Transcriptomics:

RNA-seq analysis: an overview

ANTHONY HALL

Earlham Institute

Decoding Living Systems

Evolution of transcriptomics technologies

- Northern blots: Single Genes
- RT-PCR: Multiple genes
- Microarrays: Model organism
- (NGS) RNA-seq: Any organism

What is RNA-seq?

- RNA-seq is the high throughput sequencing of cDNA using NGS technologies

- RNA-seq works by sequencing RNA molecules and profiling the expression of a particular gene by counting the number of time its transcripts have been sequenced.

- The summarized RNA-seq data is known as count data

How can RNA-seq be used?

- To look at regulation of gene expression
- To annotate a genome
- To score SNPs

Different types of RNA-seq

- **mRNA Sequencing**

Measures gene and transcript abundance and detects both known and novel features in the coding transcriptome

  - PolyA Selection, Oligo-dT, often using magnetic beads, can't be used to sequence non-polyA RNA.
  - rRNA Depletion

RNA-seq, strand specific (directional), standard illumina protocol

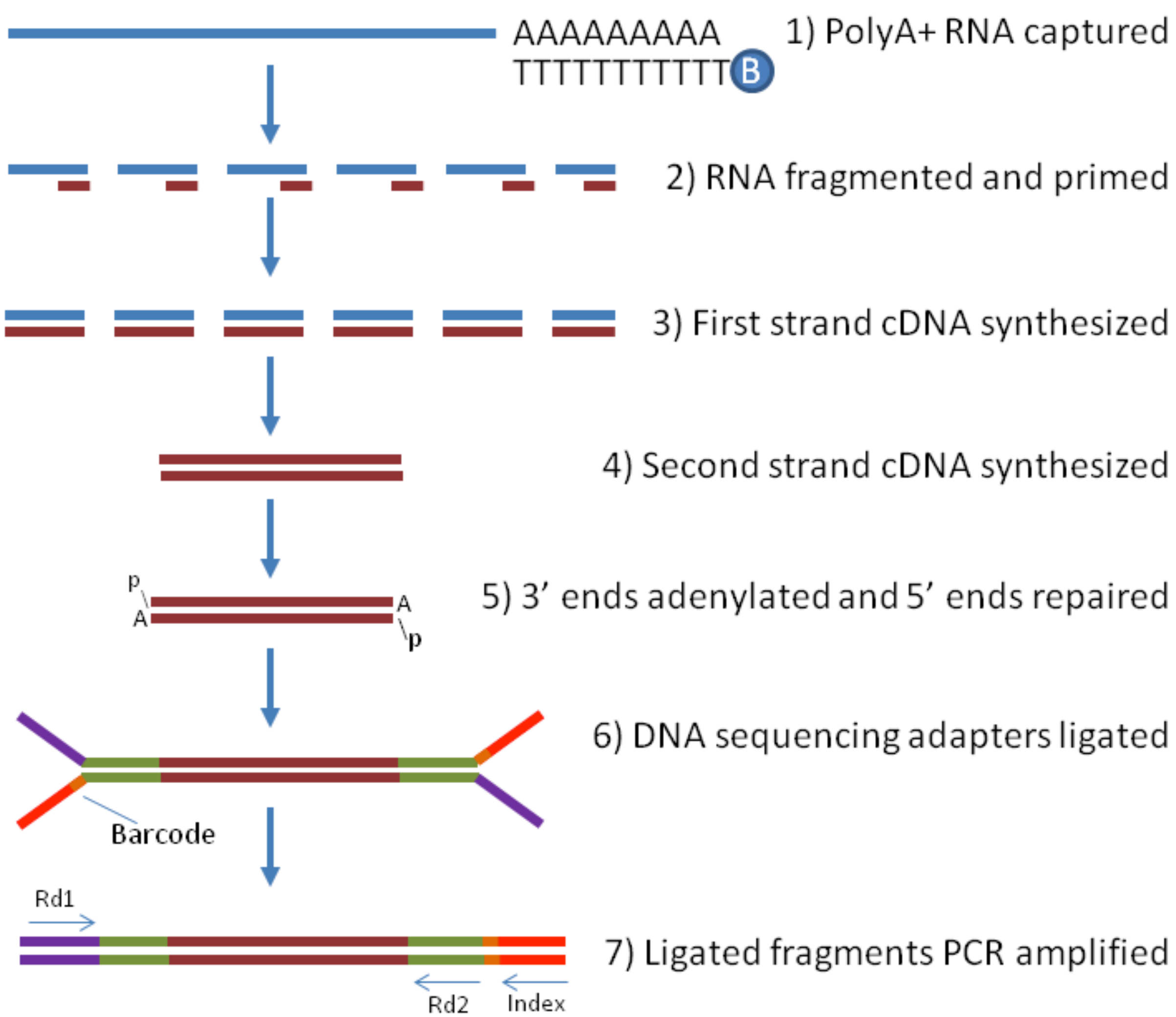
Quant seq 3'RNA-seq

- **Small RNA Sequencing**

Isolates and sequences small RNA species, such as microRNA

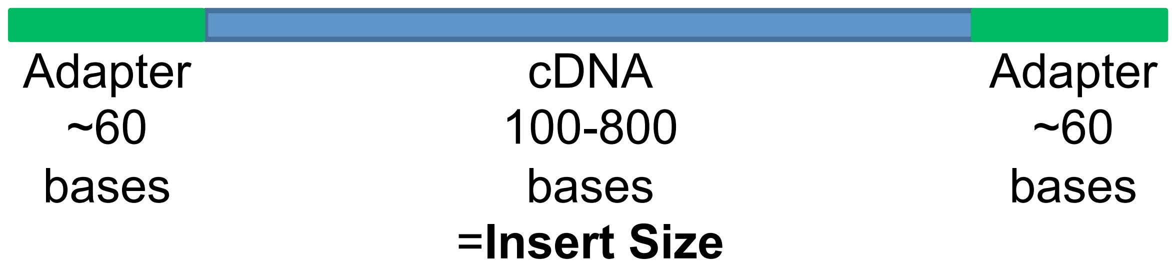
- **Single-Cell RNA-Seq** 10xGenomics, RNA-seq or 3' RNA-seq from single cells.
- Full length RNA sequencing
  - Iso-seq, Pacific Bioscience
  - Oxford nano-pore direct RNA sequencing

Creating mRNA-seq libraries

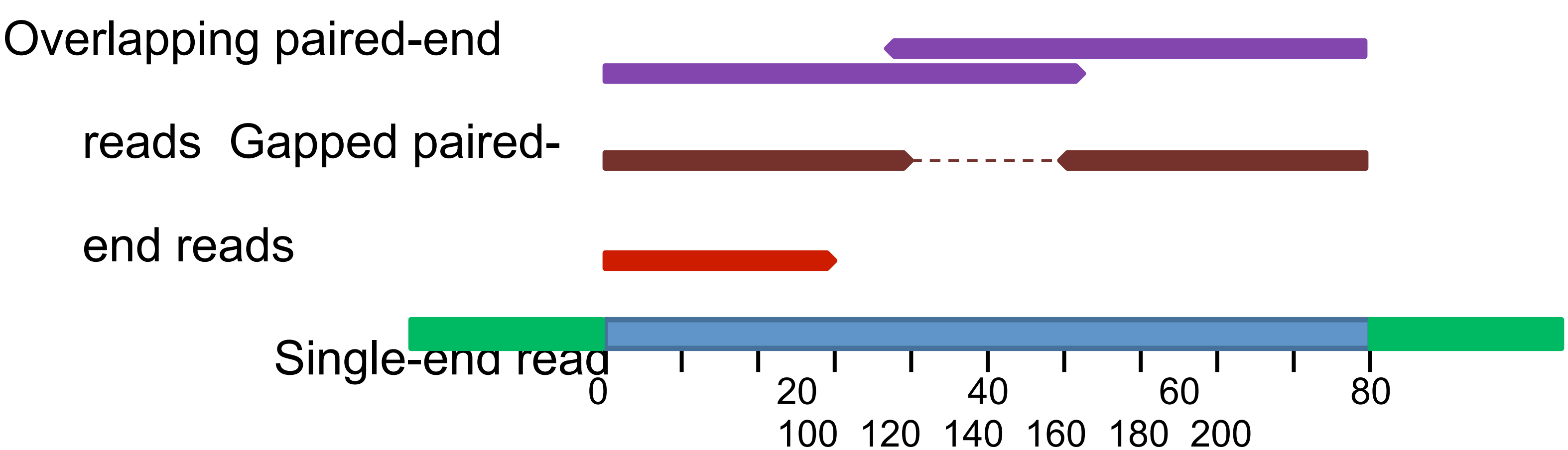


UMI- allow ID of unique transcripts verses duplicates . Barcodes allow multiplexing in 1 lane.

Creating mRNA-seq libraries



The “insert” is the cDNA (or RNA) ligated between the adapters. Typical insert size is 160-200 bases, but can be larger. Insert size distribution depends on library prep method.



Experimental considerations

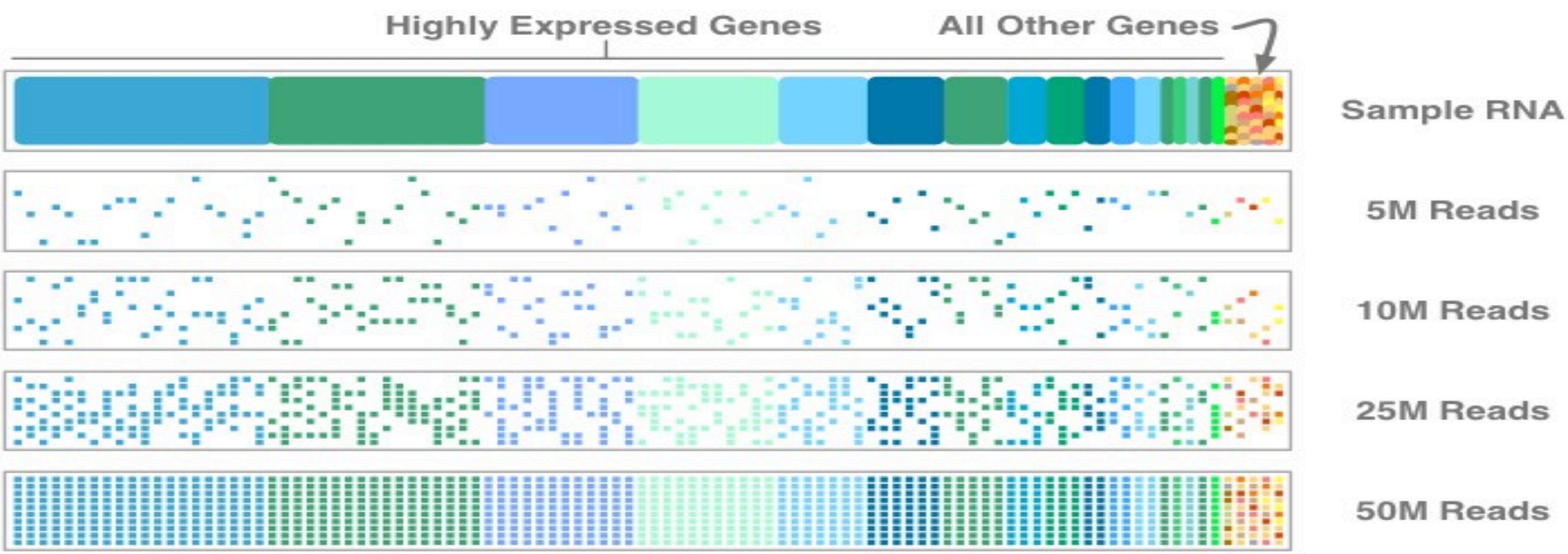
The same cDNA pool is sequenced several times: **technical replicates**.

- Account for variation in preparation
- Cost can be prohibitive
- Better to do more biological replicates

- Barcoding/pooling samples across multiple lanes
  - Batch effects, process all samples together
  - Pooling entire experiment and putting across multiple lanes removes lane effects but reduces depth. (providers often do this as mean you get data even if a lane fails!)

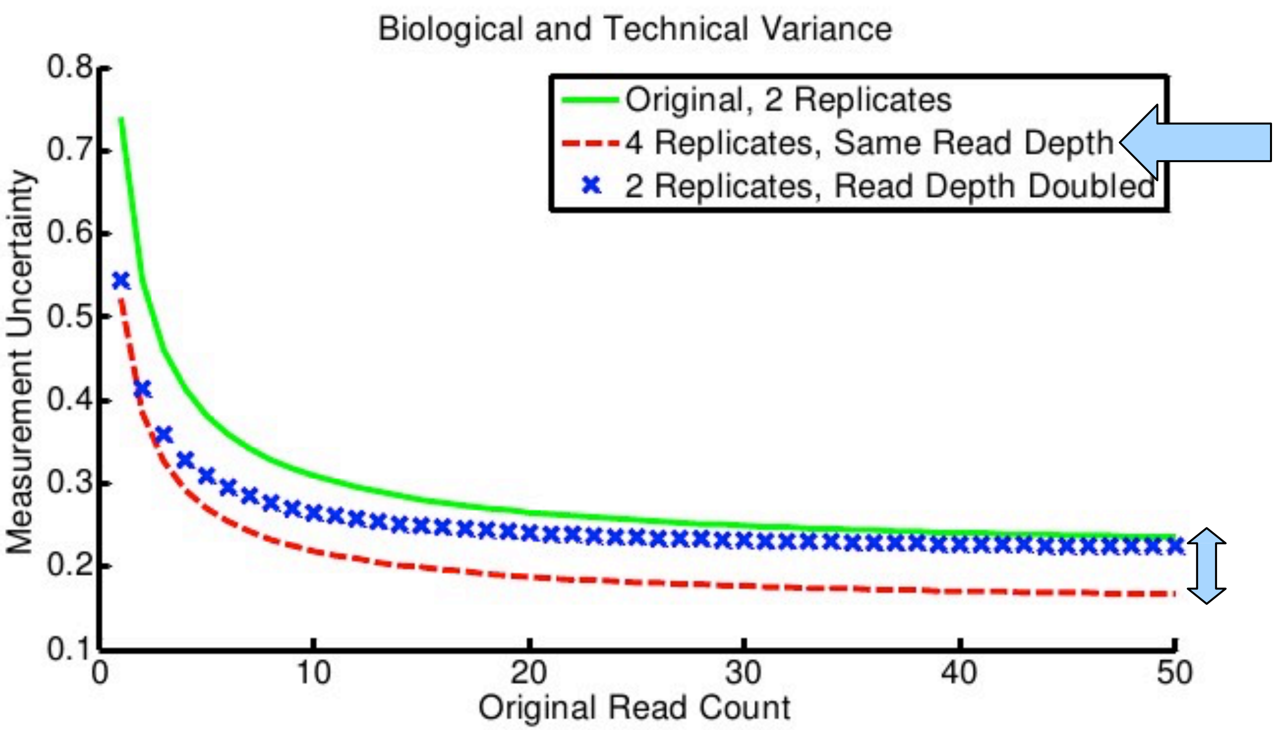
Sequencing depth

- Differential gene expression, reads/sample
  - Eukaryotes: 30+ million recommended
  - Polyploids need more wheat 60 million
  - Bacteria: 10+ million recommended
- More sequence is needed to detect rare transcripts



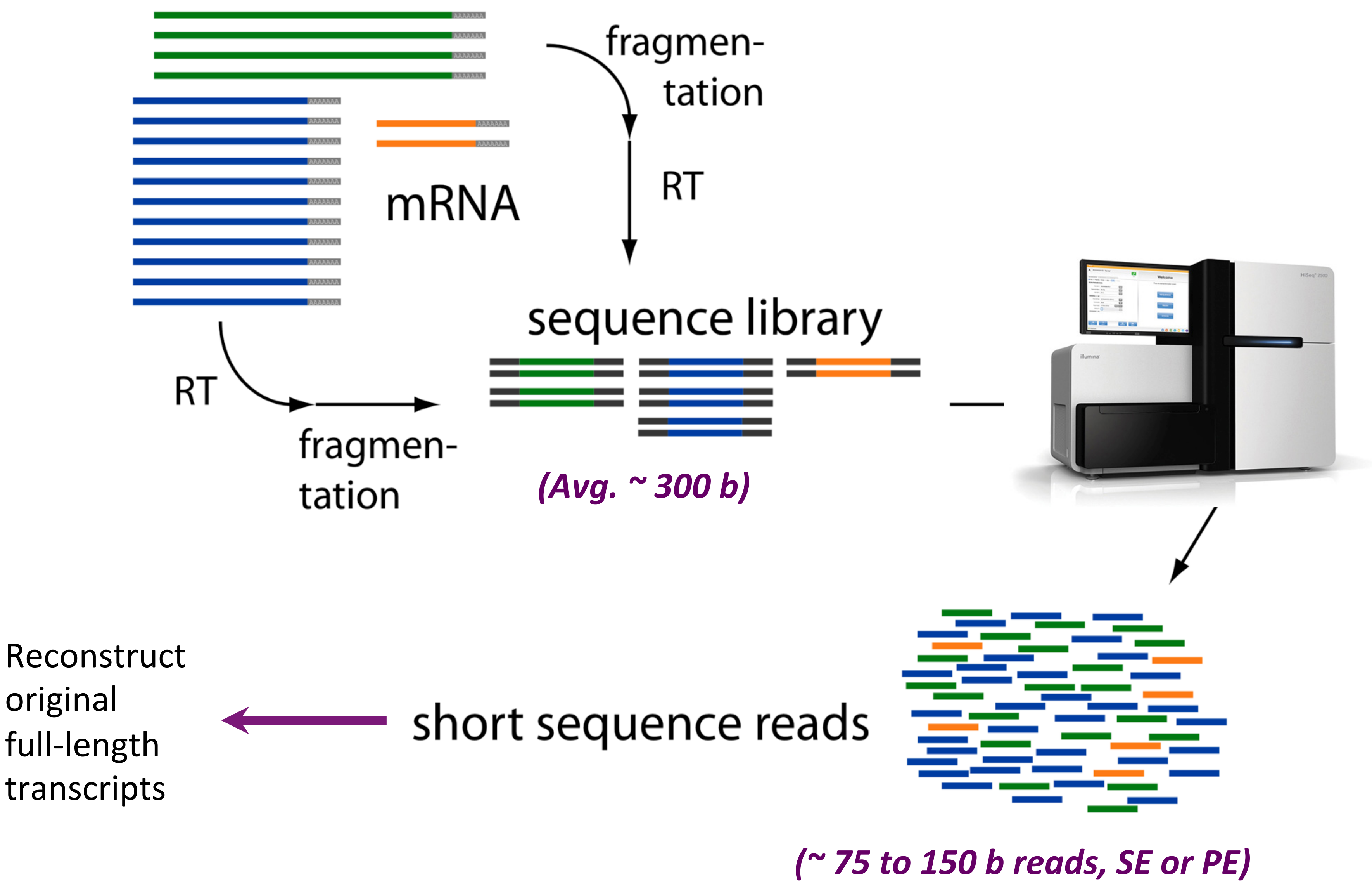
Increase sequencing depth, or replicates?

Variance will be lower with more reads: but sequencing **another biological replicate** is preferred over sequencing deeper, or technical reps.



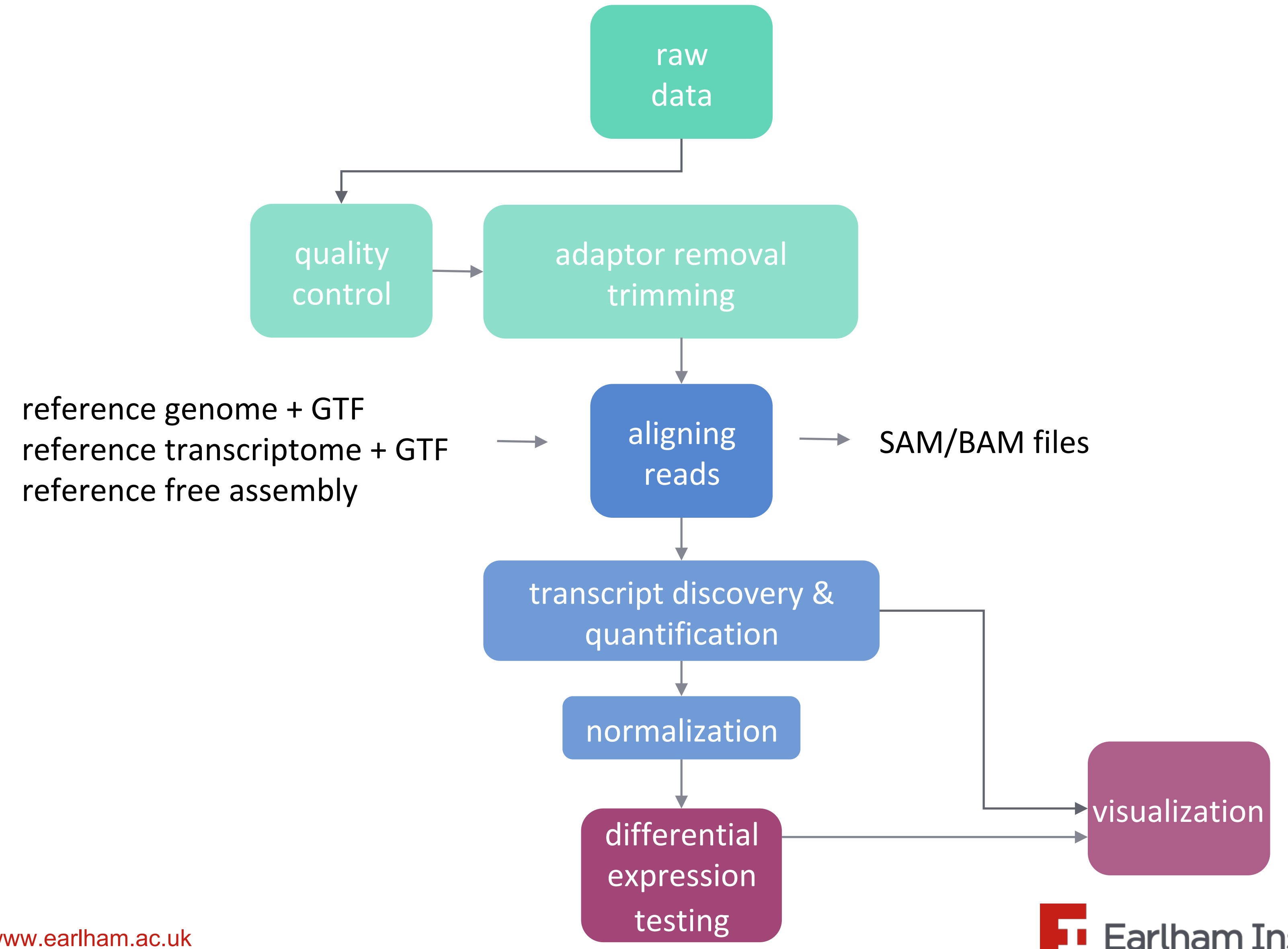
Busby et al. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. Doi: 10.1093/bioinformatics/btt015

RNA-seq challenge: Transcript Reconstruction

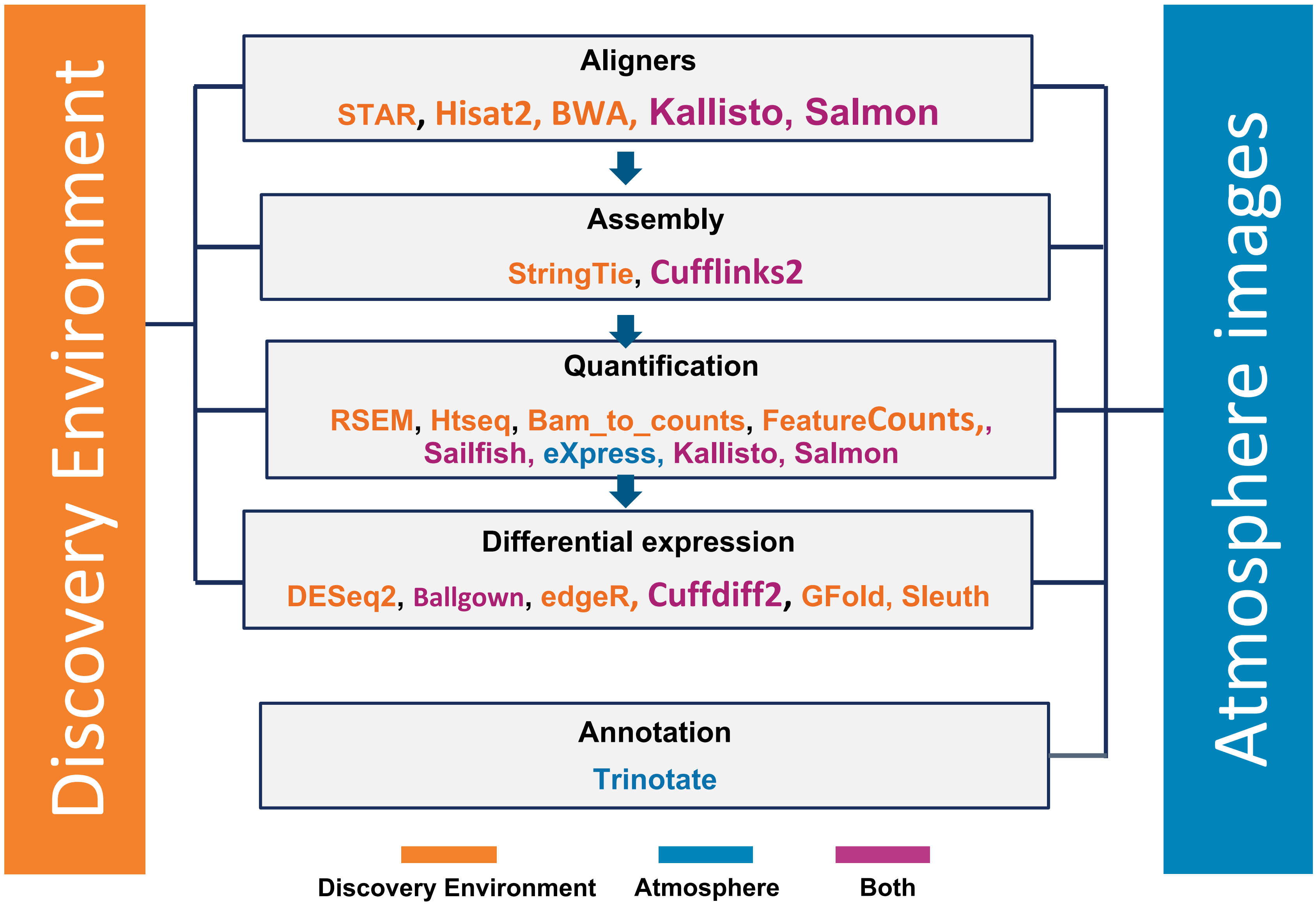


Adapted from: http://www2.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics.html

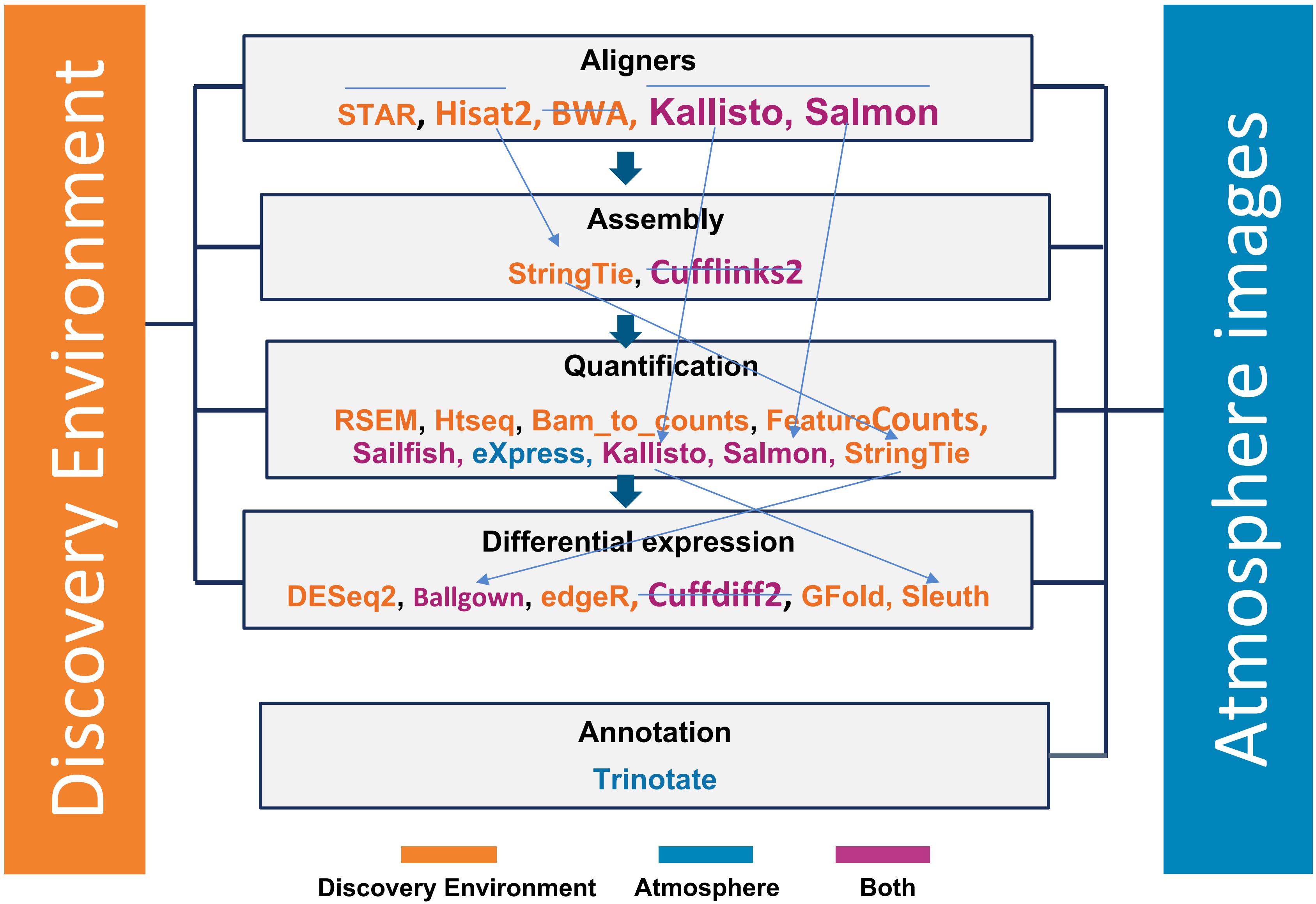
mRNA-seq analysis workflow



RNA-Seq 1 for Differential Expression



RNA-Seq 1 for Differential Expression



Common Data Formats

FASTA format:

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

FASTQ format:

@61DFRAAXX100204:1:100:10494:3070/1	
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT	Read
+	
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA	Quality values

Paired-end sequences

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

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

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

Quality control (QC) of reads

FastQC: A java based tool used to run simple quality control checks on raw sequence data.

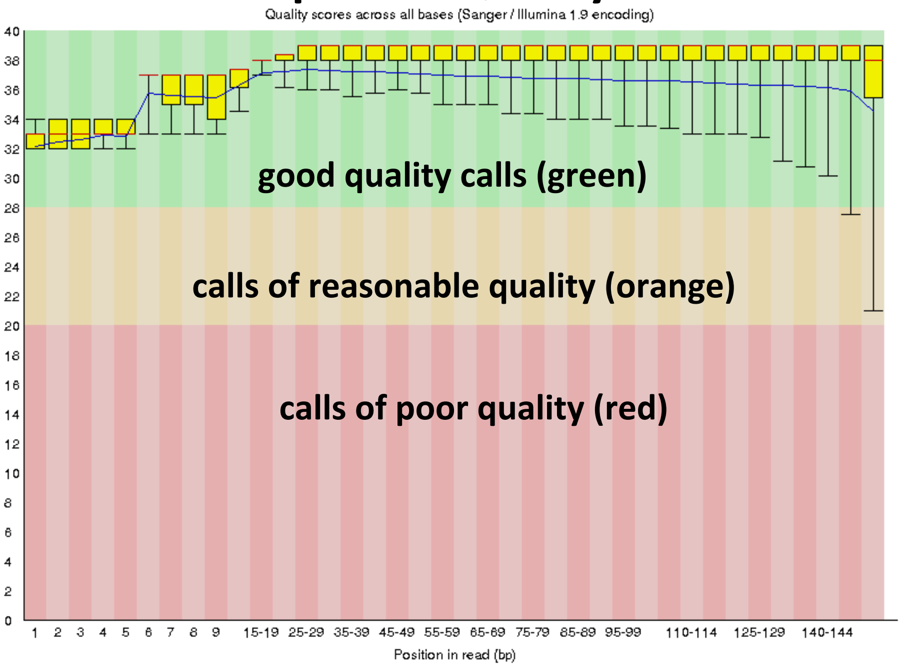
- Analysis is performed by a series of modules

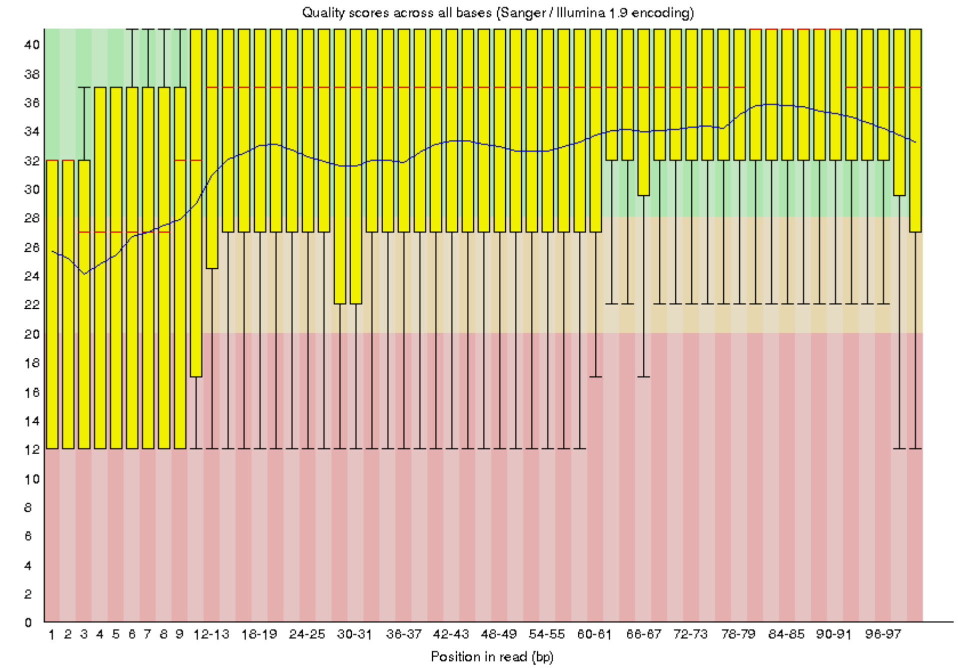
  Including: Per Base Sequence Quality and Content, GC content, Overrepresented sequences, Kmer content, Duplicate Sequences, N content
- Results are summarised graphically and are classed as ‘pass’ or ‘fail’.
- However experiments may produce libraries biased in certain ways, and this should be taken into account when interpreting the results.

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FastQC example

Per Base Sequence Quality





- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality
- The y-axis on the graph shows the quality scores.

  The higher the score the better the base call.

  The quality of calls on most platforms will degrade as the run progresses

  It is common to see base calls falling into the orange area towards the end of a read

Quality control (QC) of reads

Trimmomatic: A java based tool that performs a variety of trimming tasks of FASTQ data for Illumina PE and SE data.

- Trims adaptor sequences
- Sliding window trim: cutting once av. quality falls below threshold
- Cuts bases from start or end of read, depending on threshold quality
- Crops a specified number of bases from start of the read
- Crops read to a specified length
- Discards reads below a certain length

Sequence contamination: Align to chloroplast genome, rRNA, tRNA, ncRNA to identify % contamination of each sample.

GFF/GTF files

The GFF (General Feature Format) format consists of one line per feature, each containing 9 columns of data (fields).

The GTF (General Transfer Format) is identical to GFF version 2.

Fields must be tab-separated and all fields must contain a value; "empty" fields should be denoted with a '.'.

chr12	unknown	exon	4382902	4383401	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4383207	4383401	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	start_codon	4383207	4383209	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4385171	4385386	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4385171	4385386	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4387926	4388085	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4387926	4388085	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4398008	4398156	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4398008	4398156	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4409026	4409172	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4409026	4414522	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	stop_codon	4409173	4409175	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";

The left columns list source, feature type, and genomic coordinates

The right column includes attributes, including gene ID, etc.

Fields in the GTF files

chr12 unknown CDS 3677872 3678014 . + 2 gene_id "PRMT8"; gene_name "PRMT8"; p_id "P10933"; transcript_id "NM_019854"; tss_id "TSS4368";

Sequence Name (i.e., chromosome, scaffold, etc.) **chr12**

Source (program that generated the gtf file or feature) **unknow**

Feature (i.e., gene, exon, CDS, start codon, stop codon) **n CDS**

Start (starting location on sequence) **3677872**

End (end position on sequence) **3678014**

Score **.**

Strand (+ or -) **+**

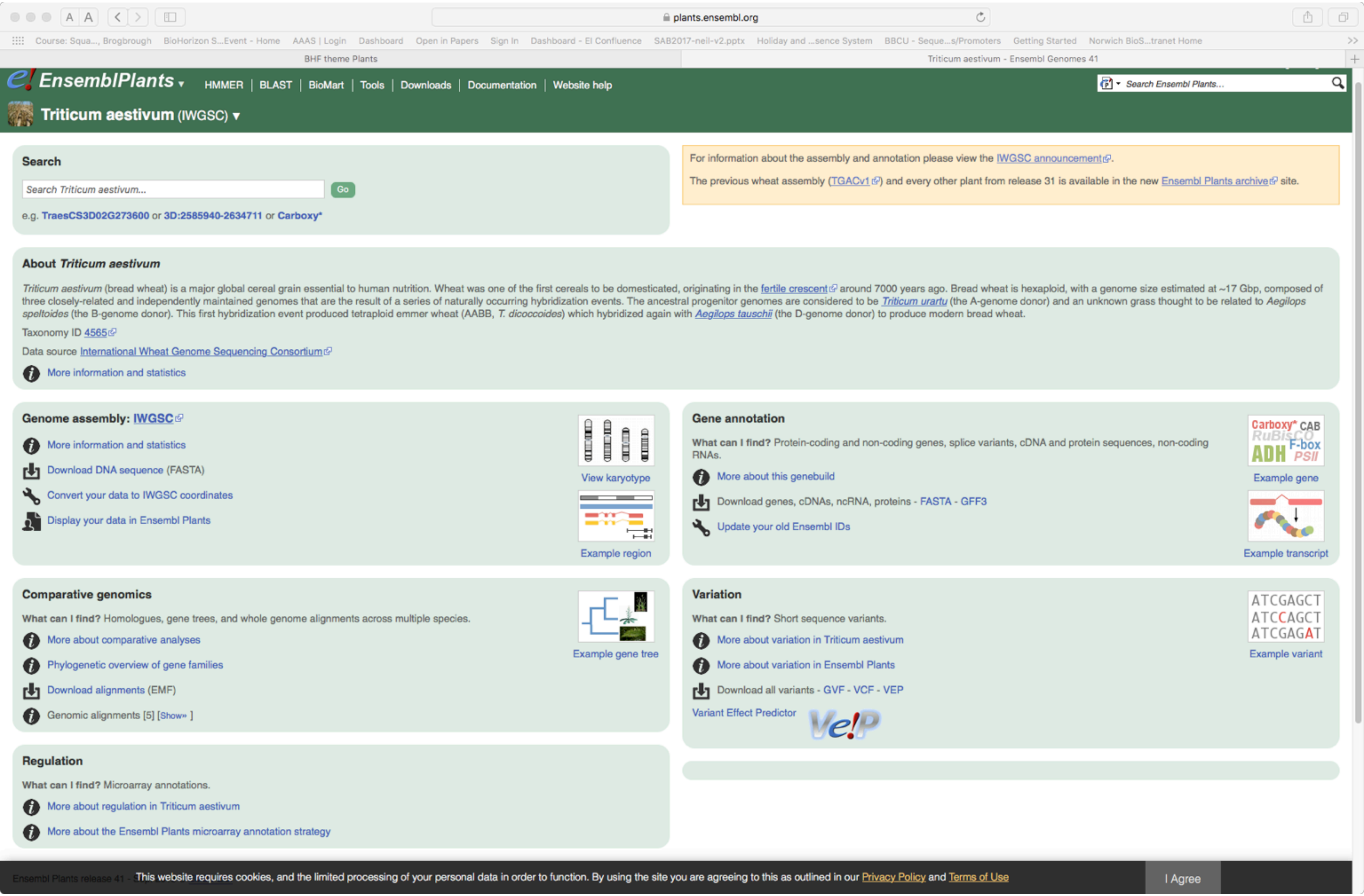
Frame (0, 1, or 2: which is first base in codon, zero-based) **2**

Attribute (";"-delimited list of tags with additional info)

This attribute provides info to HISAT2/ StringTie

gene_id "PRMT8"; gene_name "PRMT8"; p_id "P10933"; transcript_id "NM_019854"; tss_id "TSS4368";

Locating a reference genome (Ensembl)



Reference genome

What a reference looks like:

- Normally a compressed file or set of files
- Check the README file!
- May exist as scaffolds, chromosomes, or complete genome.
- FASTA format

```
>1A dna:chromosome chromosome:IWGSC:1A:1:594102056:1 REF
CTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCCTAACCCTAAACCCTAACCAAAAC
CCTAAACCCTAAACCCCTAAACCCTAAACCCTAACCTAAACCTACCTAAACCATACATTG
TTCCATTGGCCAGAGGCTGTTCACCTTGGAGACCTGATGCGGTTATGAGTACGACCGGGC
GTGAACGGTACTCATTTCGTGGAATGACCTAGGCAAGGGAAGTCGGCAAAACGGATCCGT
AACTTCGGGAAAAGGATTGGCTCTGAGGACTGGGCTCGGGGGTCGCGGATTGCTCGAGCT
GCTCACGCGGCGAGAGCGGGTCGCCGCGTGCCGGCCGGGGGACGGACCGGGAATCGCCCC
TTCGGGGGCTTTCCTCGAGCATGAAACAATCGACTCAGAACTGGTACGGACAAGGGGAAT
CCGACTATTTAATTAAAATAAAGCATTGCGATGGTCCTCGCGGATGCTGACGCAATGTGA
TTTCTGCCCAGTGCTCTGAATGTCAAAGTGAAGAAATTCAACCAAGCGCGGGTAAACGAC
```

Paired with relevant GFF for annotation

When there is no reference genome….

Post QC and read trimming carry out *de novo* transcriptome assembly using RNA-seq reads.

Assemblers: Velvet/Oases, Trinity

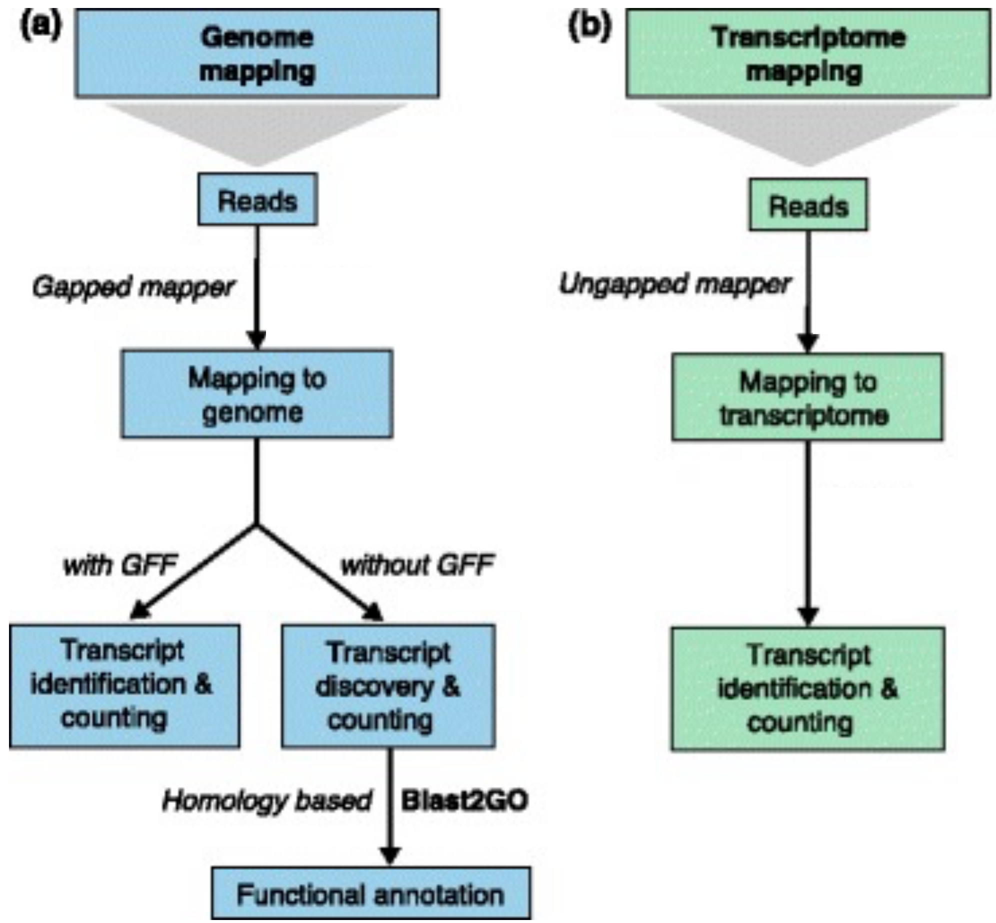
Can merge multiple assemblies to develop single consensus gene model: TACO

Align reads to consensus transcriptome

Genome vs. Transcriptome

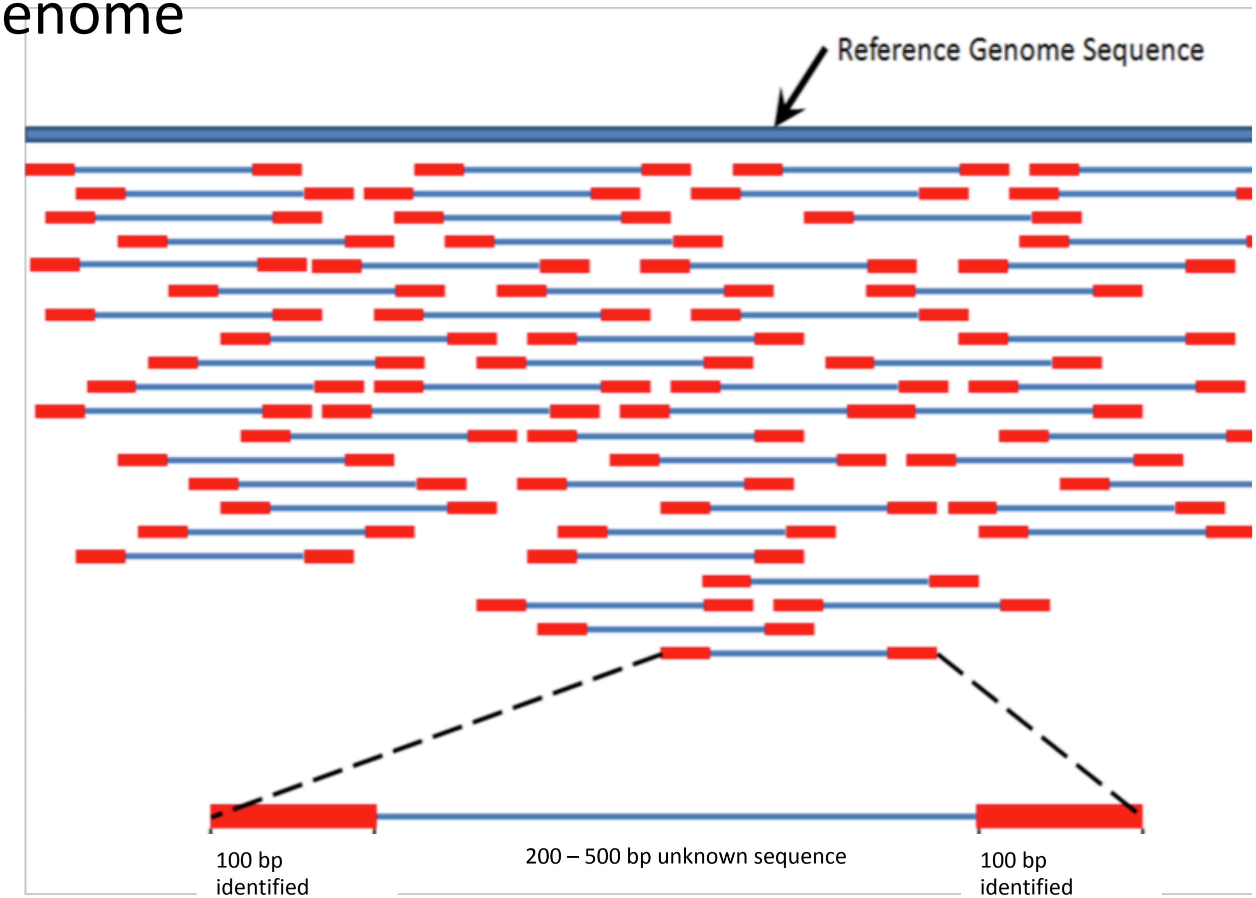
HISAT2

kallisto



Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17,** 13 (2016).

“Mapping reads to the reference” is finding where their sequence occurs in the genome



Source: Wikimedia, file:Mapping Reads.png

Read Mappers

BWA, Bowtie2, HISAT2

Are based on the Burrows-Wheeler Transformation (BWT)

- BWT: special sorting of all letters in the text (sequence)
- Similar suffixes (word ends) will be close to each other
- Easier to compress
- Good for approximate string matching (sequence alignment)
- Index (FM index) for finding the locations of matched strings (sequences) in the genome

Kallisto

- Index is created by splitting the transcriptome into k-mers (default 31bp). Constructing a transcriptome deBruijn graph of all the k-mers present in the genome.

Mapping reads to the genome/transcriptome

Requirements:

- An index corresponding to a reference genome sequence (fasta), or a reference transcriptome (fasta)
- A genome annotation file (GFF3 or GTF) for guiding the mapping.
- The trimmed/cleaned (pre-processed) reads (fastq).

Mapping reads to the genome

Transcript Reconstruction from RNA-Seq Reads

- A substantial proportion of reads will span 2 exons
- Where the anchors matching the neighbouring exons are small, some alignment tools will have difficulty

- HISAT2 creates 2 indices:

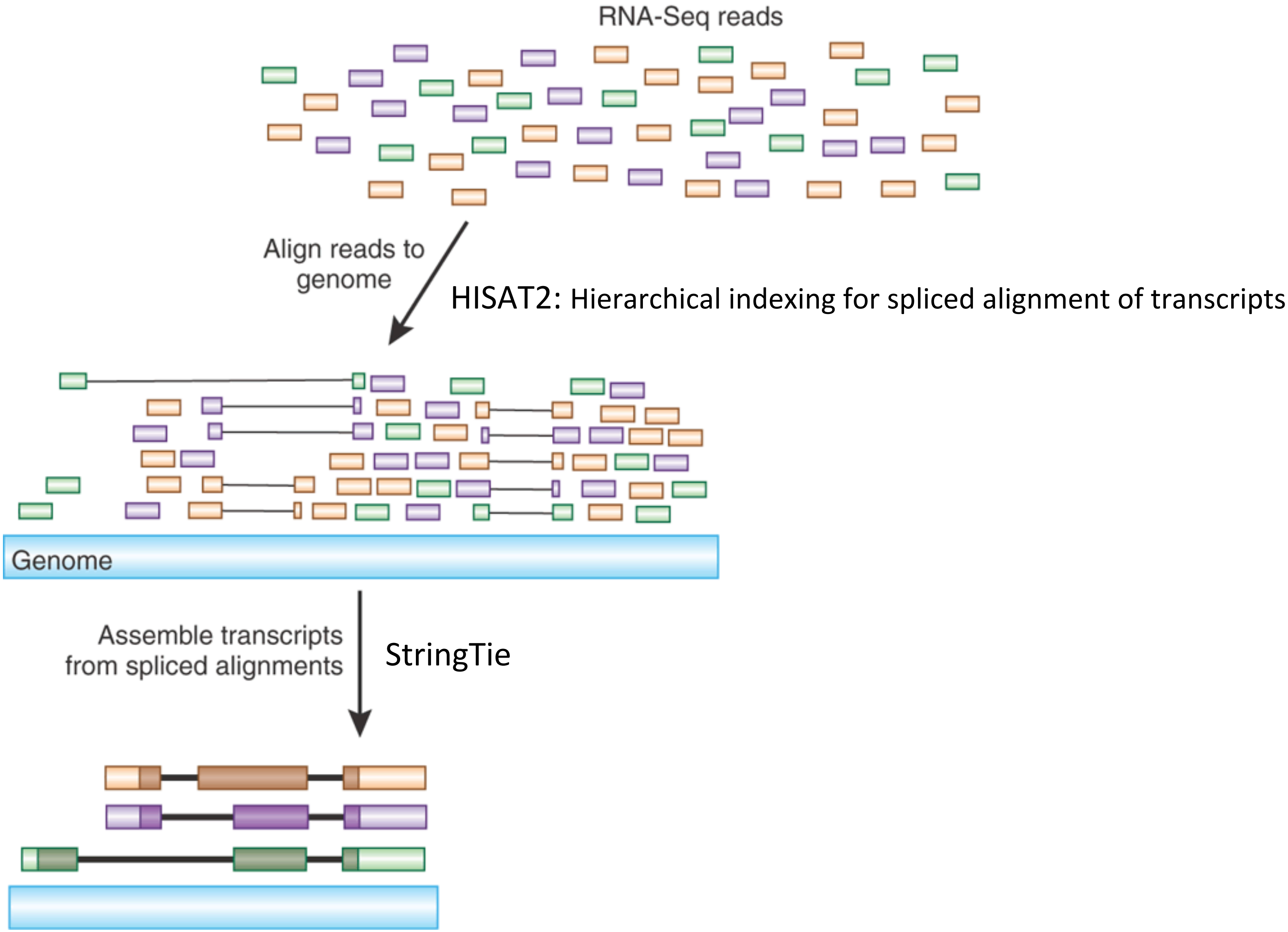
  a global index that represents the whole genome

  numerous small local indices that collectively cover whole genome

- Maps larger part of read to relevant local index then searches relevant local index rather than whole genome

- Uses splice site interpretation to do this.

Transcript Reconstruction from RNA-Seq Reads



HISAT output- PE

```
2074611 reads; of these:
  2074611 (100.00%) were paired; of these:
    640169 (30.86%) aligned concordantly 0 times
    965644 (46.55%) aligned concordantly exactly 1 time
    468798 (22.60%) aligned concordantly >1 times
    ----
    640169 pairs aligned concordantly 0 times; of these:
      60347 (9.43%) aligned discordantly 1 time – Errors or complicated
    ----
    579822 pairs aligned 0 times concordantly or discordantly; of these:
      1159644 mates make up the pairs; of these:
        1052478 (90.76%) aligned 0 times
        27580 (2.38%) aligned exactly 1 time
        79586 (6.86%) aligned >1 times
74.63% overall alignment rate
```

remaining pairs

Chloroplast?

- **Concordant**: read pair align on the same chromosome/contig, correct orientation and appropriate distance.
- **Discordant**: relative orientation not F/R, or distance too great.

How well do the reads align to reference?

Multi-mapping reads

- Some reads will align to more than one place in the reference, because:
  - Common domains, duplication, gene families
  - Paralogs, pseudogenes, etc.
  - Shared exons (if reference is transcriptome)

- This can distort counts, leading to misleading expression levels
- If a read can't be uniquely mapped, how should it be counted?
- Should it be ignored (not counted at all)?
- Should it be randomly assigned to one location among all the locations to which it aligns equally well?

This may depend on the question you're asking……and also depends on the software you use.

Transcript Reconstruction from RNA-Seq Reads

StringTie
Assembles genes from each sample separately, estimating expression levels of each gene/isoform

First groups into distinct gene loci then assembles each locus into as many isoforms as needed to explain the data

Uses gff to guide assembly

StringTie Output is a set of linked tables containing expression data, gene names and coordinates

Extras
Stringtie –merge
Looks at all the samples and the genes/isoforms identifies and creates a consistent set of transcripts across all samples.

Mapping reads to the transcriptome

kallisto

- Pseudo aligner, quasi-mapping
- Map to transcripts, not genome
- Does transcript quantifications (or gene)
- FAST and can run on most laptops
- Reports suggest differences between ‘traditional’ mappers are in the low abundance genes.

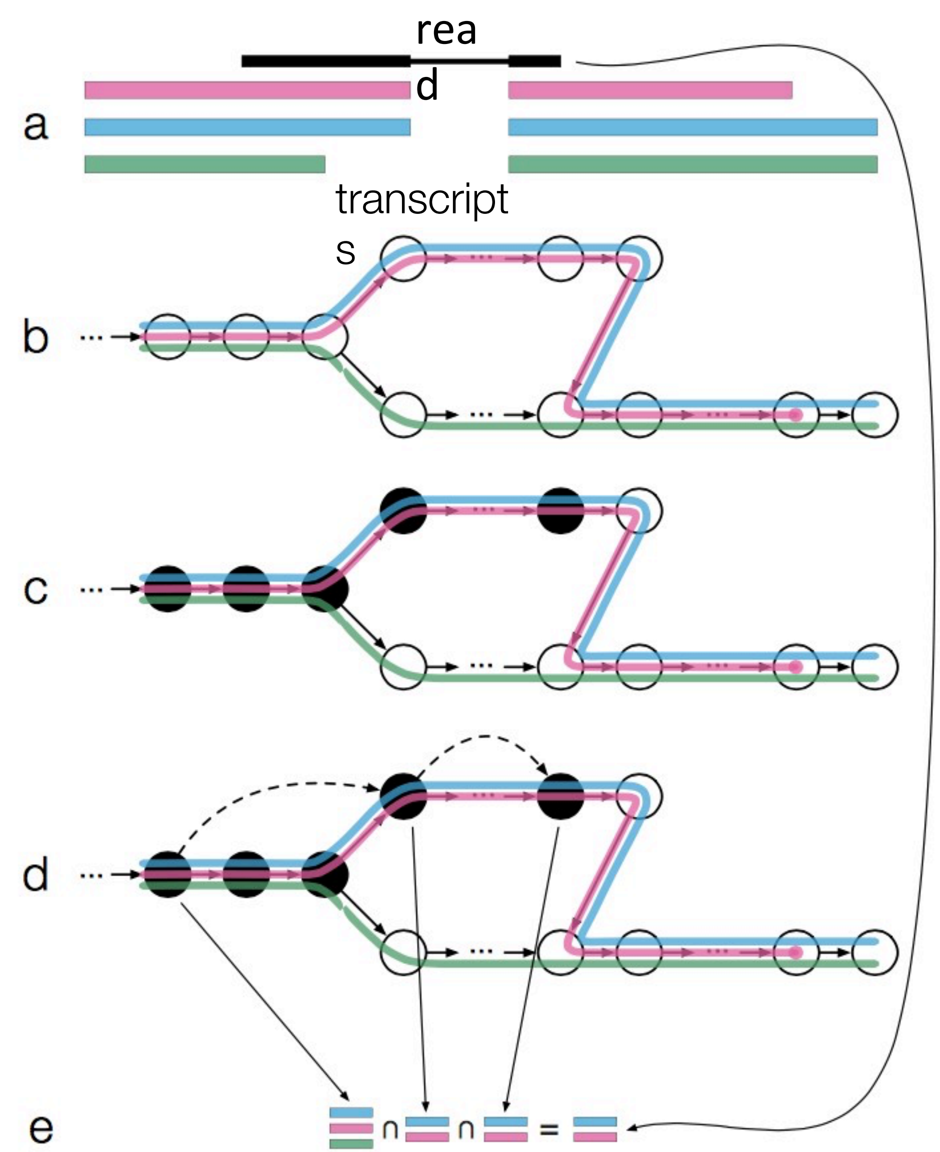
- Focuses on only identifying transcripts from which reads could have originated, rather than pinpointing exactly how they align.
- The k-mers of the reads are hashed to find a compatibility class with the k-mers of the transcriptome. k-mers with the same compatibility class can be skipped.

Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).

Which transcripts are the read k-mers *compatible with?*

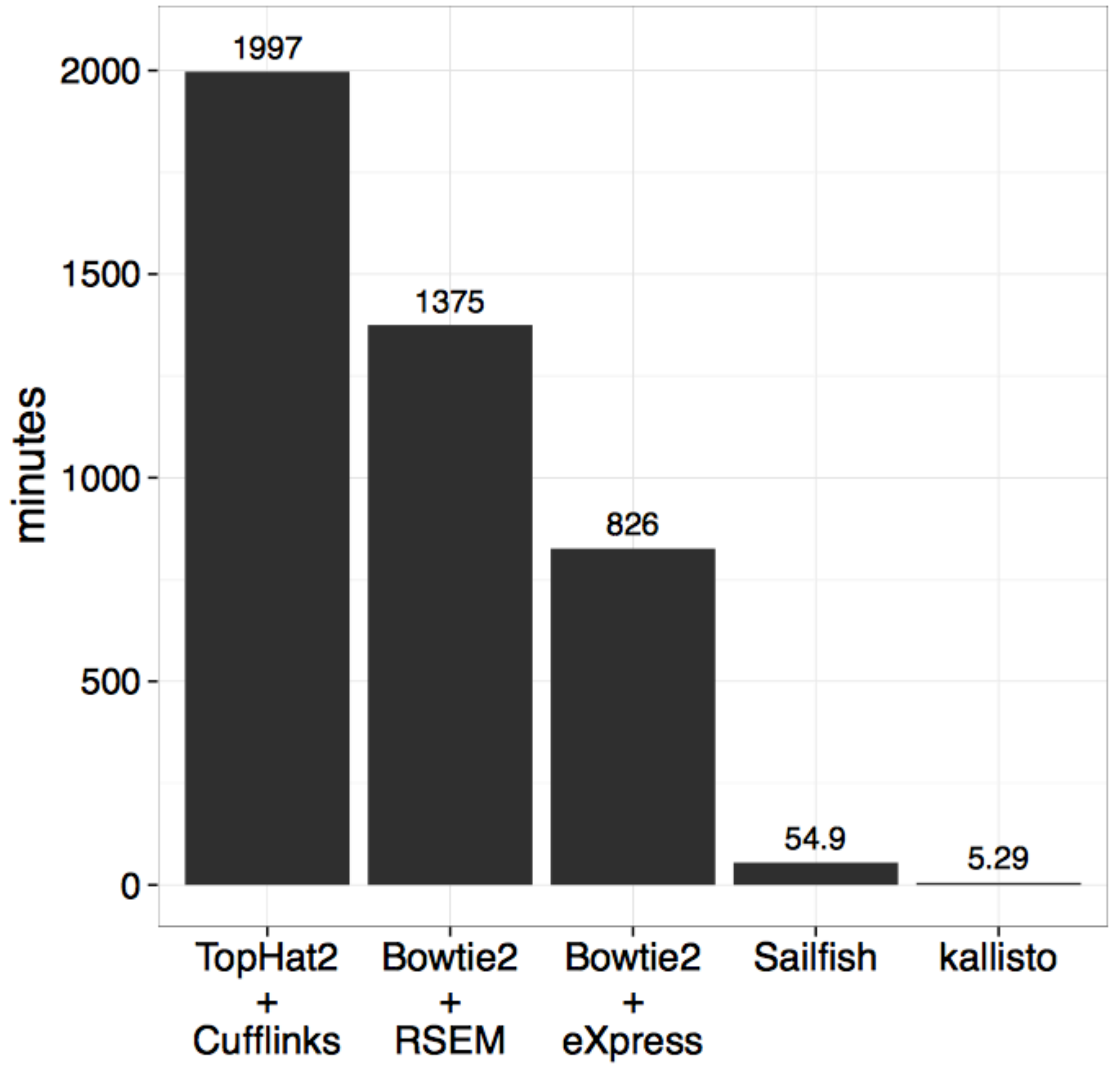
- Construct a de-Bruijn graph from the transcriptome (t-DBG)
- Comparison of reads to transcript is done using the t-DBG).

  Evaluate k-mers of reads for compatibility with the t-DBG; can skip over k-mers where compatibility doesn't change



Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).

kallisto



Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).

kallisto

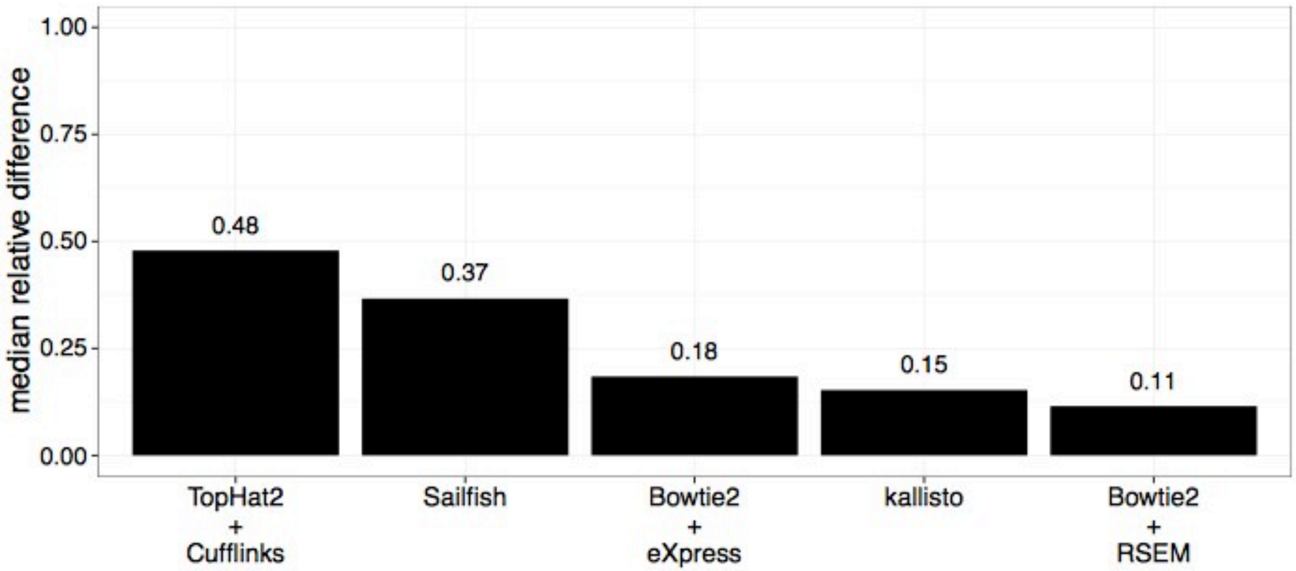


Figure 2a: Accuracy of kallisto, Cufflinks, Sailfish, eXpress and RSEM on 20 RSEM simulations of 30 million 75bp paired-end reads based on the abundances and error profile of Geuvadis sample NA12716 (selected for its depth of sequencing). For each simulation we report the accuracy as the median relative difference in the estimated read count of each transcript. Estimated counts were used to separate between the assignment of ambiguous reads and the estimation of effective lengths of transcripts. The values reported are means across the 20 simulations (the variance was too small for this plot). Relative difference is defined as the absolute difference between the estimated abundance and the ground truth divided by the average of the two.

Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).

Expression Counts, FPKM and TPM

To normalize for sequencing depth and gene length:

RPKM (reads per kilobase million) SE
Sum number reads in sample and divide by 1M (= per M scaling factor)
Divide read counts by 'per M scaling factor' (normalises for depth, gives RPM)
Divide RPM value by length of gene in kb (= RPKM)

FPKM (fragments per kilobase million) PE
Takes into account two reads can correspond to a single fragment, and therefore will not count such fragments twice.

TPM (transcripts per million)
Divide read counts by length of each gene in kb (= reads per kilobase RPK)
Sum all RPK values in a sample and divide by 1M (= per M scaling factor)
Divide RPK values by per M scaling factor (= TPM)

Because the sum of TPMs in each sample will always add up to the same number you can compare across samples, unlike RPKM or FPKM

Normalisation between samples

Name	Version	Normalization
baySeq	2.4.1	Scaling factors (quantile/ TMM/ total)
DESeq	1.22.1	DESeq size factors
EBSeq	1.12.0	DESeq median normalization
edgeR	3.12.1	TMM/ Upper quartile / RLE / None (all scaling factors are set to be one)
limma+voom	3.26.9	TMM
NOIseq	2.14.1	RPKM / TMM / Upper quartile
SAMseq (samr)	2.0	Based on the read count mean over the null features of data set.
DESeq2	1.10.1	DESeq size factors
sleuth	0.29.0	DESeq size factors (with slight modifications)

Costa-Silva, J., Domingues, D. & Lopes, F. M. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One* **12,** 1–18 (2017).

Differential expression analysis

- Differential expression analysis means taking the normalized sequencing fragment count data and performing statistical analysis to discover quantitative changes in gene expression levels between experimental groups. PAIRWISE COMPARISONS

- For example, we use statistical testing to decide whether, for a given gene, an observed difference in fragment counts between groups is significant, that is, whether it is greater than what would be expected just do to random biological variation.

Filtering reads

- Most common filter is to remove genes that are less than x normalized read counts across a certain number of samples.
- A second less used filter is for genes with minimum variance across all samples, so if a gene isn’t changing (constant expression) its inherently not interesting, therefore no need to test. NOT AWAYS RIGHT

Statistical models for count data

- The number of reads that are mapped onto a gene was first modelled using a Poisson distribution.
- Poisson distribution appears when things are counted.
- It assumes that mean and variance are the same.

- However biological variability of RNA-seq count data cannot be captured using the Poisson distribution because data shows overdispersion (ie. variance of counts larger than mean)
- Negative Binomial (NB) distribution takes into account overdispersion; hence, it has been used to model RNA-seq data
- Poisson distribution has only one parameter λ, while NB is a two-parameter distribution λ and ϕ.

Modeling for count data for DE

Different packages use different statistical models, based on known probability distributions such as Binomial, Poisson, Negative Binomial, etc.

R packages in Bioconductor:
edgeR (Robinson et al., 2010): Exact test based on Negative Binomial distribution.
DESeq (Anders and Huber, 2010): Exact test based on Negative Binomial distribution.
DEGseq (Wang et al., 2010): MA-plots based methods (MATR and MARS), assuming Normal distribution for M|A.
baySeq (Hardcastle et al., 2010): Estimation of the posterior likelihood of differential expression (or more complex hypotheses) via empirical Bayesian methods using Poisson or NB distributions.

Differential expression analysis tools compared……

Software packages for detecting differential expression

Method	Version	Reference	Normalization[a]	Read count distribution assumption	Differential expression test
edgeR	3.0.8	[4]	TMM/Upper quartile/RLE (DESeq-like)/None (all scaling factors are set to be one)	Negative binomial distribution	Exact test
DESeq	1.10.1	[5]	DESeq sizeFactors	Negative binomial distribution	Exact test
baySeq	1.12.0	[6]	Scaling factors (quantile/TMM/total)	Negative binomial distribution	Assesses the posterior probabilities of models for differentially and non-differentially expressed genes via empirical Bayesian methods and then compares these posterior likelihoods
NOIseq	1.1.4	[7]	RPKM/TMM/Upper quartile	Nonparametric method	Contrasts fold changes and absolute differences within a condition to determine the null distribution and then compares the observed differences to this null
SAMseq (samr)	2.0	[8]	SAMseq specialized method based on the mean read count over the null features of the data set	Nonparametric method	Wilcoxon rank statistic and a resampling strategy
Limma	3.14.4	[9]	TMM	voom transformation of counts	Empirical Bayes method
Cuffdiff 2 (Cufflinks)	2.0.2-beta	[10]	Geometric (DESeq-like)/quartile/classic-fpkm	Beta negative binomial distribution	*t*-test
EBSeq	1.1.7	[11]	DESeq median normalization	Negative binomial distribution	Evaluates the posterior probability of differentially and non-differentially expressed entities (genes or isoforms) via empirical Bayesian methods

[a]In case of availability of several normalization methods, the default one is underlined.

Seyednasrollah, F., Laiho, A. & Elo, L. L. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.* **16,** 59–70 (2013).

Typical output from DE analysis

	logFC	**logCPM**	**p-value**	**FDR**
TRINITY_DN876_c0_g1_i1	-7.15049572793027	10.6197708379285	0	0
TRINITY_DN6470_c0_g1_i1	-7.26777912190146	7.03987604865422	1.687485656951e-287	6.46813252309319e-284
TRINITY_DN5186_c0_g1_i1	-7.85623682454322	9.18570464327063	1.17049180235068e-278	2.99099671894011e-275
TRINITY_DN768_c0_g1_i1	7.72884741150304	9.7514619195169	4.32504881419265e-272	8.28895605240022e-269
TRINITY_DN70_c0_g1_i1	-12.7646078189688	7.86482982471445	3.92853491279431e-253	6.02322972829624e-250
TRINITY_DN1587_c0_g1_i1	-5.89392061881667	9.07366563894607	6.32919557933429e-243	8.08660221852944e-240
TRINITY_DN3236_c0_g1_i1	-7.27029815068473	8.02209568234202	3.64955175271959e-235	3.99678053376405e-232
TRINITY_DN4631_c0_g1_i1	-7.45310693639574	6.91664918183241	4.30540921272851e-229	4.1256583780971e-226
TRINITY_DN5082_c0_g5_i1	-5.33154406167545	10.6977538760467	2.74243356676259e-225	2.33594396920022e-222
TRINITY_DN1789_c0_g3_i1	10.2032564835076	7.32607652700285	1.44273728647186e-213	1.10600240380933e-210
TRINITY_DN4204_c0_g1_i1	4.81030233739325	9.88844409410644	9.27180216086162e-205	6.46160321501501e-202
TRINITY_DN799_c0_g1_i1	-4.22044475626154	6.9937398638711	1.24746518421083e-197	7.96922341846683e-195
TRINITY_DN196_c0_g2_i1	4.60597918494257	9.86878463857276	1.9819997623131e-192	1.16877001368402e-189
TRINITY_DN5041_c0_g1_i1	-4.27126549355785	9.70894399883	1.8930437900069e-185	1.03657669244235e-182
TRINITY_DN1619_c0_g1_i1	-4.47156415953777	9.22535948721718	1.76766063029526e-181	9.03392426122899e-179
TRINITY_DN899_c0_g1_i1	-4.90914328409143	7.93768691394594	1.11054513767547e-180	5.32089939088761e-178
TRINITY_DN324_c0_g2_i1	4.87160837667488	6.84850312231775	2.20092562166991e-179	9.92487989160089e-177
TRINITY_DN3241_c0_g1_i1	-4.77760618069256	7.94111259715689	1.60585457735621e-173	6.83915621667372e-171
TRINITY_DN4379_c0_g1_i1	3.85133572453294	7.23712813663389	3.48140532848425e-164	1.4046554341137e-161
TRINITY_DN1919_c0_g1_i1	4.05998814332136	6.95937301668582	1.8588621194715e-161	7.12501850393425e-159
TRINITY_DN2504_c0_g1_i1	-6.92417817059644	6.20370039359785	2.42022459856956e-160	8.83497227268296e-158

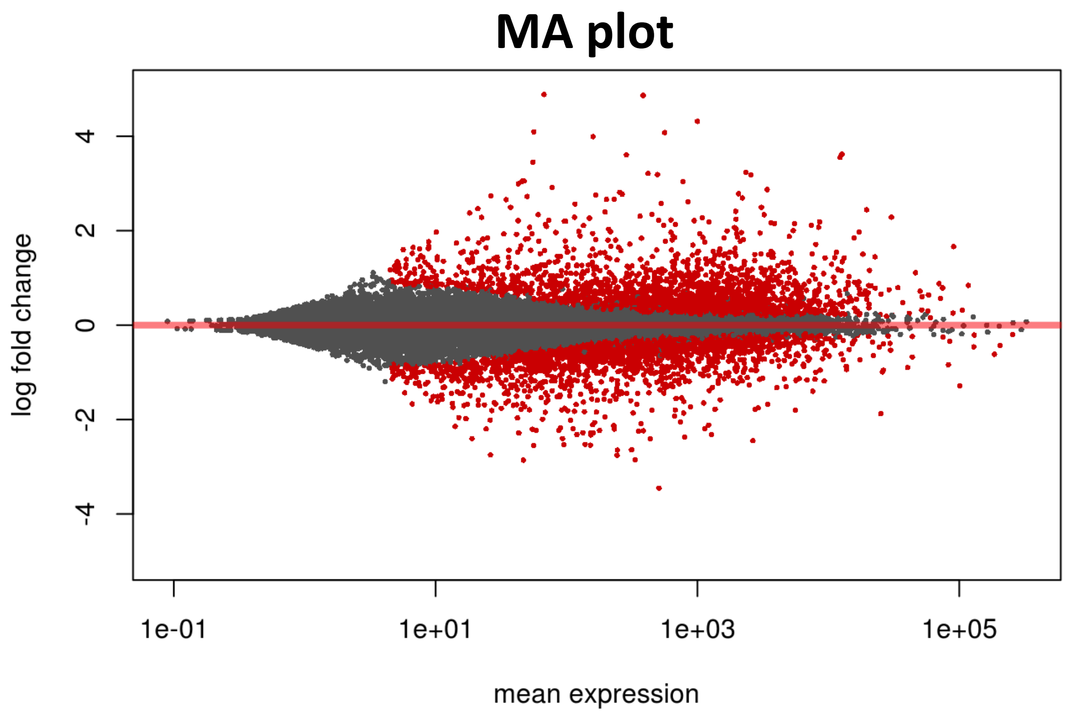
…

Up vs. Down regulated

Avg. expression level

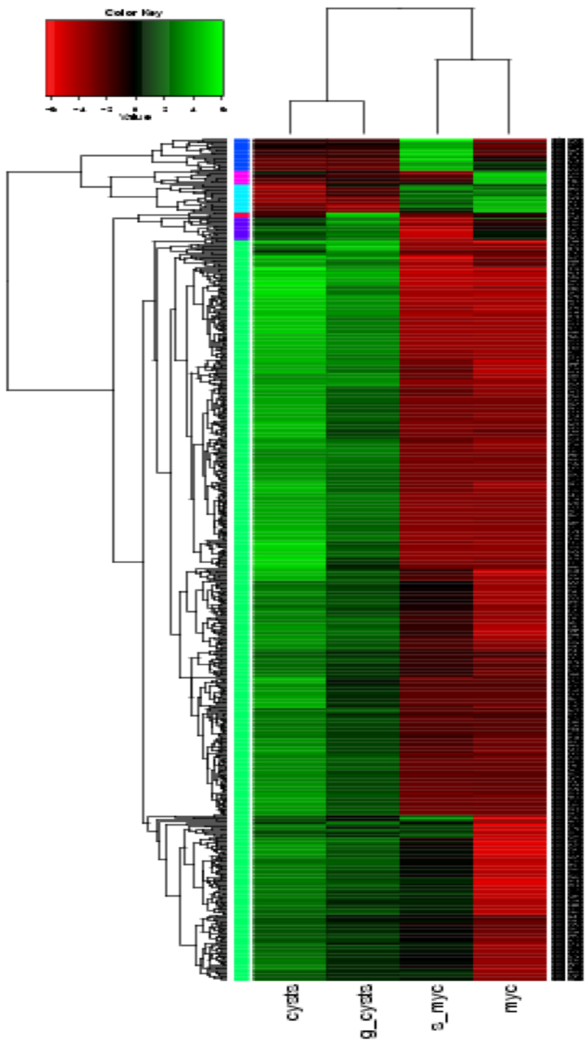
Significance

Plotting Pairwise Differential Expression Data



- Each gene is represented with a dot
- Genes with an adjusted p value below a threshold (here 0.1) shown in red
- This plot demonstrates that only genes with a large average normalized count contain sufficient information to yield a significant call

Comparing Multiple Samples



Heatmaps provide an effective tool
for navigating differential expression across
multiple samples.

Clustering can be performed across both axes:
-cluster transcripts with similar
expression
patters.
-cluster samples according to similar
expression values among transcripts.

Going Beyond Gene Lists

Additional analysis:

- GO (gene ontology) terms
- Identifying co-expressed genes and network inference

- Results of RNA-seq data can be integrated with other sources of biological data to establish a more complete picture of gene regulation. Such as:

  Genotyping data identify genetic loci responsible for variation in gene expression between individuals

  Epigenomic information (transcription factor binding, histone modification, methylation)

A Guide to Annotating Eukayotic Genomes

DAVID SWARBRECK

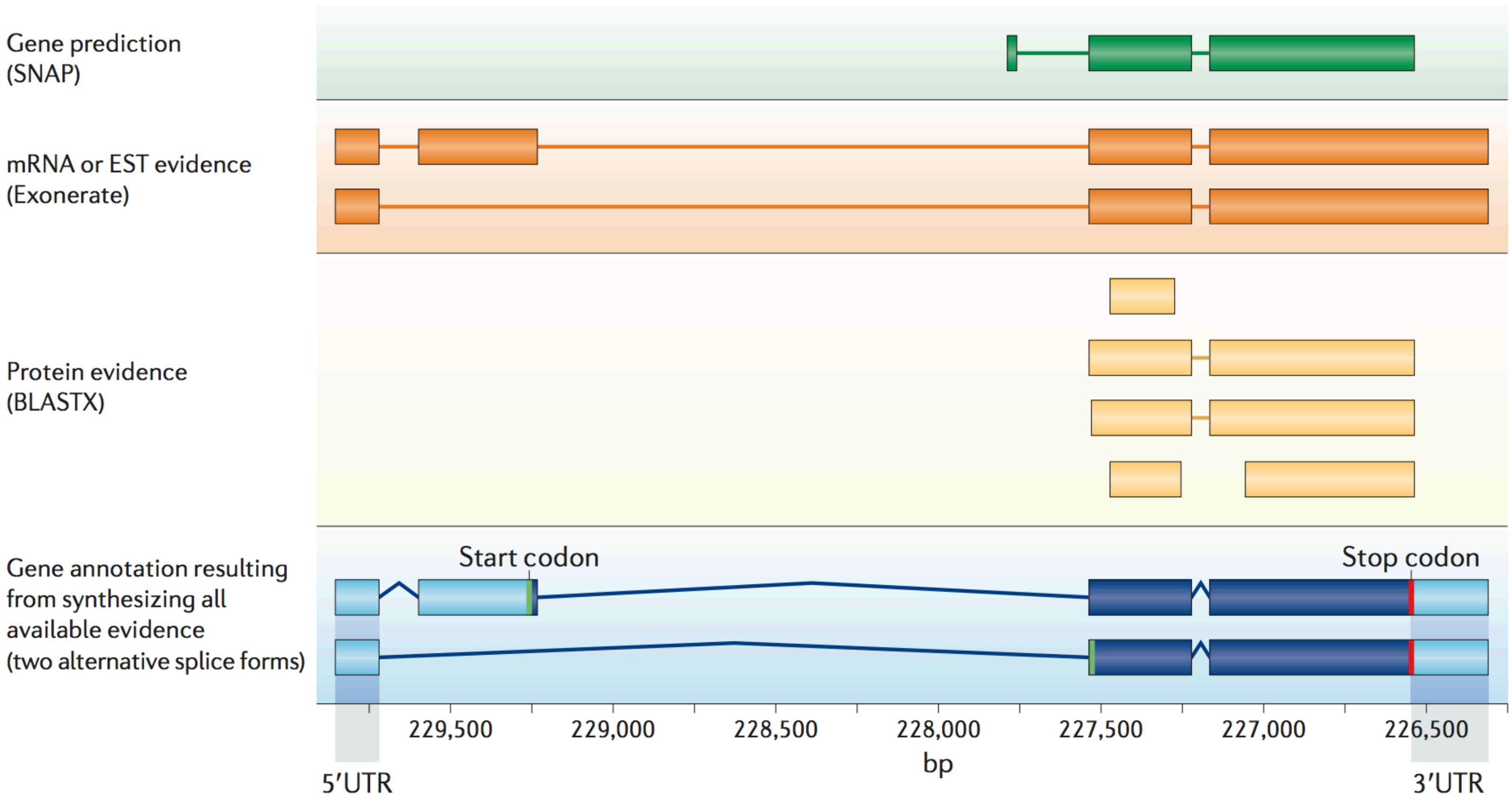
Head of Core Bioinformatics

Earlham Institute

Decoding Living Systems

What are annotations

- Annotations are descriptions of features of the genome
  - Structural: exons, introns, UTRs, CDS, stop, start, splice variants, regulatory seqs
  - Coding & non-coding genes
  - Repeats, transposons



Nat Rev Genet. 2012 Apr 18;13(5):329-42. doi: 10.1038/nrg3174.
A beginner's guide to eukaryotic genome annotation.
Yandell M1, Ence D.

Challenges in annotating eukaryotic genomes

- Genomes come in a variety of sizes **big** and **small**
- Can be highly repetitive, contain TE related genes
- Contain non functional genes (pseudogenes)
- Assembly errors and fragmentation
- Incomplete evidence
- Many different tools and choices
- Methods differ in accuracy
- Each method has it's own pros and cons

Alternative approaches to gene prediction / annotation

Ab initio

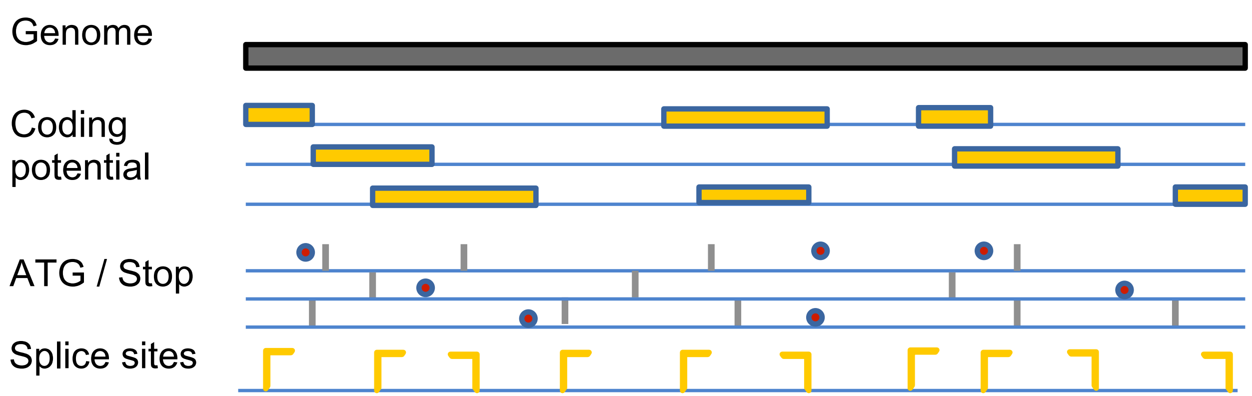
Identify genes based on intrinsic factors

- Require no external evidence
- Fast
- Require training
- Low transcript level accuracy

Ab initio gene prediction is based on gene content and signal detection

Signals e.g. splice sites, start / stop codons
Content e.g. statistical properties of coding sequence, intron / exon sizes

Gene finders use probabilistic models (e.g. HMMs) to combine signal and content measurements

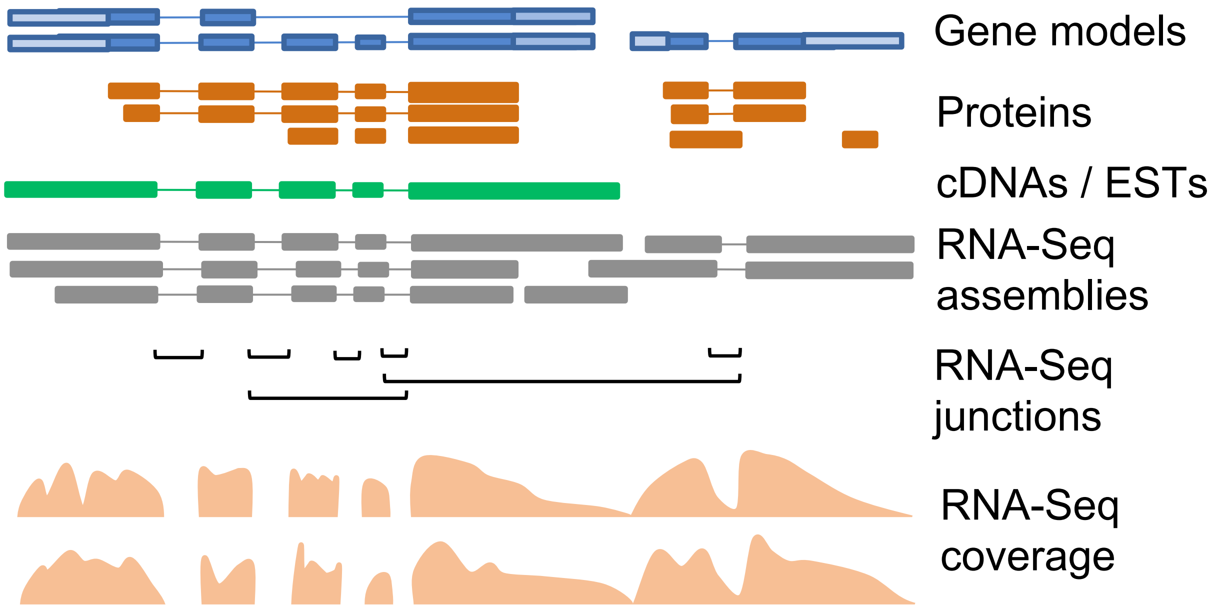


Alternative approaches to gene prediction

Alignment driven

Genes built based on aligned EST, cDNA, RNA-Seq or proteins

- Provides clear supporting evidence
- Less tolerant to lower quality data
- Dependent on available samples



- Gene models are as accurate as the alignments they are based on
- Different evidences can suggest different structures

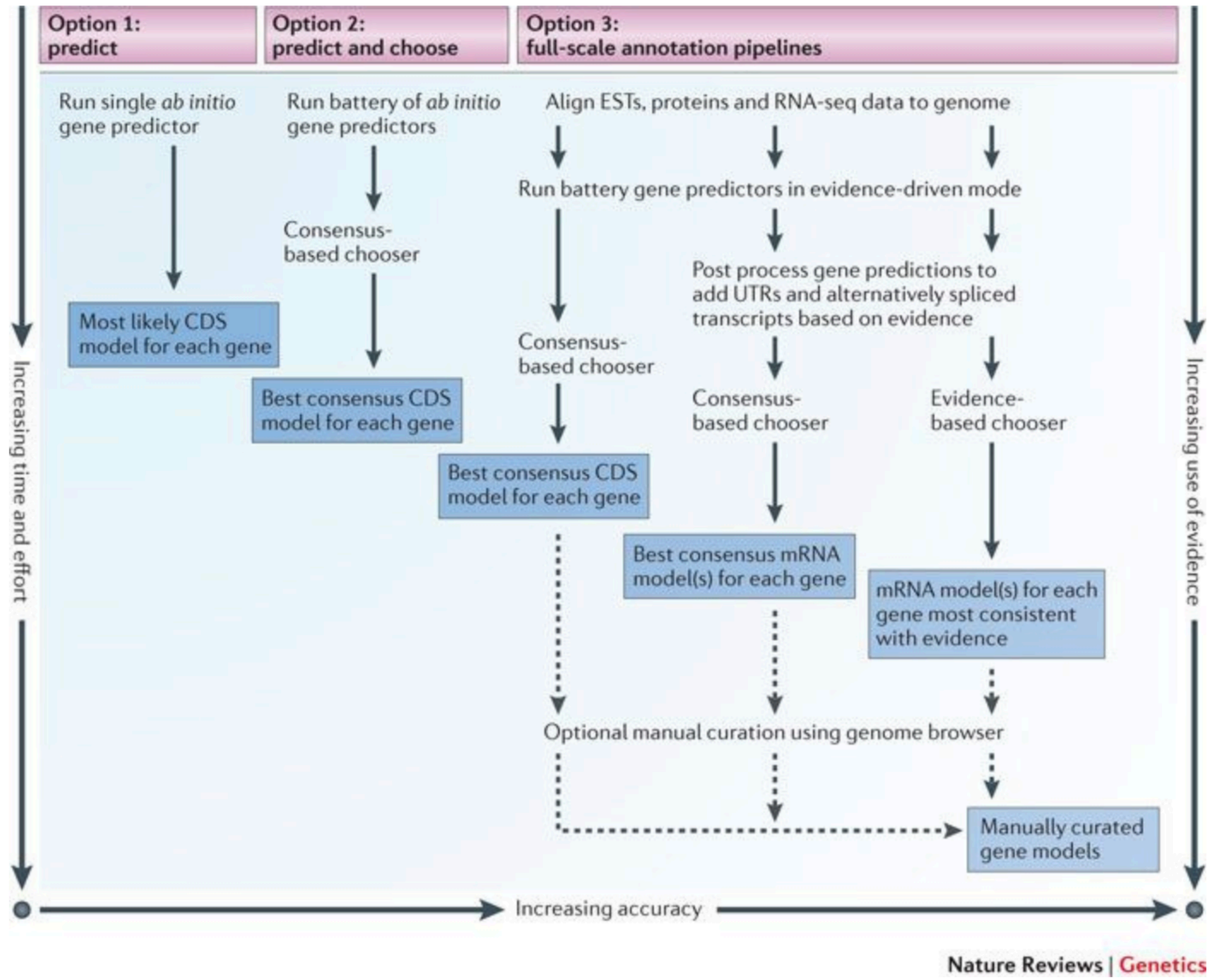
Alternative approaches to gene prediction

Ab initio	Alignment driven	Evidence guided
Identify genes based on intrinsic factors • Require no external evidence • Fast • Require training • Low transcript level accuracy	**Genes built based on aligned EST, cDNA, RNA-Seq or proteins** • Provides clear supporting evidence • Less tolerant to lower quality data • Dependent on available samples	**Use external evidence to improve the accuracy of their predictions** • Mix ab initio and alignment approaches • Can incorporate high and low quality data • More complex to run and require optimization

Which approach should I use ?

- Genome annotation is not a point and click exercise
- Balance between effort and accuracy

- Will depend on type of organism
- Available compute resources
- Purpose for generating the annotation.

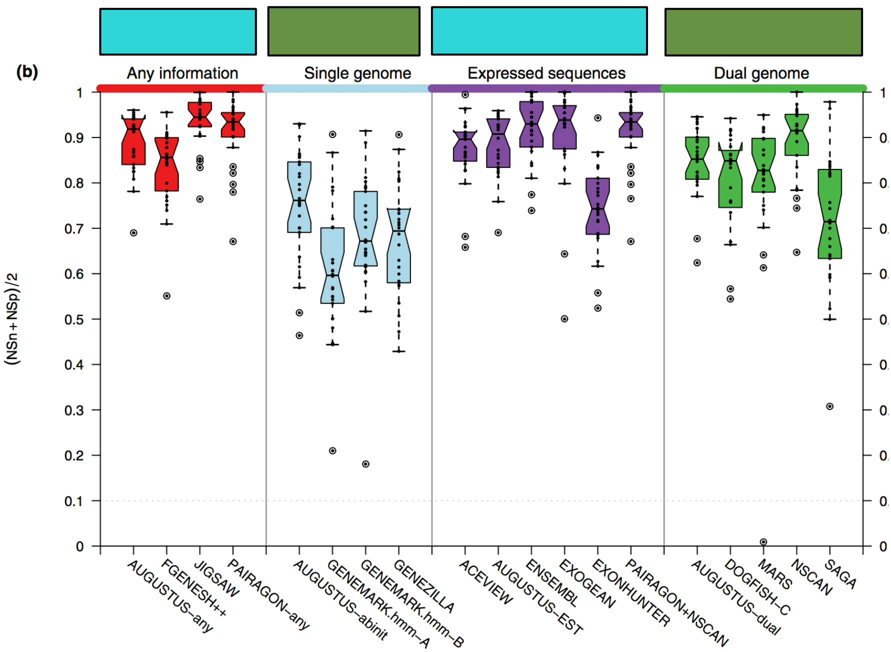


Nat Rev Genet. 2012 Apr 18;13(5):329-42. doi: 10.1038/nrg3174.
A beginner's guide to eukaryotic genome annotation.
Yandell M1, Ence D.

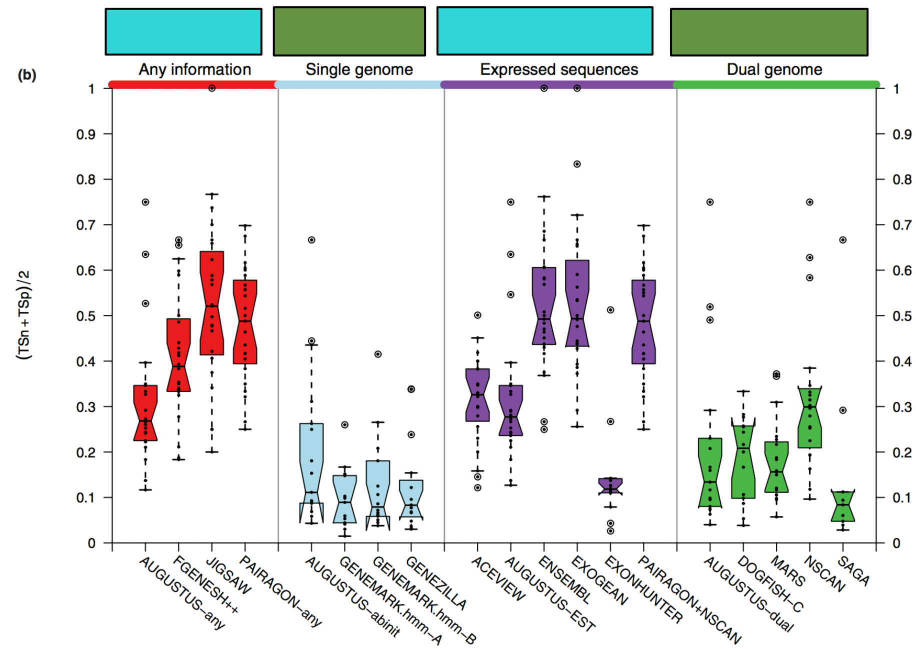
Which approach should I use ?

- Evidence based methods improve accuracy
- Gene finder performance varies
- Predicting fully correct transcripts remains challenging

Nucleotide level Accuracy



Transcript level Accuracy



Evidence based predictors

ab initio / genome based predictors

Guigó R, Flicek P, Abril JF, et al. EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol*. 2006;7 Suppl 1(Suppl 1):S2.1-31.

Earlham Institute

How to approach Eukaryotic Genome Annotation

- Annotation pipelines typically
  - incorporate both ab initio and evidence guided approaches
  - comprise of many distinct steps and incorporating a variety of evidence types

Typical components of evidence guided pipeline

Repeat Identification	**Evidence Alignment**	**Training *ab initio* gene predictors**
Genebuild	**Evidence assignment / confidence scores**	**Functional Annotation**

The choice of tools / parameters will have a substantial impact on the gene models that are generated.

What data do I need?

- RNA-Seq
  - Number of reads depends on transcriptome complexity
  - Aim to capture a wide range of tissues / conditions
  - Preferably use strand specific data
  - Long reads (cDNAs, Pacbio, nanopore)

- Cross species proteins
  - Level of similarity/ Identity to my species of interest
  - Quality of gene models

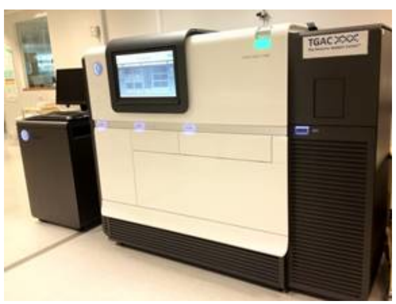
- Repeat library
  - Check public repositories e.g. repbase

RNA-Seq a key data type to aid gene annotation

Choice over what technology, aligner and assembler to use?

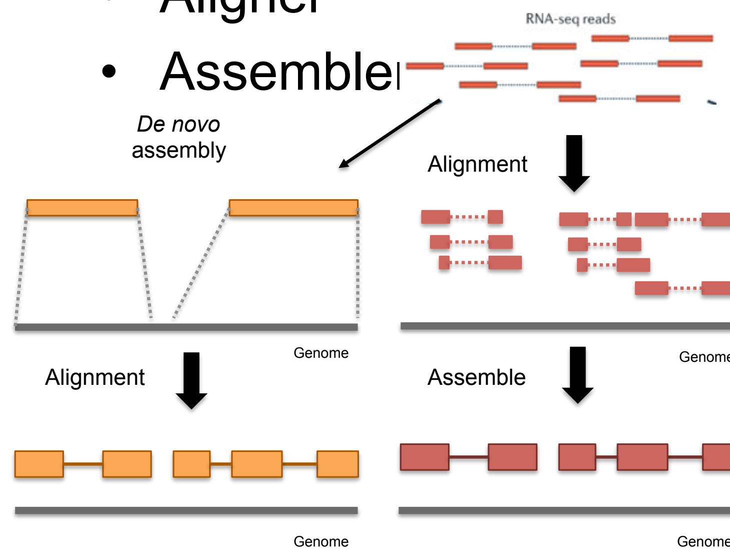
Technology / protocol

- Illumina HiSeq / PacBio
- Strand specific
- Poly A / ribodepleted
- Library insert size
- DSN normalization



Computational approach

- Aligner
- Assembler



Oases
TransAbyss
Trinity

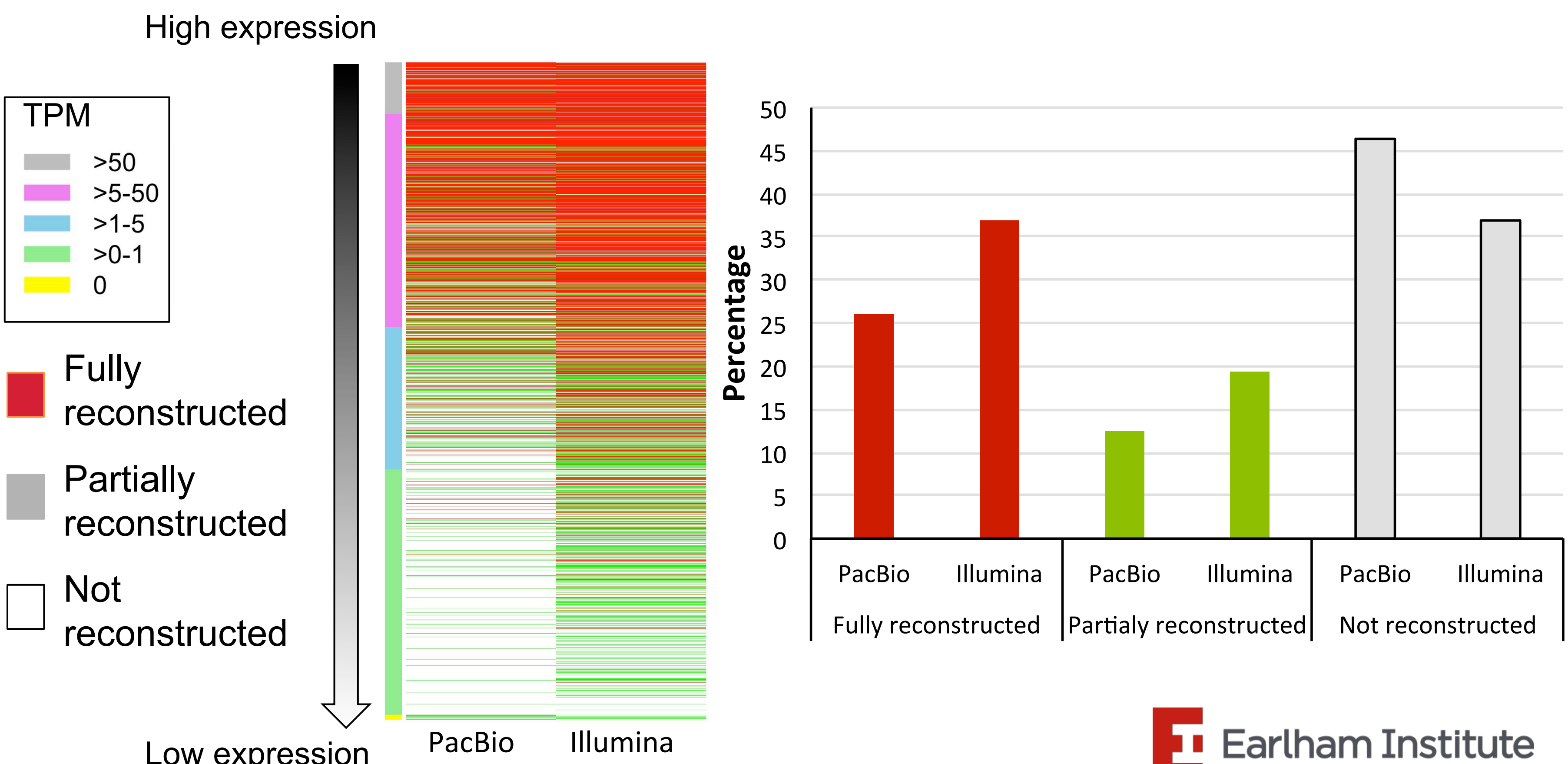
Cufflinks
Stringtie
CLASS
iReckon

These choices can have a substantial impact on the transcript assemblies generated and the gene models built using them.

PacBio and Illumina provide complementary approaches

Human reference genes reconstructed by Pacbio and Illumina assemblies

- A larger proportion of reference genes could be fully reconstructed by Illumina reads
- Illumina data allowed genes to be reconstructed across a larger expression range

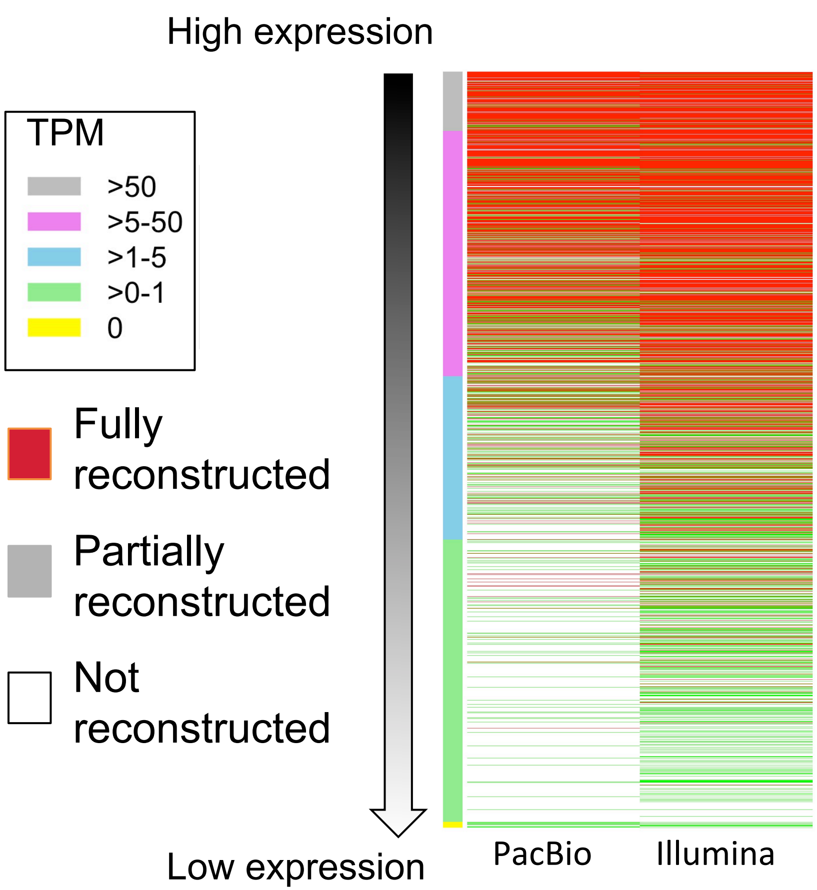


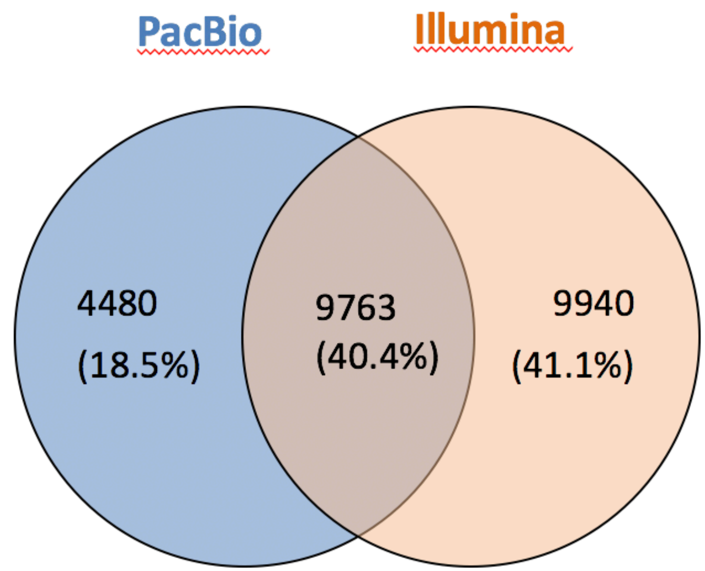
Earlham Institute

PacBio and Illumina provide complementary approaches

Human reference genes reconstructed by Pacbio and Illumina assemblies

- A larger proportion of reference genes could be fully reconstructed by Illumina reads
- Illumina data allowed genes to be reconstructed across a larger expression range

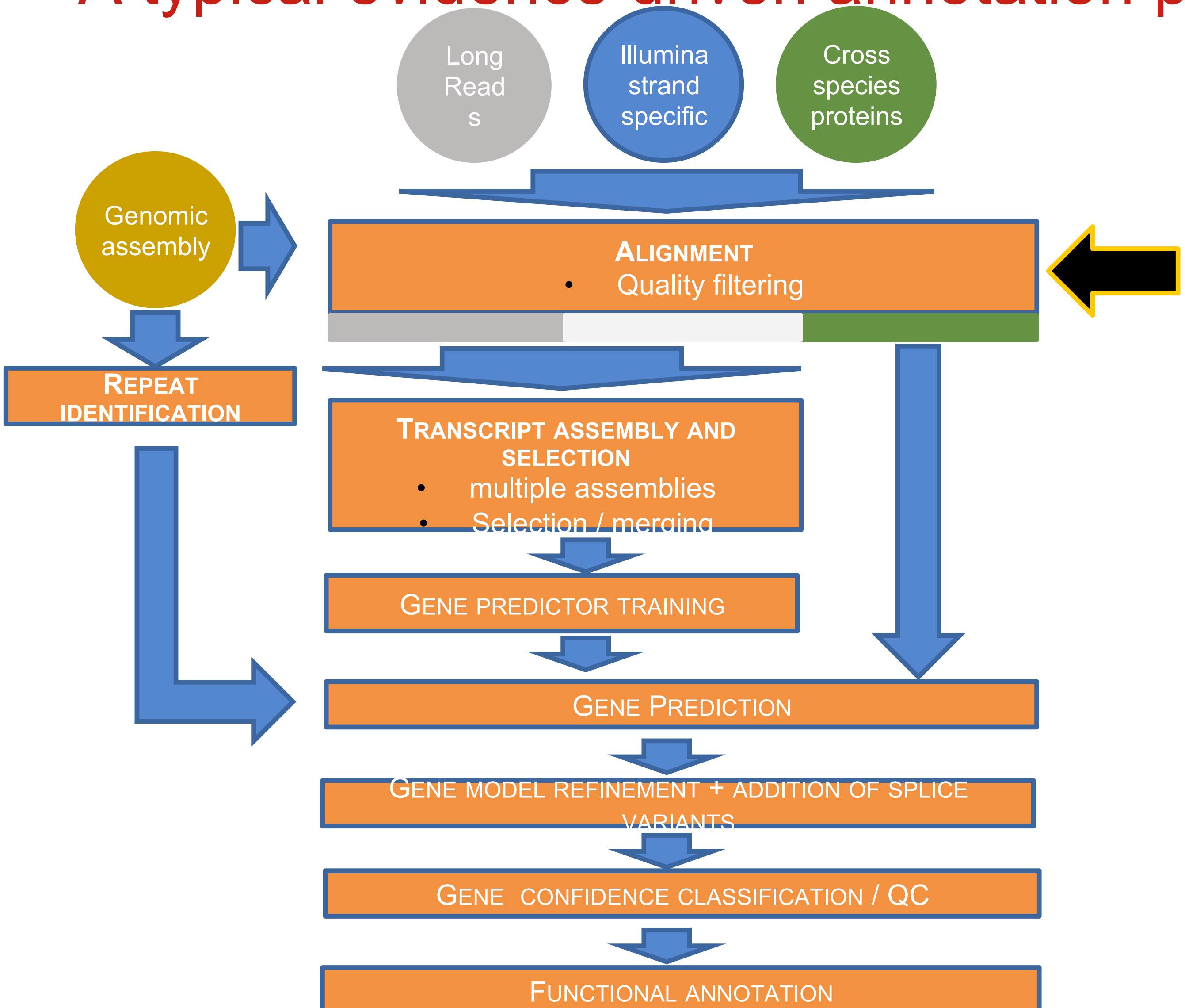




Reference genes fully reconstructed by PacBio or Illumina assemblies

Earlham Institute

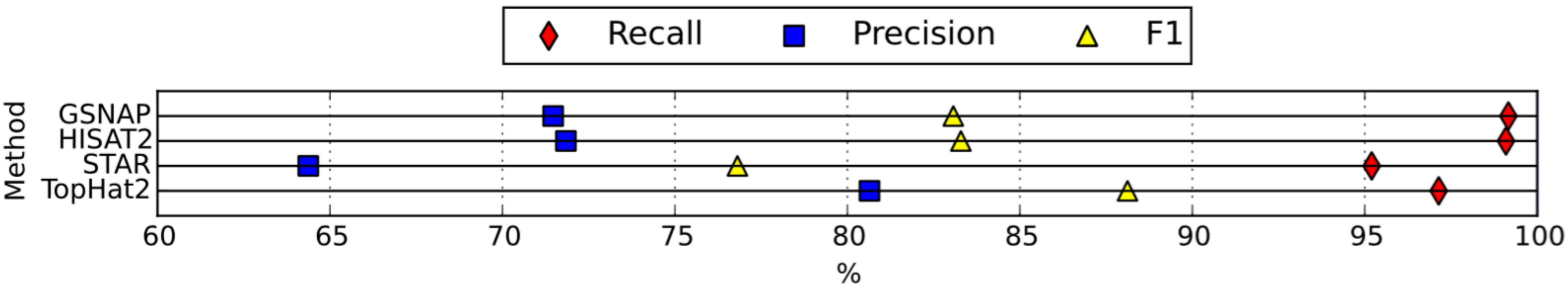
A typical evidence driven annotation pipeline



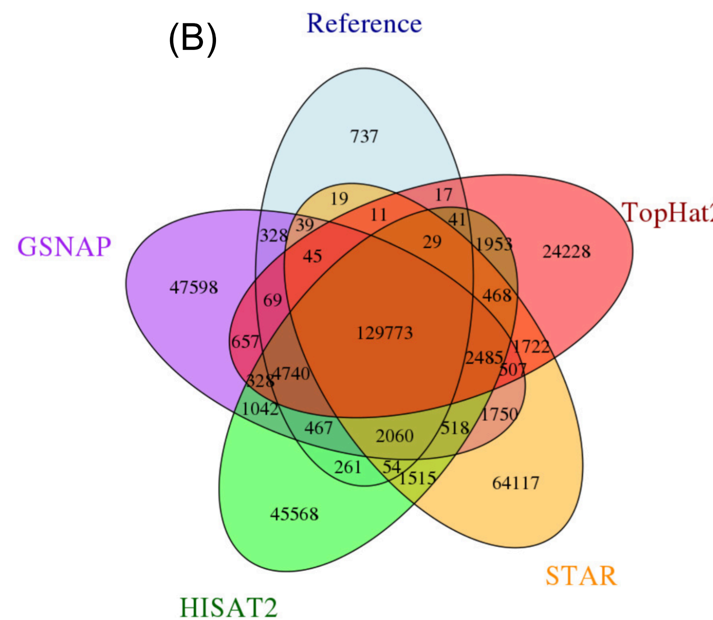
- Short (spliced) read aligners
  - Hisat2
  - Star
  - GSNAP
- cDNA / long read aligners
  - GMAP
  - Exonerate
  - GenomeThreader
  - minimap2
- Protein alignments
  - Exonerate
  - GenomeThreader

RNA-Seq aligners vary in performance

(A)



(B)



Recall=(TP)/(TP+FN)

Precision=(TP)/(TP+FP)

- All aligners generated a substantial proportion of false junctions (i.e. low precision) **(A)**
- The majority of false junctions were specific to one aligner **(B)**

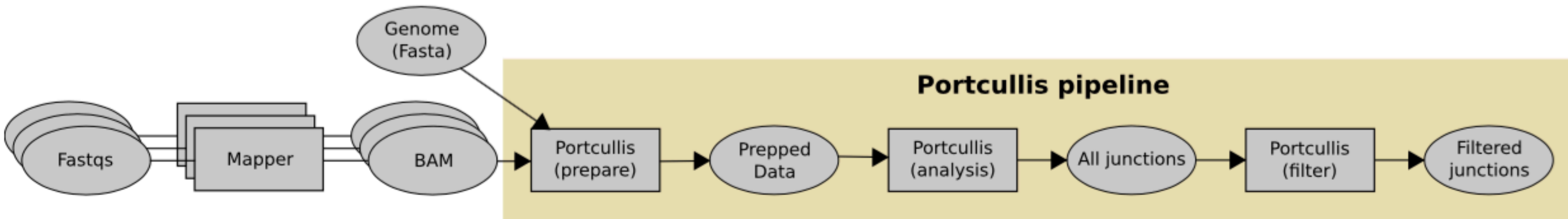
Earlham Institute

Improved splice junction detection

- Portcullis - Fast, reliable and accurate splice junction prediction from RNAseq data

PORTCULLIS

https://github.com/EI-CoreBioinformatics/portcullis



- Using information derived from both genomic and RNA-Seq mapping information, portcullis utlilises a machine learning approach to classify genuine and false positive junctions to a high-degree of accuracy.

Spanki

Splicing Analysis Kit

Spanki is a set of tools to facilitate analysis of alternative splicing from RNA-Seq data.
http://www.cbcb.umd.edu/software/spanki/

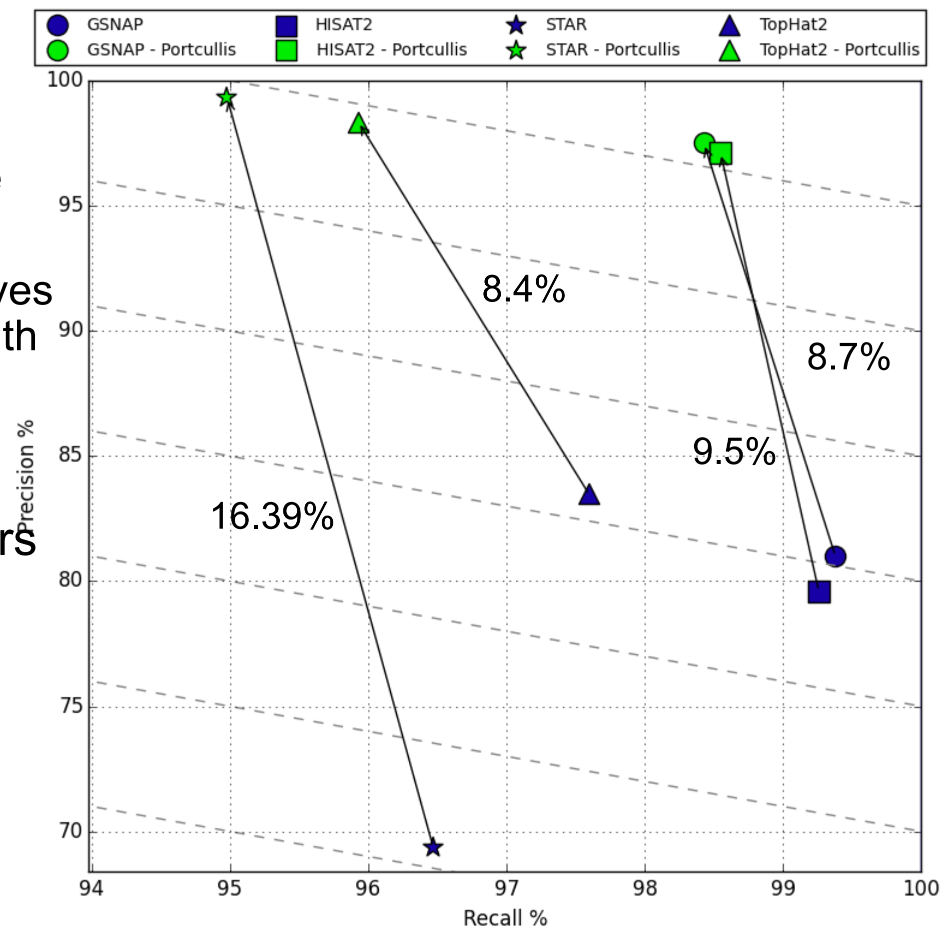
Improved splice junction detection

- Portcullis - Fast, reliable and accurate splice junction prediction from RNAseq data

PORTCULLIS

https://github.com/EI-CoreBioinformatics/portcullis

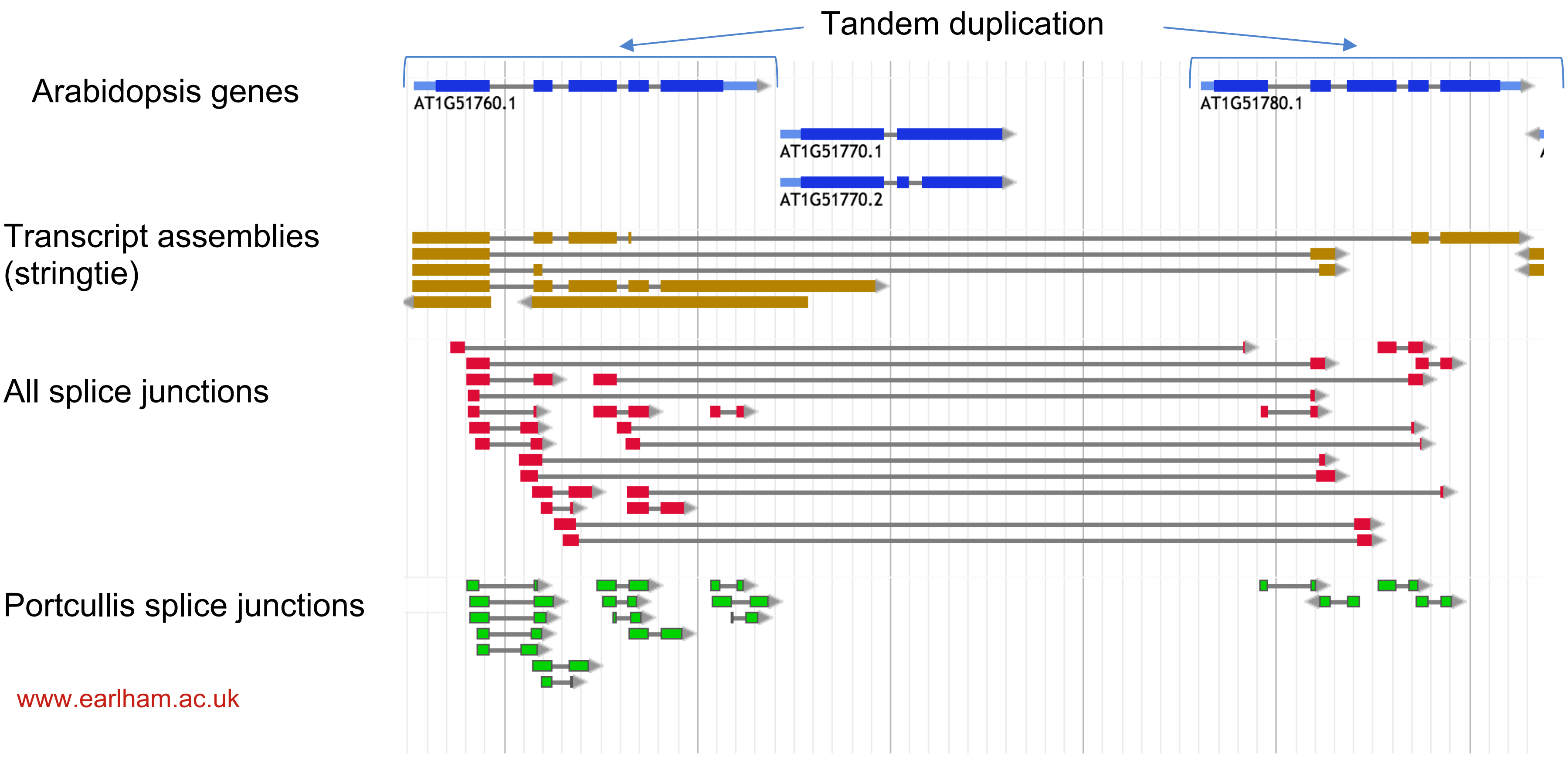
- Portcullis improved aligner performance by up to 16% F1*
- Removing large numbers of false positive splice junctions
  - e.g. using tophat portcullis removes 24,262 false positive junctions with only a lose of 2317 true positive junctions
- As read counts increase junction precision decreases for all aligners



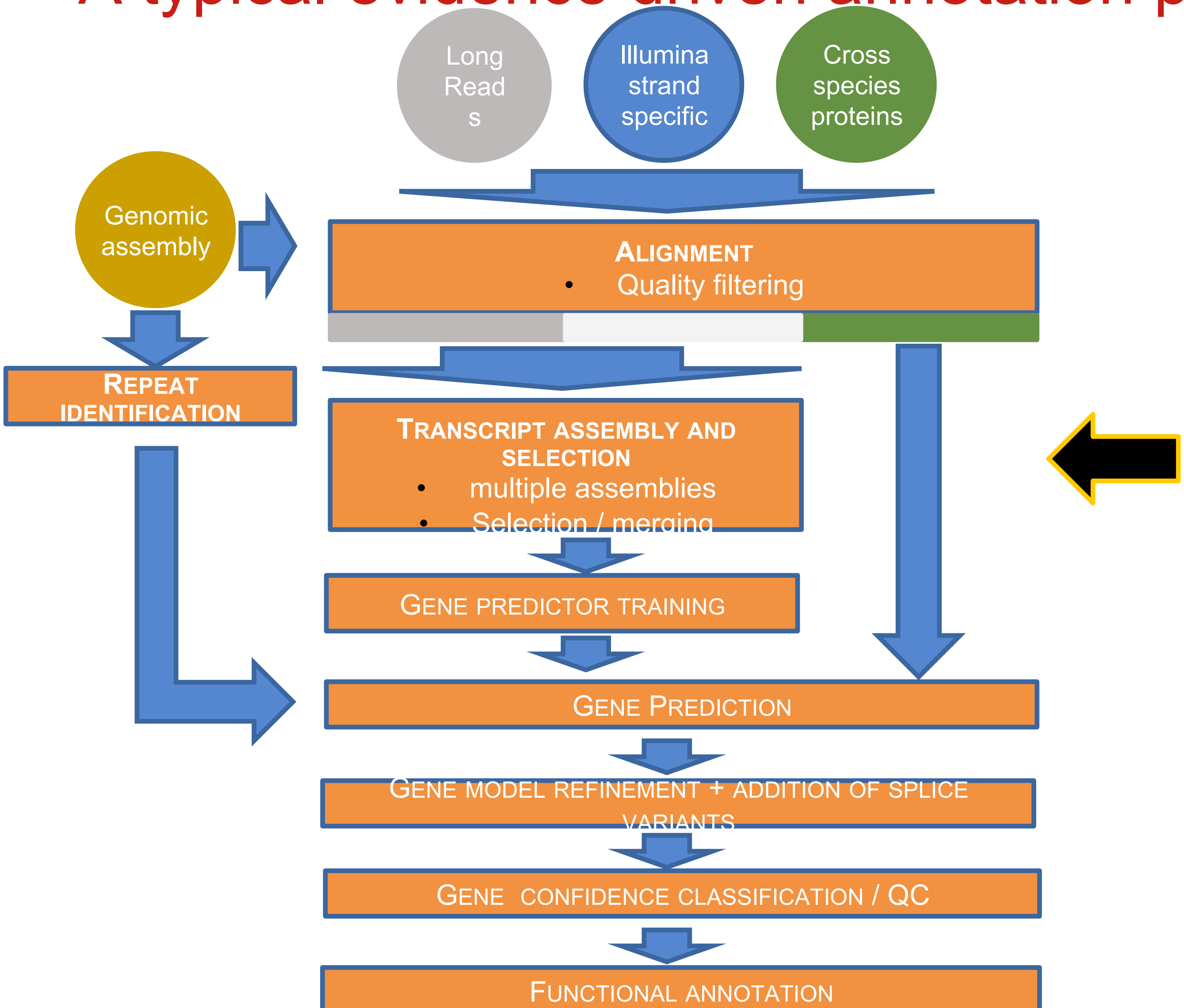
*based on simulated human data ~76million 101bp paired-end unstranded reads

Splice junction filtering can improve transcript reconstruction

- Example region from Arabidopsis showing a tandem duplication, with incorrectly aligned chimeric reads linking the duplicated genes.
- Chimeric reads are incorporated into transcript assemblies creating chimeric transcripts
- Chimeric reads do not pass portcullis filtering enabling the chimeric transcripts to be identified.

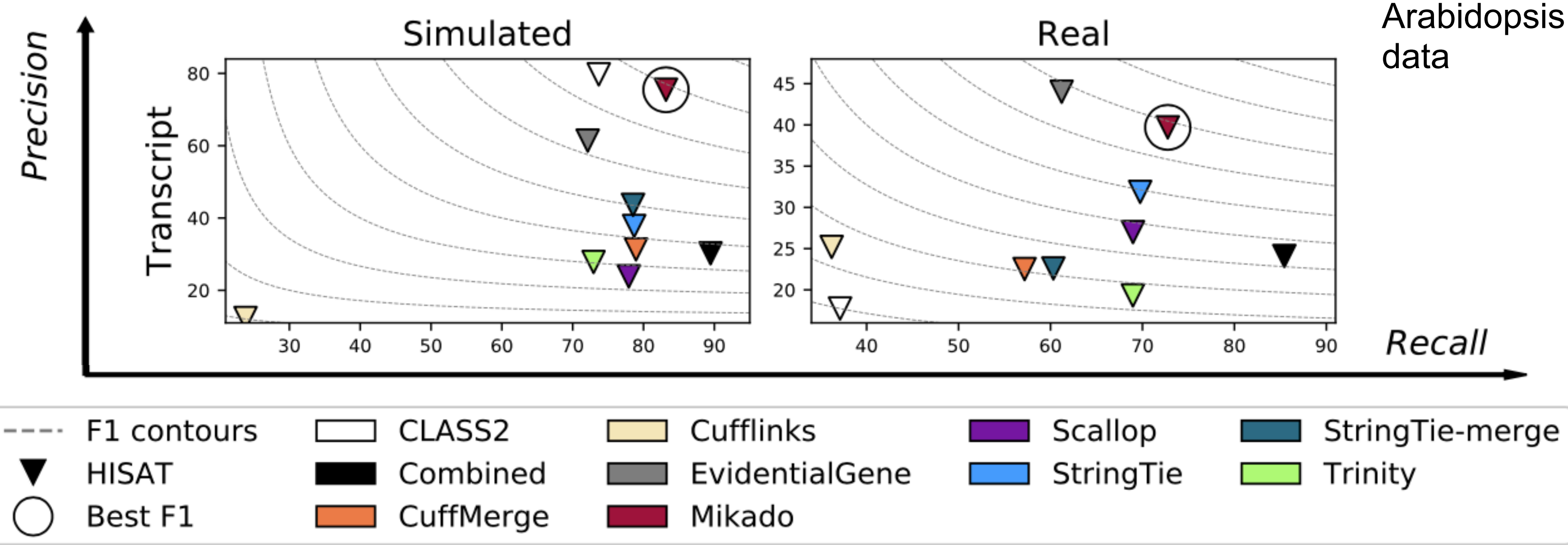


A typical evidence driven annotation pipeline



- Reference guided short read assembly
  - StringTie
  - Scallop
  - Cufflinks
- De novo assembly
  - Trinity
  - rnaSPAdes
- Selection / merging
  - Cuffmerge
  - StringTie merge
  - Mikado
  - EvidentialGene

RNA-Seq assembly approaches vary in accuracy



Choice of assembly method and read aligner can greatly impact transcript reconstruction

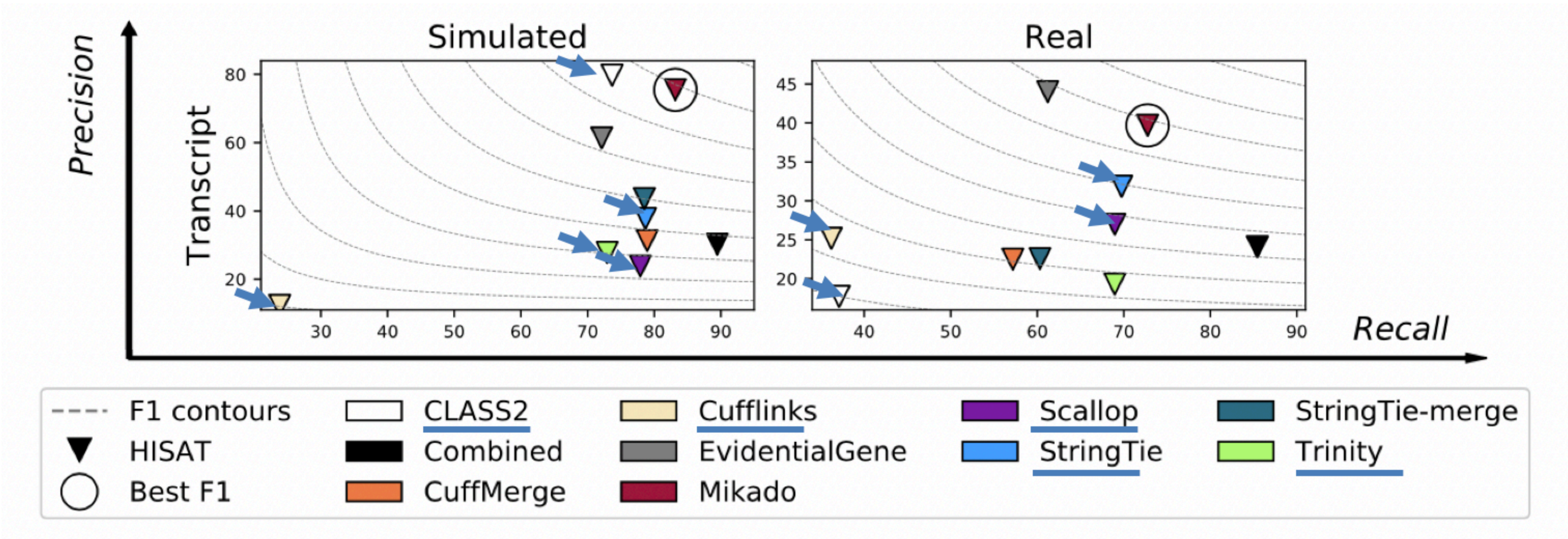
There is substantial variation in recall and precision e.g. on *A.thaliana* (real) data recall (% of correctly assembled transcripts) ranges between 38% and 72%

Recall = the fraction of reference transcripts that are matched by assembled transcripts
Precision = the fraction of assembled transcripts that accurately match a reference transcript

Leveraging multiple transcriptome assembly methods for improved gene structure annotation
L Venturini, S Caim, GG Kaithakottil, DL Mapleson, D Swarbreck
GigaScience 7 (8), giy093

Earlham Institute

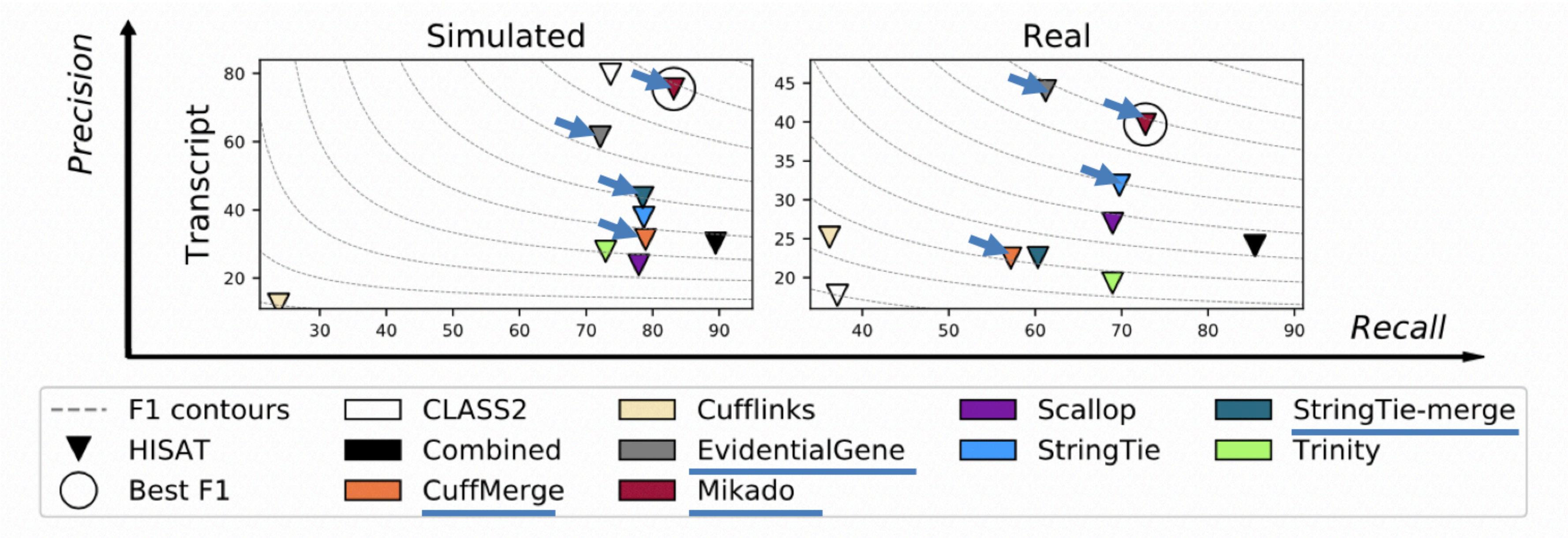
RNA-Seq assembly approaches vary in accuracy



Single sample assemblers

Recall = the fraction of reference transcripts that are matched by assembled transcripts
Precision = the fraction of assembled transcripts that accurately match a reference transcript

RNA-Seq assembly approaches vary in accuracy



Tools for meta-assembly or selecting from a pool of transcripts

- Integrating transcript assemblies from multiple methods / can provide a more comprehensive and accurate set of transcripts

Recall = the fraction of reference transcripts that are matched by assembled transcripts
Precision = the fraction of assembled transcripts that accurately match a reference transcript

Earlham Institute

Integrating sets of transcript assemblies

- There is considerable variation in transcript reconstruction performance between the methods tested.
- Different transcript reconstruction approaches complement each other.
- It's difficult to determine which approach (aligner, assembler) would be best for a given project.
- A strategy that integrates multiple methods has the potential to improve transcript reconstruction, but it has to deal with the additional noise and redundancy introduced by pooling assemblies.

Cuffmerge
- meta-assembler

StringTie Merge
- meta-assembler

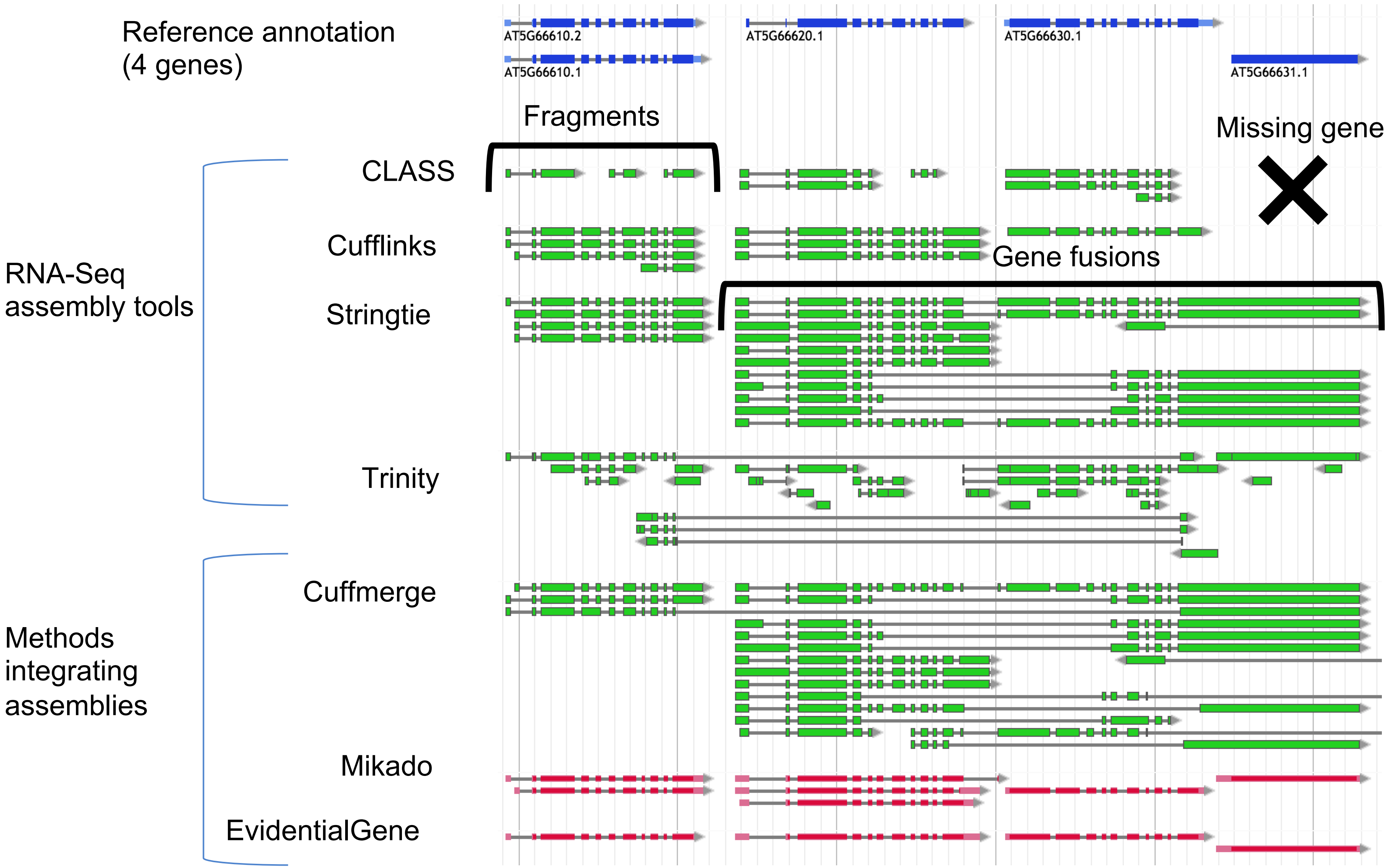
MIKADO

https://github.com/EI-CoreBioinformatics/mikado

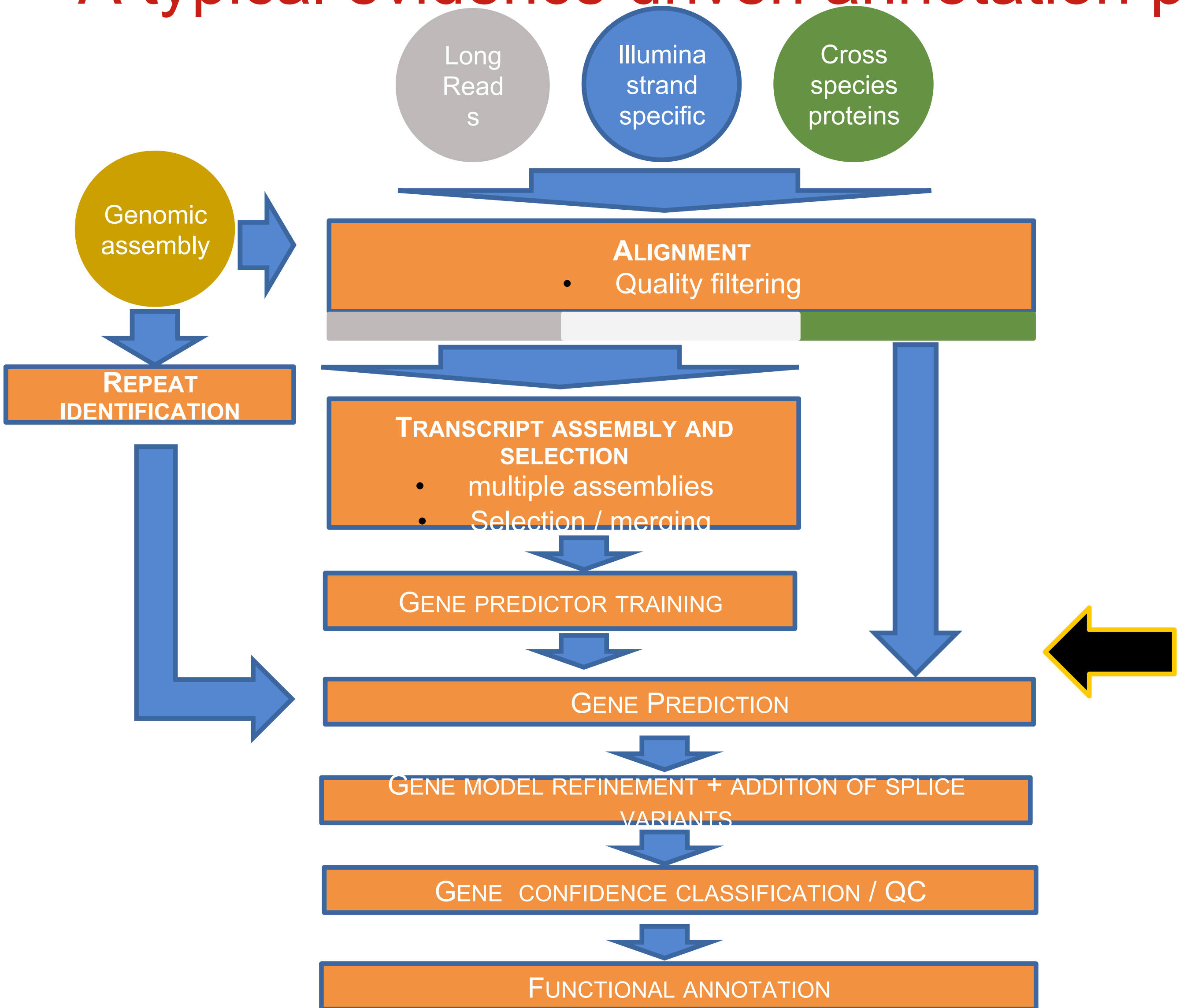
Framework for selecting transcripts from a pool of transcript assemblies.
- Generates a range of intrinsic metrics
- Can incorporate cross species similarity

EvidentialGene
- Selects from pool of transcripts (CDS alignment)

Example – Differences between assembly tools

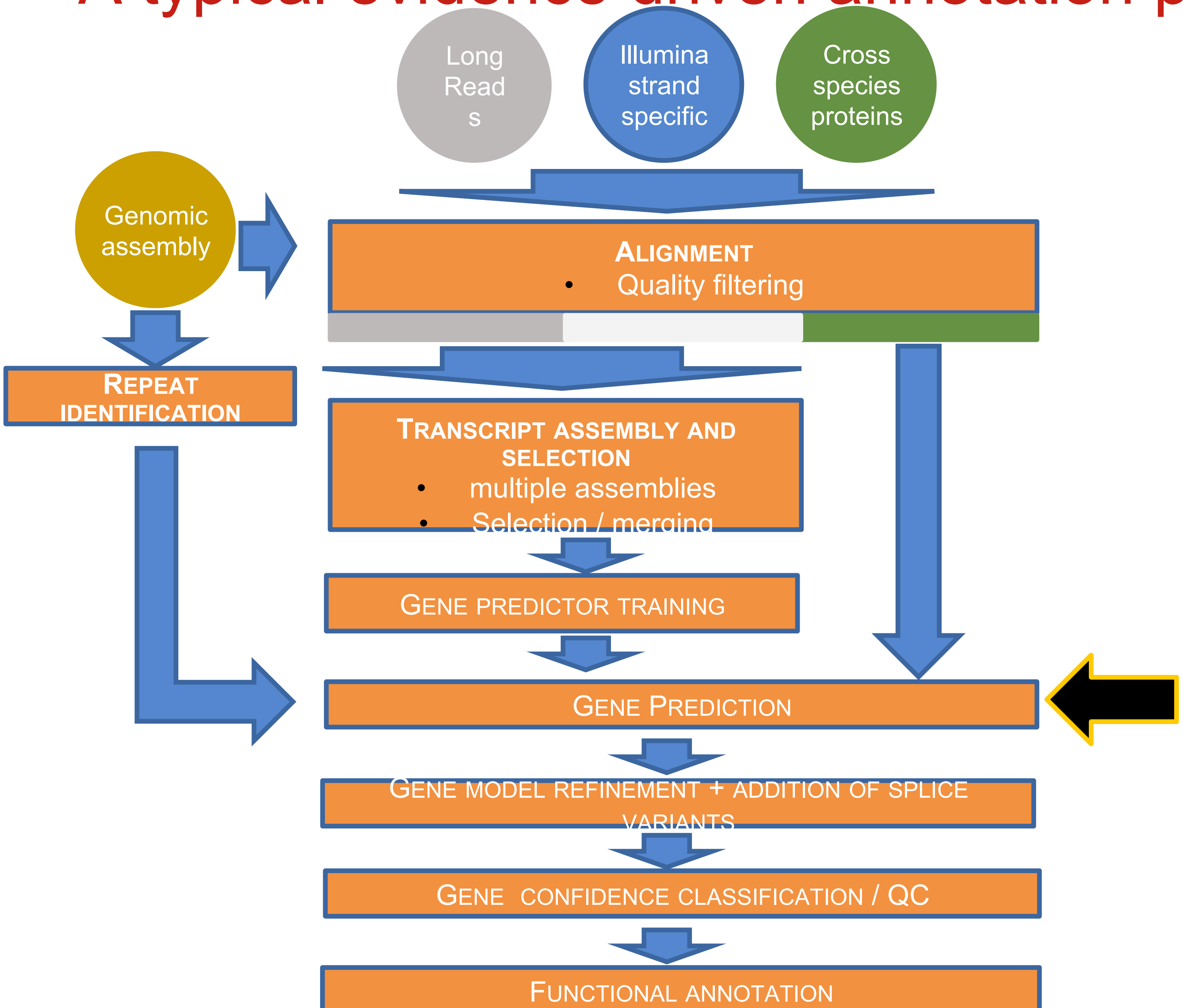


A typical evidence driven annotation pipeline



- Training gene predictors
  - Use a subset of the transcript assemblies
  - Annotate CDS features
  - Identify CDS with homology support
- *Ab initio*
  - SNAP
  - Augustus
- Evidence based
  - Maker
  - PASA
  - Augustus

A typical evidence driven annotation pipeline



Multiple gene models can be created using different tools, parameters or evidence

Approaches are available to choose from these gene models based on identifying:

1) the gene model that represents the best consensus intron-exon structure (JIGSAW, EVidenceModeler, GLEAN)

2) the gene model that is best supported by the evidence (EVidenceModeler, Maker, Mikado)

Example – Differences between annotation runs

Augustus Run 1
(weighted towards transcript assemblies and with RNA-Seq)

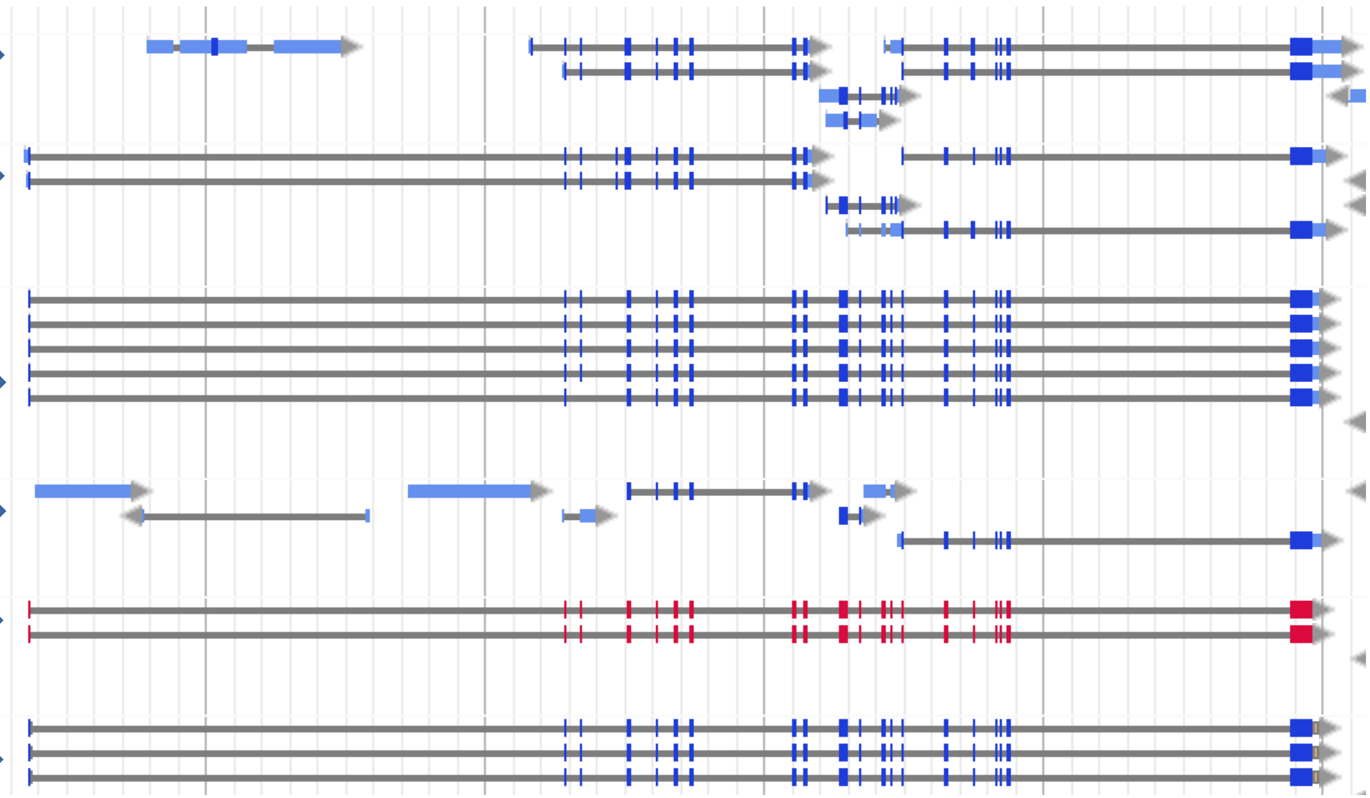
Augustus Run 2
(weighted towards transcript assemblies and **NO** RNA-Seq)

Augustus Run 3
(weighted towards protein alignments)

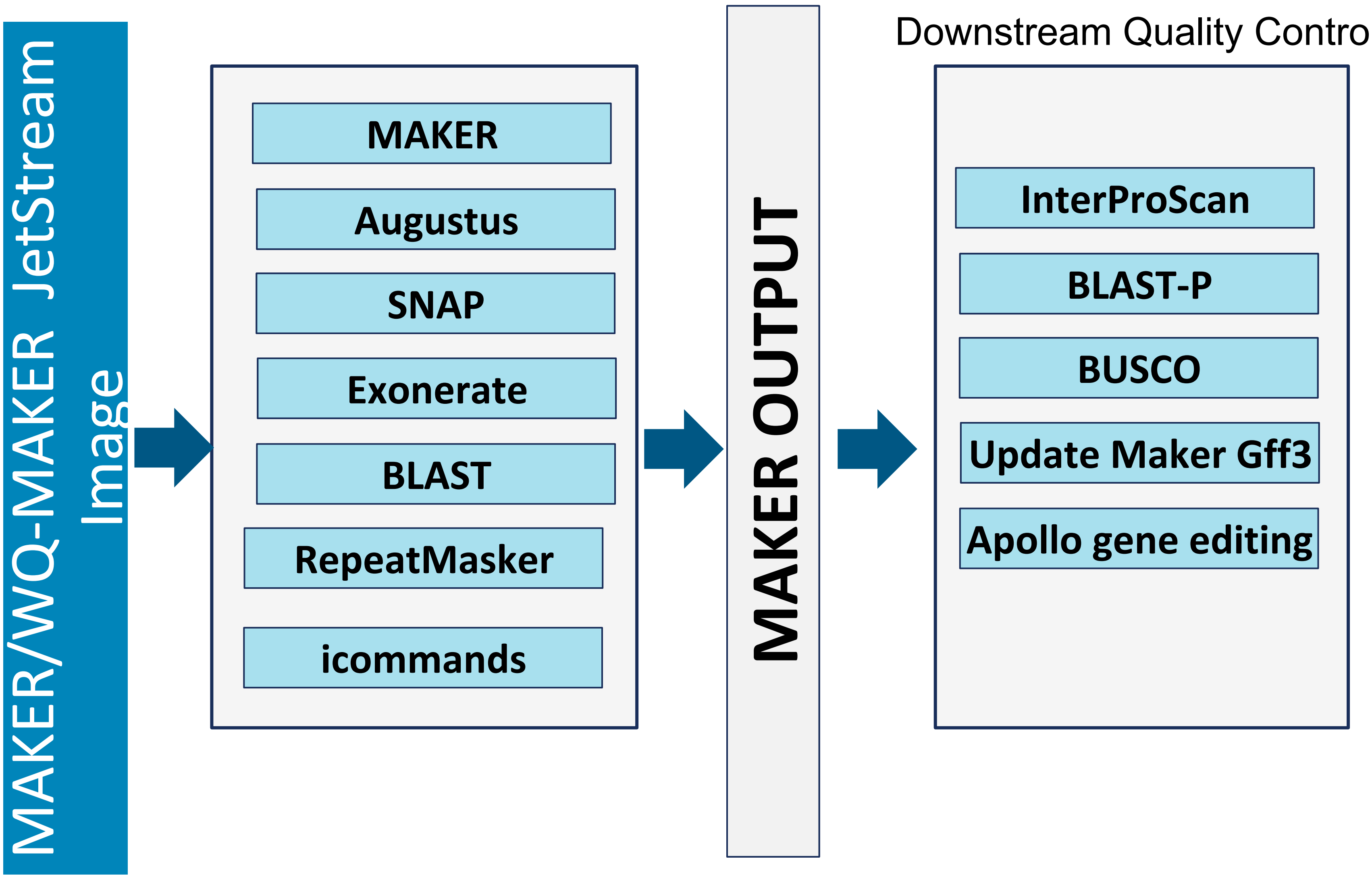
Transcript assemblies

Protein alignments

Final selected models



Genome Annotation: MAKER/WQ-MAKER in JetStream

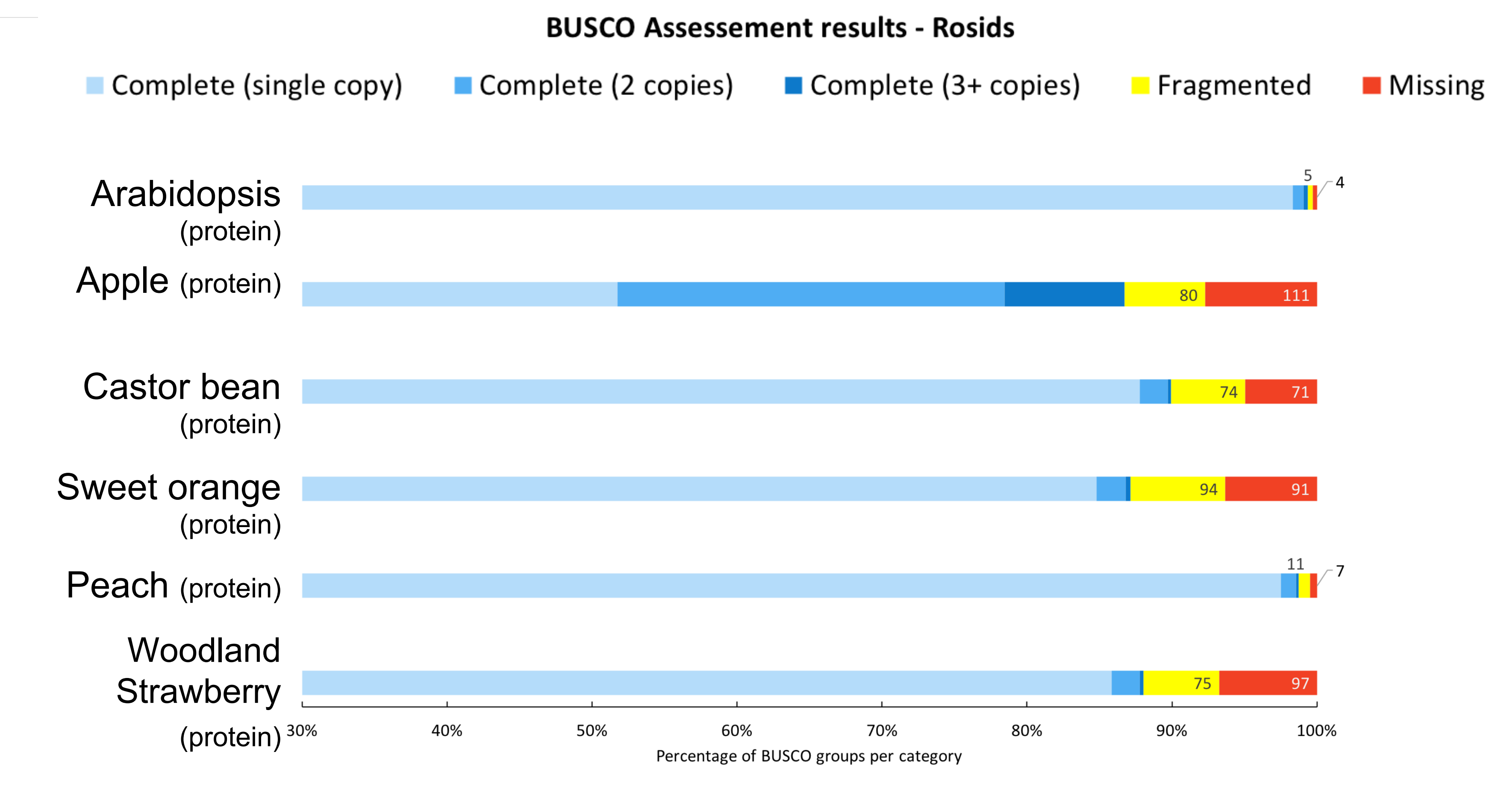


Assessing the quality of the annotation

- Check percentage of genes with known domains (PFAM, InterProScan)
  - D. melanogaster, C. elegans, A. thaliana and S. cerevisiae proteomes varies between 57% and 75%
- Check against BUSCO, set of conserved single copy genes for specified taxonomic groups

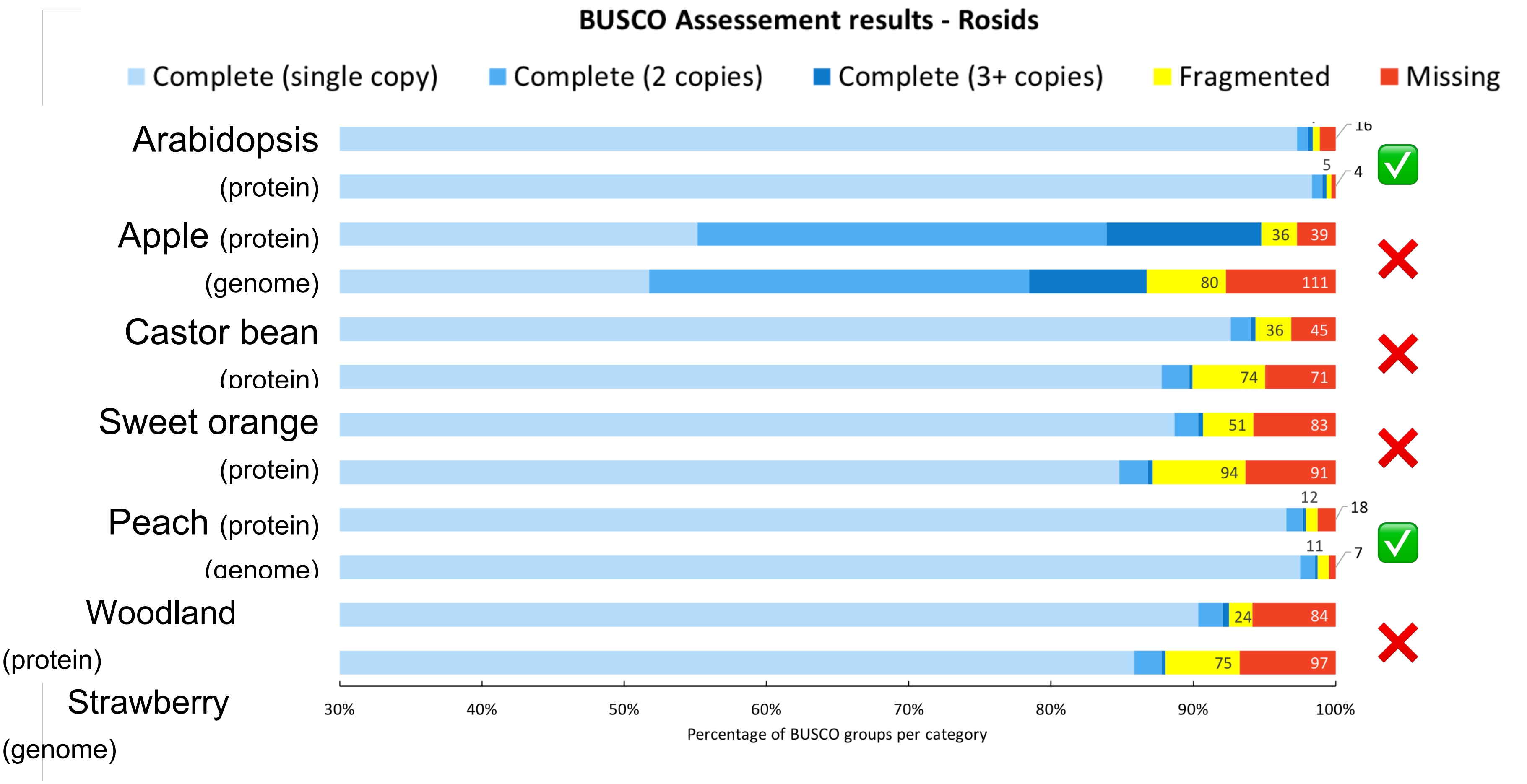
Variation in annotation quality

- Looking at 1440 “single copy” conserved genes



Variation in annotation quality

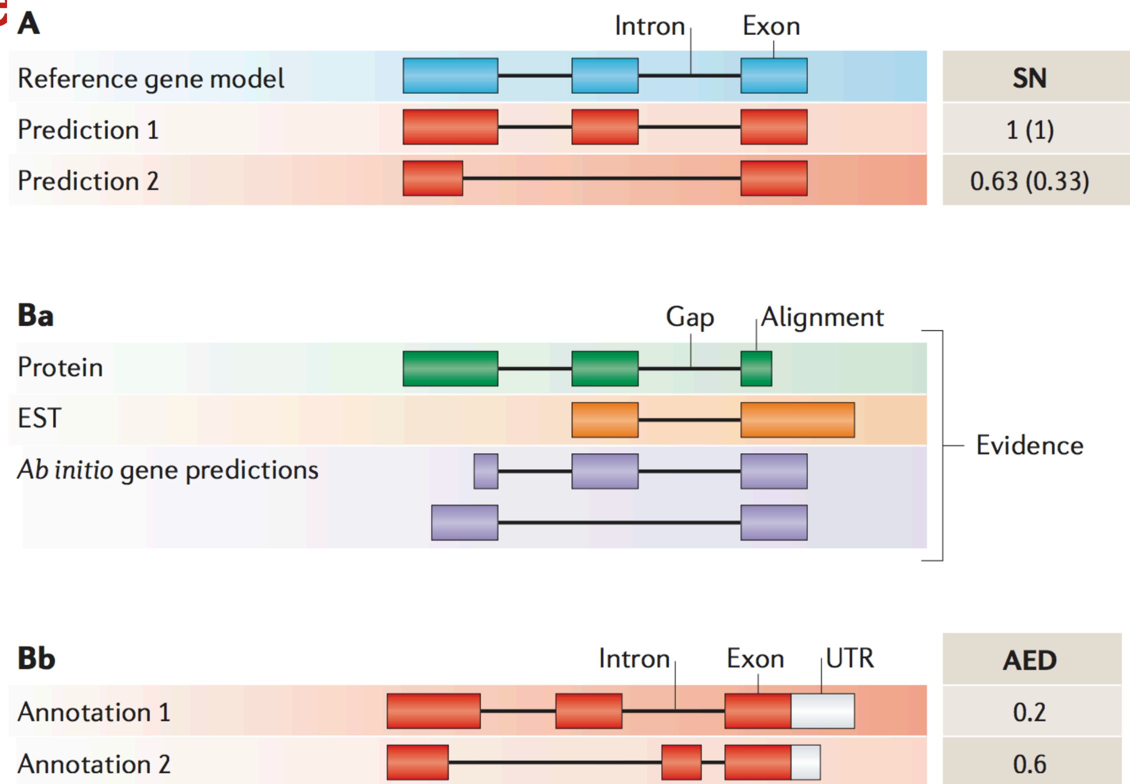
- Results should show less fragmented and missing genes in the predicted gene set (protein) compared to the genome



Assessing the quality of the annotation

- Check percentage of genes with known domains (PFAM, InterProScan)
  - D. melanogaster, C. elegans, A. thaliana and S. cerevisiae proteomes varies between 57% and 75%
- Check against BUSCO, set of conserved single copy genes for specified taxonomic groups
- Does evidence (RNA-Seq, Proteins) support or contradict the annotated intron-exon structure
  - Blast against know proteins to check support (% coverage) for CDS
  - AED = Annotation Edit Distance

Measures of gene finder performance



	SN	SP	AC
Reference gene model			
Prediction 1	1 (1)	1 (1)	1 (1)
Prediction 2	0.63 (0.33)	1 (0.5)	0.81 (0.42)

SN = TP / (TP + FN)
SP = TP / (TP + FP)
AC = (SN + SP) / 2

	AED
Annotation 1	0.2
Annotation 2	0.6

AED = 1 – AC

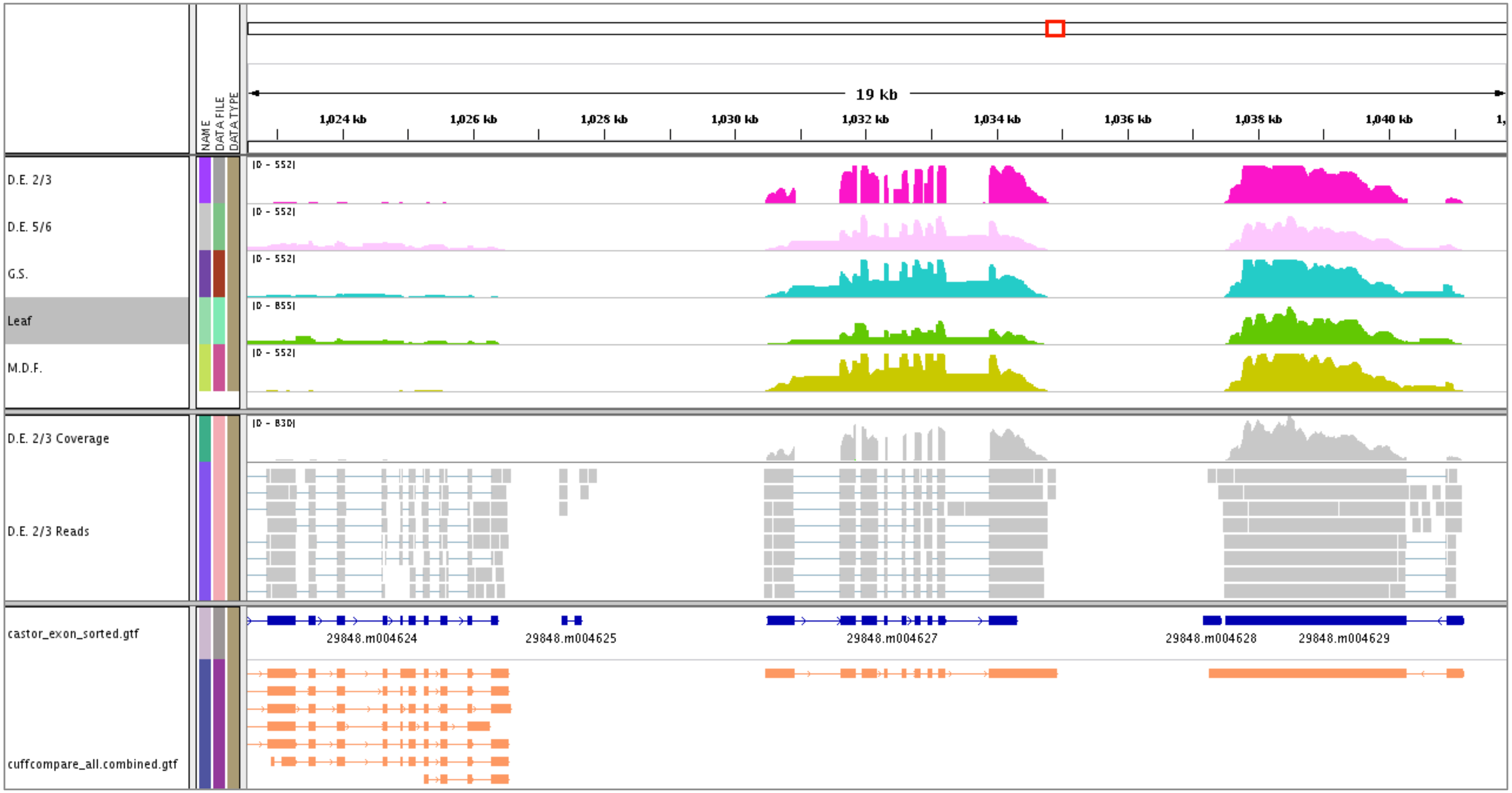
An AED of 0 indicates that the annotation is in perfect agreement with its evidence, whereas an AED of 1 indicates a complete lack of evidence support for the annotation.

Nat Rev Genet. 2012 Apr 18;13(5):329-42. doi: 10.1038/nrg3174.
A beginner's guide to eukaryotic genome annotation.
Yandell M1, Ence D.

Earlham Institute

Visualizing the annotation data

- Load gene models and evidence to a browser and review



Integrative Genomics Viewer

http://www.broadinstitute.org/igv/

Earlham Institute

Summary

- Annotation approach will depend on resources and aims
- Generally this isn't a point and click exercise
- A full annotation pipeline has many steps
    - The choice of tools and QC will impact the overall accuracy and completeness of the annotation
- Incorrect and incomplete annotations poison downstream experiments that make use of them.
- Pipelines like the Maker pipeline available in CyVerse make it easier for investigators to tackle genome annotation
    - Still requires careful QC of the data (bad data in will lead to poor annotations).

Earlham Institute

Decoding Living Systems