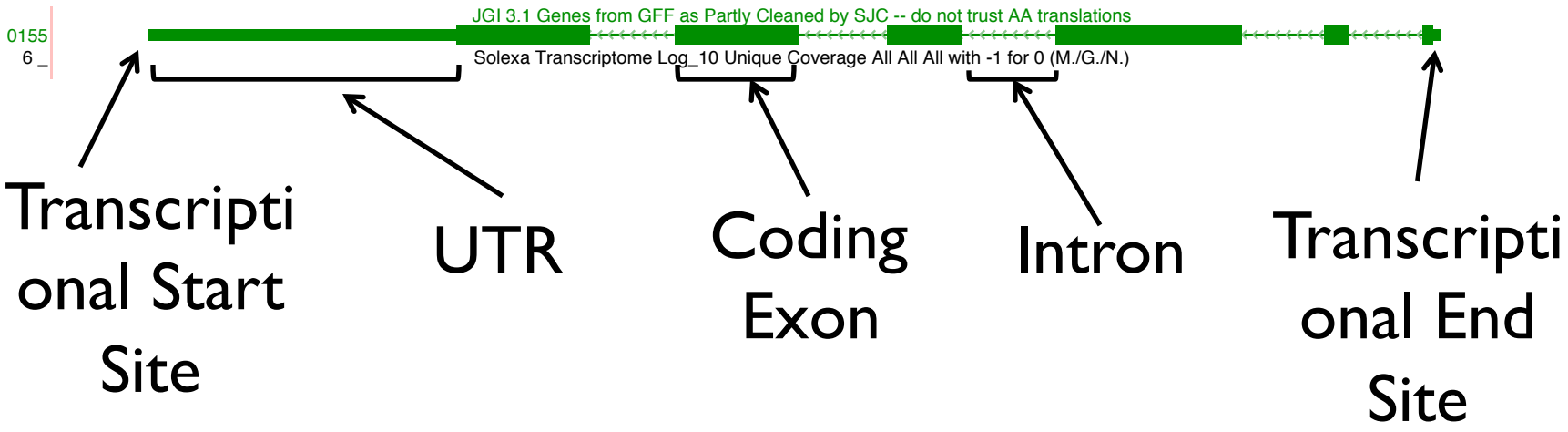
Annotation of Plant Genomes using RNA-seq

Matteo Pellegrini (UCLA)

In collaboration with Sabeeha Merchant (UCLA)

What is Annotation

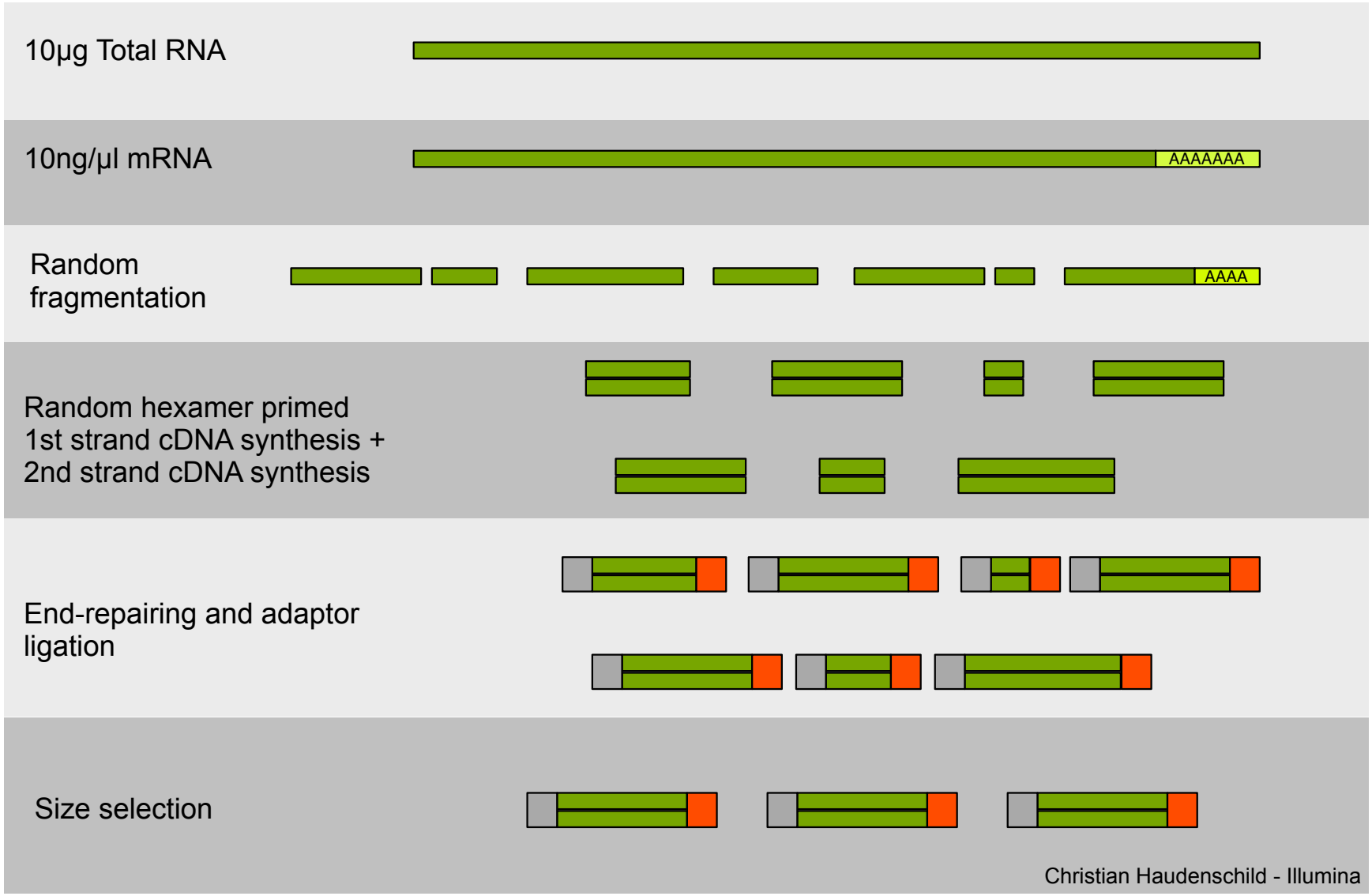
- Reconstruction of gene structure within genome



How are Genomes Annotated?

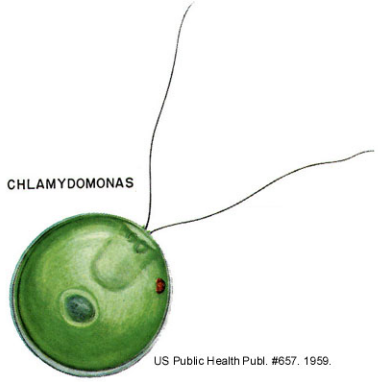
- Traditional Approaches use:
  - Using information from expressed sequence tags (ESTs)
  - Conservation across organisms
  - Prior knowledge of sequence motifs (e.g. splice junctions)
  - Do not take advantage of data generated from next-generation sequencers
- Challenge: Develop data-driven annotation using RNA-seq data

Whole-genome Transcriptome Analysis (WTA)

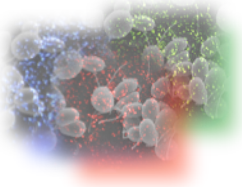


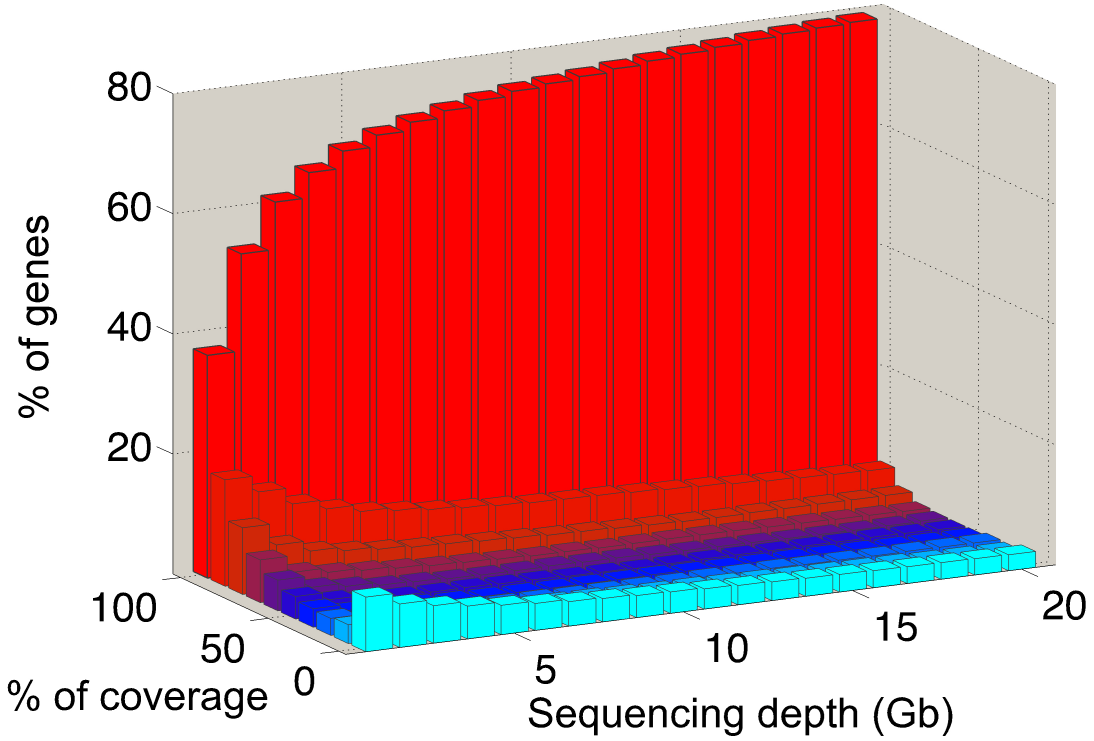
Plant Genomes

- Chlamydomonas is a model algae with a sequenced genome and still incomplete annotation
  - Currently being used for biodiesel studies
- Arabiodpsis is a model plant with a very high quality genome and nearly complete annotation
  - Genetically tractable organism



Limitations of Current Chlamydomonas Annotation from Augustus Models





Even at high coverage more than 20% of predicted genes have no RNA-seq evidence

Two Approaches for Annotating Genomes using RNA-seq

1. First Approach
    - Align reads to genome
    - Concatenate reads that map to overlapping bases on the genome
2. Second approach
    - Assemble reads directly before mapping to genome
    - Use Assembly tools such as ABySS

Method I - Alignment of Reads to Genome

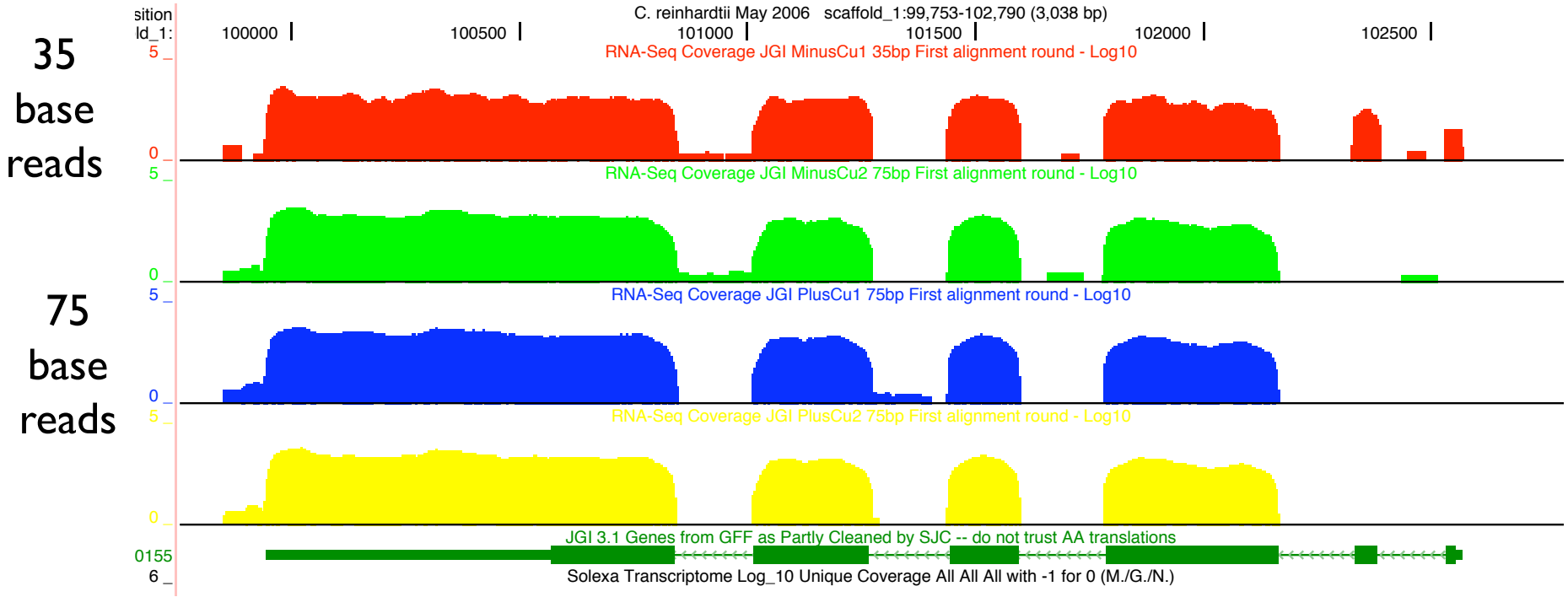
- First perform ungapped alignments using a fast aligner (e.g. Novoalign or Bowtie)

- The reads that do not map are mapped using a gapped alignment protocol (e.g. BLAT or TopHat)

    - The gaps identify splice junctions

- We compute the number of reads that align to each base in the genome

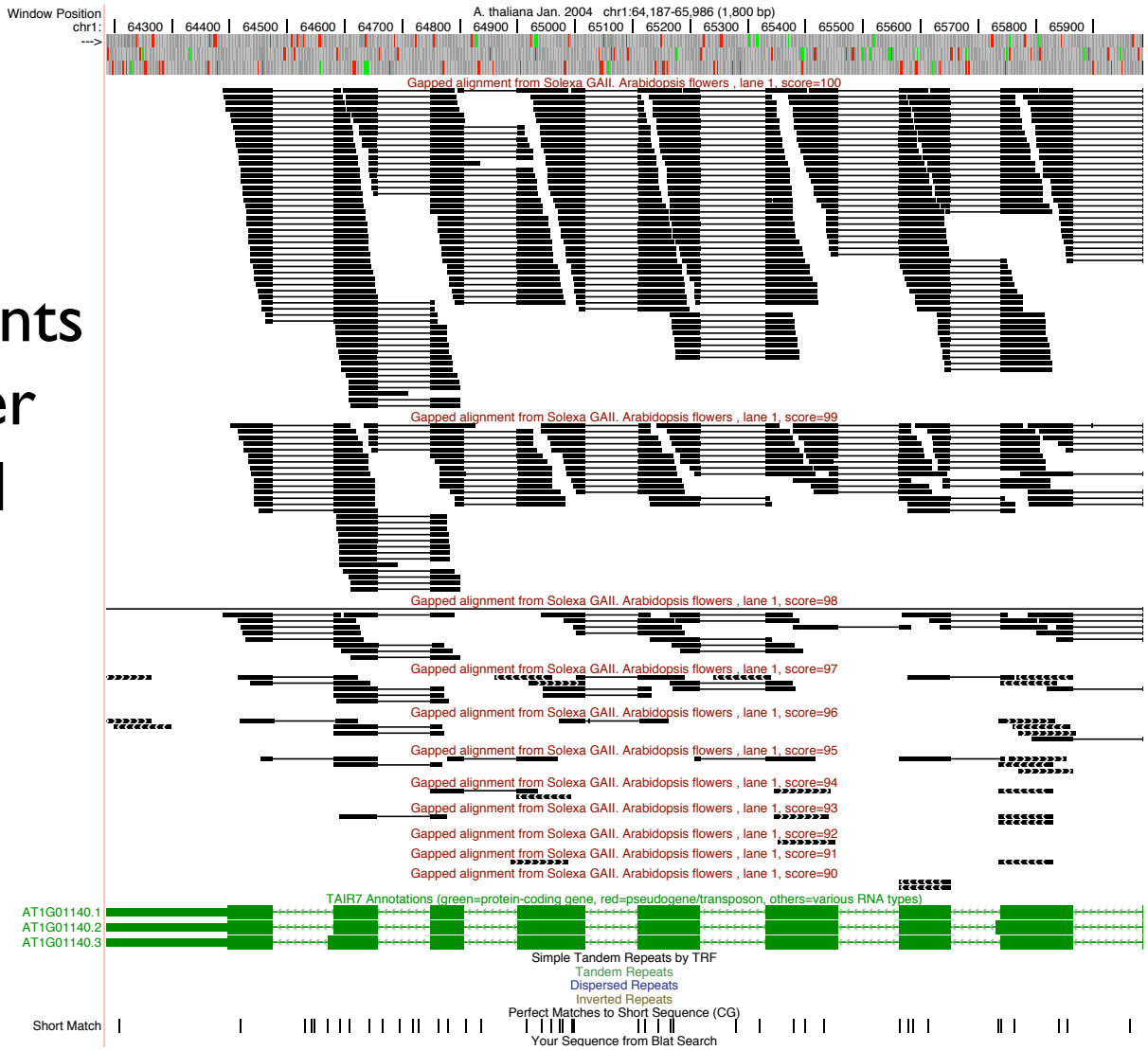
Read Counts Across Chlamydomonas Genome



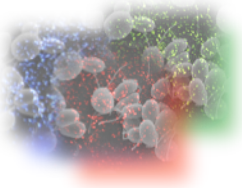
Long Reads are not aligned across short exons in ungapped alignments

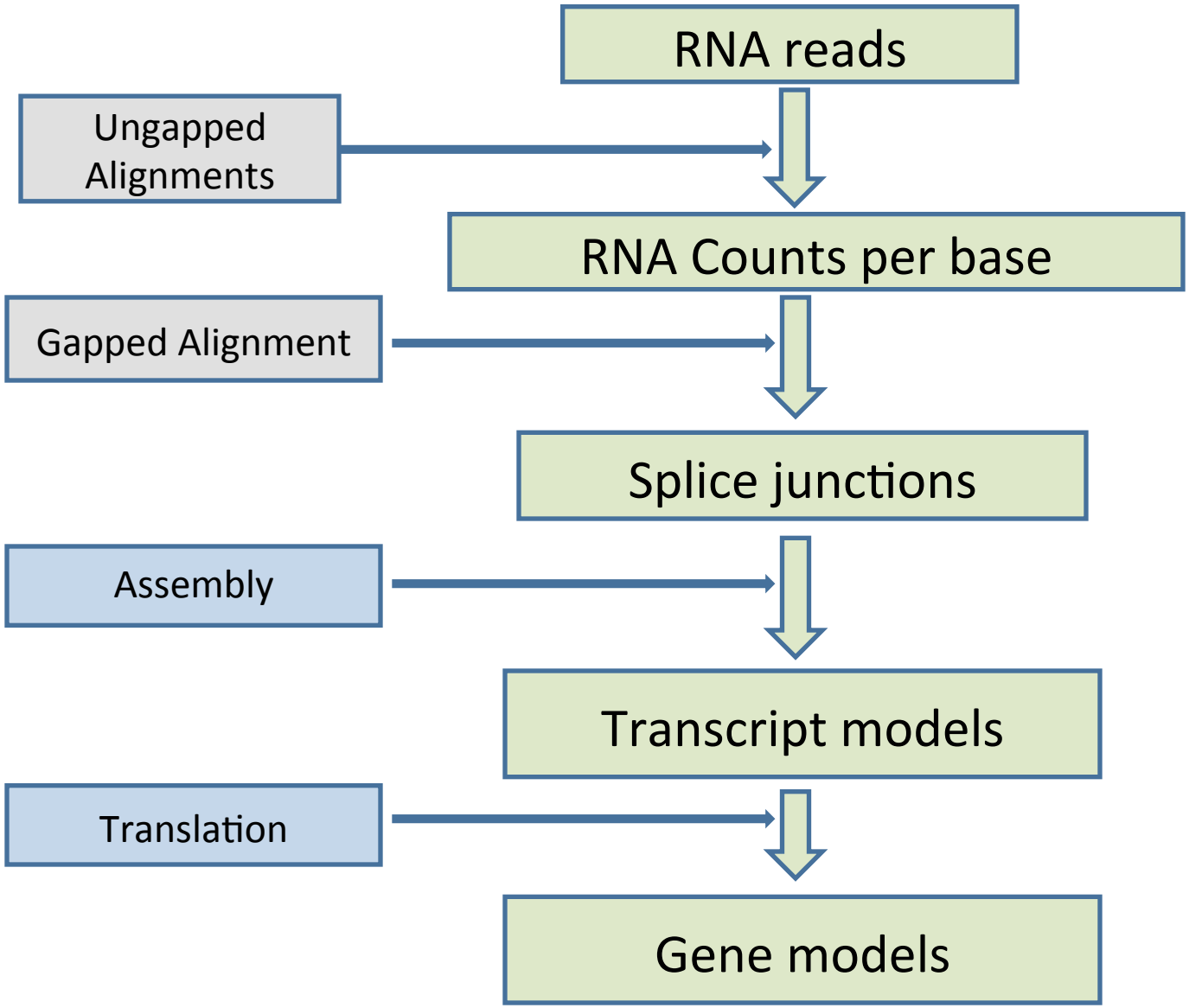
Gapped Alignments in Arabidopsis Genome

Gapped alignments allow us to cover short exons and define splice junctions

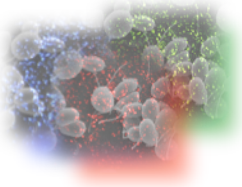


Assembly of Mapped Reads





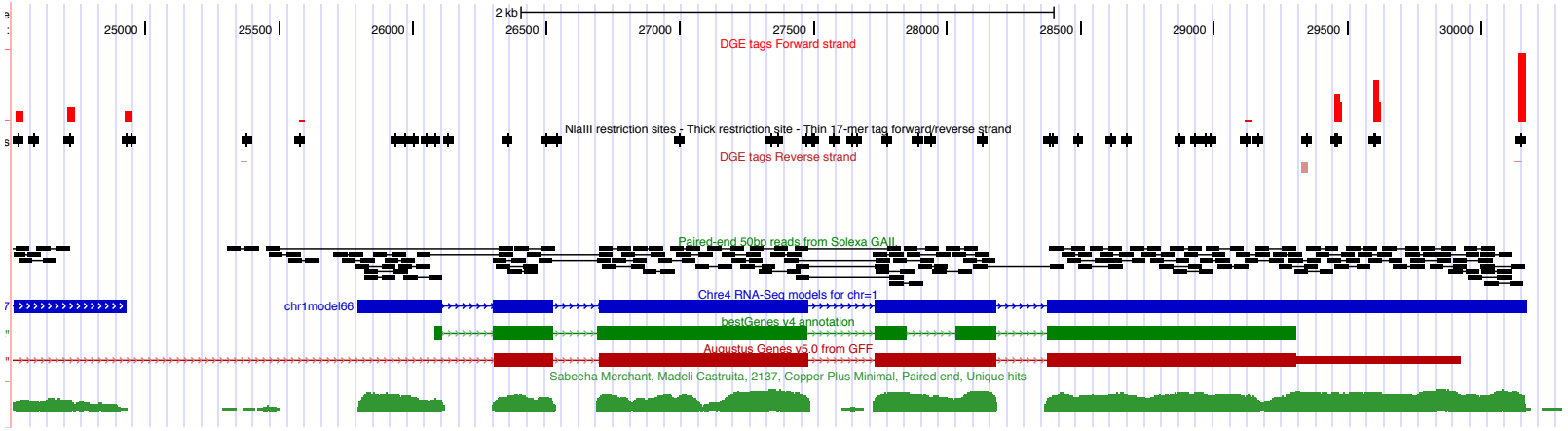
Assembly



Our Models (blue)

JGI (green)

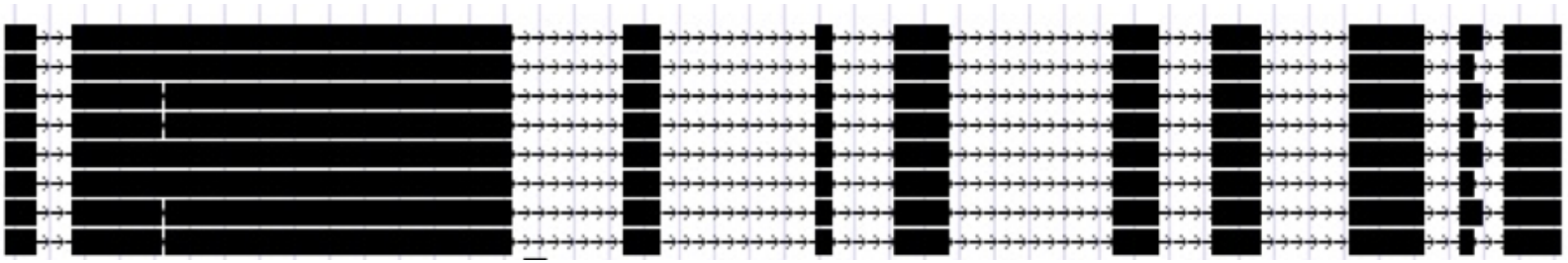
Augustus (red)



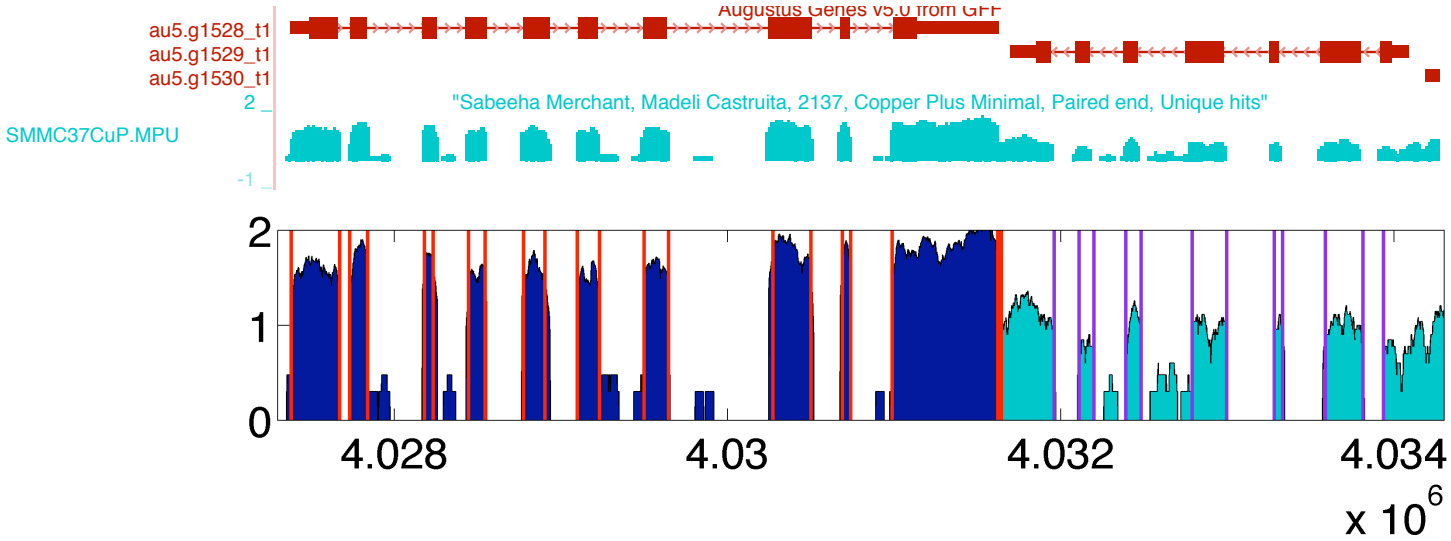
Reads that map to overlapping bases on the genome are concatenated into contigs (blue)

Alternative Splicing

- The same locus can generate multiple transcripts due to alternative splicing, TSS and TTS sites
- Our Assembly generates multiple models that represent different combinations of splice sites



Segmentation

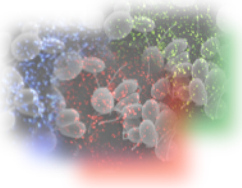


•Regions where two genes overlap show continuous counts in RNA-seq data

•Discontinuities in read counts may be used to define the boundaries

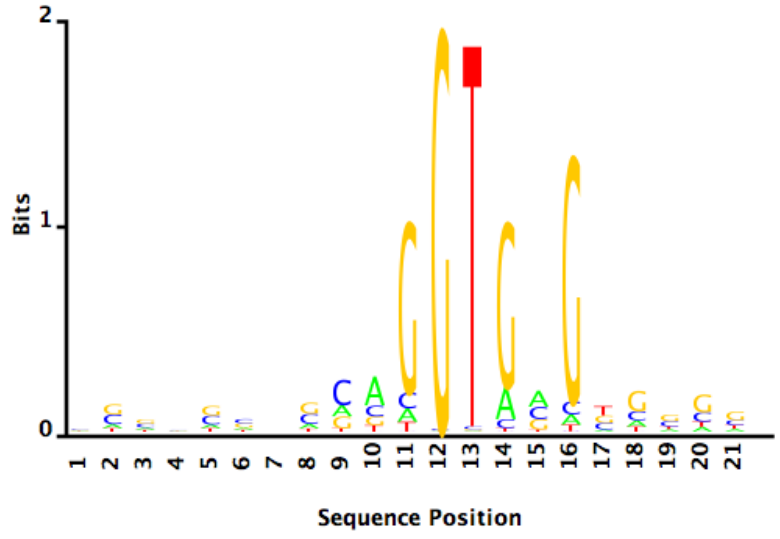
•We use Dynamic Programming approaches to efficiently segment count profiles

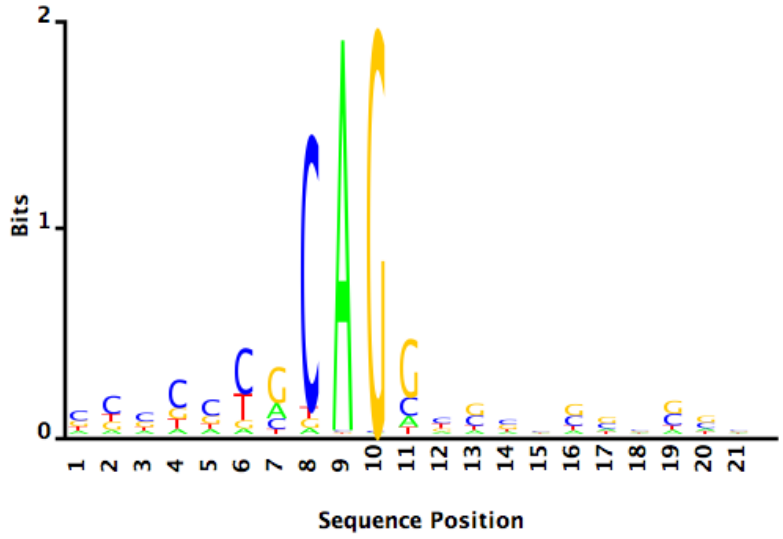
Refining Splice Junctions



Splicing Junction Motifs

Donor — Acceptor

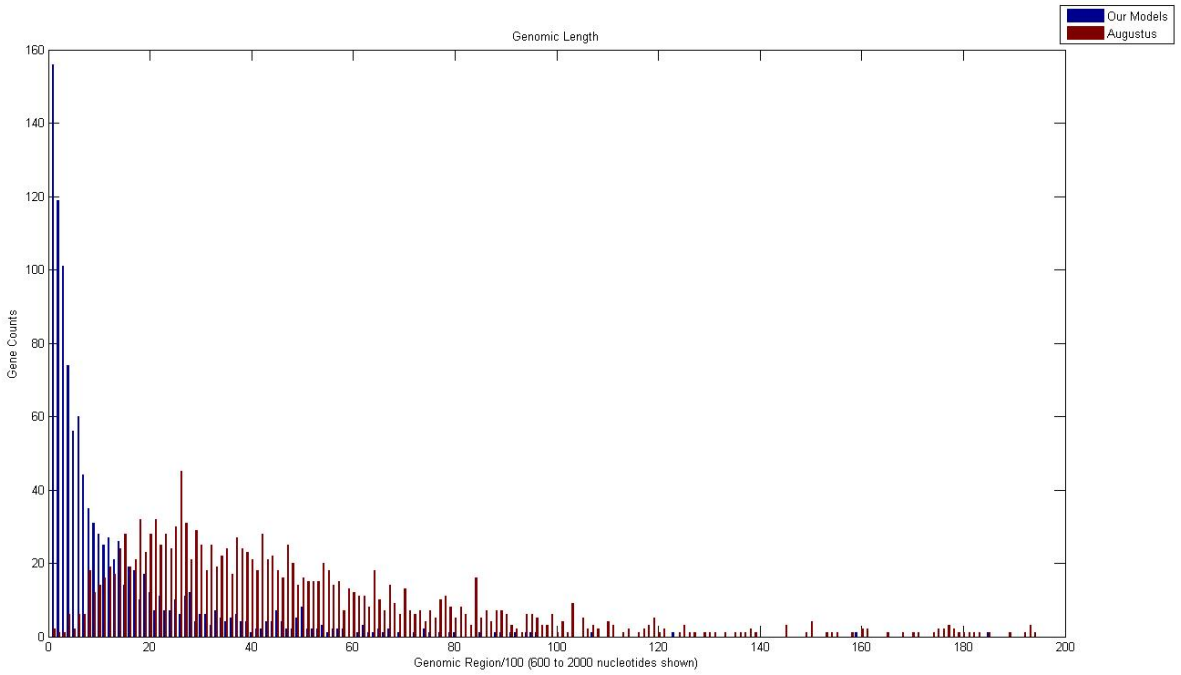




Splice Junction motifs are computed and used to refine ambiguous gapped alignments

Preliminary Results

- 77% of the bases in our models overlap with Augustus
- 76% of Augustus models overlap our models
- Our models are often limited by poor RNA-seq coverageof genes which results in the generation of gene fragments rather than complete transcripts



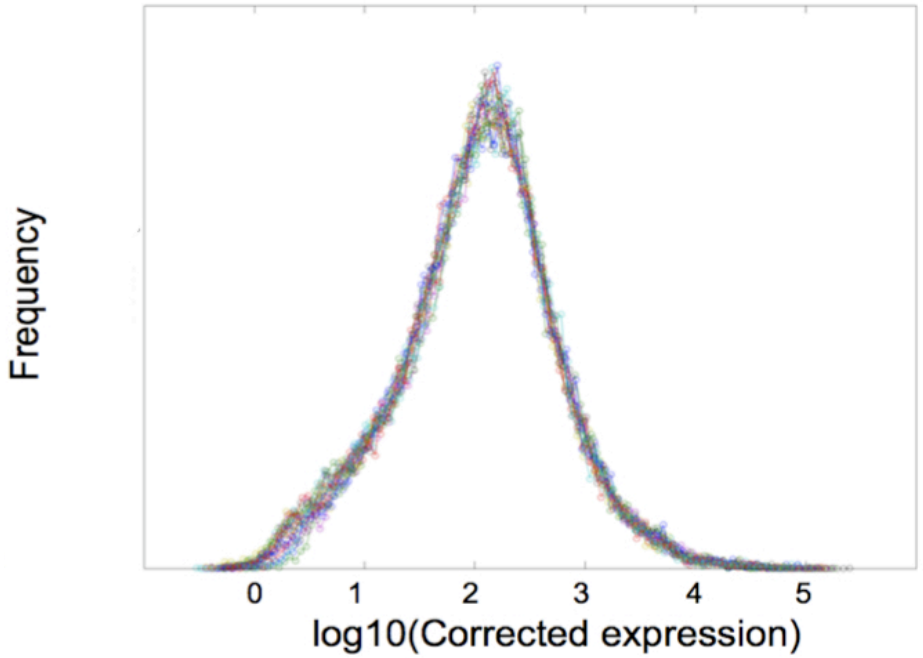
De Novo Transcriptiome Assembly with ABySS

- Transcriptome assembly
  --May be used when a genome seqeunce is not available
  --Not biased by errors in genome sequence

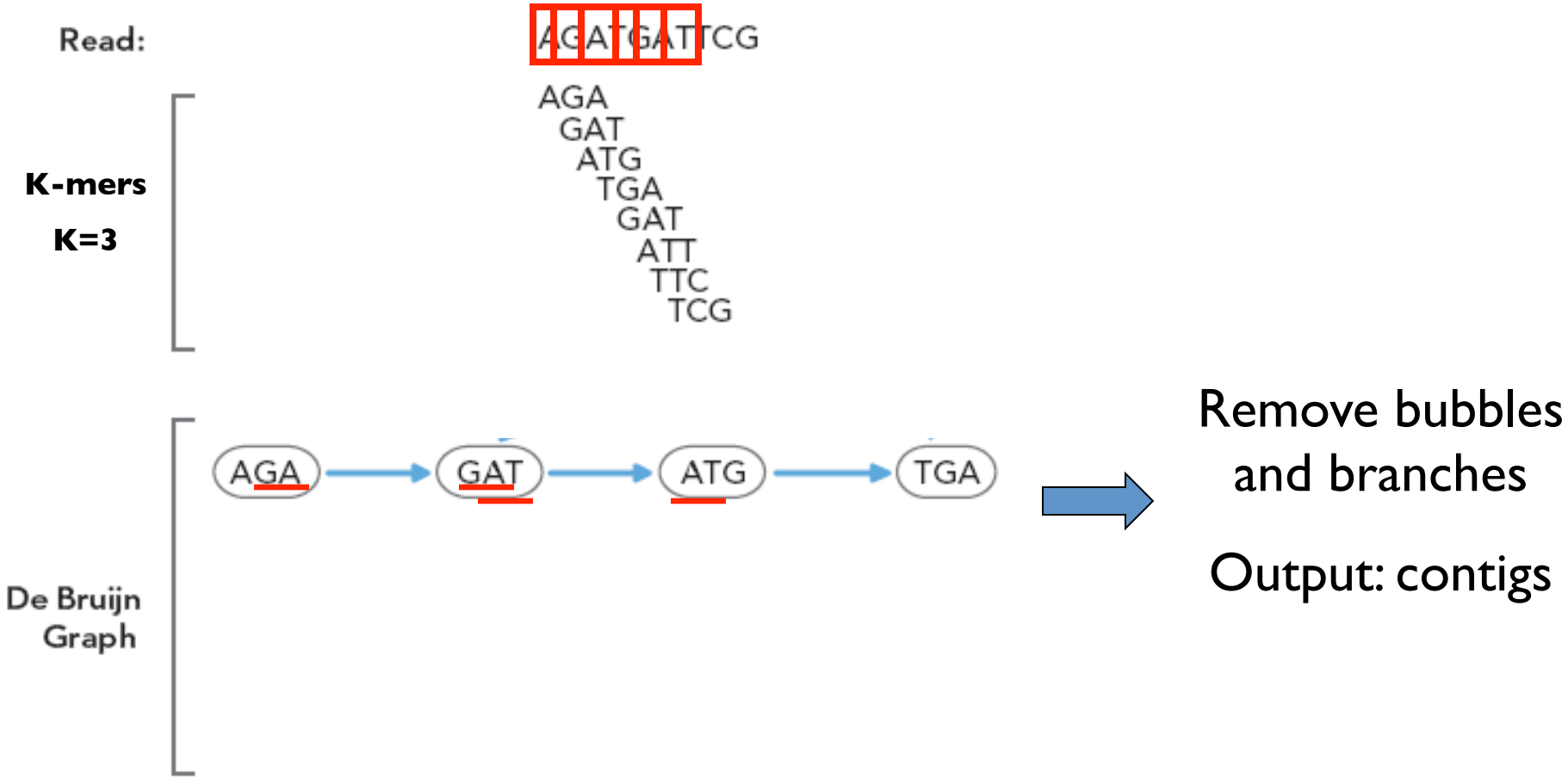
- De novo assembly—ABySS Assembler
  --Assembly By Short Sequence
  --Assembly basis: de Bruijn graph

Differences Between Genome and Transcript Assemblies

- Transcript have a large dynamic range of abundances



De Novo Transcriptiome Assembly with ABySS



The k-mers are connected if the overlap is k-1=2

Blue arrows indicate the order of the k-mers and their overlaps

De Novo Transcriptiome Assembly with ABySS

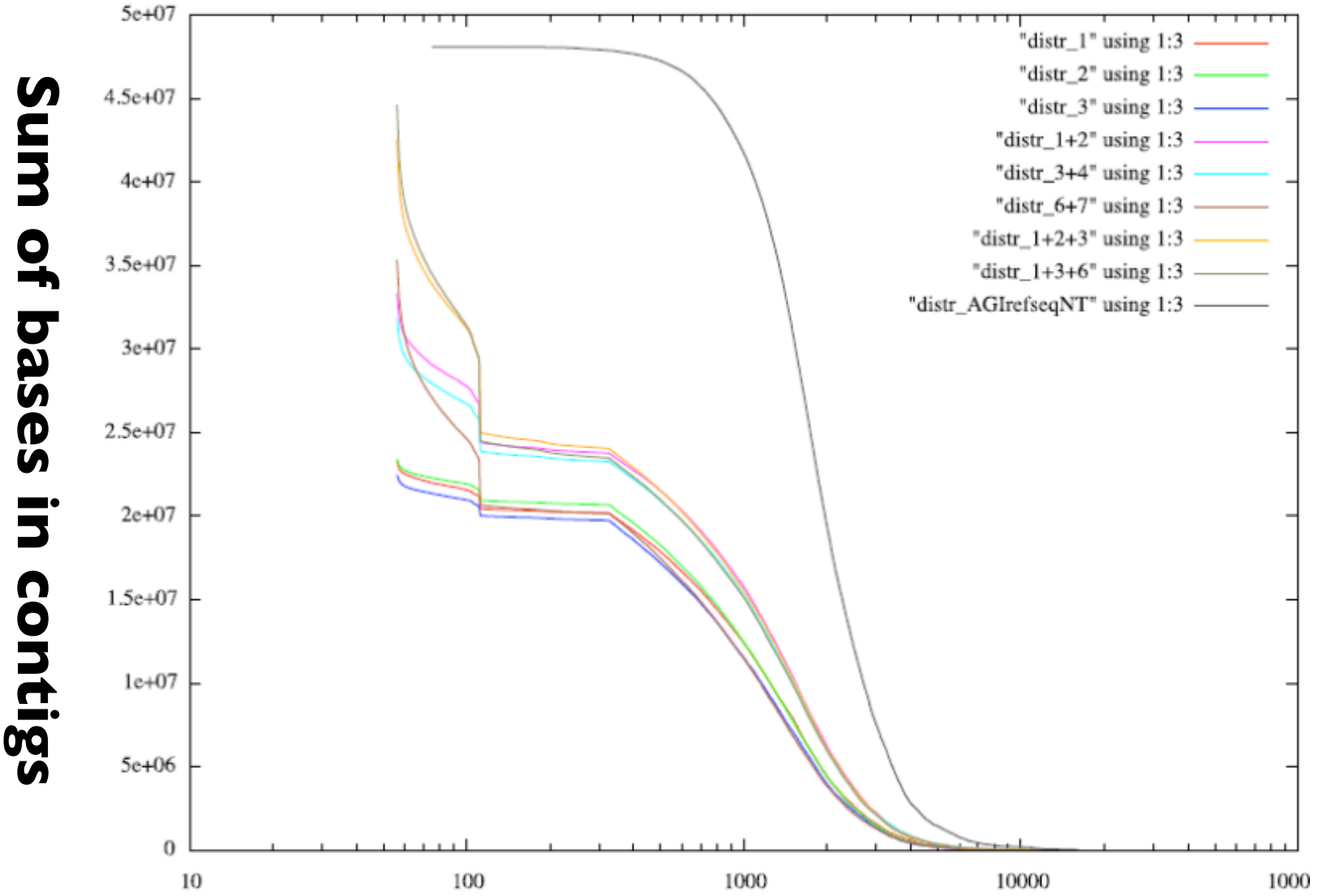
- Generated RNA-seq library from Arabidopsis flowers
- Sequenced 20 million reads, 100 bases long
- Reads were assembled using ABySS

ABySS: Parameter Search for Optimal *k* value

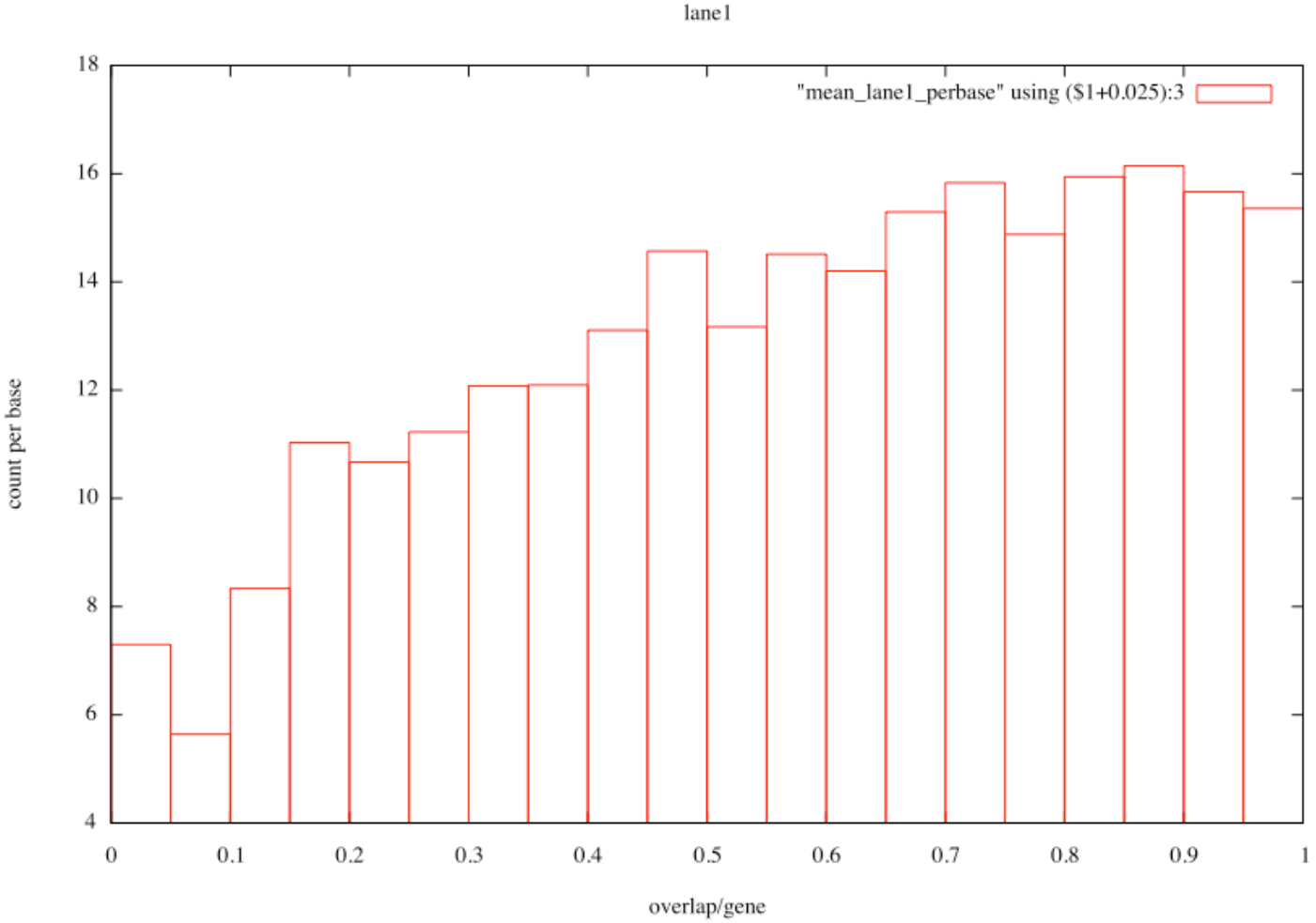
Assembly	Birol et al	Arabipodisis txscriptome (14 mill 100mer reads)							
k-mer value	28	28	61	**60**	**59**	**58**	**57**	**56**	55
#contigs	812,300	1,700,453	40,603	42,365	45,491	47,319	51,115	54,541	59,672
#contigs >= 100	95,080	37,352	29,074	29,468	30,153	30,347	30,891	31,548	32,467
#contigs >= N50	N/A	8,621	5,828	5,818	5,805	5,771	5,775	5,780	5,803
median (bp)	N/A	170	503	498	485	483	475	463	447
mean (bp)	N/A	249	707	703	692	690	680	669	653
N50 (bp)	481	308	1,106	1,116	1,131	1,143	1,148	1,154	1,155
max (bp)	7,386	3,495	8,539	11,911	11,911	11,911	11,911	11,911	8,373
sum (Mbp)	29.0	9	21	21	21	21	21	21	21

- Stats for contigs >= 100bp (except #contigs)
- N50: contigs of size >= N50 make up 50% of assembly's bases
- Opted for k = 56 because highest N50, max contig, & total Mbp

ABySS：Assembly effienciency improved when adding more reads



How Much Coverage do We Need to Generate Full Length Transcripts?



15 counts per base are sufficient to assemble full length transcripts for most genes

ABySS: Assembly Coverage

– To determine ABySS assembly quality
  - Aligned contigs to TAIR mRNA ref seq w/ BLAST
  - Perl script to calculate coverage: only alignments w/ 98% identity (2% MM & 0 gaps)

BLAST of Contigs against Refseqs

#Contigs	Qual Hits	Low Qual Hits	No Hits
31,548	27,530	965	3,053

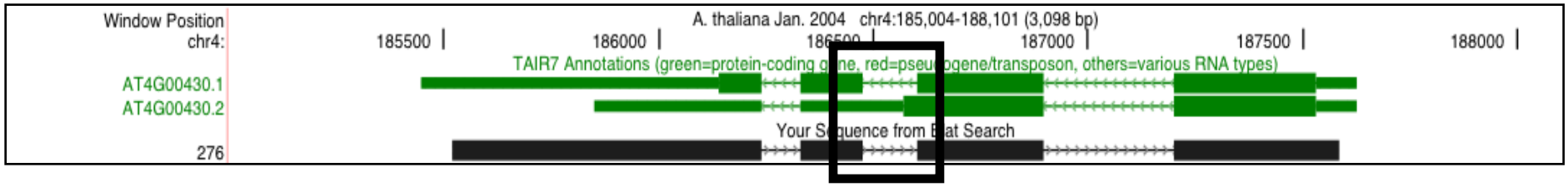
Coverage	Total	Covered	%Covered
Queries	31,548	27,530	87%
Bases	21,124,429	19,575,778	93%

BLAST of Refseqs against Contigs

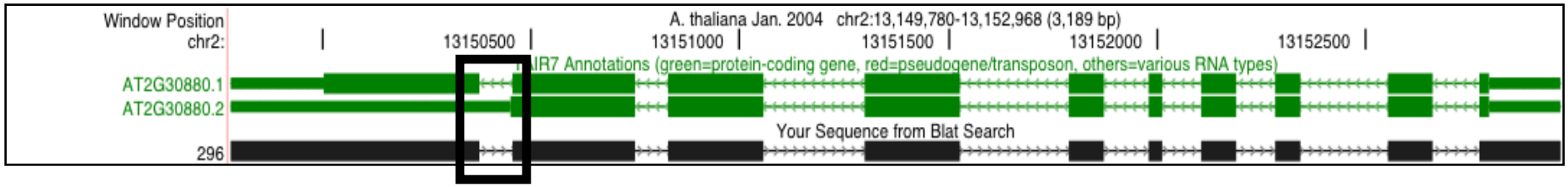
#Refseqs	Qual Hits	Low Qual Hits	No Hits
31,770	19,189	5,146	7,435

Coverage	Total	Covered	%Covered
Queries	31,770	19,189	60%
Bases	48,103,124	23,494,369	49%

Example ABySS Contigs In Black

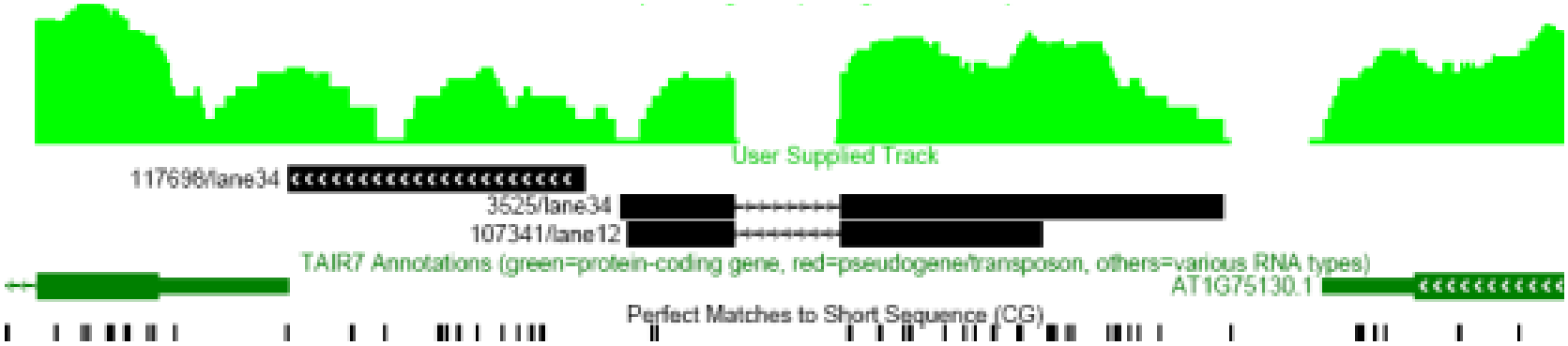


ABySS contigs often underestimate TSS and TTS



ABySS contigs only capture a single transcript

Able to Predict new genes that are not in Annotated



- Were able to identify 414 novel transcripts with no matches to existing annotation
- 90 of these had hits to protein database

Acknowledgements

- Pellegrini Lab
  - David Casero Diaz-Cano
  - Stephen Douglass
  - Darren Kessner
- Sabeeha Merchant Lab
  - Steven Karpowicz
  - Madeli Castruita
  - Janette Kropat
- Sequencing done at JGI