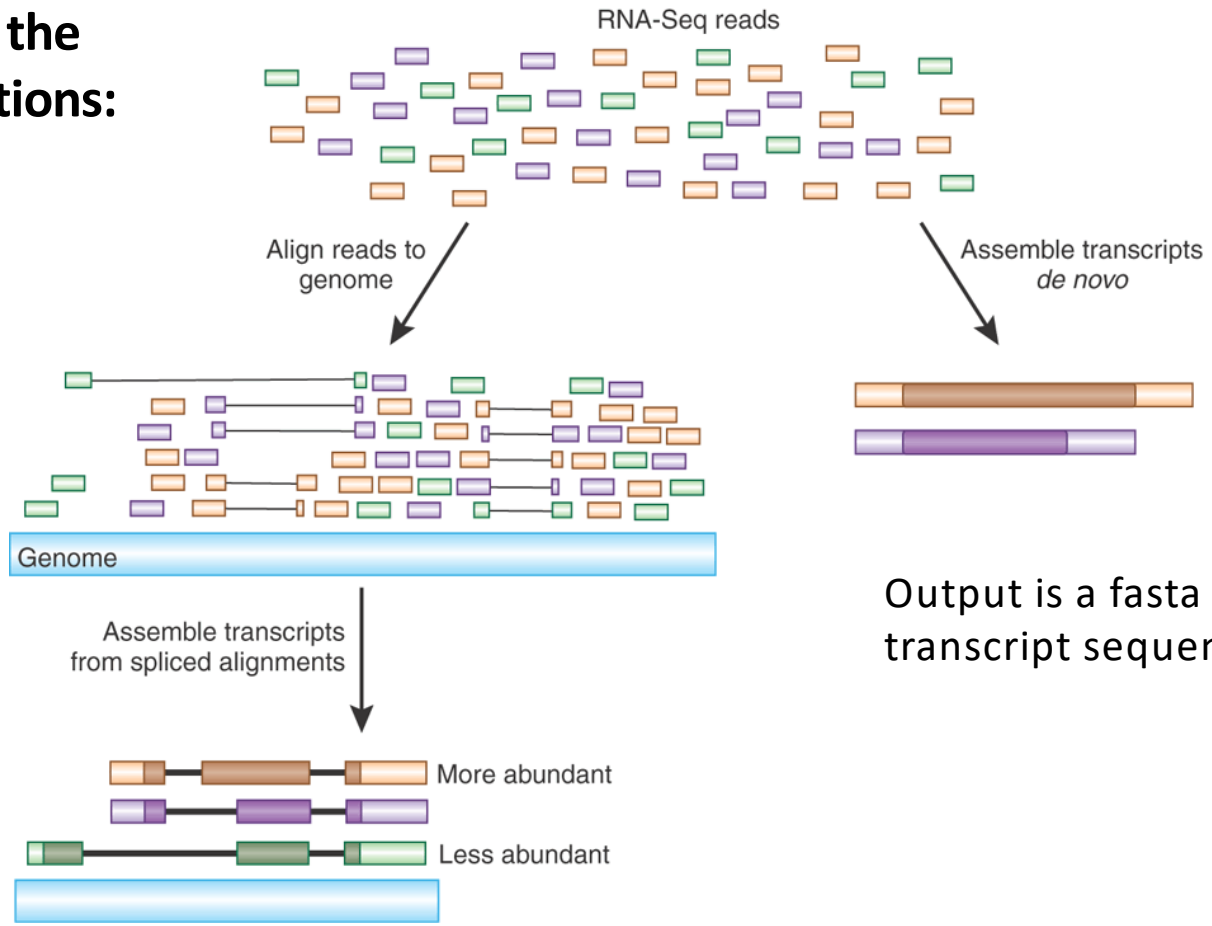
If we have a "less than model" organism, how can we find the genes?

How can we determine the function of the products of these genes?

Is an assembly always the answer?

Gene Construction (Alignment) vs. De Novo Assembly

For several years, these were the primary options:



Genome-Sequenced Organisms

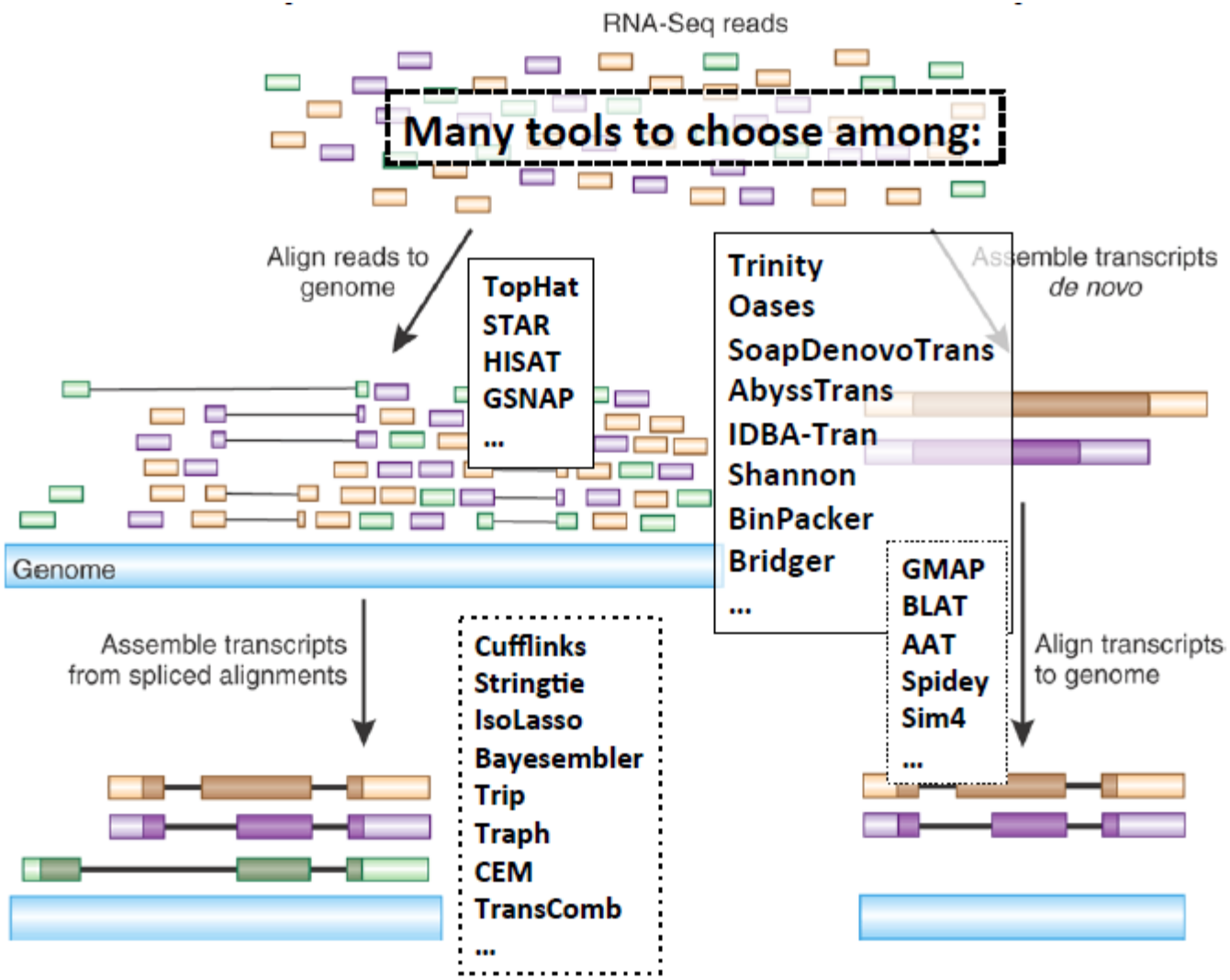
Novel or Non-Model Organisms

Output is a fasta file of transcript sequences

Output is a GTF file

Haas and Zody (2010) Nat. Biotech. 28:421-3

Lots of Software Choices



From: https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/tree/master/docs

Gene / Transcriptome Construction

- Genome is needed (can be “drafty”)
  - Alignments parsed to find putative exons
  - Spliced reads used to join exons into transcripts
  - *Caution*: Nearby or overlapping genes can be inappropriately combined.
- Annotation can be improved – even for well-annotated model organisms
  - Identify **all** expressed exons (and retained introns)
  - Combine expressed exons into genes
  - Find all splice variants for a gene
  - Discover novel transcripts
- For newly sequenced organisms
  - Validate ab initio annotation
  - Comparison between different annotation sets

Transcriptome Assembly With Trinity

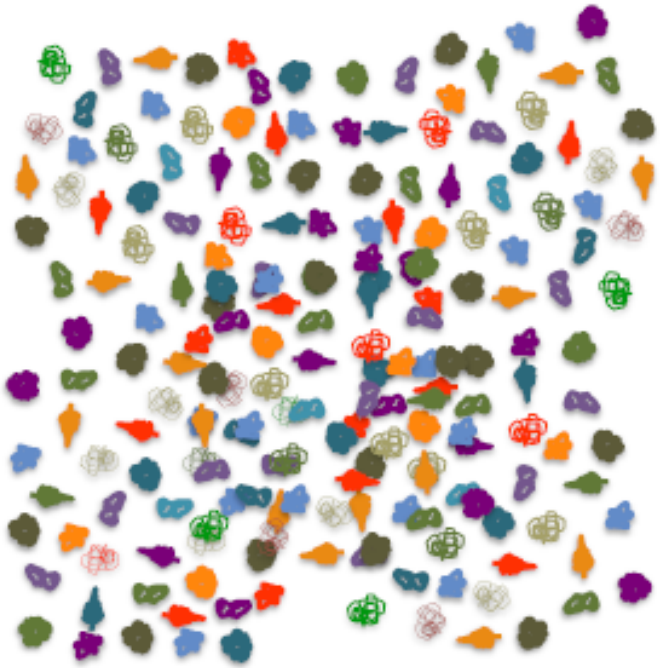
Genome Assembly

Single Massive Graph

Entire chromosomes represented.

Trinity Transcriptome Assembly

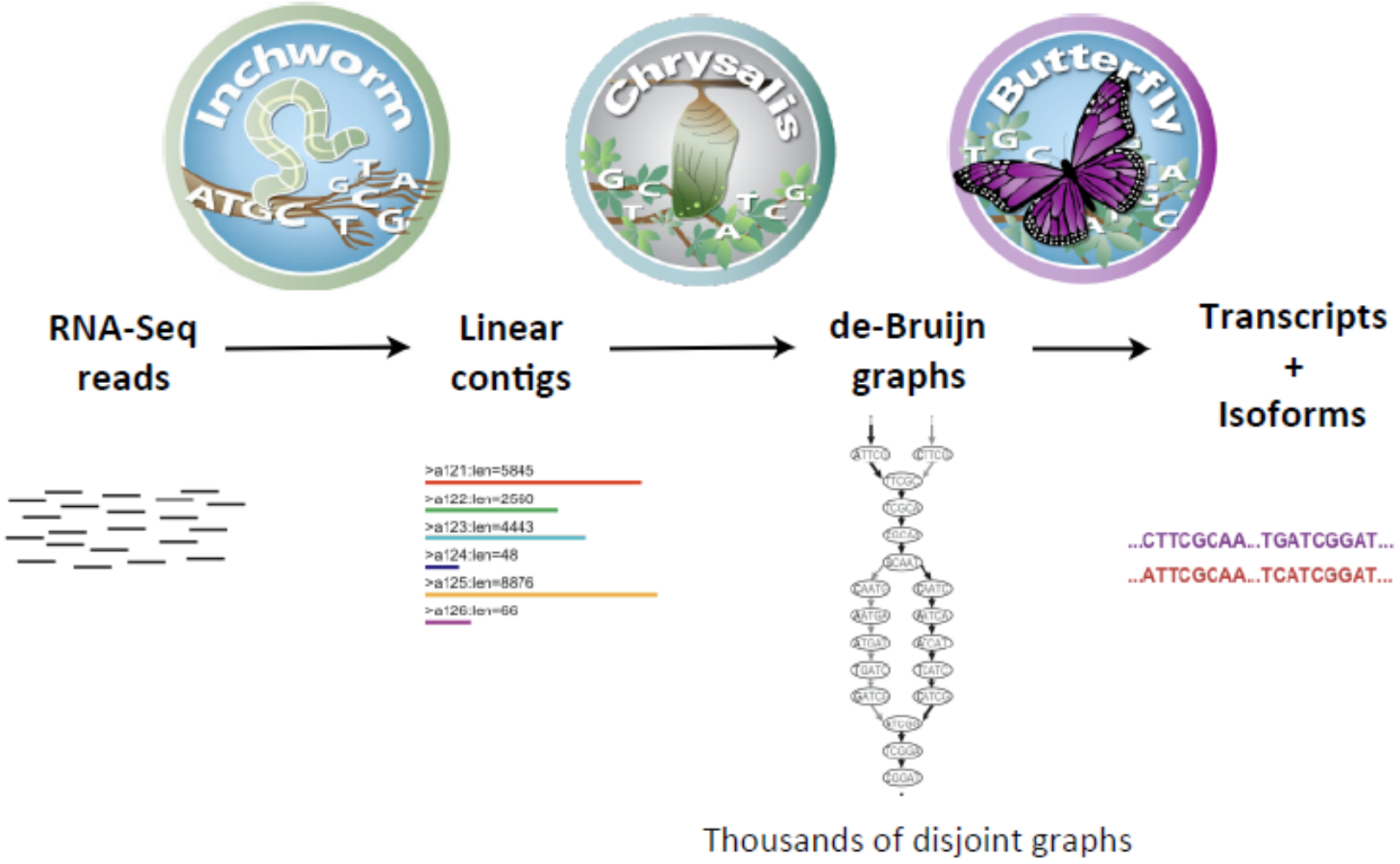
Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

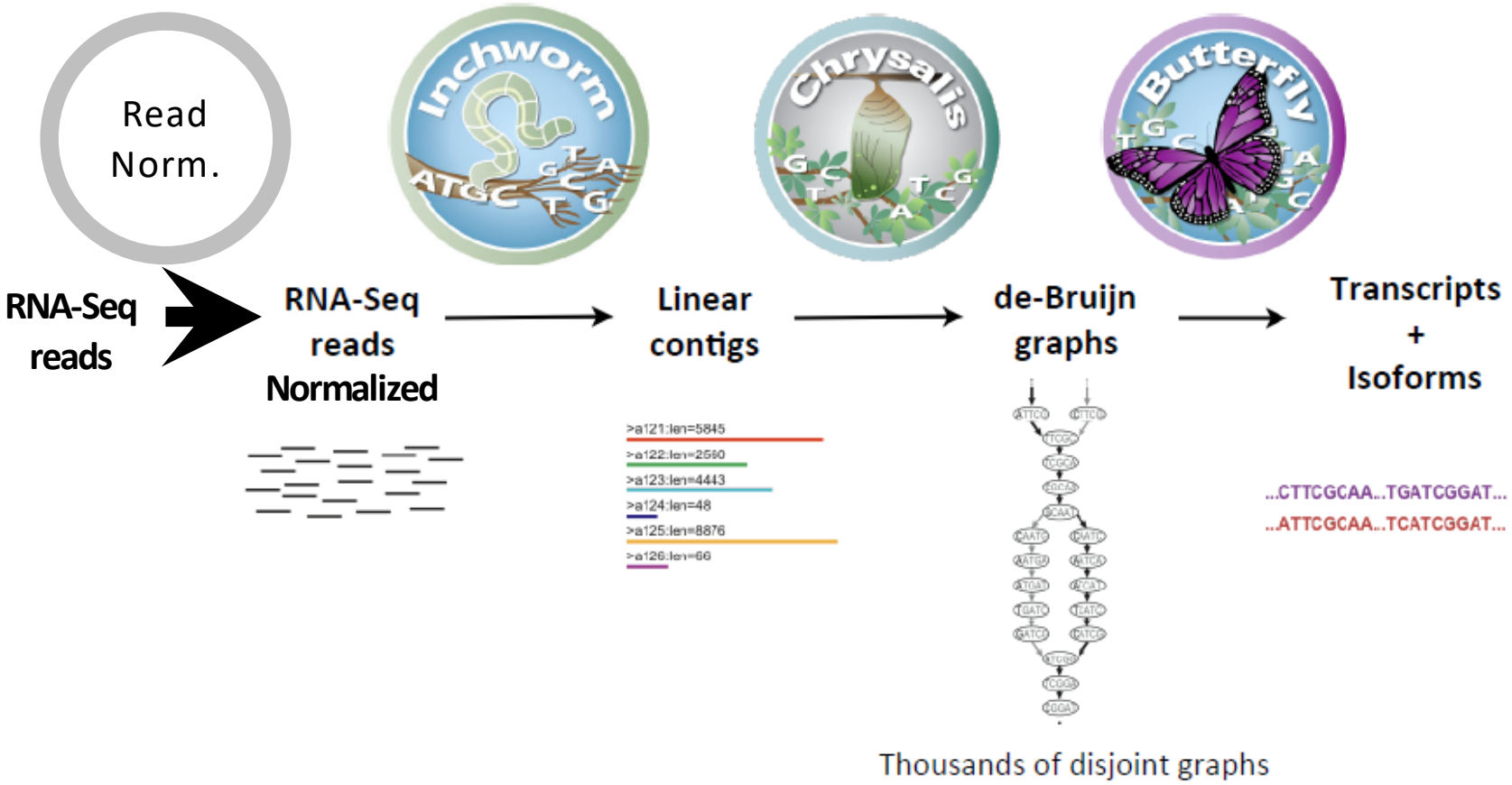
From: https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/tree/master/docs

Transcriptome Assembly With Trinity



From: https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/tree/master/docs

There are Really Four Steps to Trinity Assemblies



From: https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/tree/master/docs

Recommendations for Transcriptome Assemblies

Strand-specific RNA-Seq is **very useful**. This allows precise construction of sense and anti-sense transcripts.

Always use paired-end reads.

If possible, generate a library from all possible tissues/stages/conditions, prepared with a wide range of insert sizes, and sequenced on a MiSeq.

Aggressively trim raw reads to remove all traces of adapters and polyA tails.

Only generate one final assembly per organism. (Don't run separate assemblies for different tissues/stages/conditions.)

(Try to avoid ever doing a hybrid 454-Illumina assembly.)

Trinity Results: Why do I have so many contigs???

It's common to get far "too many" transcript contigs (100k to over 1 million)

Reducing this large set of contigs to a manageable gene set is the ***real*** work of transcriptome assembly.

The first step is to align all the original (trimmed) reads to the raw transcriptome. Over 90% of the input should align!

Note: It's best to only count those reads that align in proper (concordant) pairs, which will help to minimize chimeric contigs, and may reduce the number of short contigs.

Then comes filtering.

Filtering Trinity Assemblies – this is NOT “Cookbook”

Steps may include:

- Protein Prediction/ORF-Calling (Transdecoder)
- Isoform abundance (RSEM, eXpress, Kallisto, or Salmon)
- Annotation (Trinotate, Blast2GO, etc.)
- Contaminant screening (rRNA, PhiX, bacteria, etc.)

Goal is generally a set of transcripts to which >75% of the original trimmed reads align.

Filtering is not the same as "discarding". You can always go back to the unfiltered transcripts if later you want to look for something new!

Transcriptome Assembly Challenges

Contamination in original RNA sample (other genomes represented)

Paralogs vs. splice variants

Coverage (highly expressed vs. low expressed genes) – mitigated by read normalization step

Repetitive sequences (not as much a problem as with genome assembly)

Chimeric contigs

Is There Another/Better Way?

Both gene construction and transcriptome assemblies have drawbacks.

Those methods are based on *algorithmically* combining information from short reads to determine the sequence of a longer transcript/gene.

But now it is possible to sequence an entire transcript for a single cDNA molecule:

PacBio IsoSeq

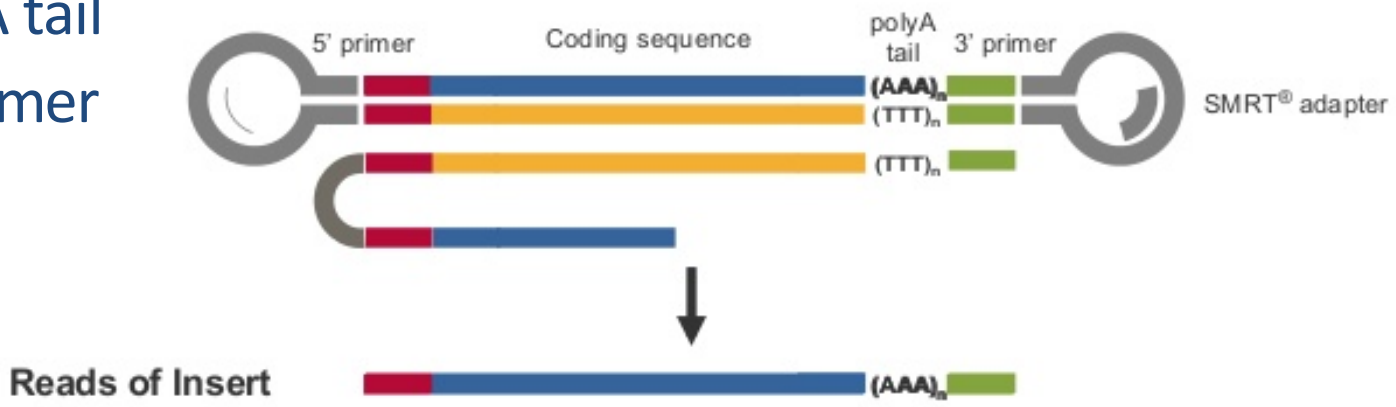
Nanopore

These technologies also promise to be able to directly sequence RNA, with no need to convert to cDNA.

PacBio IsoSeq: Full-Length (FL) Read Identification

A read is “Full-Length” if these have been identified:

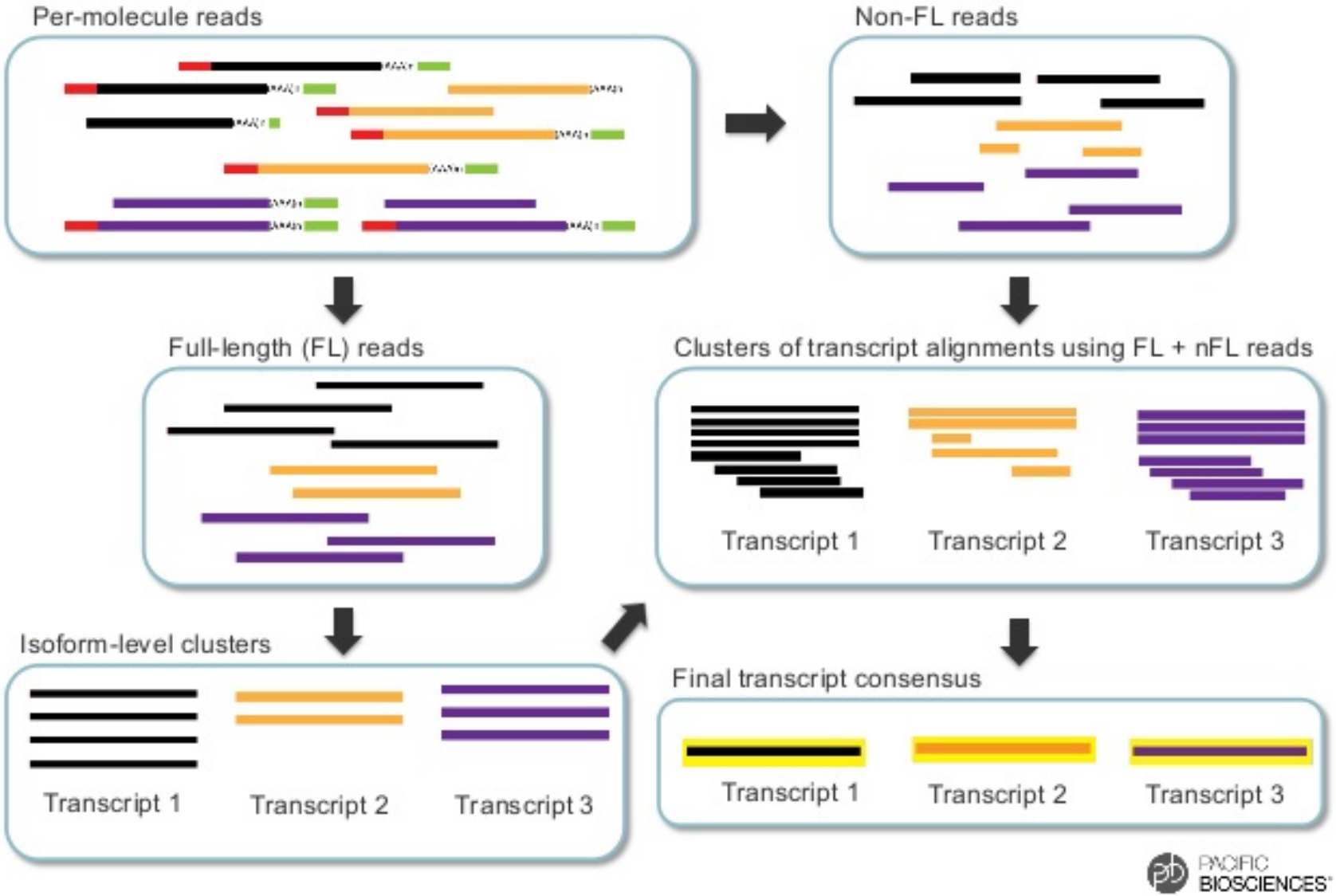
- 5’ primer
- polyA tail
- 3’ primer



5’ and 3’ primers and polyA/T tail are identified and removed.

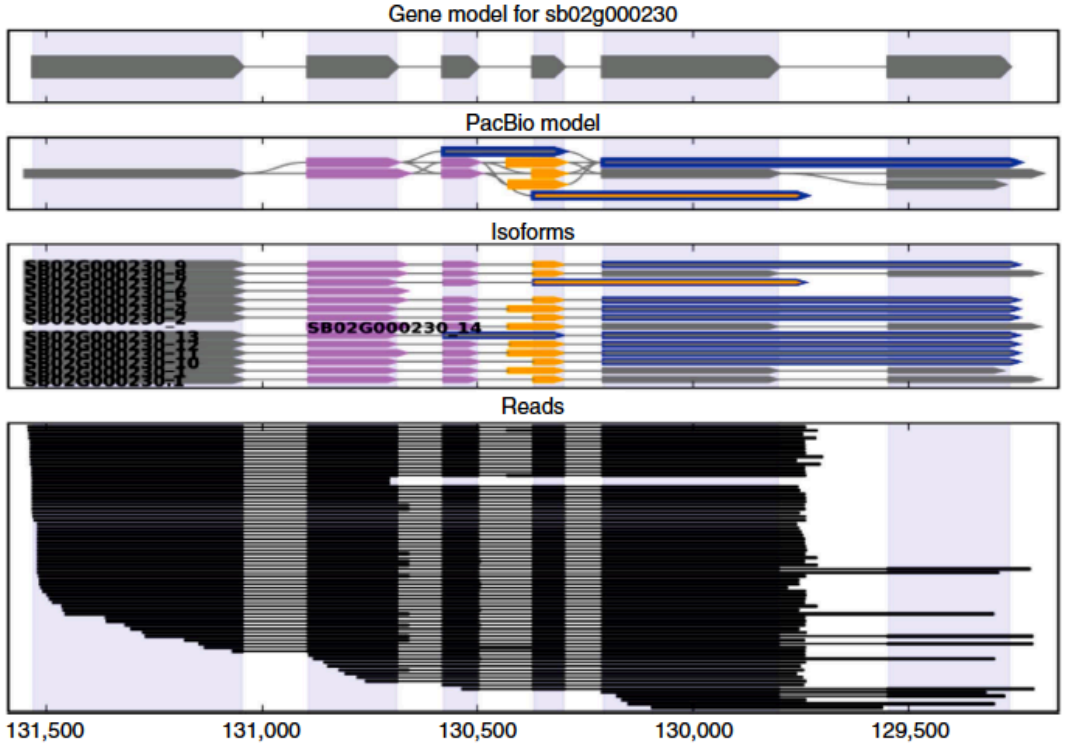
Transcript strandedness is identified

PacBio Isoseq Informatics Pipeline



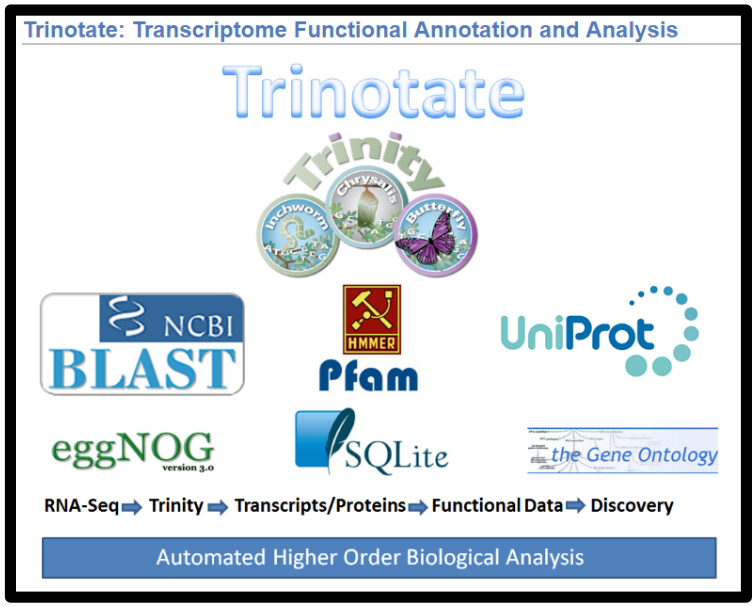
Alignment to Genome

- Current best practice is to use minimap2 or GMAP
- These alignments can be used to generate a gtf annotation file and infer isoforms within gene models
- In development: stringtie update to include support for isoseq



Abdel-Ghany et al, 2016,
Nature Communications

Functional Annotation: Trinotate and Blast2GO



- Integrates well with Trinity
- Some tweaking needed for non-Trinity fastas
- Uses high-confidence annotation databases
- Runs on command-line but has GUI visualizations
- FREE

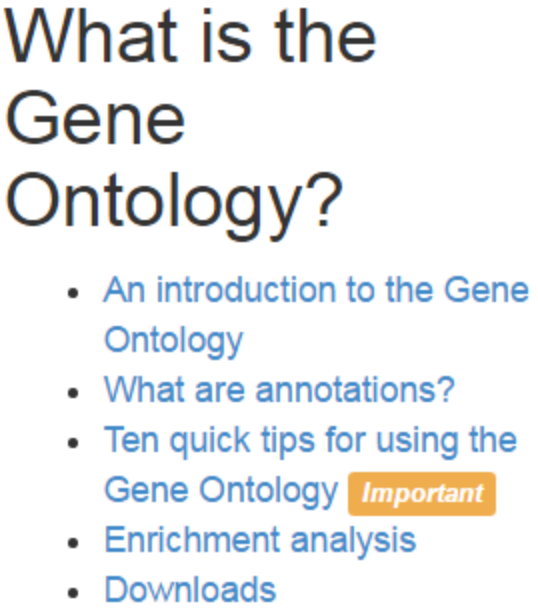
- Suitable for any sequence fasta
- Databases are highly customizable.
- Has GUI and command-line versions
- Free version is slow and has limited functions
- Paid version is $$$

Gene Ontology (www.geneontology.org)

Gene Ontology provides a controlled vocabulary of terms that allow genes within organisms and across organisms to be compared and grouped.

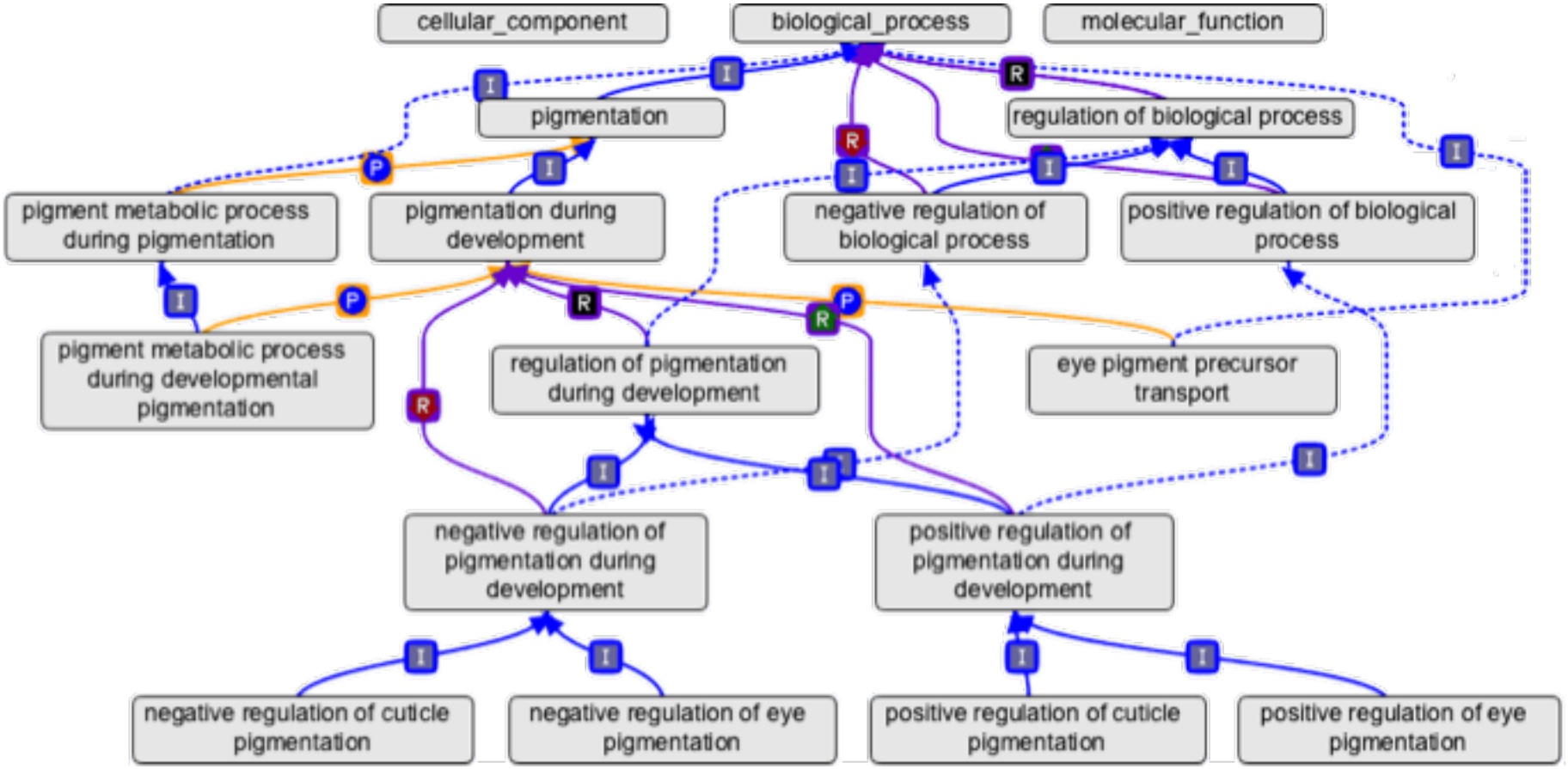
Three categories of GO:

- Molecular Function: molecular activities of gene products
- Biological Process: pathways and larger processes made up of the activities of multiple gene products.
- Cellular Component: where gene products are.



Gene Ontology (www.geneontology.org)

GO terms often have a complex parent-child relationship:



From: http://geneontology.org/page/ontology-structure

KEGG Pathways (www.genome.jp/kegg/)

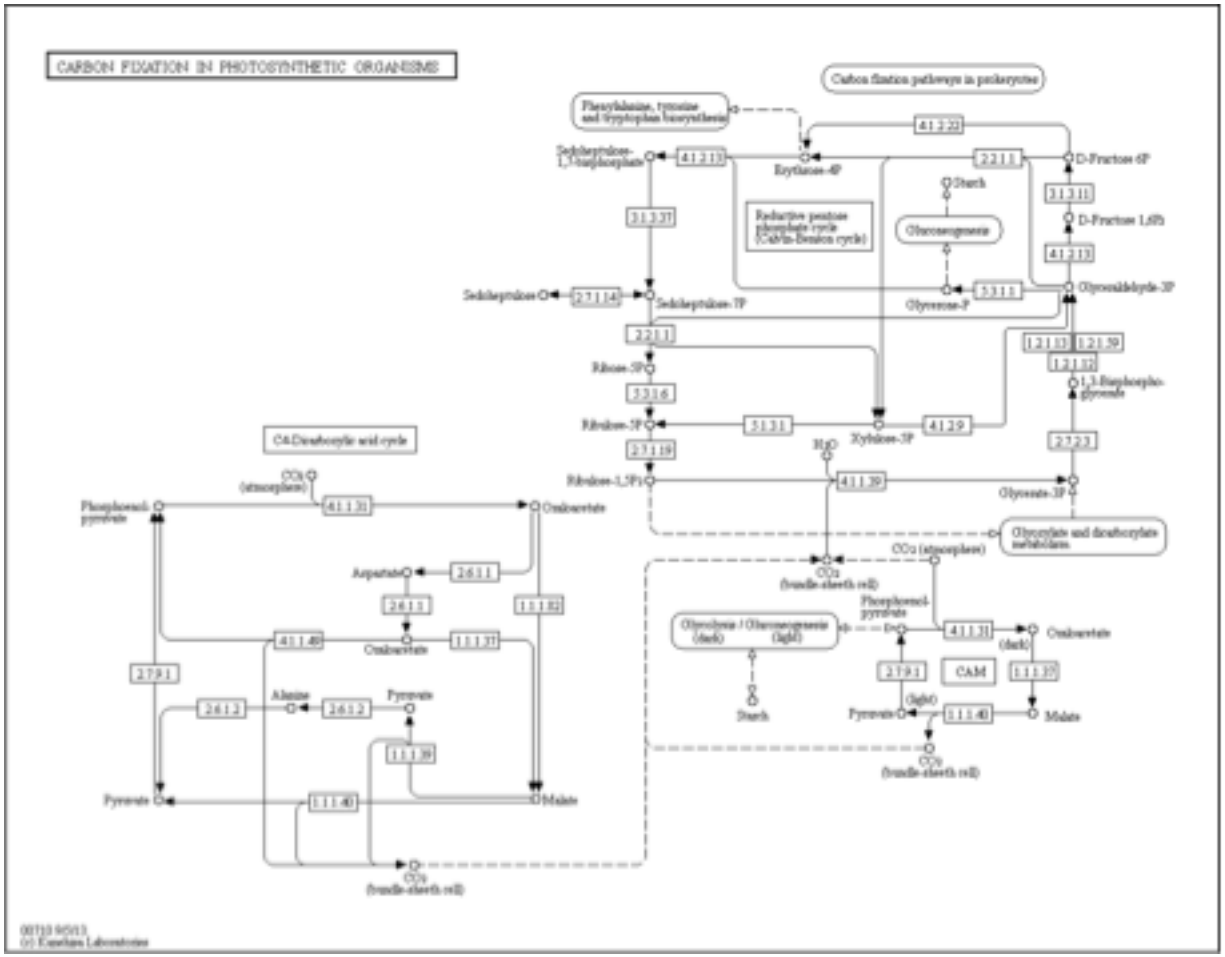
KEGG (Kyoto Encyclopedia of Genes and Genomes) integrates sixteen underlying databases.

Category	Database	Content	Color
Systems information	KEGG PATHWAY	KEGG pathway maps	
	KEGG BRITE	BRITE hierarchies and tables	
	KEGG MODULE	KEGG modules	
Genomic information	KEGG ORTHOLOGY (KO)	Functional orthologs	
	KEGG GENOME	KEGG organisms (complete genomes)	
	KEGG GENES	Genes and proteins	
	KEGG SSDB	GENES sequence similarity	
Chemical information	KEGG COMPOUND	Small molecules	
	KEGG GLYCAN	Glycans	
	KEGG REACTION	Biochemical reactions	
	KEGG RCLASS	Reaction class	
	KEGG ENZYME	Enzyme nomenclature	
Health information	KEGG DISEASE	Human diseases	
	KEGG DRUG	Drugs	
	KEGG DGROUP	Drug groups	
	KEGG ENVIRON	Health-related substances	

Chemical information category is collectively called KEGG LIGAND
Health information category integrated with drug labels is called KEGG MEDICUS

KEGG Pathways (www.genome.jp/kegg/)

An example of a KEGG pathway:



A Note About Databases

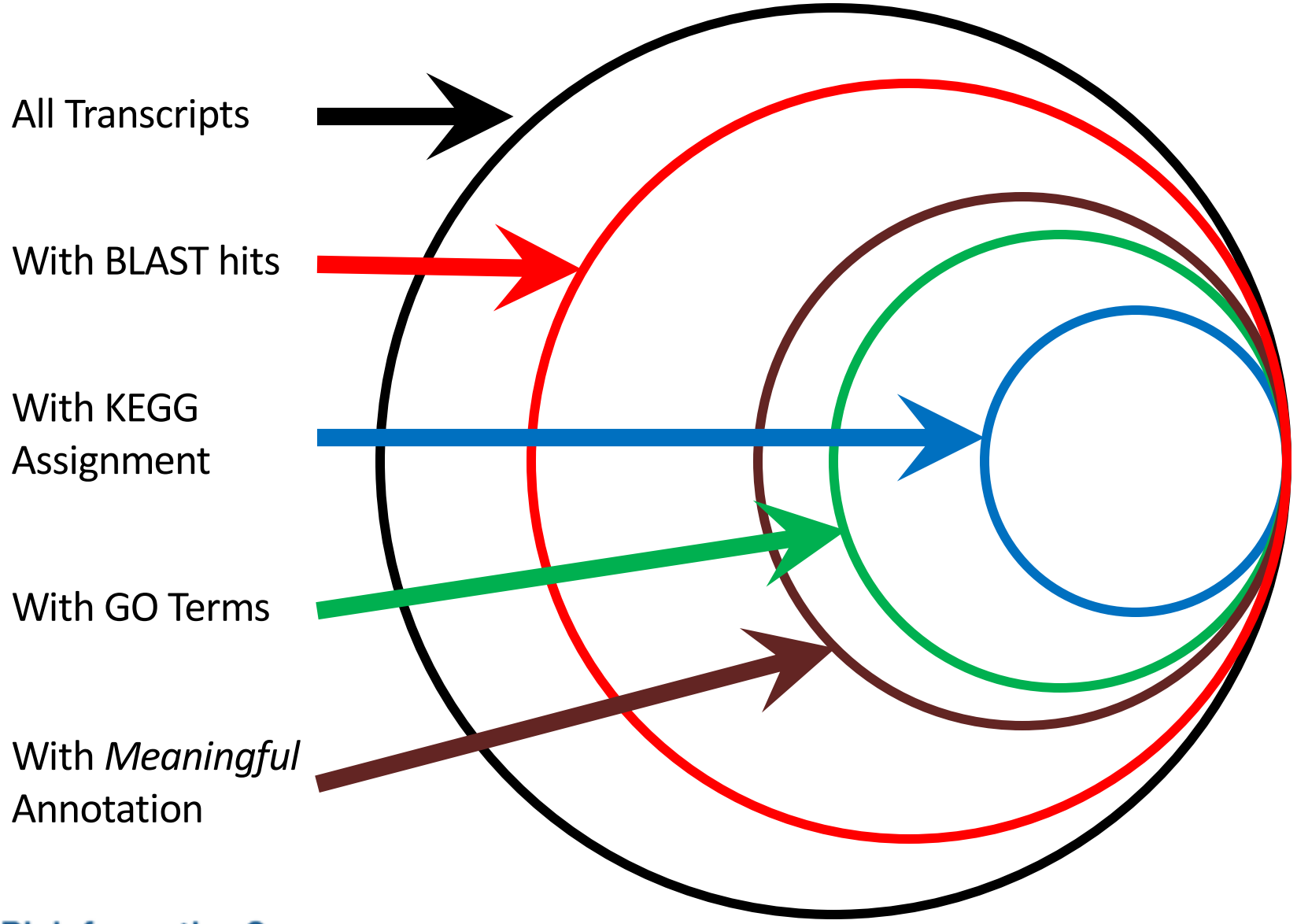
All annotation packages rely on databases that are used with BLAST (or other such programs)

Trinotate and B2G have customized databases that linked to Gene Ontology terms and KEGG enzyme codes/pathways.

You can use other databases, which can greatly inform on the function, but may not have Gene Ontology/pathway info (yet).

I often use an iterative blasting approach, where I first use the protein/gene sets of related well-annotated organisms, and then widen the databases for transcripts that don't have good hits in the first round.

Annotation Summary Might Look Like:



Be Cautious and Skeptical When Interpreting Annotation

Emsembl Biomart (www.ensembl.org/biomart/martview) is a great resource for gene IDs, GO Terms, and annotation for many organisms. Ensembl staff curate gene annotations.

But nothing is perfect …

In 2016 we were working on a horse RNA-Seq project.

Blythe was running a Gene Ontology enrichment analysis, and noticed an unexpected GO term was showing up as statistically significant:

GO:0015995 (chlorophyll biosynthetic process)

Be Cautious and Skeptical When Interpreting Annotation

Did we miss the announcement that photosynthesis was discovered in horses?

Be Cautious and Skeptical When Interpreting Annotation

I compared the GO annotation of the horse genes with their orthologs in human and mouse. The human/mouse genes were not annotated as photosynthetic. We concluded that either:

A) Photosynthesis has evolved in horses, but isn't present in other mammals, including human and mouse; or

B) There's an error in the Ensembl/GO annotation

Be Cautious and Skeptical When Interpreting Annotation

This error affected six genes in a gene family in horse, and potentially involved four GO Terms:

GO:0015979 photosynthesis
GO:0015995 chlorophyll biosynthetic process
GO:0016787 hydrolase activity
GO:0016851 magnesium chelatase activity

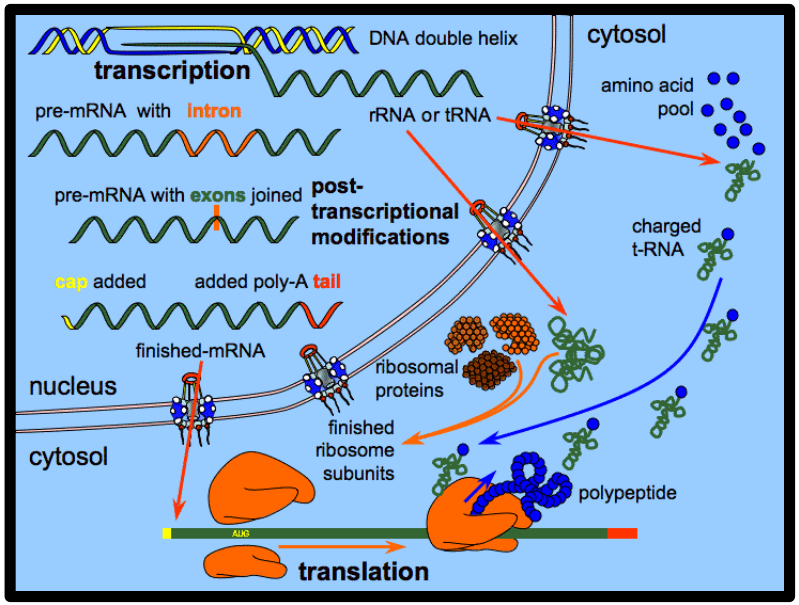
The first two GO Terms are obviously in error. It's more difficult to tell if the third and fourth GO Terms are incorrect.

Ensembl agreed with our assessment. However, their release schedule meant that the corrections would take six months to "go live".

Some mRNA-Seq Applications

- Differential gene expression analysis
- Transcriptional profiling

Assumption:
Changes in transcription/mRNA levels correlate with phenotype (protein expression)



- Identification of splice variants
- Novel gene identification
- Transcriptome assembly
- SNP finding
- RNA editing