Capturing the chicken transcriptome with PacBio long read RNA-seq data

OR

Chicken in awesome sauce: a recipe for new transcript identification

Gladstone Institutes
Pacific Biosciences

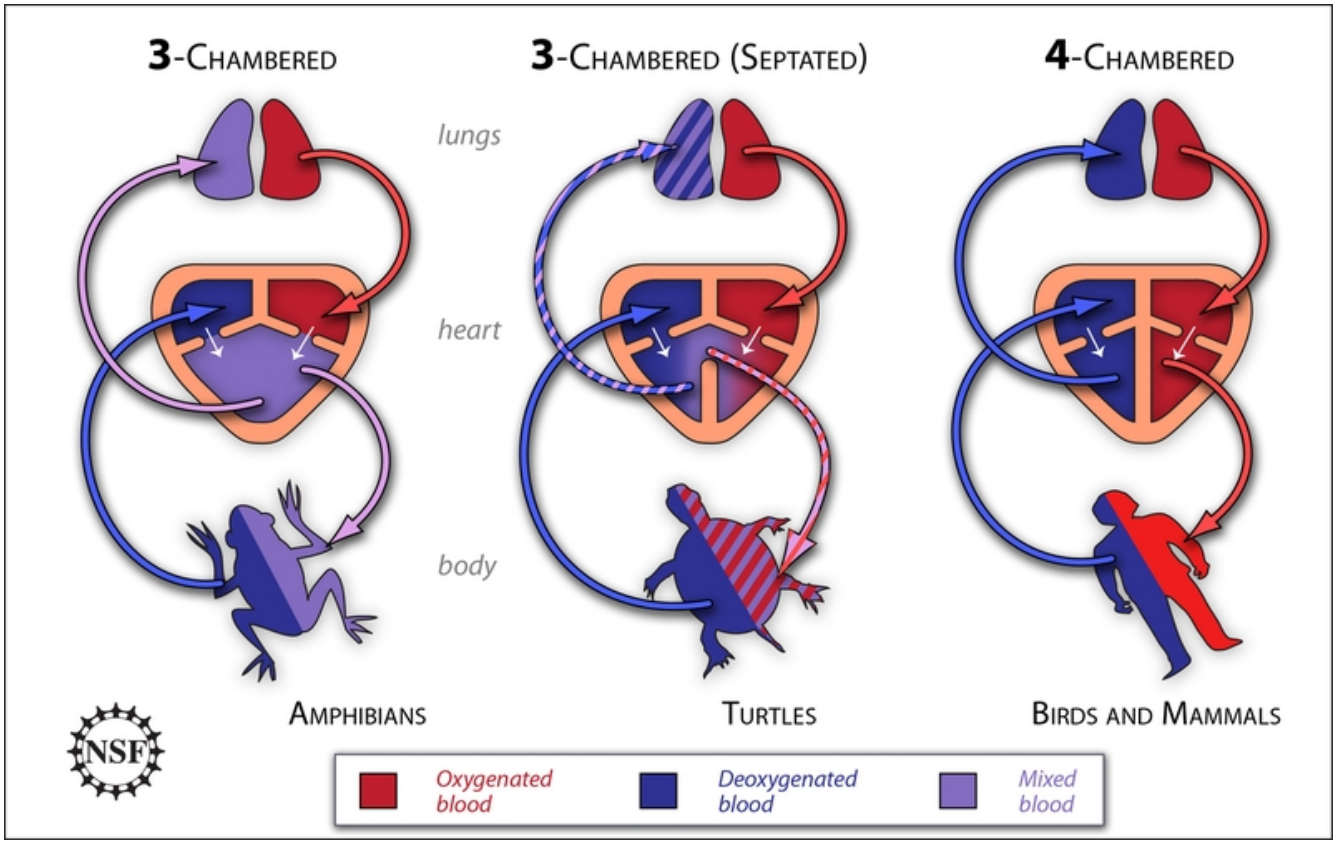
Sean Thomas and Alisha Holloway

Motivation

- Overarching goal: understand gene regulation during heart development and why children are born with congenital heart defects
- Accelerate discovery to clinical practice by fostering collaborations of basic, translational and clinical researchers
- www.benchtobassinet.org

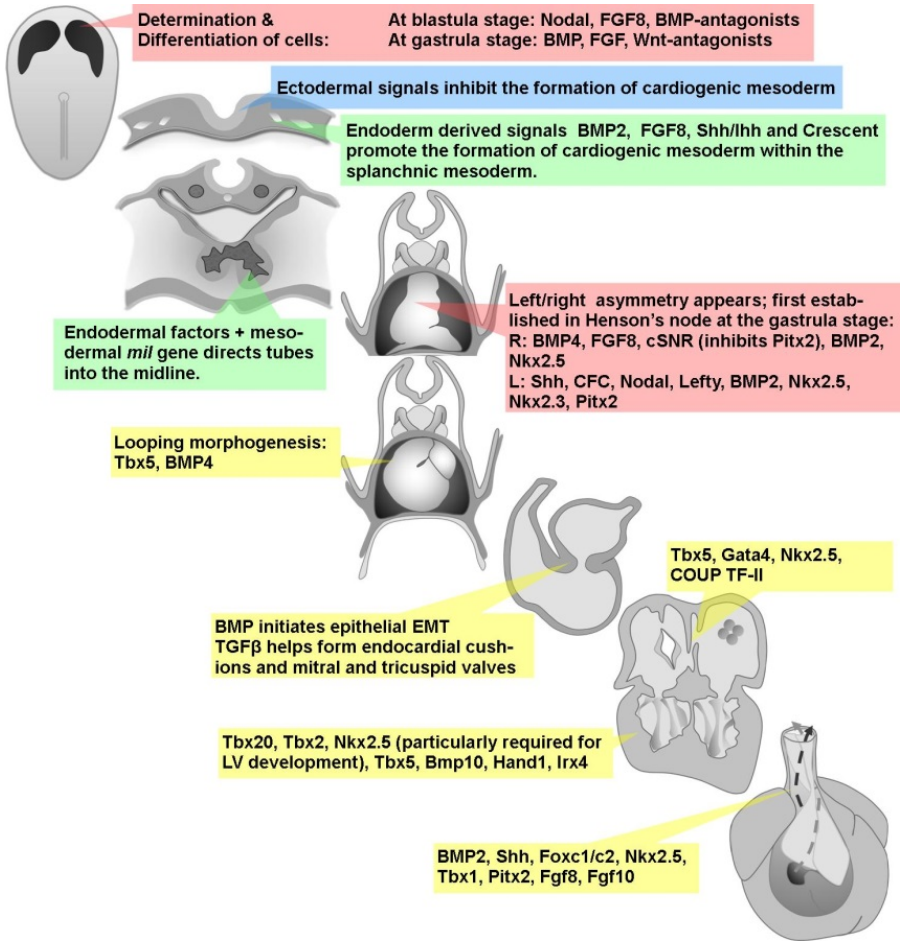
Motivation

Chicken hearts are being used as models of cardiac development



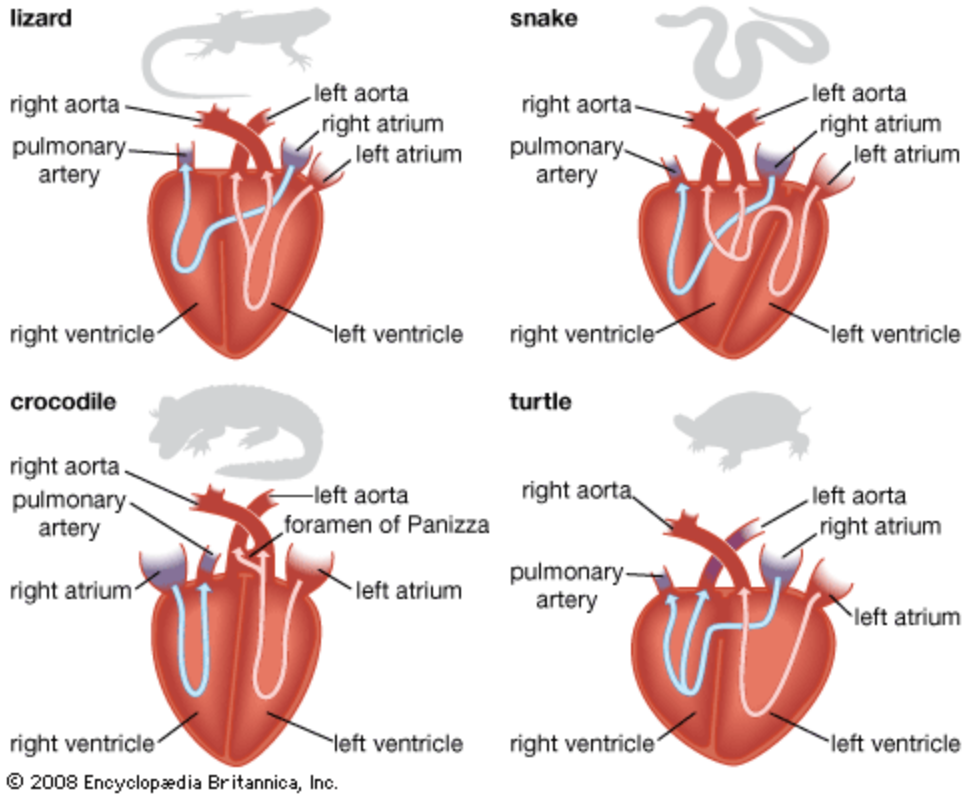
Motivation

Functional genomics studies of the molecular mechanisms behind cardiac development require solid genome and transcript annotations.



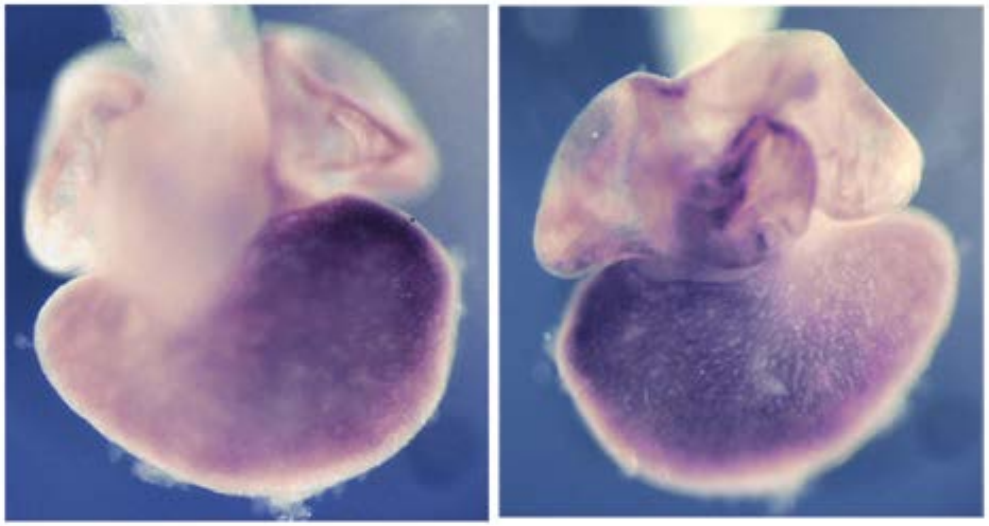
Motivation

Poor annotations are common for many model organisms that could be useful for understanding heart development and evolution



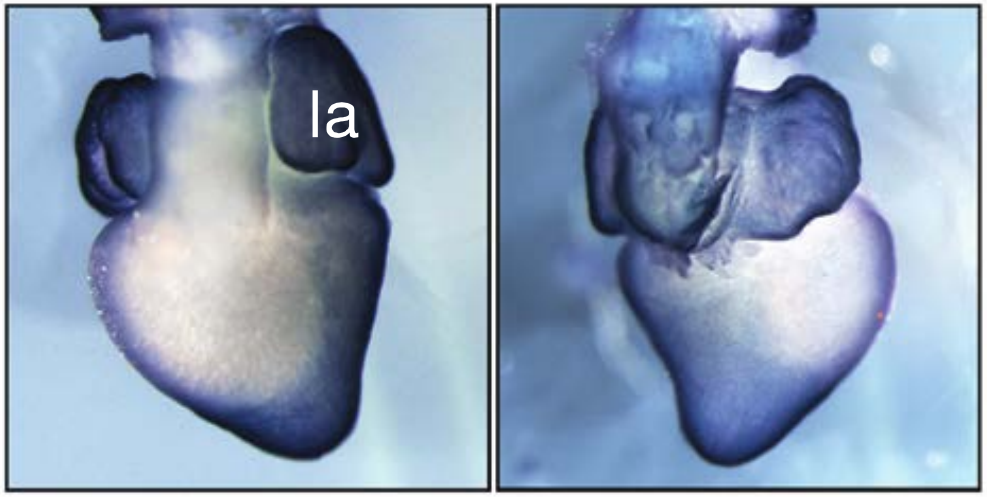
Motivation

Turtle



Tbx5
expression

Chicken



Koshiba-Takeuchi et al. 2009, Nature

Motivation

Current best chicken annotations, as of 2012*: Ensembl and refSeq

refSeq annotation contained only 6,459 transcripts, but were well-polished

Ensembl annotation contained ~20k transcripts but with many errors

mouse and chicken have similar genome sizes and numbers of genes, but Ensembl annotation for mouse has ~95k transcripts

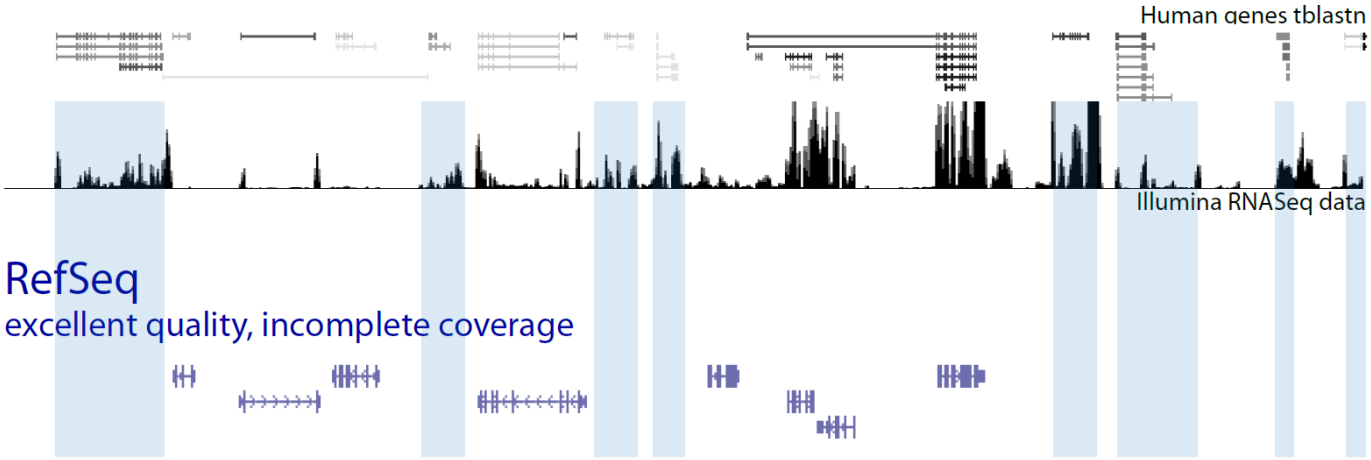
*galGal3 assembly

Motivation

available annotations were unreliable

Current annotations

Strengths and shortcomings

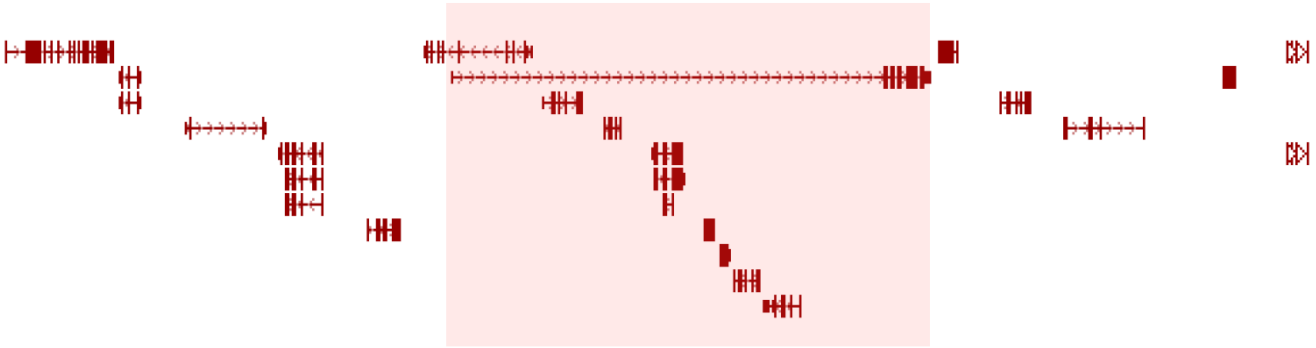


RefSeq

excellent quality, incomplete coverage

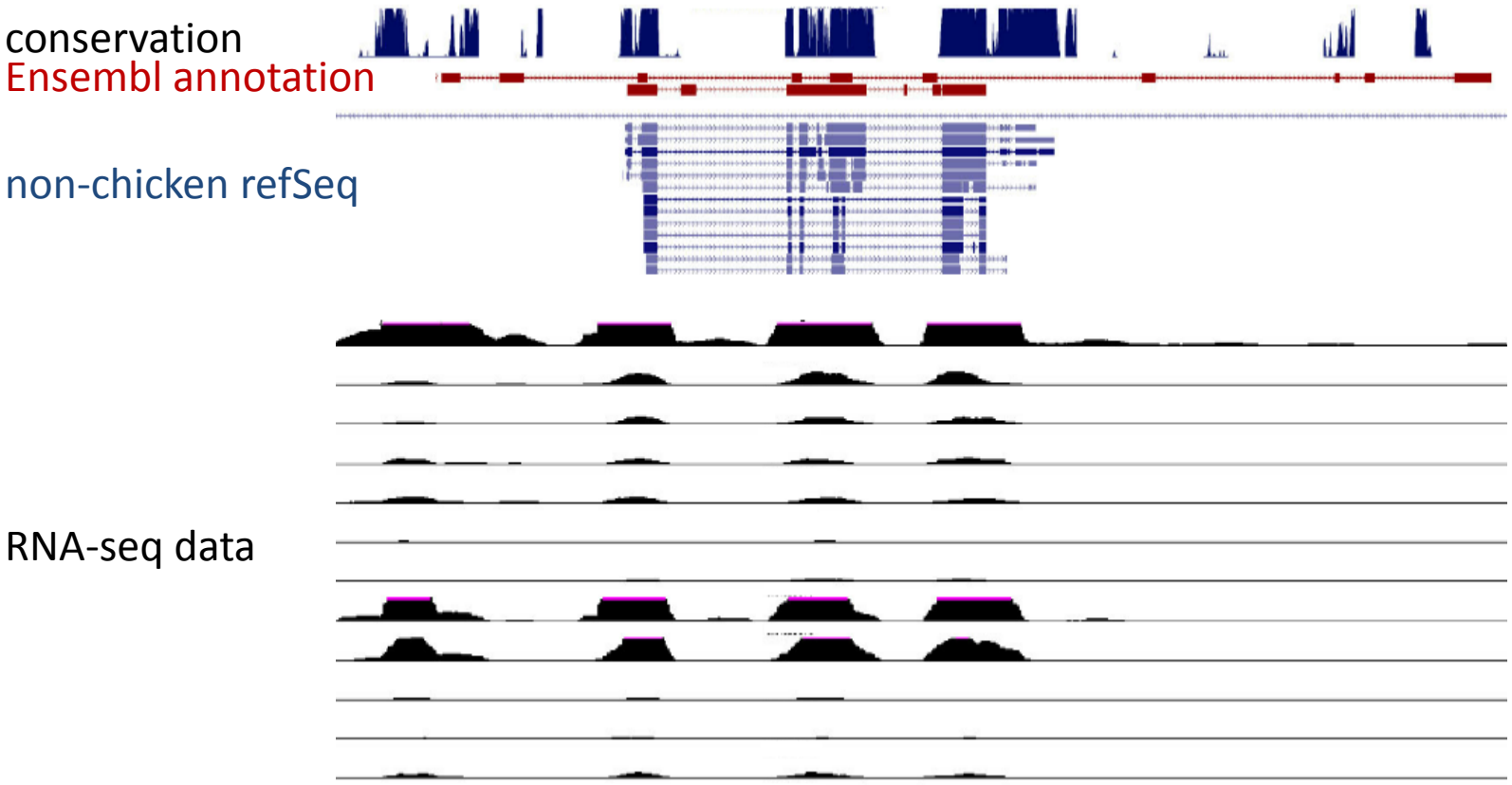
Ensembl

good exon coverage, hit-and-miss isoform assembly



Motivation

available annotations were unreliable



Znf503, a likely regulator of heart development

Solution? *De novo* transcriptome assembly of short read data and EST data

Acquired deep short read data from many tissue types

- Illumina data – 150 million uniquely mapping fragments
- Tissues –brain, cerebellum, heart, kidney, liver, testicle

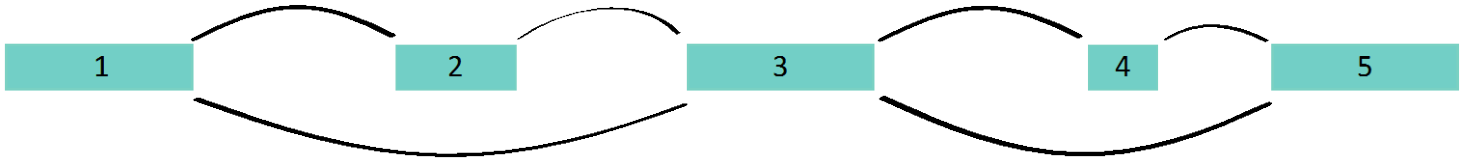
Acquired EST data from existing databases

Employed *de novo* transcriptome assembly tools to generate annotation

- Trinity
- MAKER

Solution?

Assembly of exons is possible with short reads but assembling isoforms is trickier…

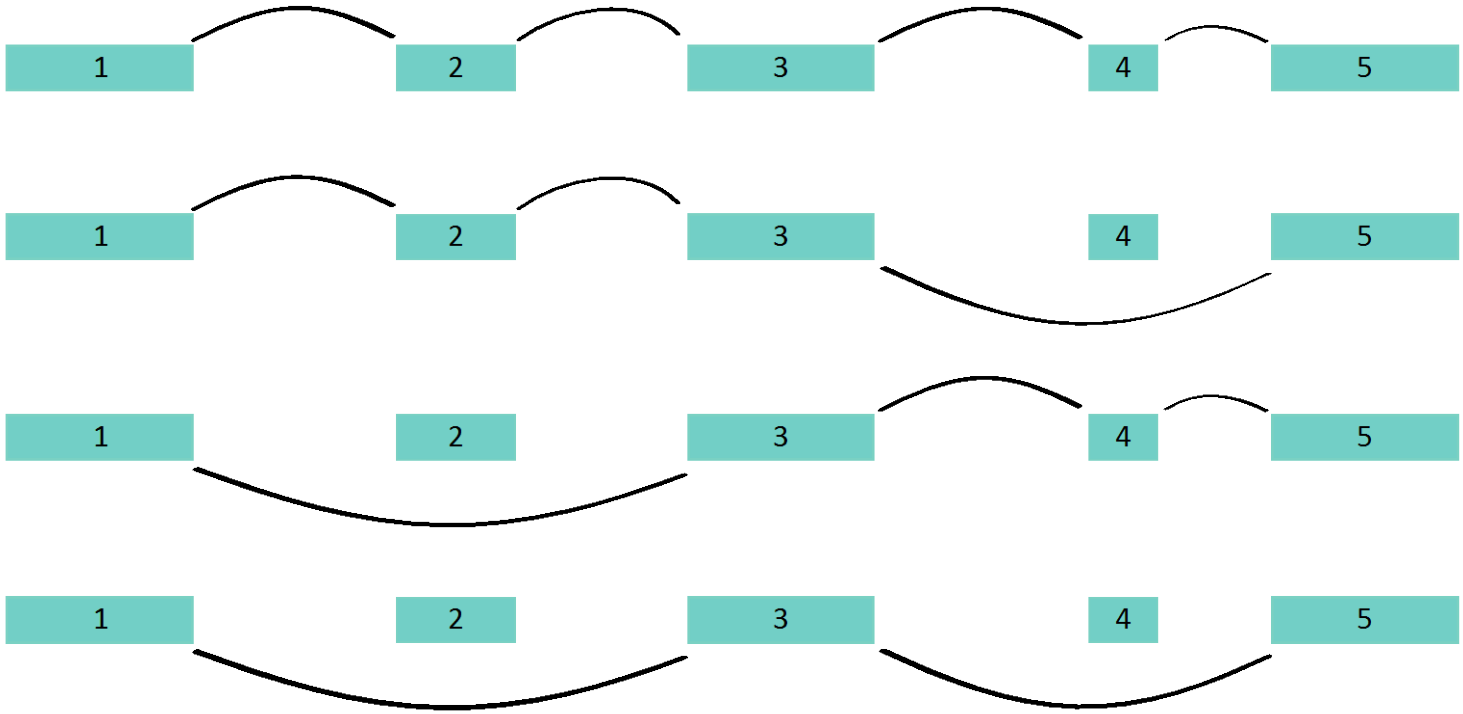


Blue boxes are exons

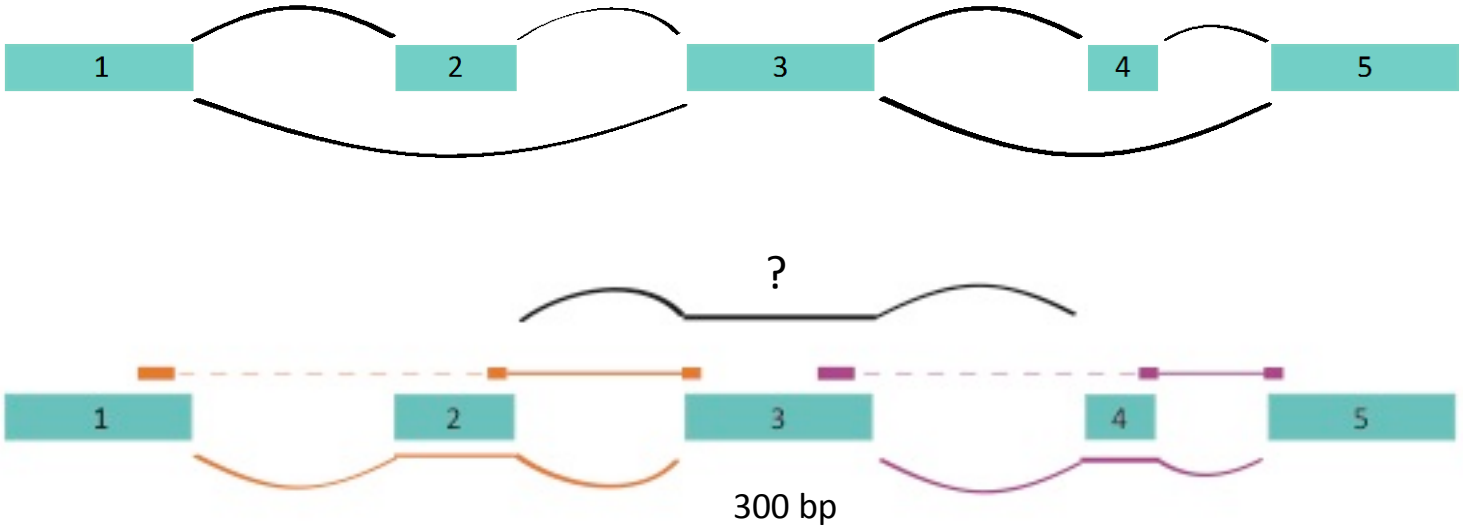
Black lines show exons joined by:

1. exon spanning reads
2. paired-end reads

Solution?



Solution?



Three exons can be joined by:

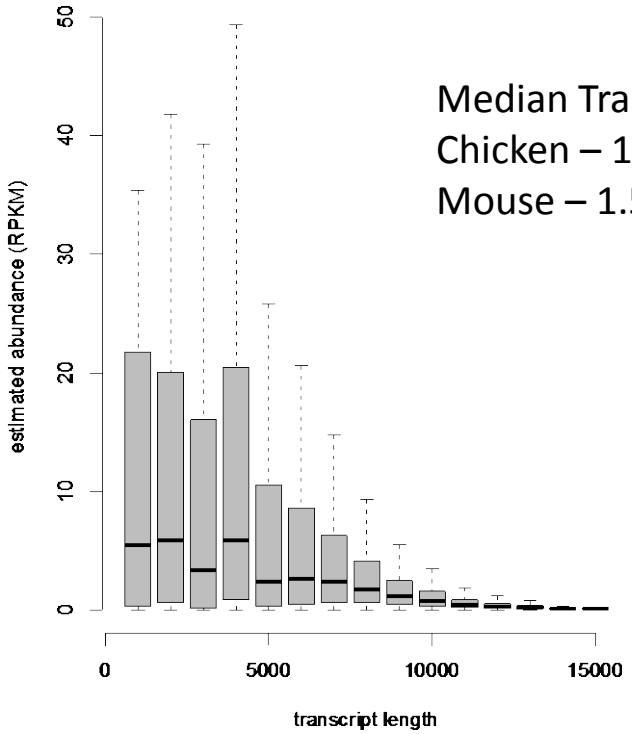
1. one end of a pair mapping to exon 1
2. other end spanning exons 2 & 3

Can’t join 2 - 3 - 4 because exon 3 longer than insert size

Solution?

Assembly of exons is possible with short reads but assembling isoforms is trickier…

Assembly of Illumina reads yielded 120k distinct contigs with average length of ~600bp, well below median transcript length



Median Transcript Length
Chicken – 1.2kb
Mouse – 1.5kb

Improving model organism genome annotations

i. sequence full length transcripts

embryonic chicken heart RNA

↓

full length cDNA

↓

PacBio SMRTBell libraries

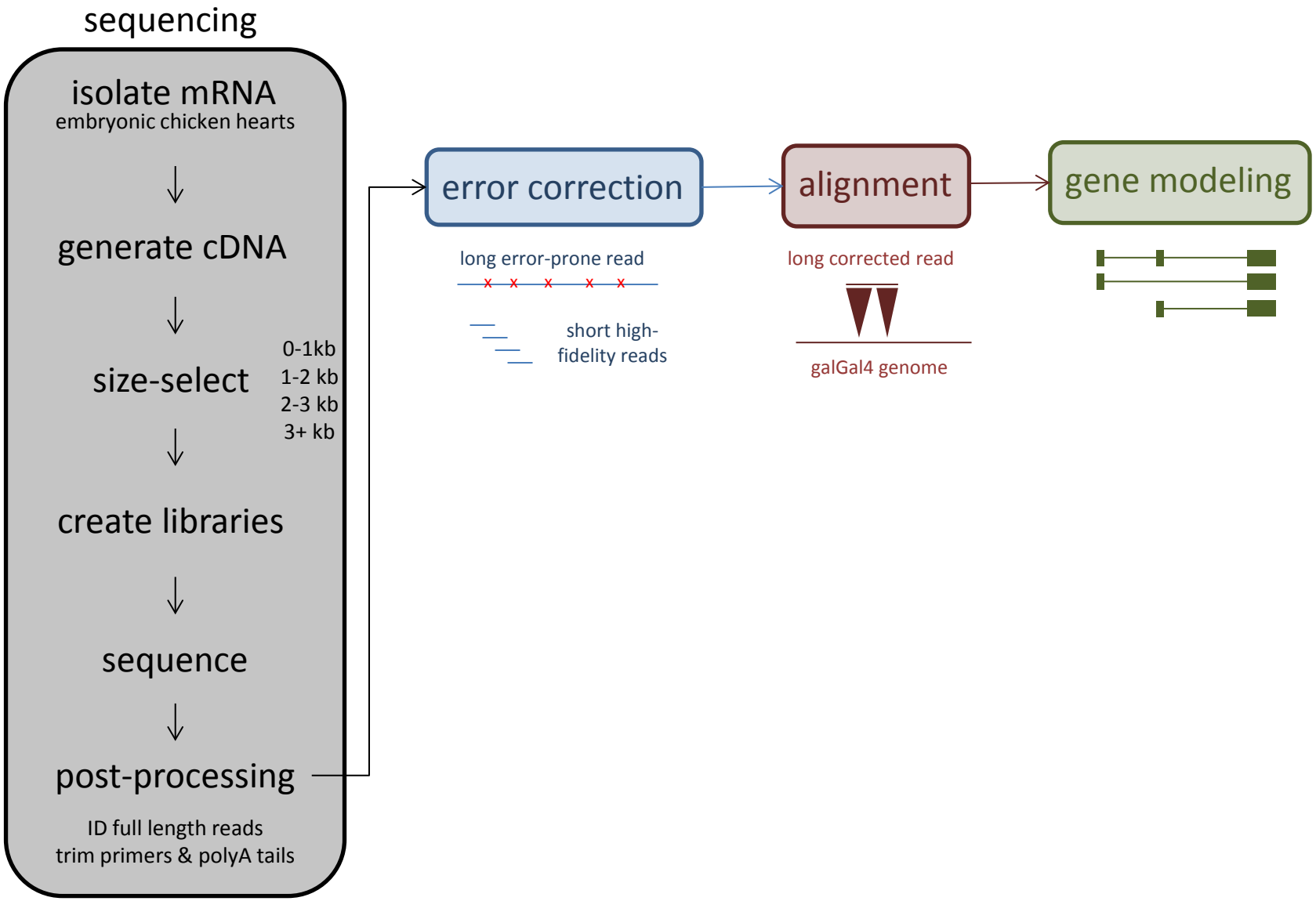
↓

long read sequencing (PacBio)

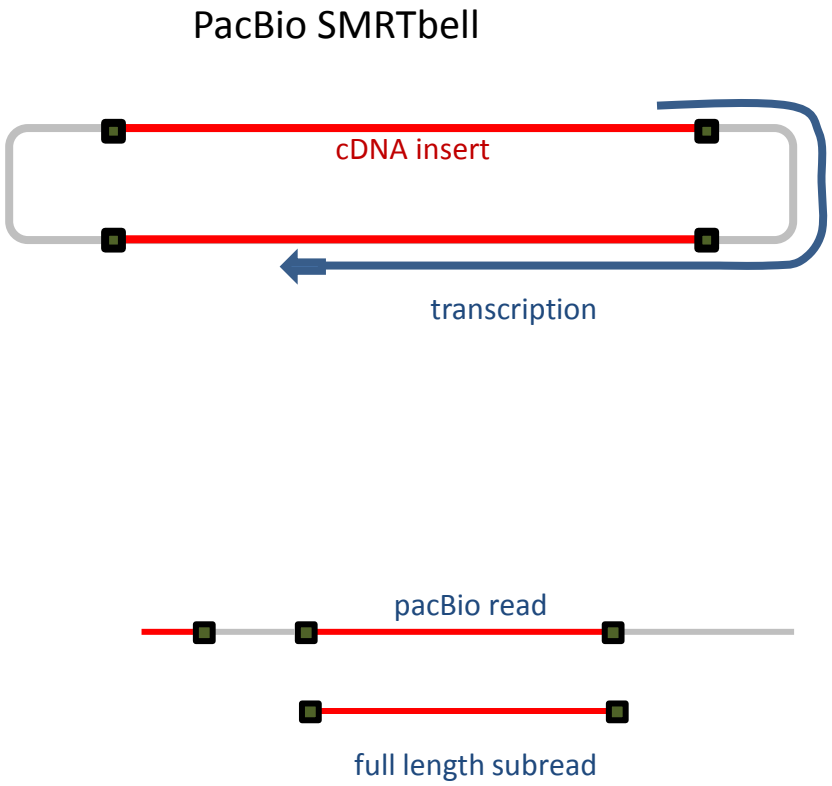
ii. compare new data to current annotations *

iii. improve existing annotations

methodology

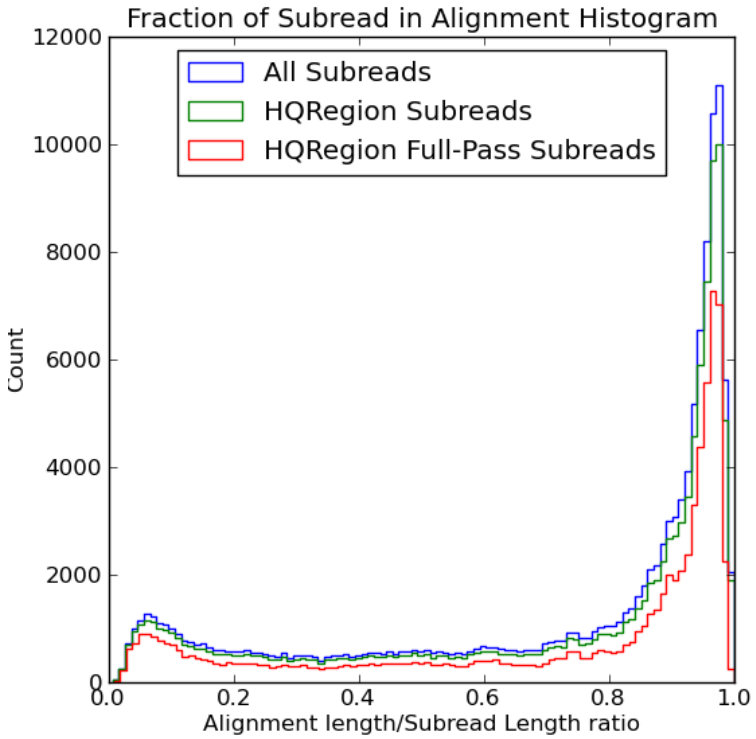


Terminology



1,508,184 subreads mapped uniquely (~72%) to galGal4 assembly

Most reads cover full length of transcript

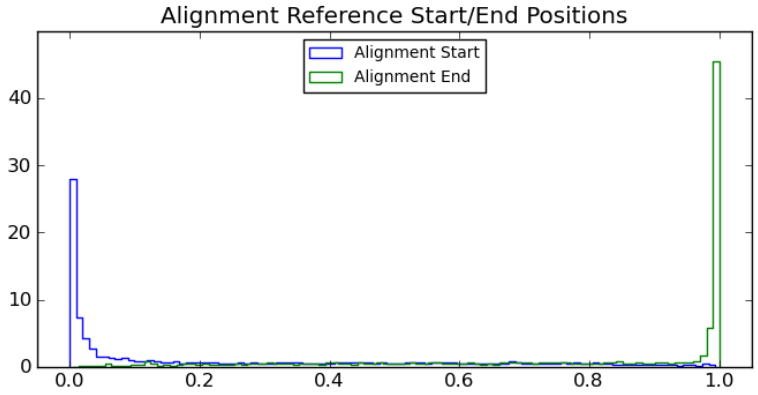


All Subreads – includes incompletely sequenced transcript

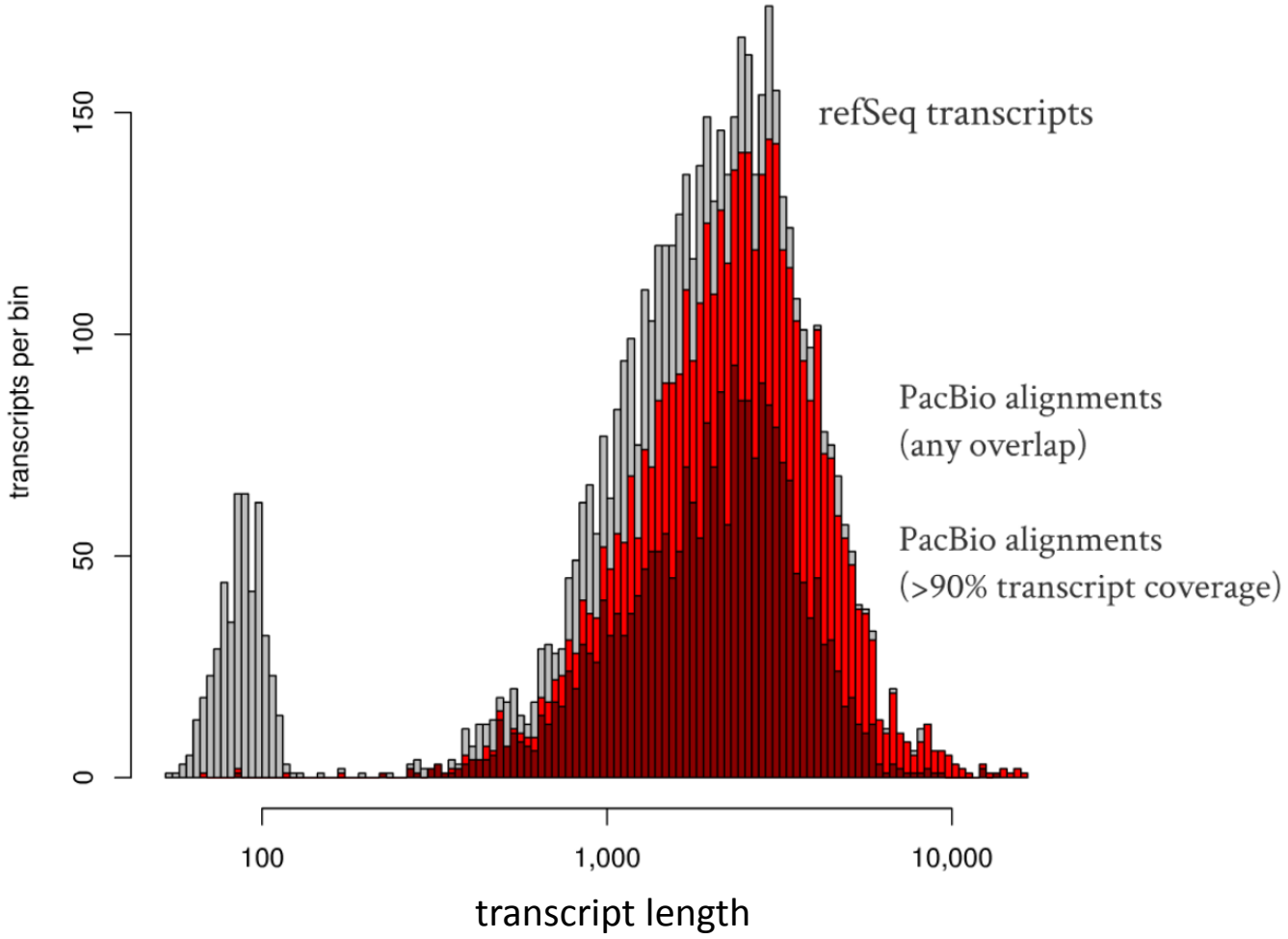
HQRegion Subread – completely sequenced >= 1x

HQRegion Full-Pass Subreads – completely sequenced >= 2x

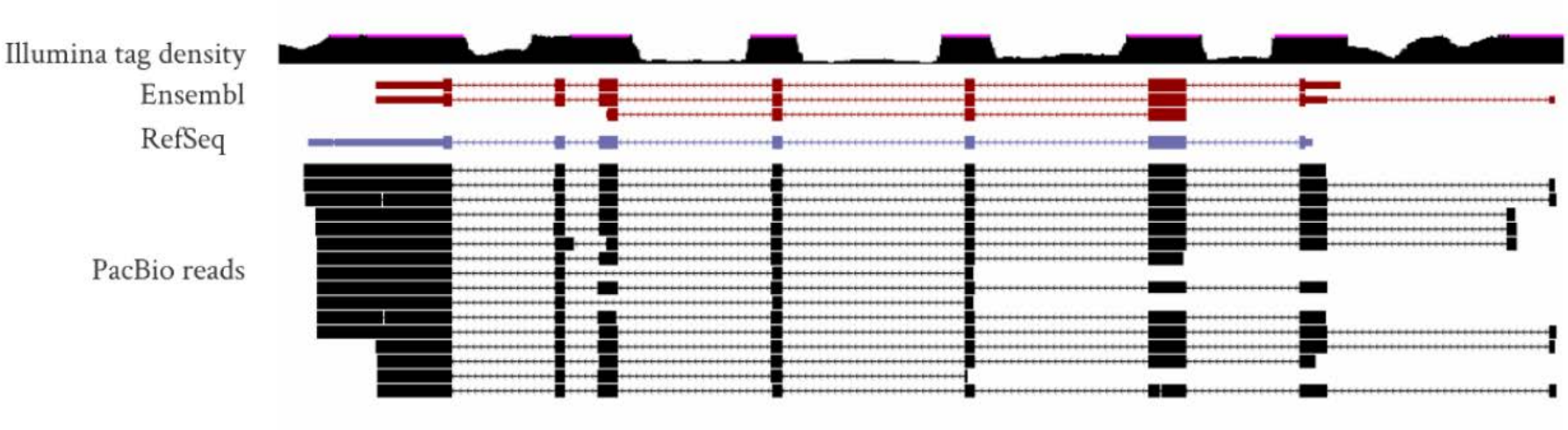
Most reads begin at the 5’ end of transcripts and end at the 3’ end



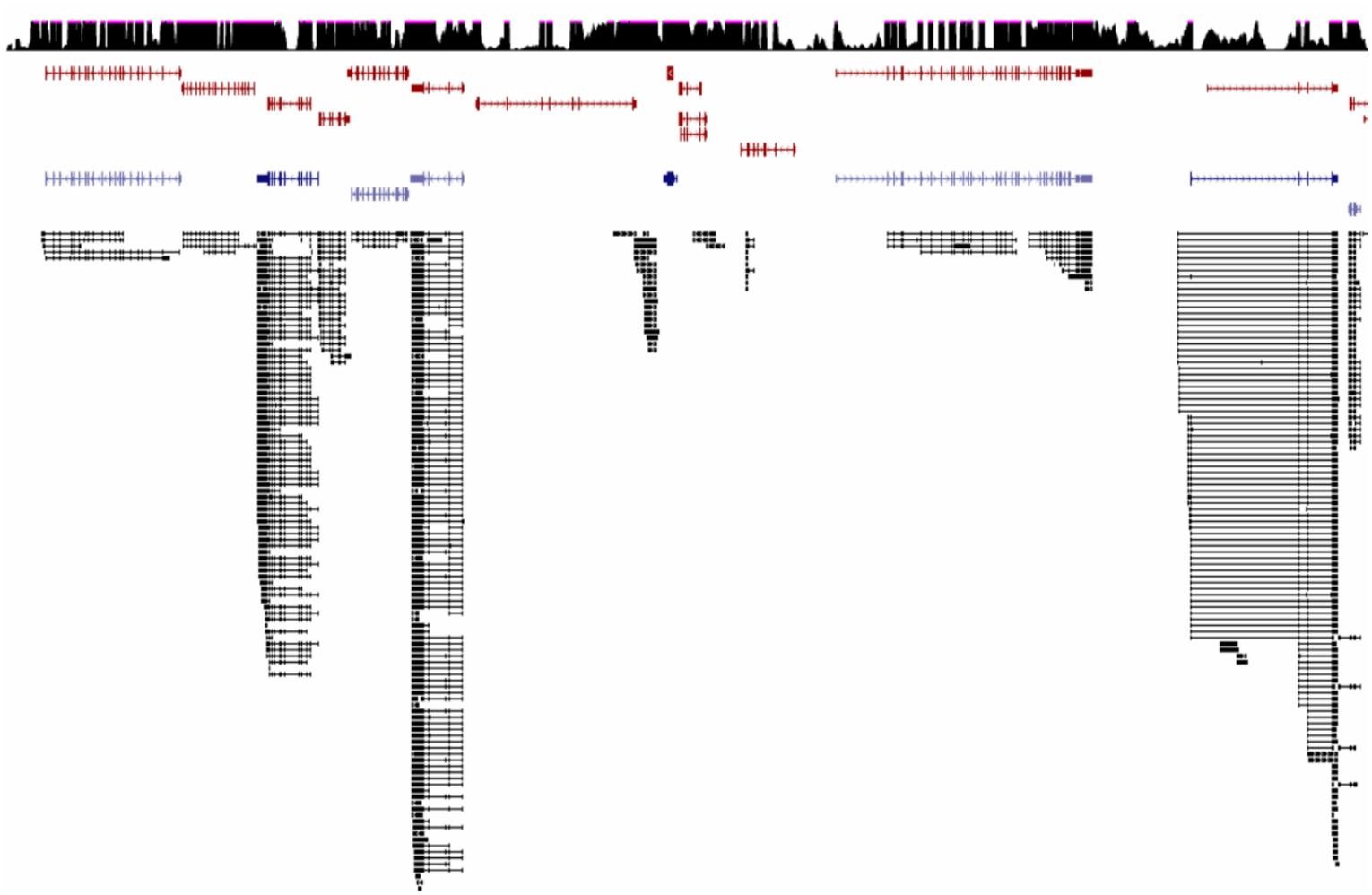
Coverage of refSeq transcripts



Example of data…good existing annotation



Example of current coverage…

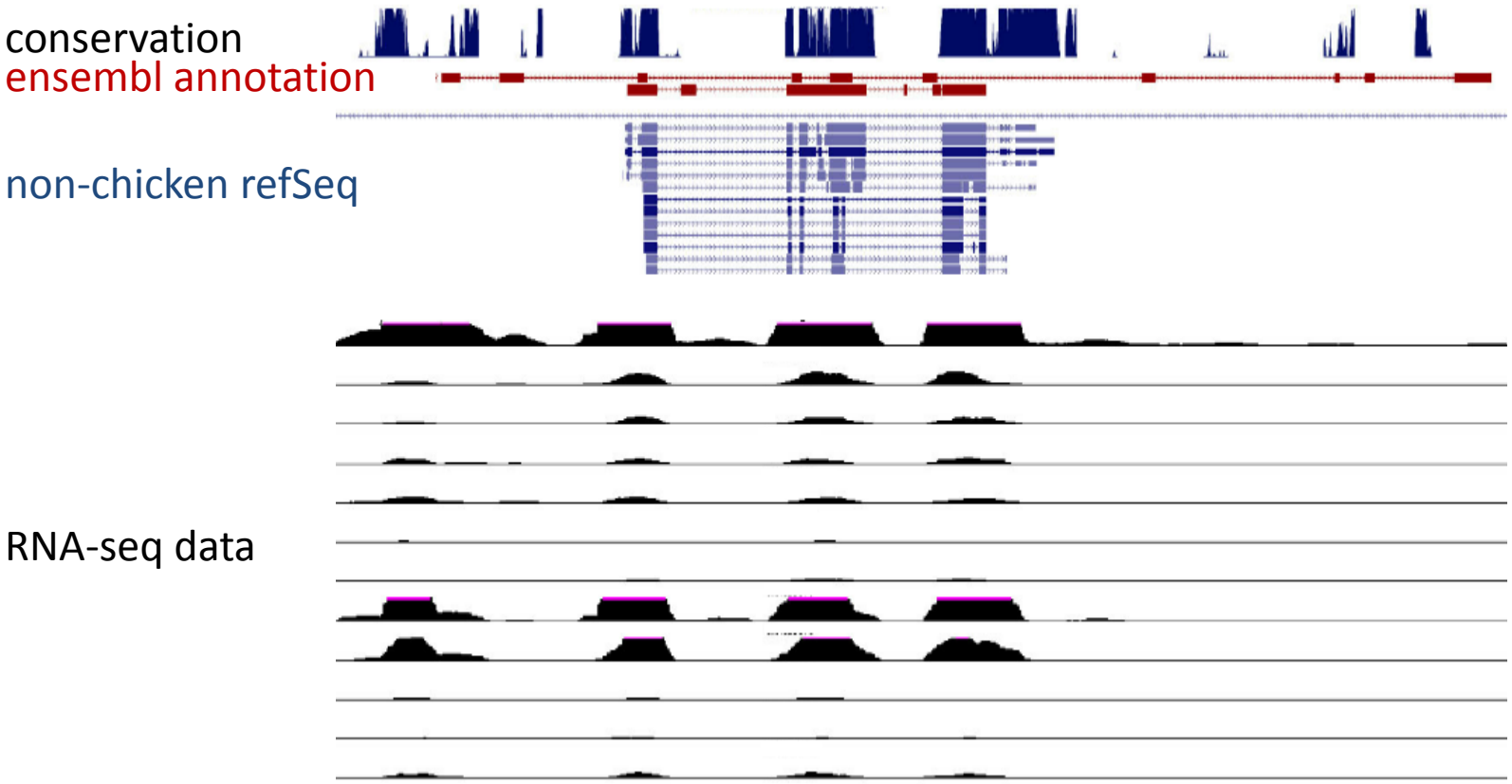


New genome assembly, new annotation

New ensembl annotation based on galGal4 fixed many of the issues that motivated our efforts

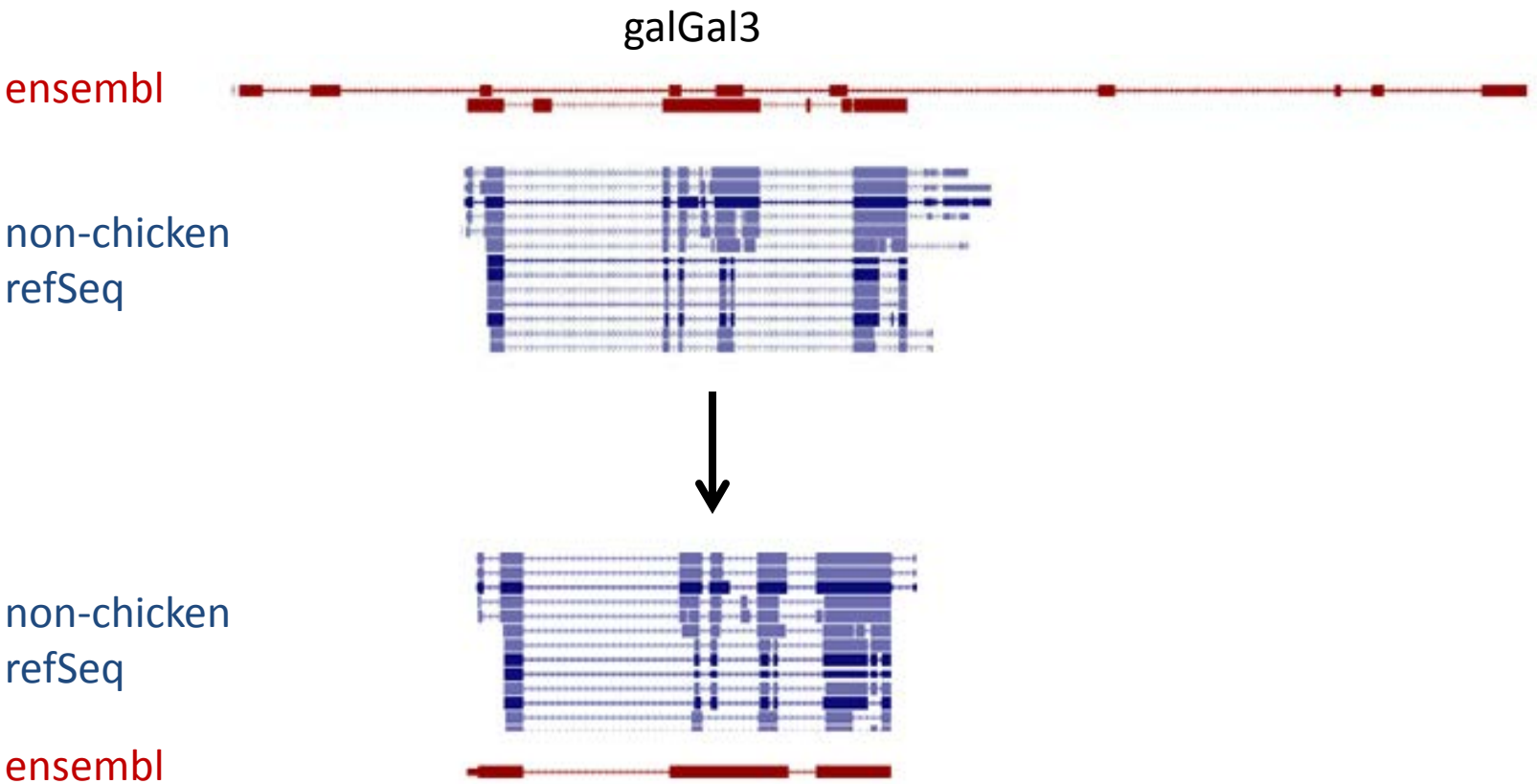
Remember this gene?

available annotations were unreliable



Znf503, a likely regulator of heart development

Annotation improvement



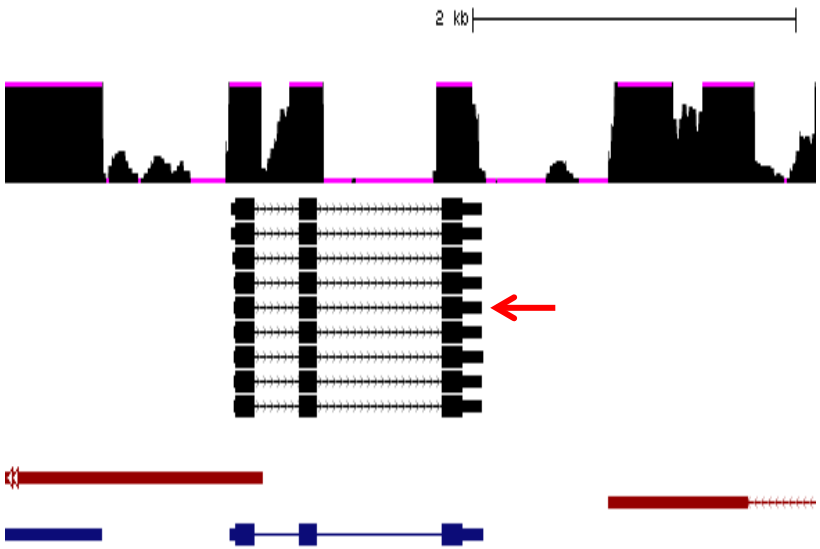
Znf503, a likely regulator of heart development

New genome assembly, new annotation

New ensembl annotation based on galGal4 fixed many of the issues that motivated our efforts

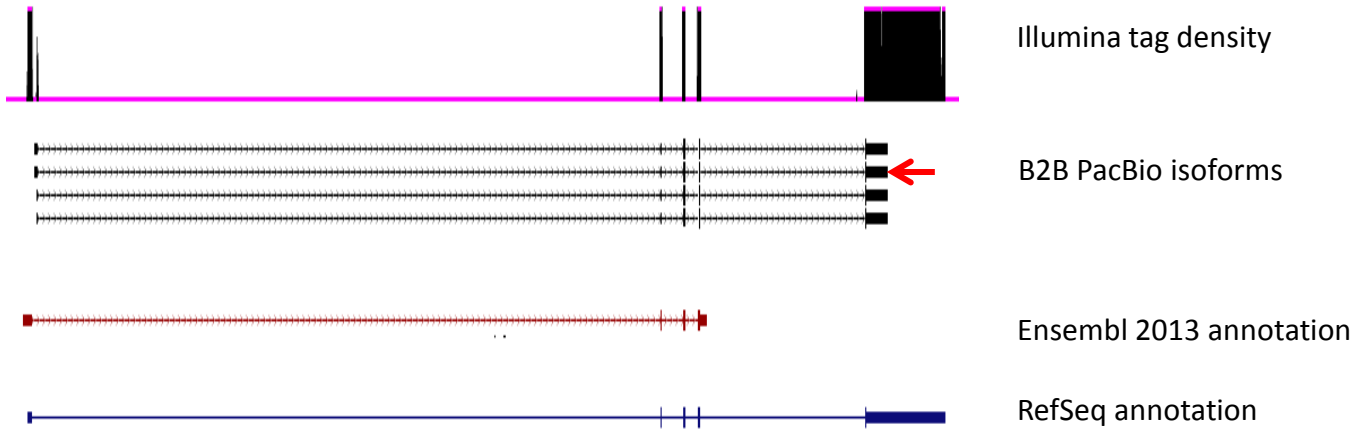
However, the PacBio data contains ~2,000 transcripts that represent improvements to even this newest annotation

Categories of annotation improvements…



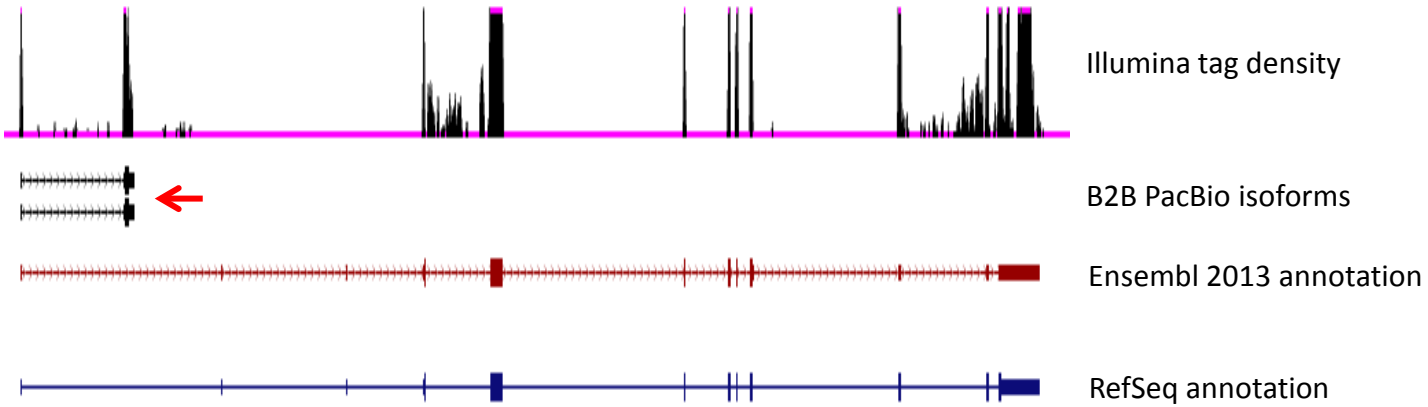
corrected genes missing from Ensembl

Categories of annotation improvements…



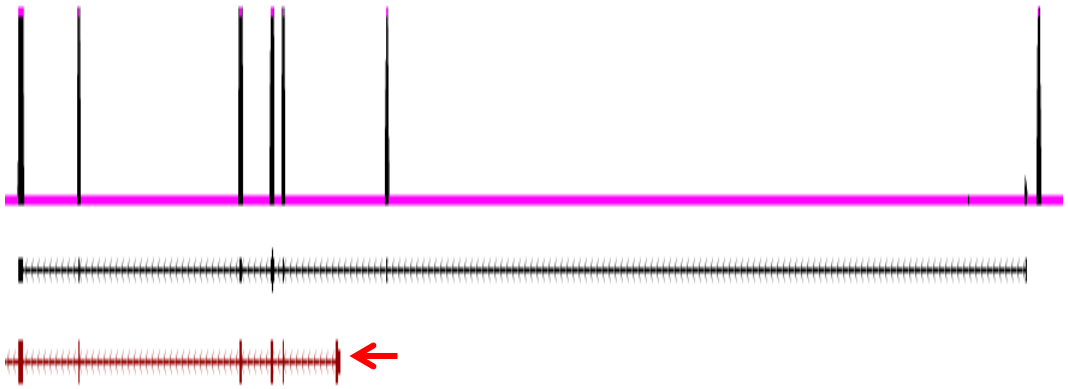
corrected exons missing from Ensembl

Categories of annotation improvements…



identify completely new isoforms

Categories of annotation improvements…



Illumina tag density

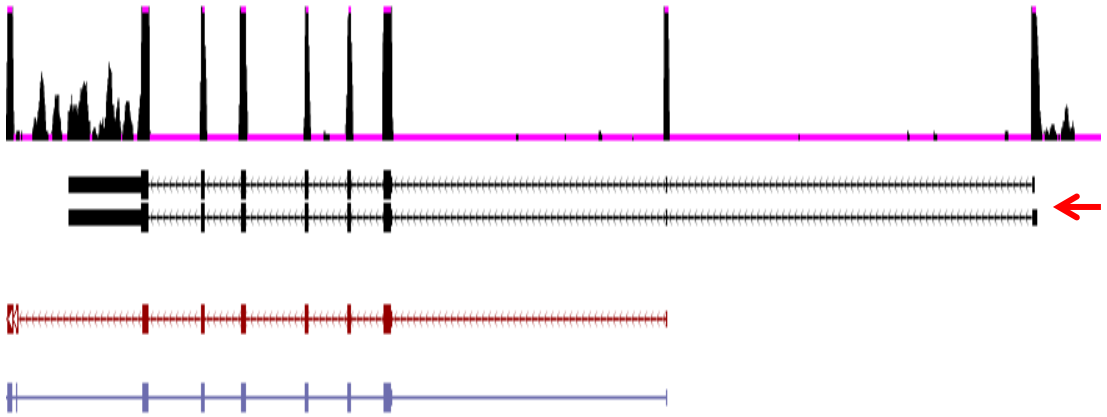
B2B PacBio isoforms

Ensembl 2013 annotation

RefSeq annotation

corrected false exons

Categories of annotation improvements…



Illumina tag density

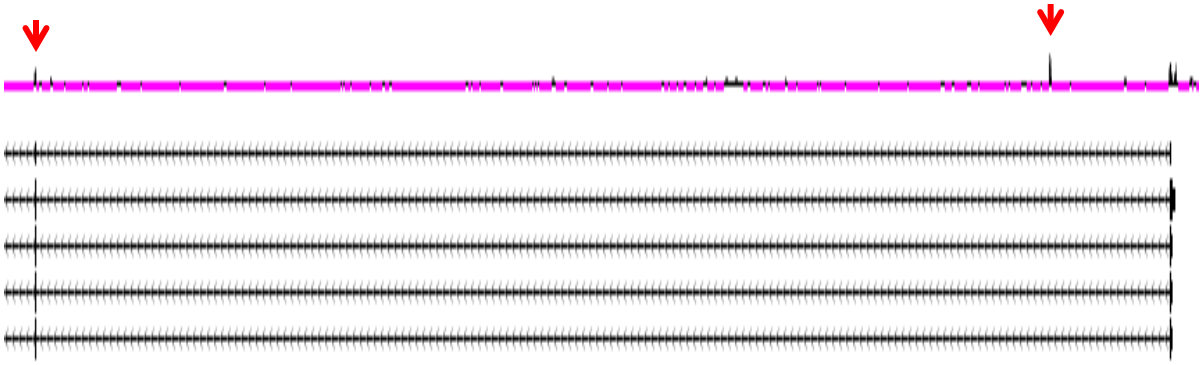
B2B PacBio isoforms

Ensembl 2013 annotation

RefSeq annotation

identify new transcription start sites

Categories of annotation improvements…

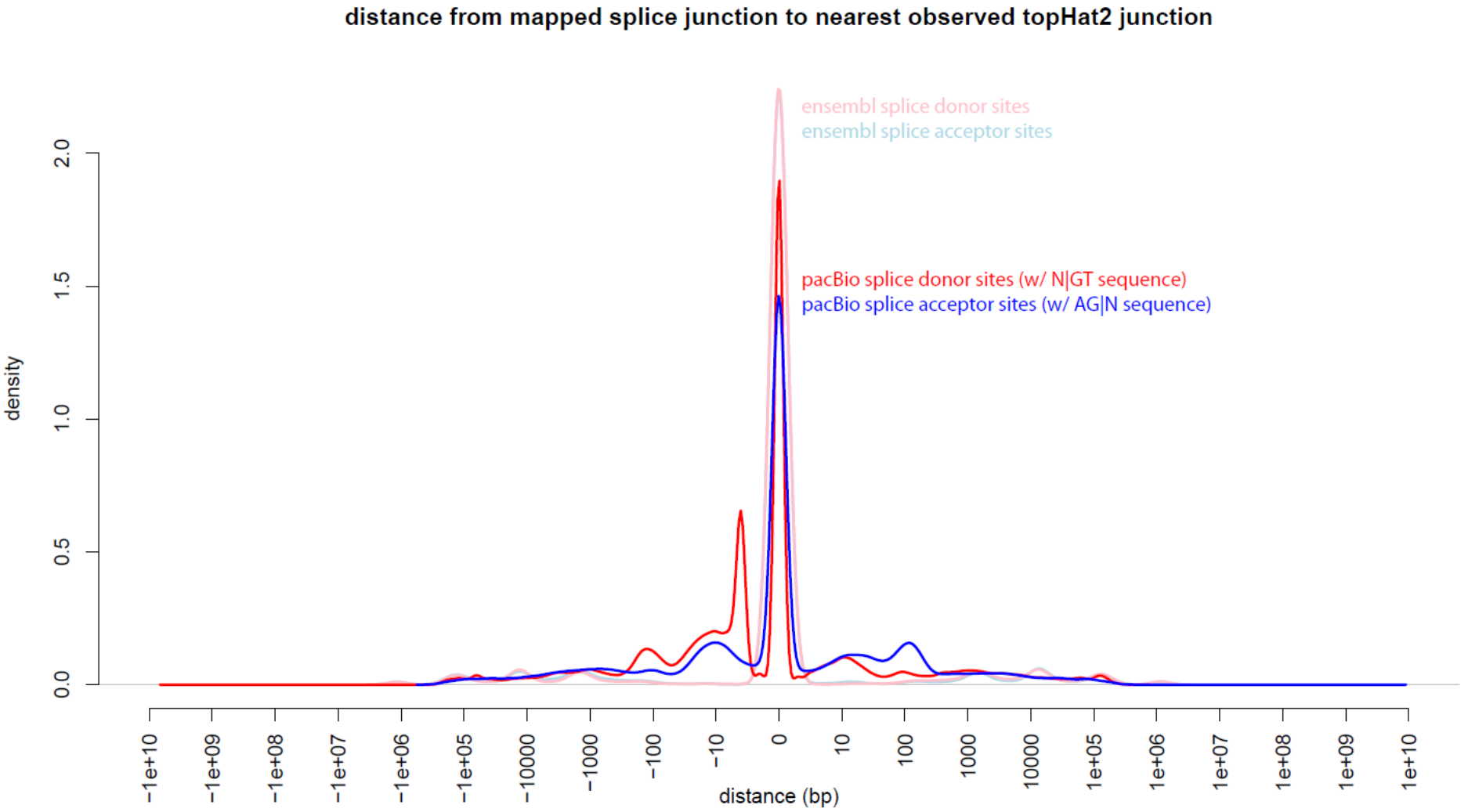


Illumina tag density

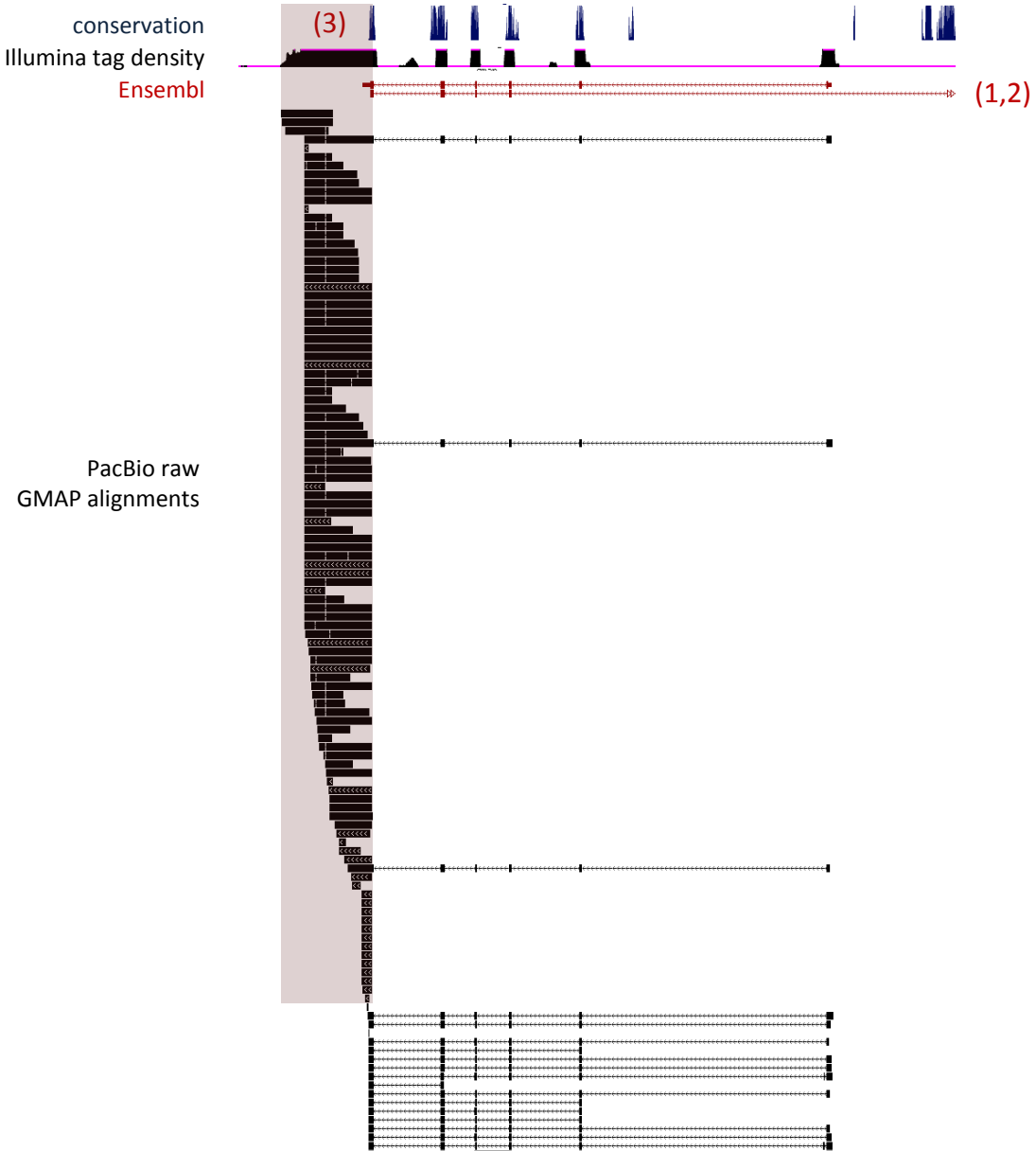
B2B PacBio isoforms

identify new low-abundance genes/exons

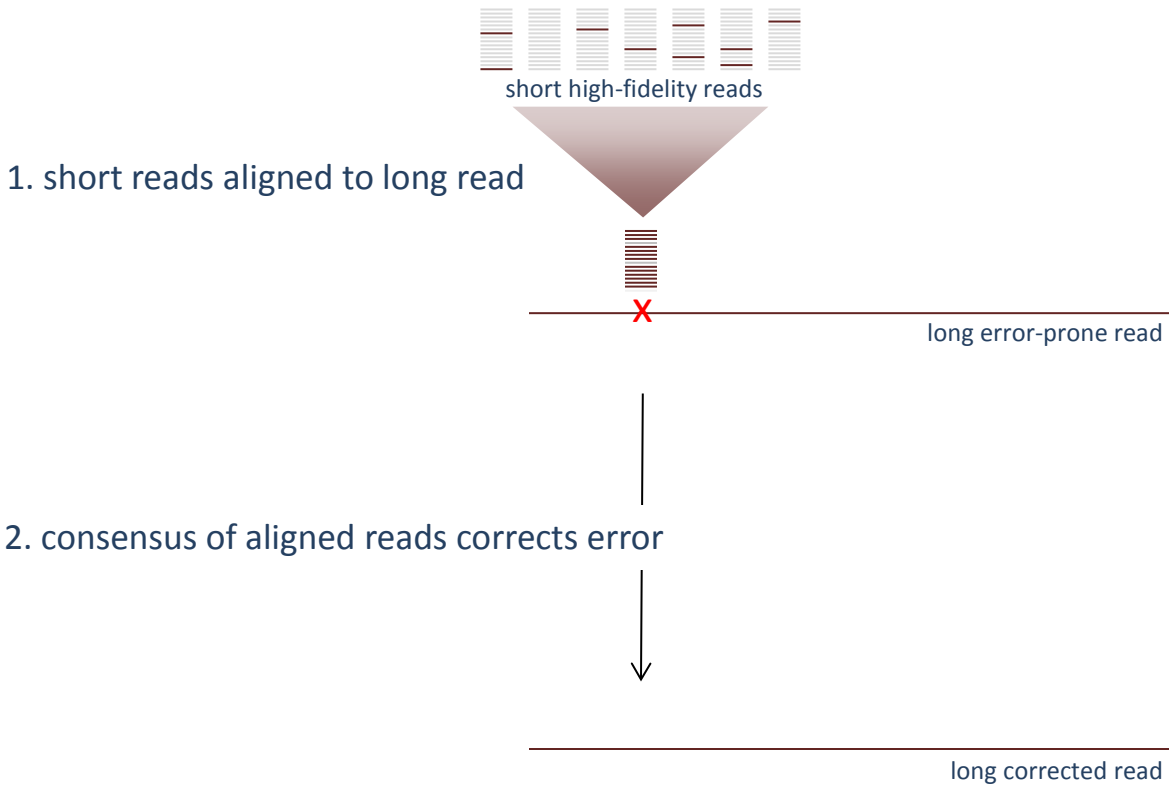
Mapped ends of PacBio reads (GMAP) exhibit systematic splice donor site errors



Peculiar buildup at 3’ end of reads…



Error correction wasn't really useful in this case (good underlying genome build)



Summary and recommendations

1. New Ensembl annotation fixed many problematic transcripts
2. **PacBio data added another 2,000 transcripts to the set expressed in embryonic chicken hearts**

Recommendations for others with similar projects

1. Select mRNAs with mature 5'cap and poly-A tail to ensure full length transcript
2. Perform normalization using double stranded nuclease to get greater coverage
3. Don't worry about error correction if you've got a good reference genome
4. Be aware of some of the systematic errors associated with mapping results

Acknowledgements

PACIFIC BIOSCIENCES®

Jason Underwood
Elizabeth Tseng
Luke Hickey

GLADSTONE INSTITUTES

Alisha Holloway
Sean Thomas