

Chicken or the egg: Iso-Seq library preparation to analysis

Richard Kuo





### Outline



- Big picture
- Library prep
  - 5' cap selection
  - normalization
- Analysis
  - Full pipeline
  - Different tools
- TAMA
  - Collapse
  - Merge
  - TAMA-GO





@GenomeRIK #tamatools



Iso-Seq Webinar:

### Big Picture



- What are you trying to find?
  - Whole transcriptome
  - Specific genes
  - Alternative splicing
  - Transcription start/termination sites
  - Rare genes/transcripts
  - Transcriptome without genome
- Need to design experiment according to your goals
  - Number of samples
  - Types of samples
  - Number of SMRT cells
  - Barcoding/multiplexing
  - 5' cap selection
  - Normalization
  - Targeted sequencing
  - Depth



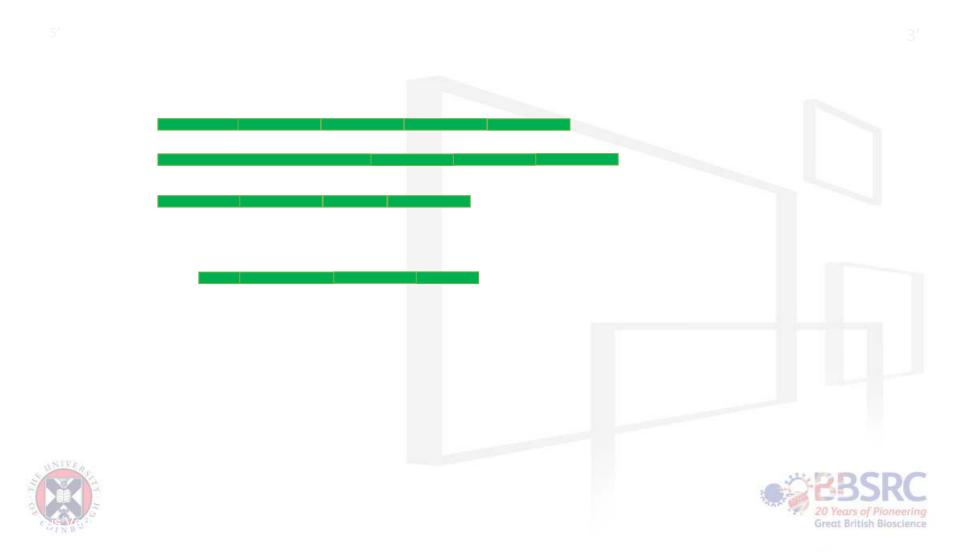
Chicken or the egg Iso-Seq planning: None of the steps come first in planning. They are all dependent on each other.





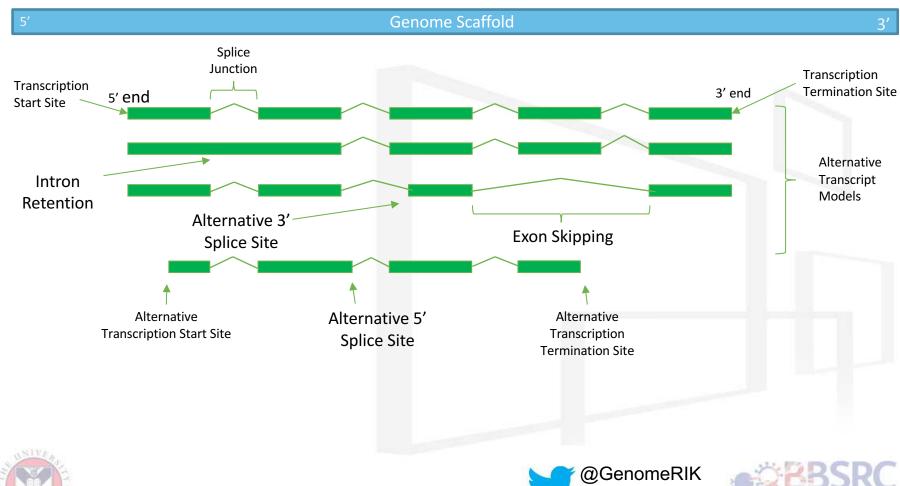
# Transcripts without genome





# Transcript Models











# Starting from nothing



- No prior information
  - No genome
  - No transcriptome
- With the least amount of starting information, you will need to do the most work to get good results
  - High depth/many SMRT cells
  - 5' cap selection
  - Short read error correction
- Harder to identify
  - Splice junctions
  - Rare genes/transcripts
  - Gene groups
  - Paralogs









# With a genome



- Only a genome assembly available
- Can map to genome assembly
- Limited to assembly quality
- Transcript model focused/reference based transcriptomes
- Only need exon starts and ends to be accurate
- Want to make a transcriptome annotation
  - 5' cap selection
  - Normalization
  - Many SMRT cells
  - Short read error correction





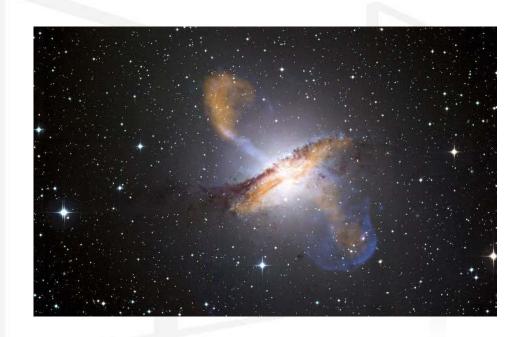




# Genome and Transcriptome



- Genome and Transcriptome
- Can map to assembly
- Limited to assembly quality
- Transcript model focused/reference based transcriptomes
- Only need exon starts and ends to be accurate
- Want to improve a transcriptome annotation
  - Depends on what you want





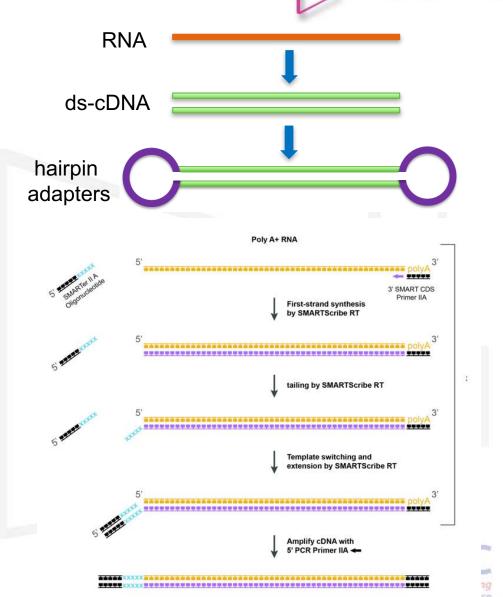




### **Standard Library Preparation**

PROSLIN

- Extract and purify RNA
- Create cDNA
  - Oligo-dT primer
  - 5' end adapter ligation
- Attach hairpin adapters
- 5' degradation?
- Overly abundant genes?



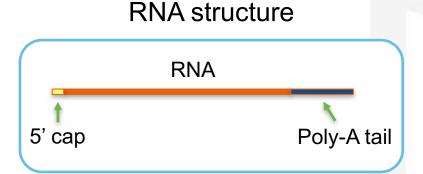


### 5' Degradation

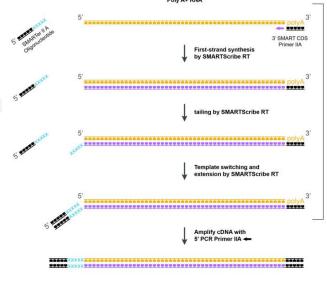


• 5' degradation?

# Standard Library Preparation



RNA from same transcript



Resulting models

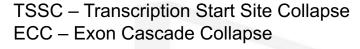


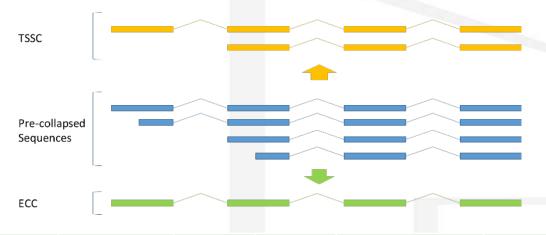
### 5' Cap Selection



Great British Bioscience

- Does 5' cap selection make a difference?
  - Collapsed using Iso-Seq Tofu Collapse tool
  - Used both methods of collapsing to compare





	Pre-collapsed	TSSC	ECC	TSSC % decrease	ECC % decrease
No Cap	199,560	80,814	55,932	59.50%	72.00%
5' Cap	11,881	9,368	8,468	21.20%	28.70%

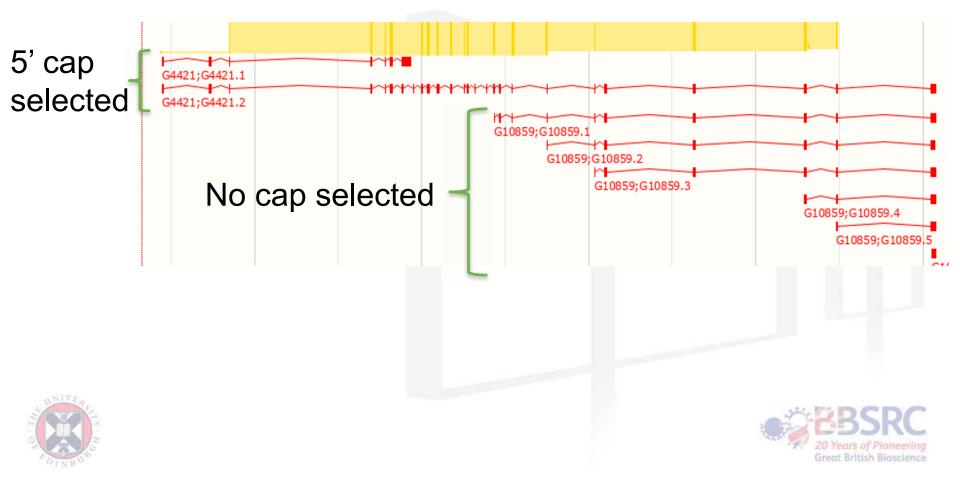


Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human anomics biomedcentral com/articles/10.1186/s12864-017-3691-

### 5' Cap Comparison







### 5' Cap Teloprime Kit



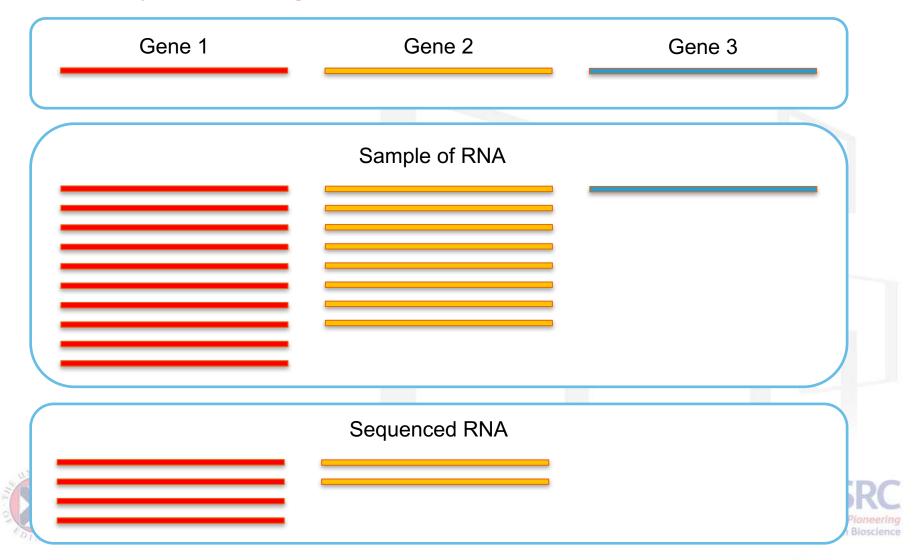




### Over abundant genes/transcripts



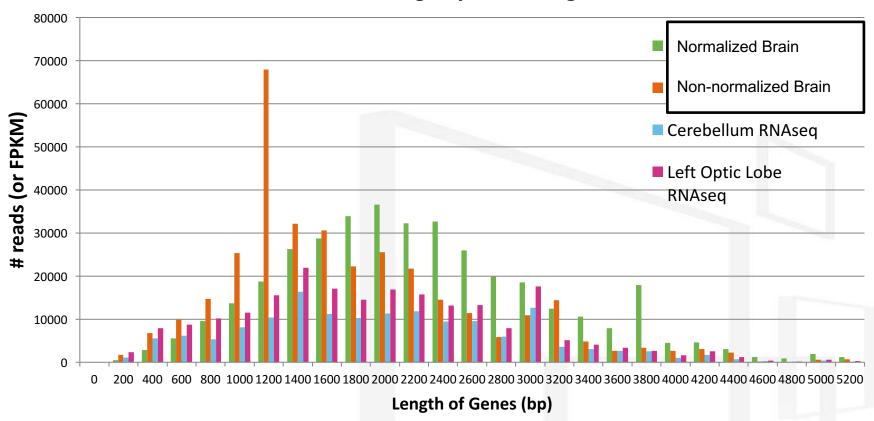
Overly abundant genes?



### Normalization



#### **Read Coverage by Gene Length**



- 2 genes in 1200bp bin for Sequel run associated with 37,679 reads
- Roughly 10% of sequencing spent on only 2 genes





### Normalization results



	CCS	FLNC	Genes	Transcripts	Genes/FLNC	Trans/FLNC
Non-Norm.	566,307	197,544	11,934	39,909	0.06	0.20
Normalized	145,527	58,567	19,849	49,465	0.34	0.84

- >5x genes per FLNC with normalization
- >4x transcripts per FLNC with normalization
- Additional genes are mostly lncRNA





### Normalization Methods



rare transcript

TTTT

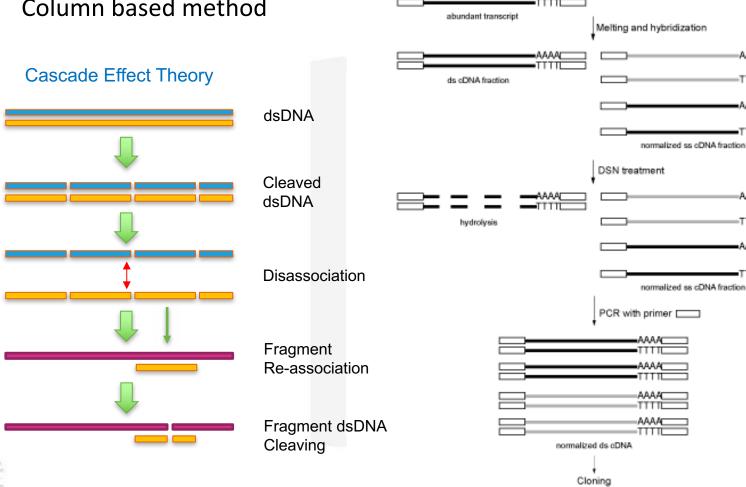
-AAAA

-TTTT •AAAA

AAAA

-TTTT

- **DSNase** method
- Column based method



ds cDNA flanked by known adapter sequences

### Normalization Methods



20 Years of Ploneering Great British Bioscience

- DSNase method
- Column based method

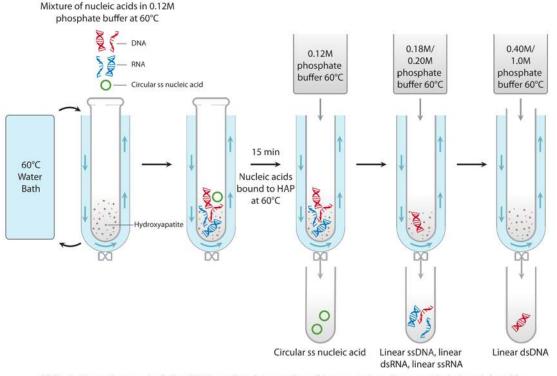


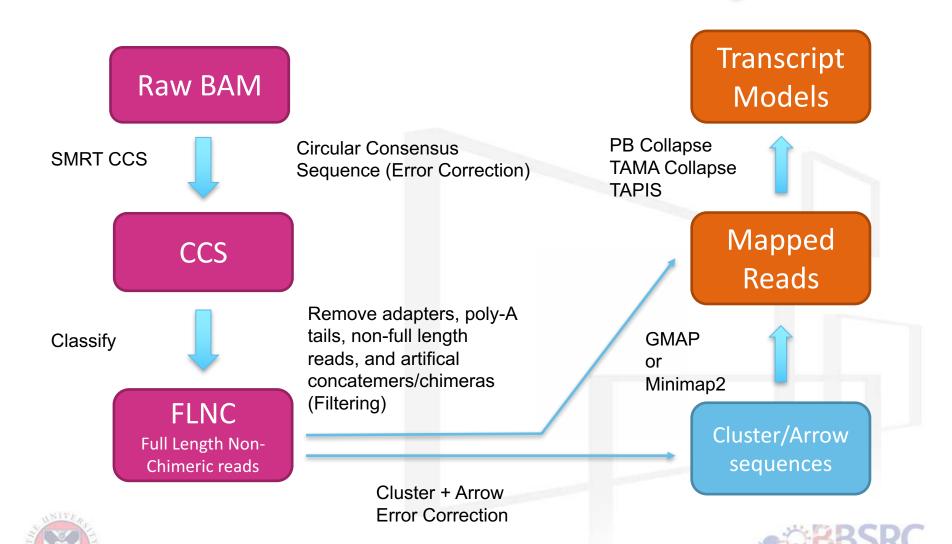
FIG. 1. Flow diagram depicting HAP-mediated separation of known and environmental viral nucleic acids.



# Iso-Seq Analysis Pipeline

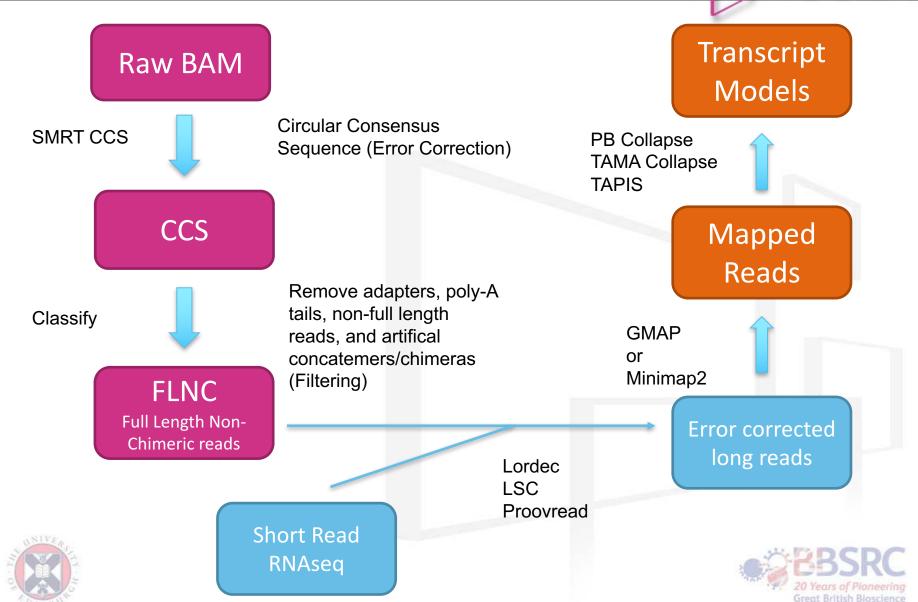


Great British Bioscience



### Iso-Seq pipeline w/ RNAseq





### TAMA



- Transcriptome Annotation by Modular Algorithms
- TAMA Collapse
- TAMA Merge
- TAMA-GO







# Collapse/Annotation



**Transcript** 

Models

Mapped

Reads

- Converting alignment files into annotation files (ie gtf, gff, bed)
- Filtering out bad alignments
- Identifying transcript model features (ie transcription start and end, splice junctions)
- Collapsing redundant transcripts

GMAP or Minimap2

**TAPIS** 

PB Collapse TAMA Collapse

ICE cluster sequences

#### **FLNC**

Full Length Non-Chimeric reads

> Cluster + Arrow Error Correction

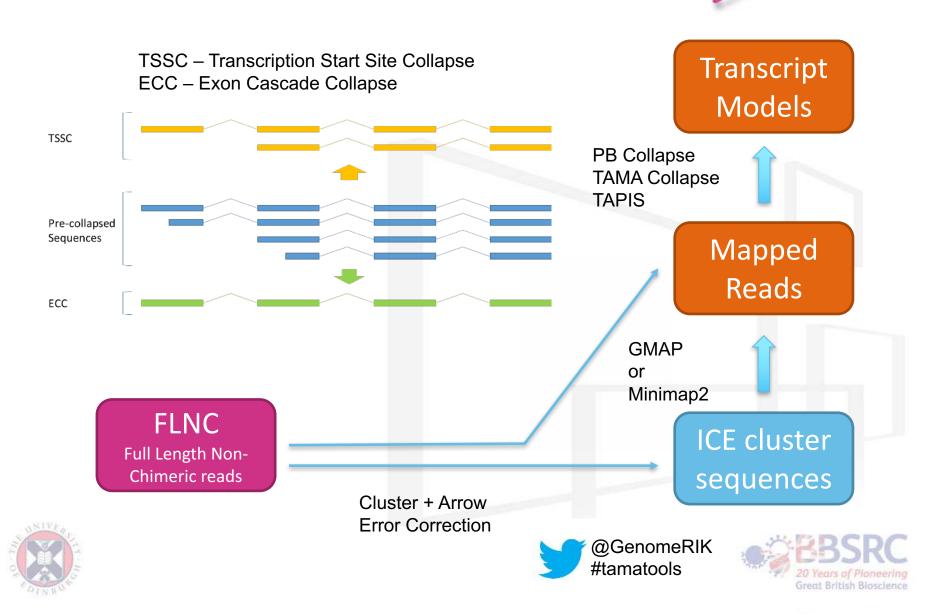






### PacBio Collapse





### TAMA Collapse

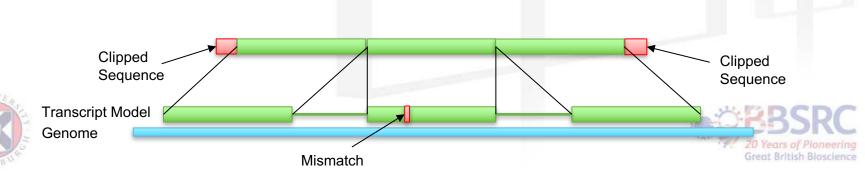


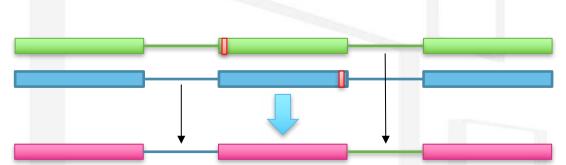
Control over transcript collapsing

T.A.M.A.G.

10 bp

- Manages 5' cap selected and non cap selected sequencing data
- Provides source information for all predicted events
  - Support for each final model
  - Support for each transcript feature (TSS/TTS, splice junctions)
- Flags uncertainties
  - Poly A truncation
  - Variation
  - Wobble
- Splice junction priority
  - Uses mapping mismatch information near splice junctions to choose best evidence





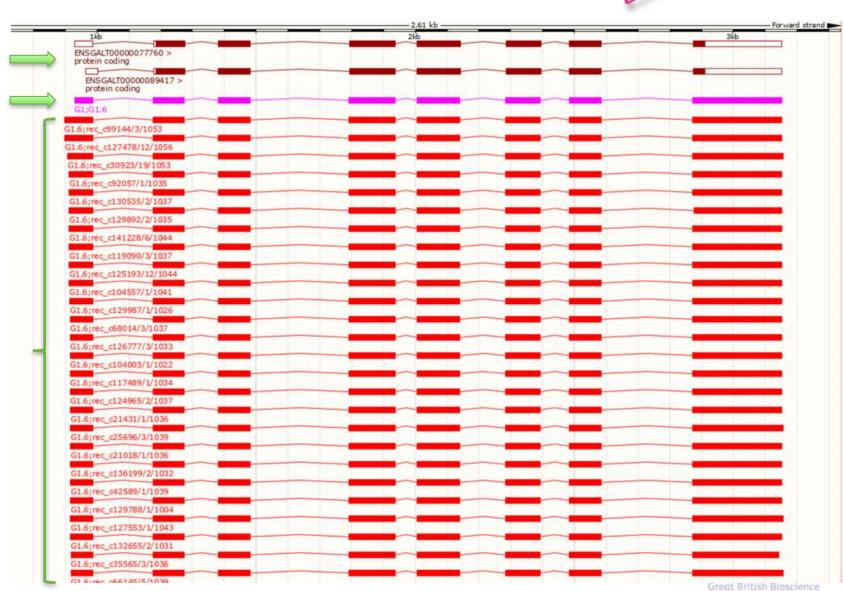
# Using TAMA Collapse



Ensembl TAMA Collapse

Mapped FLNC





# TAMA Collapse trans\_report

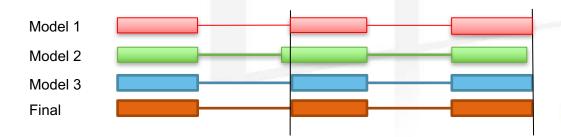


Line from trans\_report.txt:

#### Column/Field identities

transcript_id	G1.6
num_clusters	47
high_coverage	100
low_coverage	99.3
high_quality_percent	99.62
low_quality_percent	93.33
start_wobble_list	52,0,0,0,0,0,4
end_wobble_list	0,0,0,0,0,0,20
collapse_sj_start_err	0,0,0,0,0,0,1
collapse_sj_end_err	0,0,0,0,0,2,0
collapse_error_nuc	0>0;0>0;0>0;0>0;0>0;0>0;10.G.A_1D_5M>0-10.G.A>0

This is the interesting stuff!!



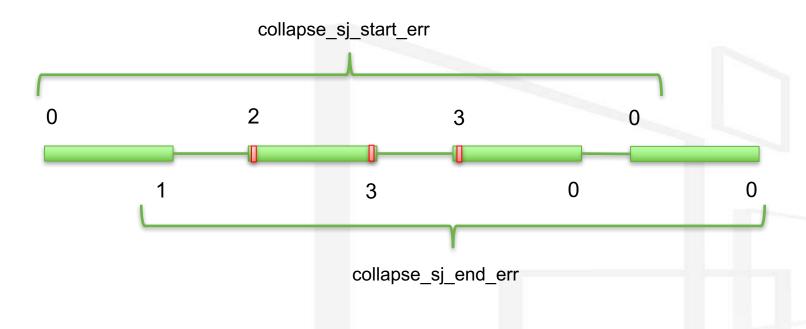




# TAMA Collapse SJ Error



#### Column/Field identities



- O No mismatches on either side of the splice junction
- 1 One mismatch on the other side of the splice junction
- One mismatch on the same side of the splice junction
- There are mismatches on both sides of the splice junction



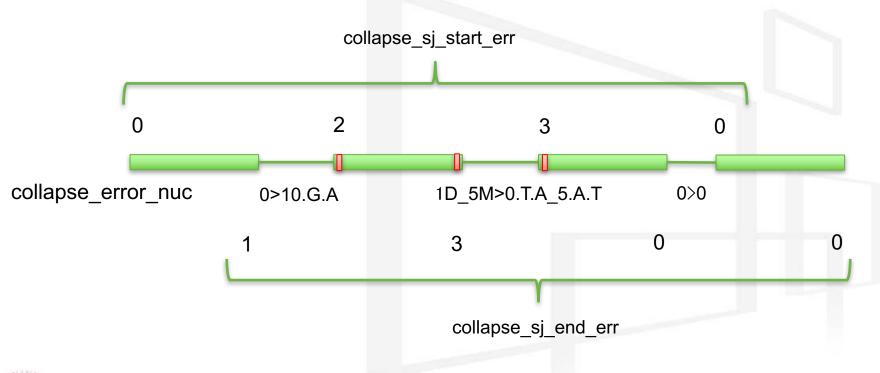


# TAMA Collapse Error Nuc



#### Column/Field identities

collapse_sj_start_err	0,2,3,0
collapse_sj_end_err	1,3,0,0
collapse error nuc	0>10.G.A; 1D_5M>0.T.A_5.A.T;0>0

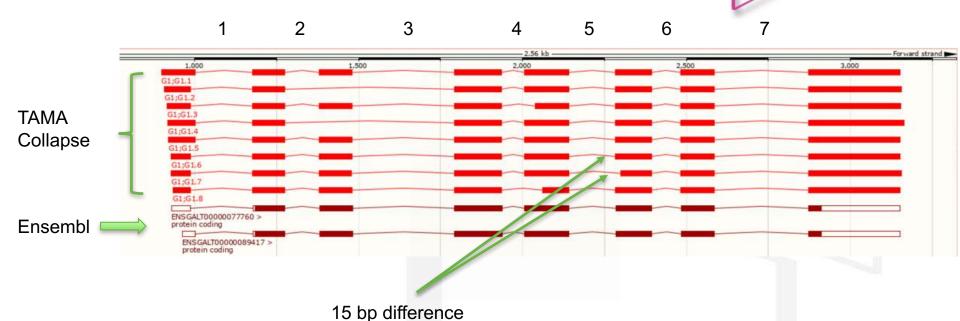






### Local density error





1	3.G.C>9M_2I
2	0>0
3	0>0.T.A_5.A.T_6.T.A
4	1D_3M_1D_1M>0.T.C_2.C.T
5	7.A.C>9I_1M_2D
6	1D_6M>0
7	10.G.A>10M_1D





# TAMA Merge



- Allows merging of Iso-Seq, RNA-seq, and public annotations
- Provides control over merging thresholds
- Allows user defined priority of transcript features from different sources
  - Use transcription start and end sites from Iso-Seq and splice junctions from RNAseq
- Tracks all merging events and outputs it in report files
- https://github.com/GenomeRIK/tama



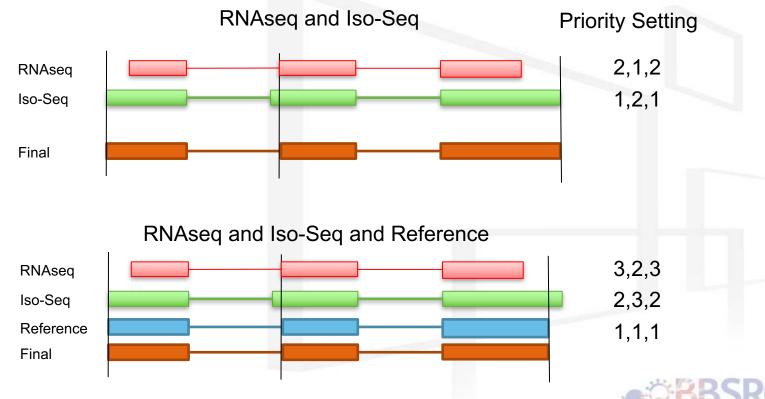




# Using TAMA Merge



- Similar algorithm for merging transcripts as TAMA collapse
- Some nuanced (but important!) differences





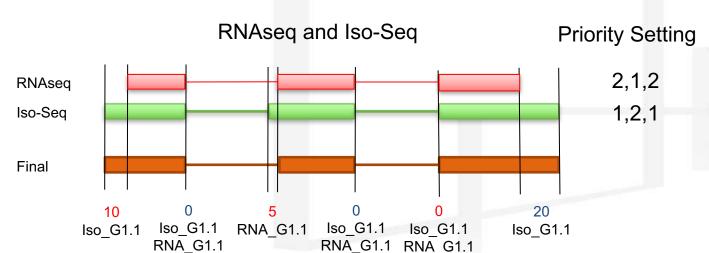


### TAMA Merge trans\_report



G1.1 1 Iso,RNA 10,5,0 0,0,20 Iso\_G1.1;RNA\_G1.1; Iso\_G1.1,RNA\_G1.1 Iso\_G1.1,RNA\_G1.1; Iso\_G1.1,RNA\_G1.1; Iso\_G1.1,RNA\_G1.1

```
start_wobble_list 10,5,0
end_wobble_list 0,0,20
exon_start_support Iso_G1.1;RNA_G1.1; Iso_G1.1,RNA_G1.1
exon_end_support Iso_G1.1,RNA_G1.1; Iso_G1.1,RNA_G1.1; Iso_G1.1
```







### TAMA-GO ORF/NMD



- 1. Convert bed to fasta
- 2. Get open reading frames (ORF)
- 3. Blast amino acid sequences against the Uniprot/Uniref
- 4. Parse the Blastp output file for top hits
- Create new bed file with CDS regions and NMD predictions

### Example BED12 output line

1 481182 484817 G28;G28.23;none;5prime\_degrade;no\_hit;NMD1;F2 40 - 482403 484816 0,200,255 4 831,81,113,127 0,1198,2386,3508







### TAMA-GO



 Suite of tools for various transcriptome annotation needs



- NMD/ORF predictions
- Format convertors
- More to come!

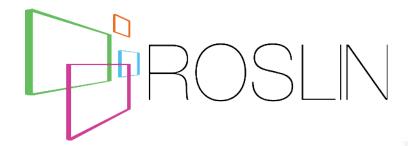
P.S. If you need a tool, please contact me.
I may have it but just haven't uploaded it yet.
If I don't have it, I may be able to make it for you.
Also if you want to contribute to the repo contact me!
GenomeRIK@gmail.com





# Acknowledgement







**Professor Dave Burt** 

**Professor Alan Archibald** 

Jacqueline Smith

Katarzyna Miedzinska

**Bob Paton** 

Lel Eory

Elizabeth Tseng

# edinburgh genomics.

Karim Gharbi

Marian Thomson







### Contact



- You can reach me at <u>GenomeRIK@gmail.com</u>
- I also tweet updates for TAMA and Iso-Seq: @GenomeRIK
- TAMA tools: <a href="https://github.com/GenomeRIK/tama">https://github.com/GenomeRIK/tama</a>
- Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human: <a href="https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3691-9">https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3691-9</a>
- Iso-Seq Webinar: <a href="https://www.youtube.com/watch?v=Pwx\_uEBuhZc&t=1071s">https://www.youtube.com/watch?v=Pwx\_uEBuhZc&t=1071s</a>



