ROSLIN

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

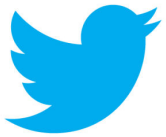
Richard Kuo

THE UNIVERSITY of EDINBURGH

BBSRC
20 Years of Pioneering
Great British Bioscience

Outline

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



@GenomeRIK
#tamatools

Iso-Seq Webinar:
https://www.youtube.com/watch?v=Pwx_uEBuhZc&t=1071s

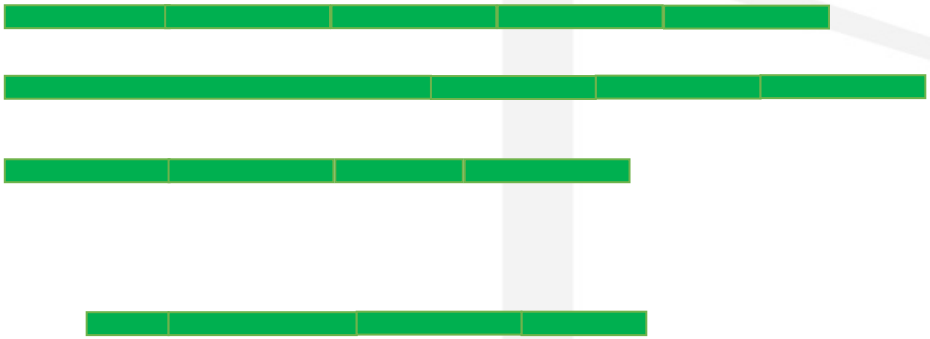
Big Picture

- 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

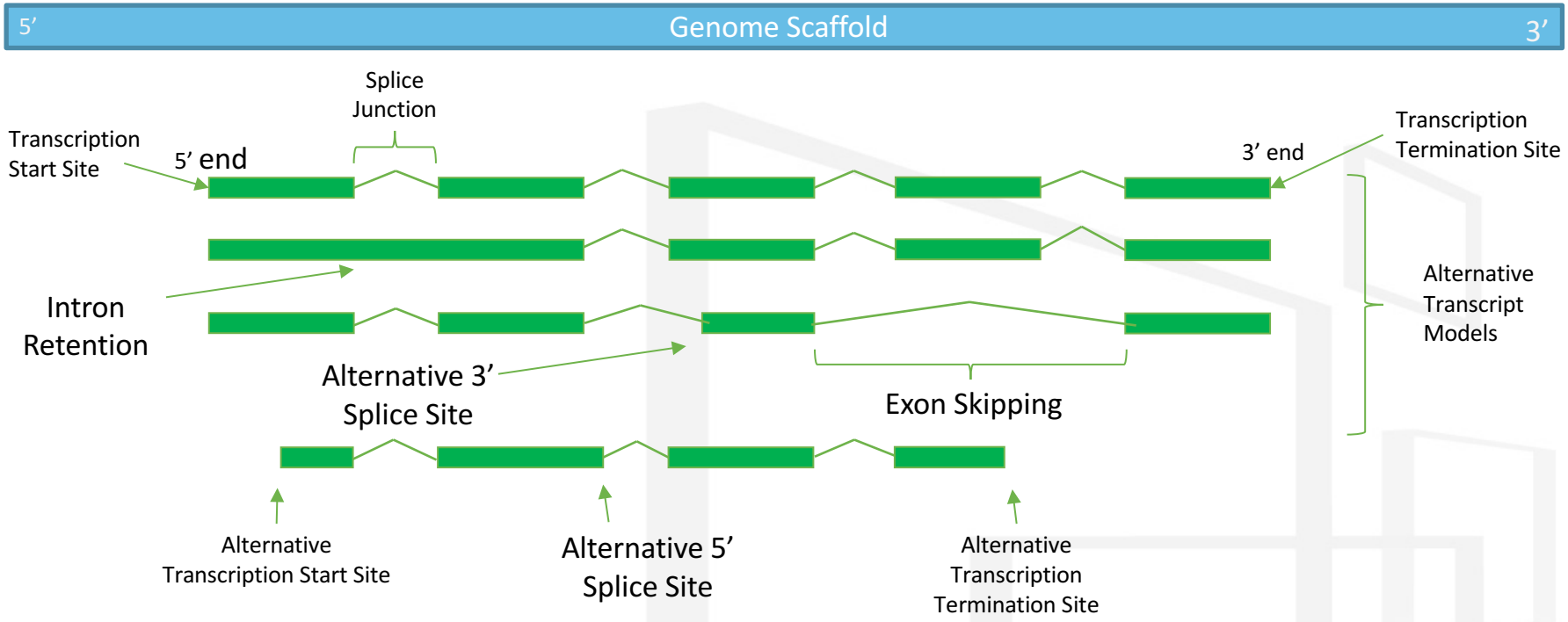
Chicken or the egg Iso-Seq planning:
None of the steps come first in planning.
They are all dependent on each other.

@GenomeRIK
#tamatools

Transcripts without genome



Transcript Models



@GenomeRIK
#tamatools

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

@GenomeRIK
#tamatools

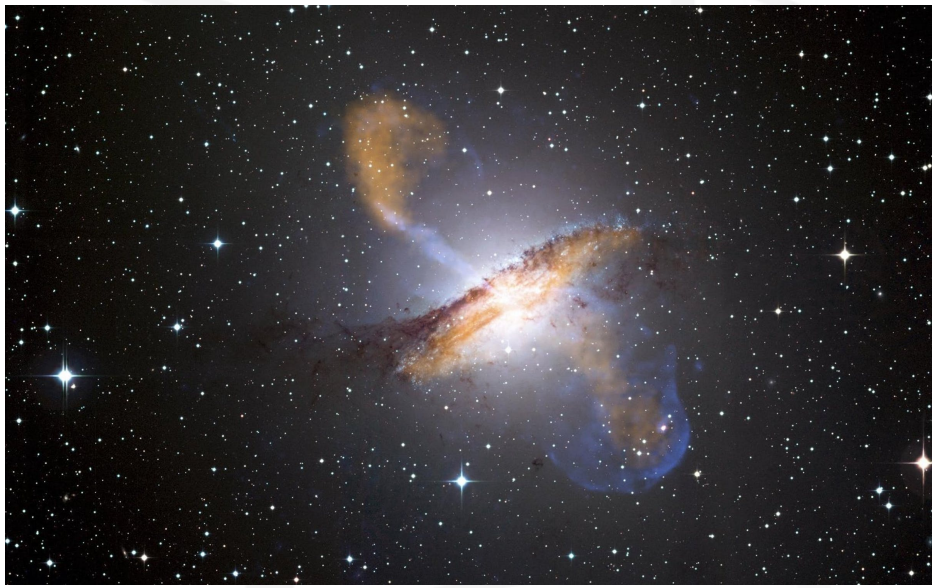
With a genome

- 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

@GenomeRIK
#tamatools

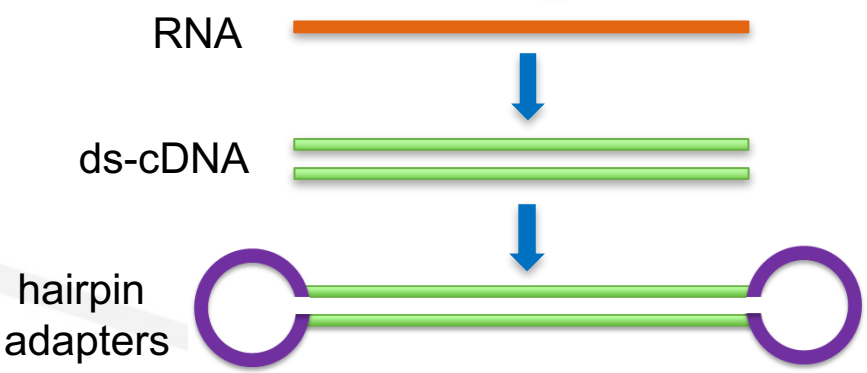
Genome and Transcriptome

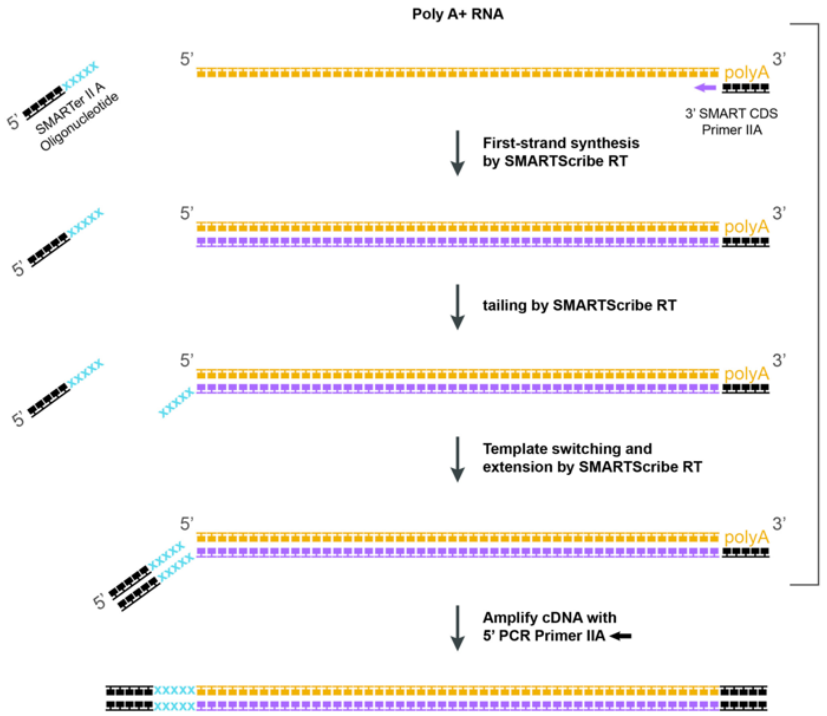
- Genome and Transcriptome
- Can map to assembly
- Limited to assembly quality
- Transcript model focused/reference based transcriptomes
- Only need exon starts and ends to be accurate
- Want to improve a transcriptome annotation
  - Depends on what you want



Standard Library Preparation

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

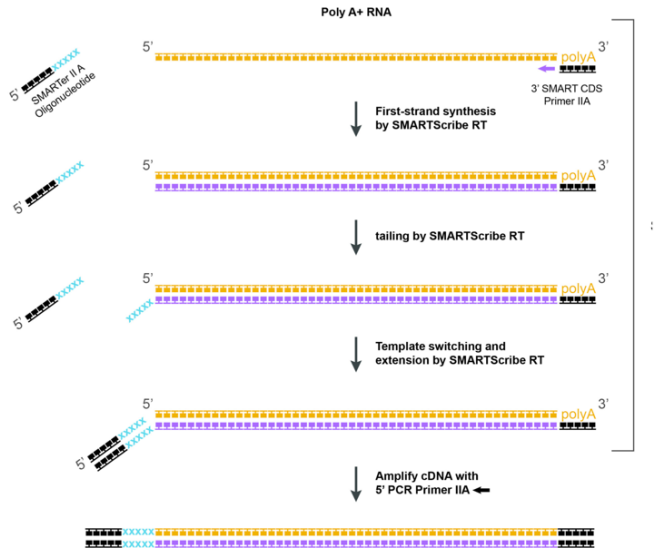




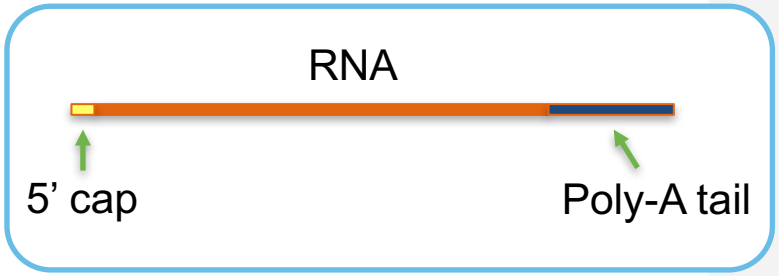
5’ Degradation

- 5’ degradation?

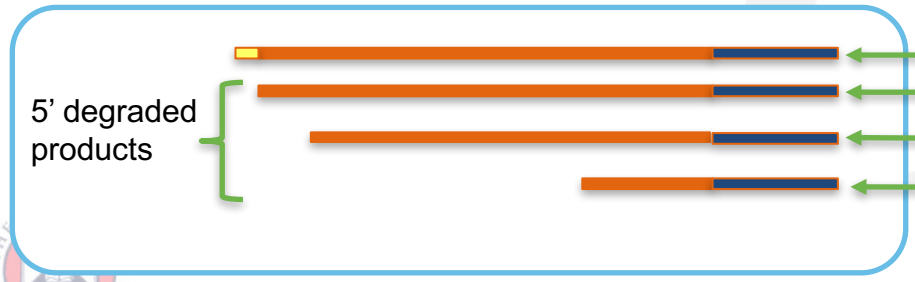
Standard Library Preparation



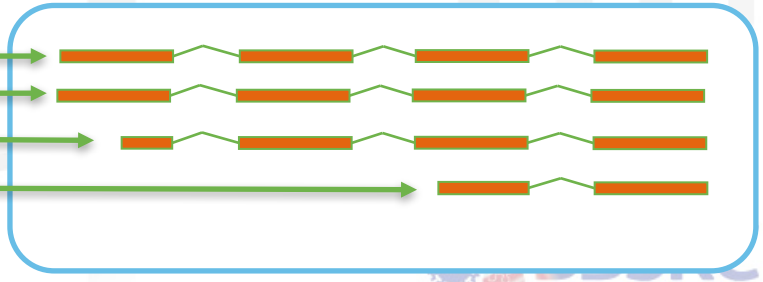
RNA structure



RNA from same transcript



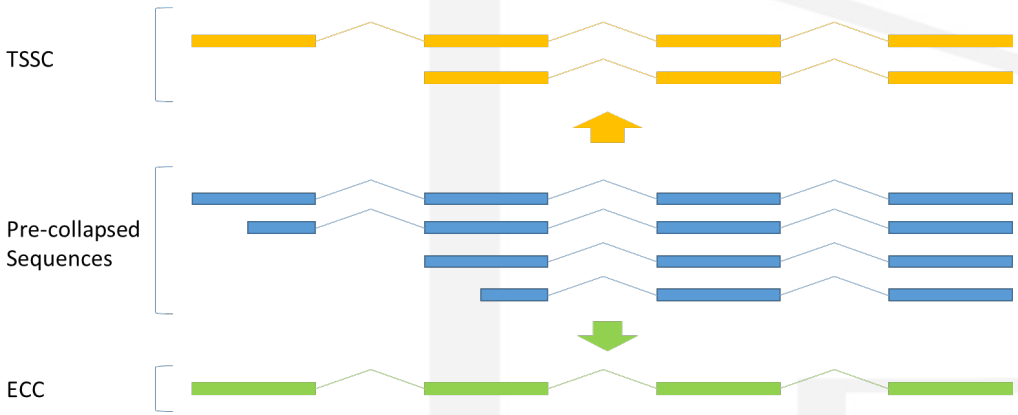
Resulting models



5’ Cap Selection

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

TSSC – Transcription Start Site Collapse
ECC – Exon Cascade Collapse



	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
https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3691-9

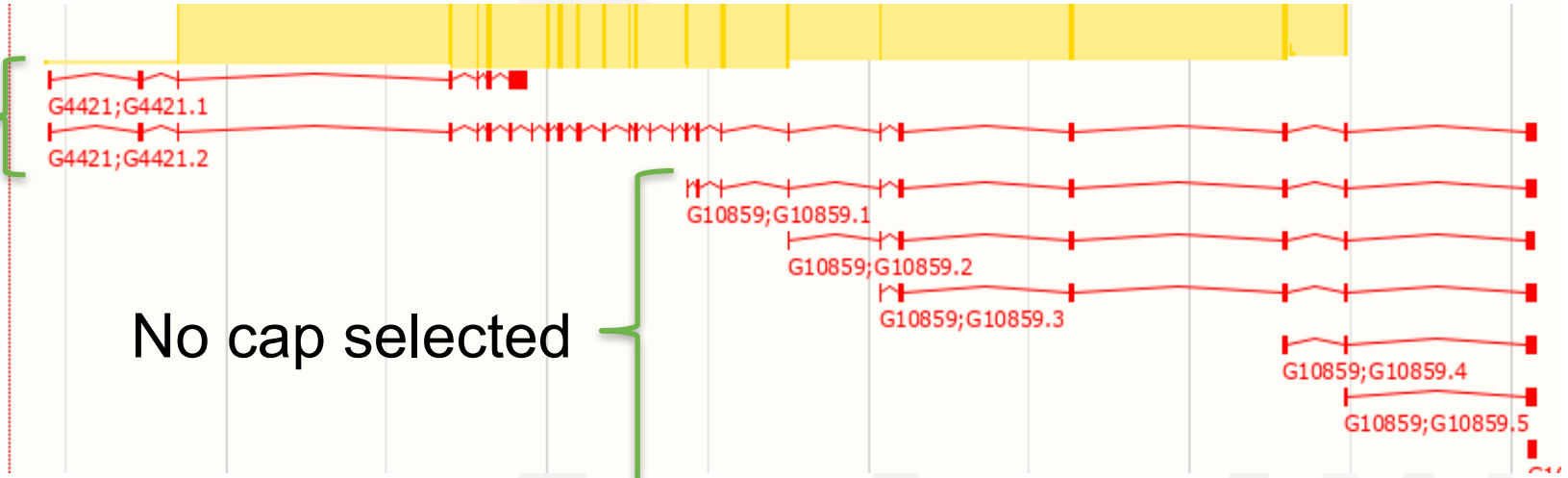
5’ Cap Comparison

ROSLIN

LEXOGEN
Enabling complete transcriptome sequencing

5’ cap selected

No cap selected



THE UNIVERSITY OF EDINBURGH

BBSRC
20 Years of Pioneering Great British Bioscience

5’ Cap Teloprime Kit

Step 2: Double-Strand Specific Ligation

Capped Poly(A) RNA

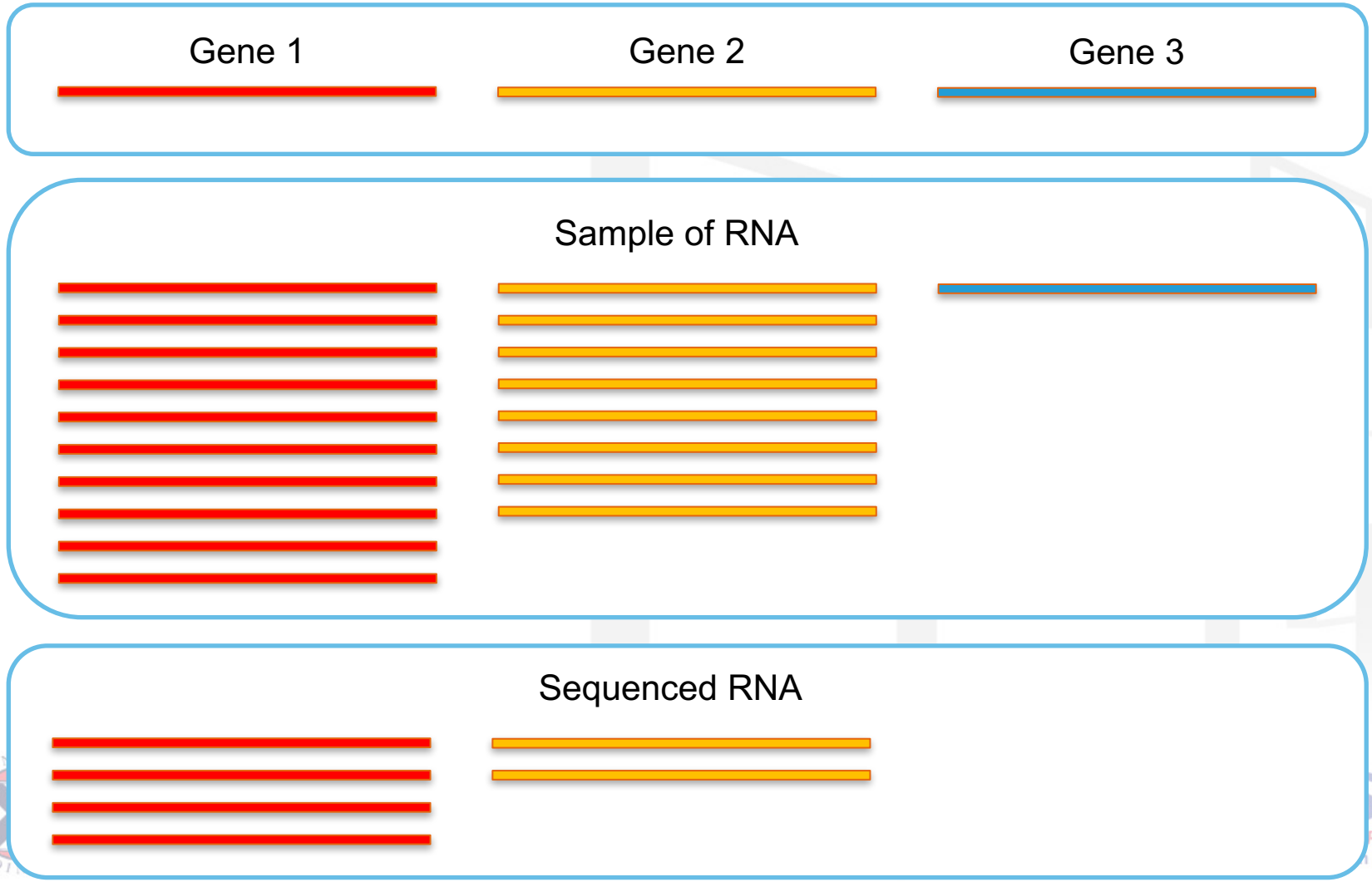
Degraded RNA

Enabling complete transcriptome sequencing

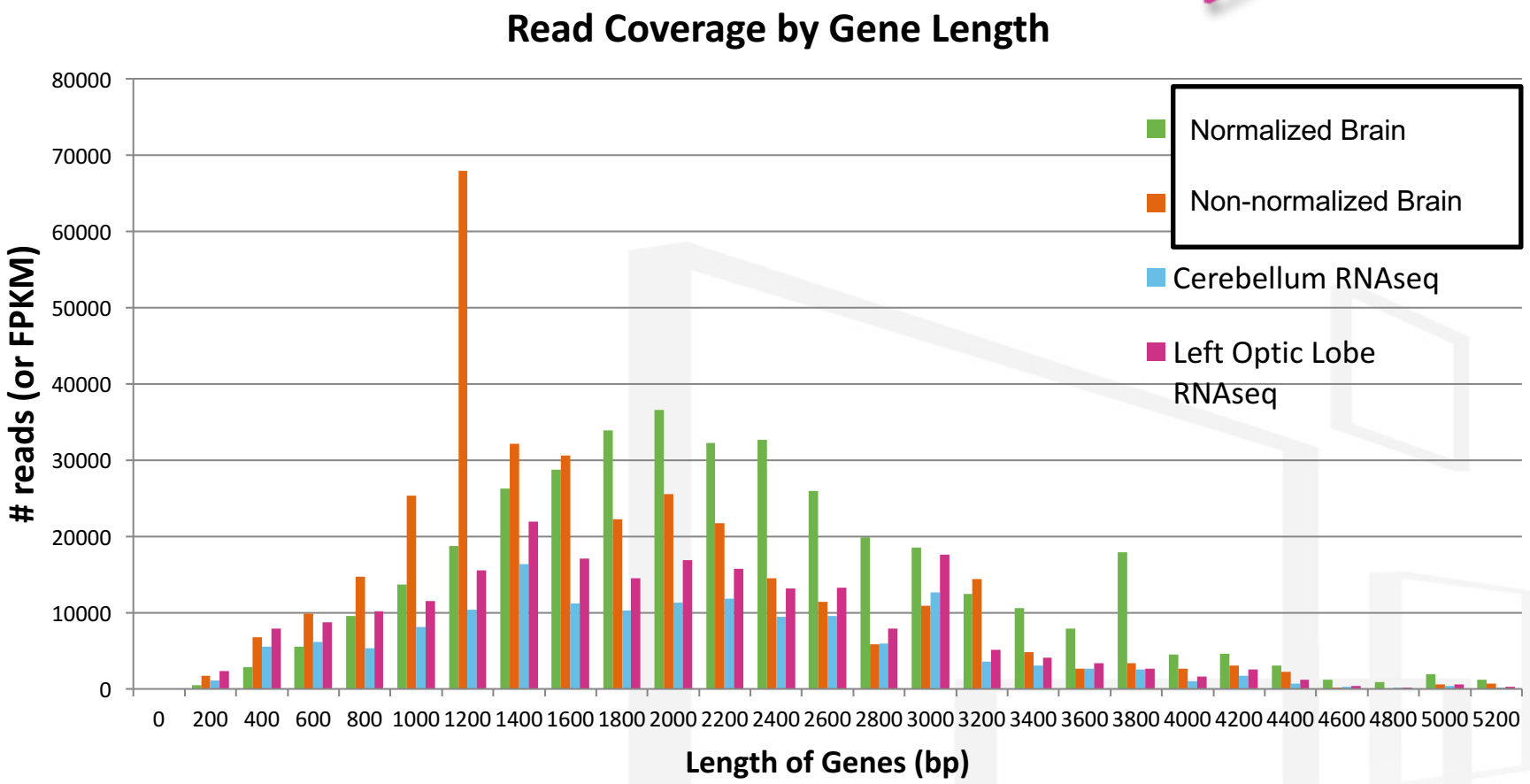
20 Years of Pioneering Great British Bioscience

Over abundant genes/transcripts

- Overly abundant genes?



Normalization



- 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

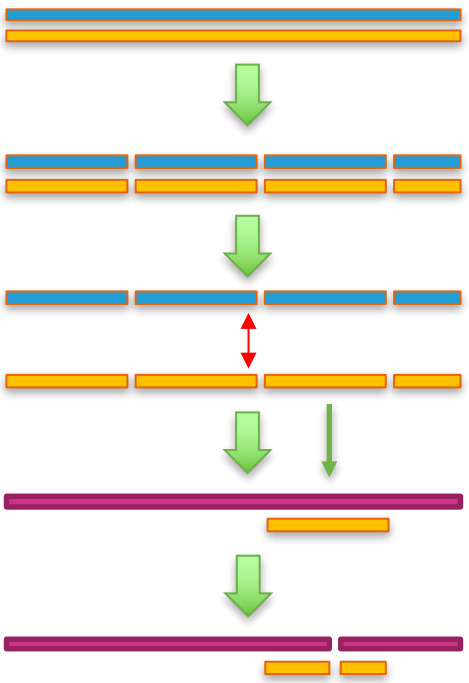
- >5x genes per FLNC with normalization
- >4x transcripts per FLNC with normalization
- Additional genes are mostly lncRNA

FLNC – Full length non-chimeric reads

Normalization Methods

- DSNase method
- Column based method

Cascade Effect Theory



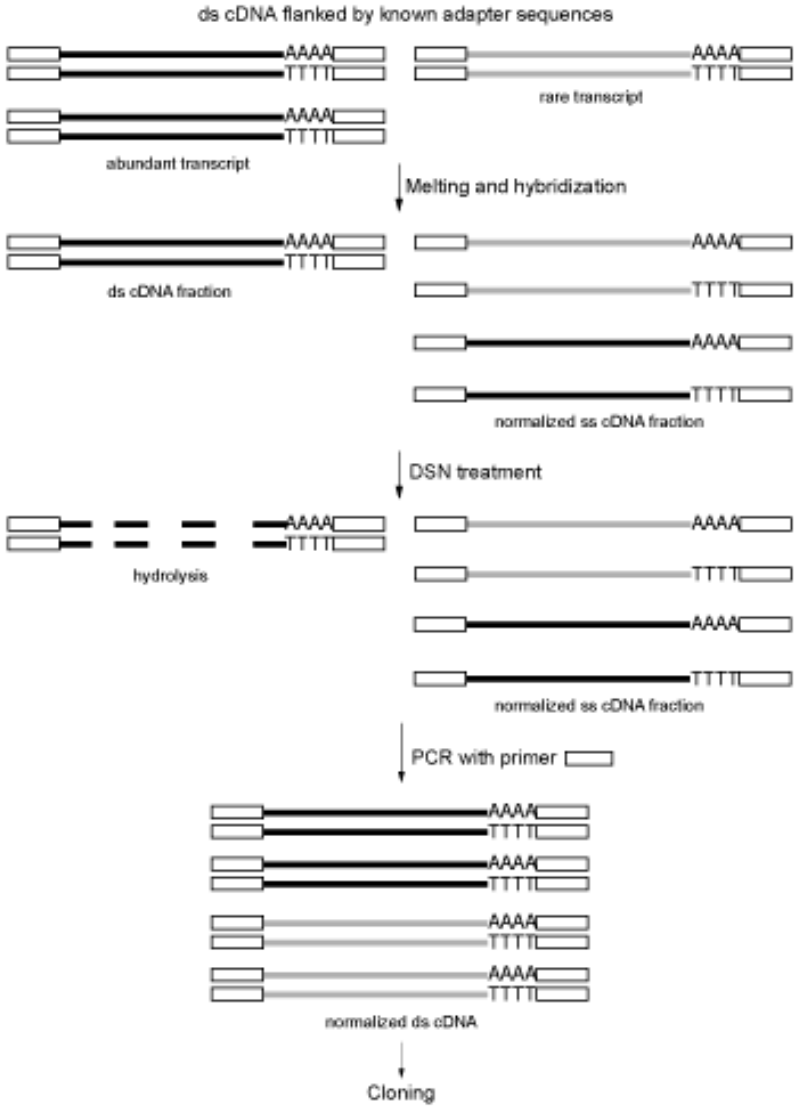
dsDNA

Cleaved dsDNA

Disassociation

Fragment Re-association

Fragment dsDNA Cleaving



From Evrogen website

Normalization Methods

- DSNase method
- Column based method

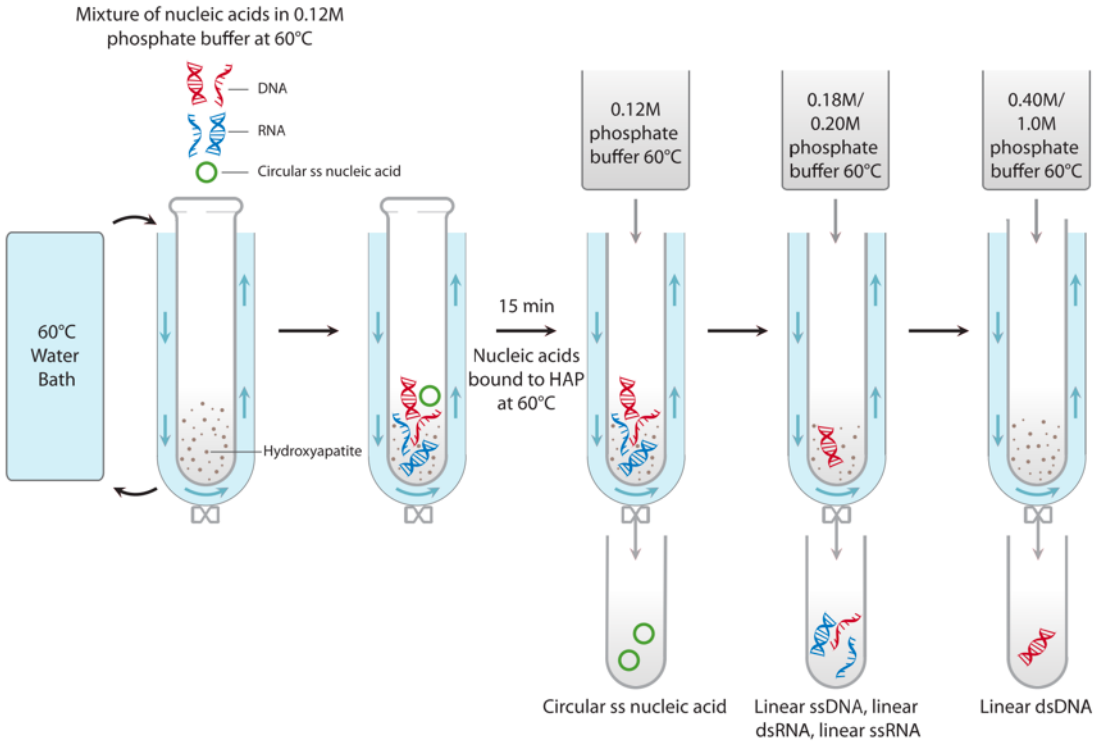
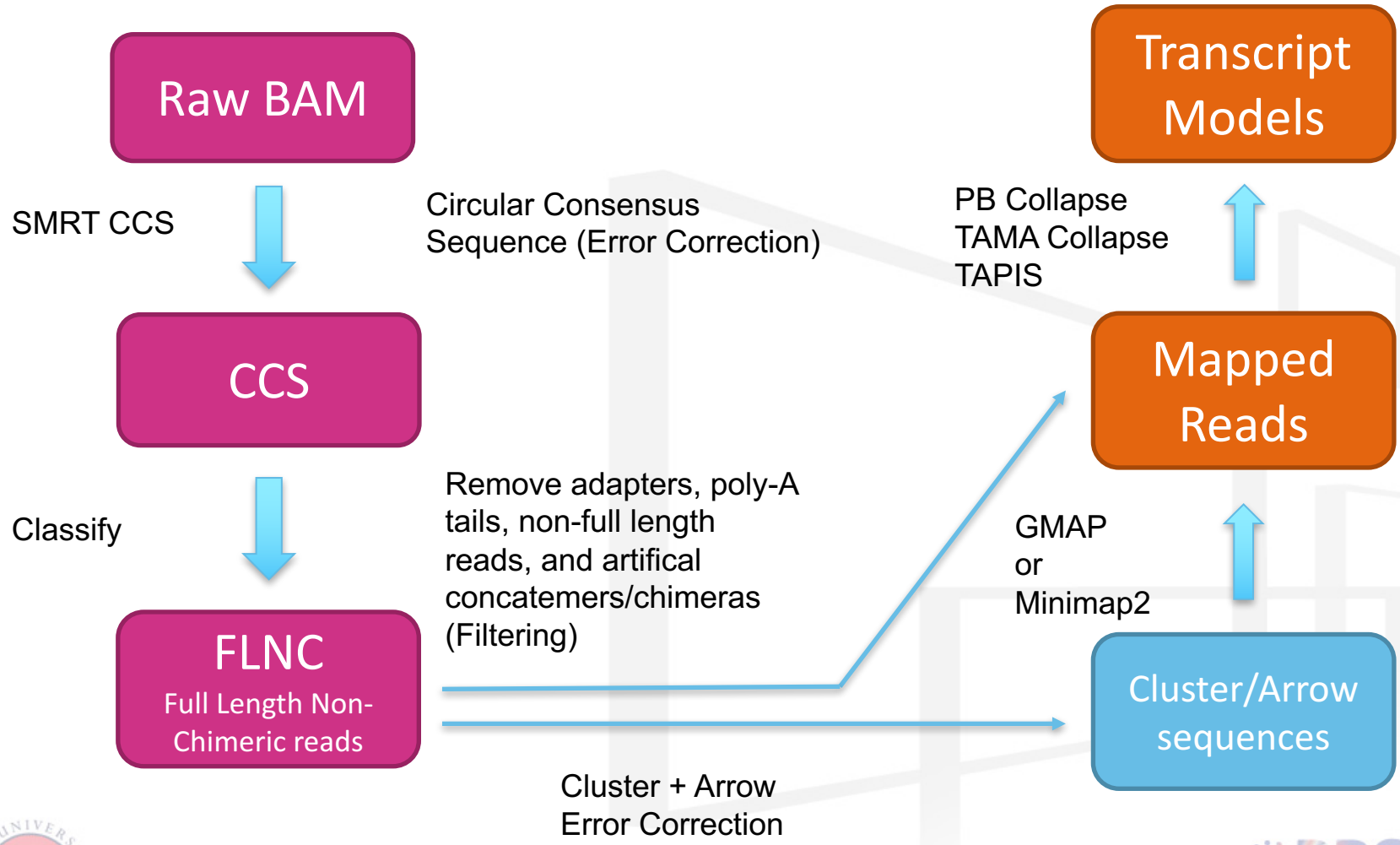


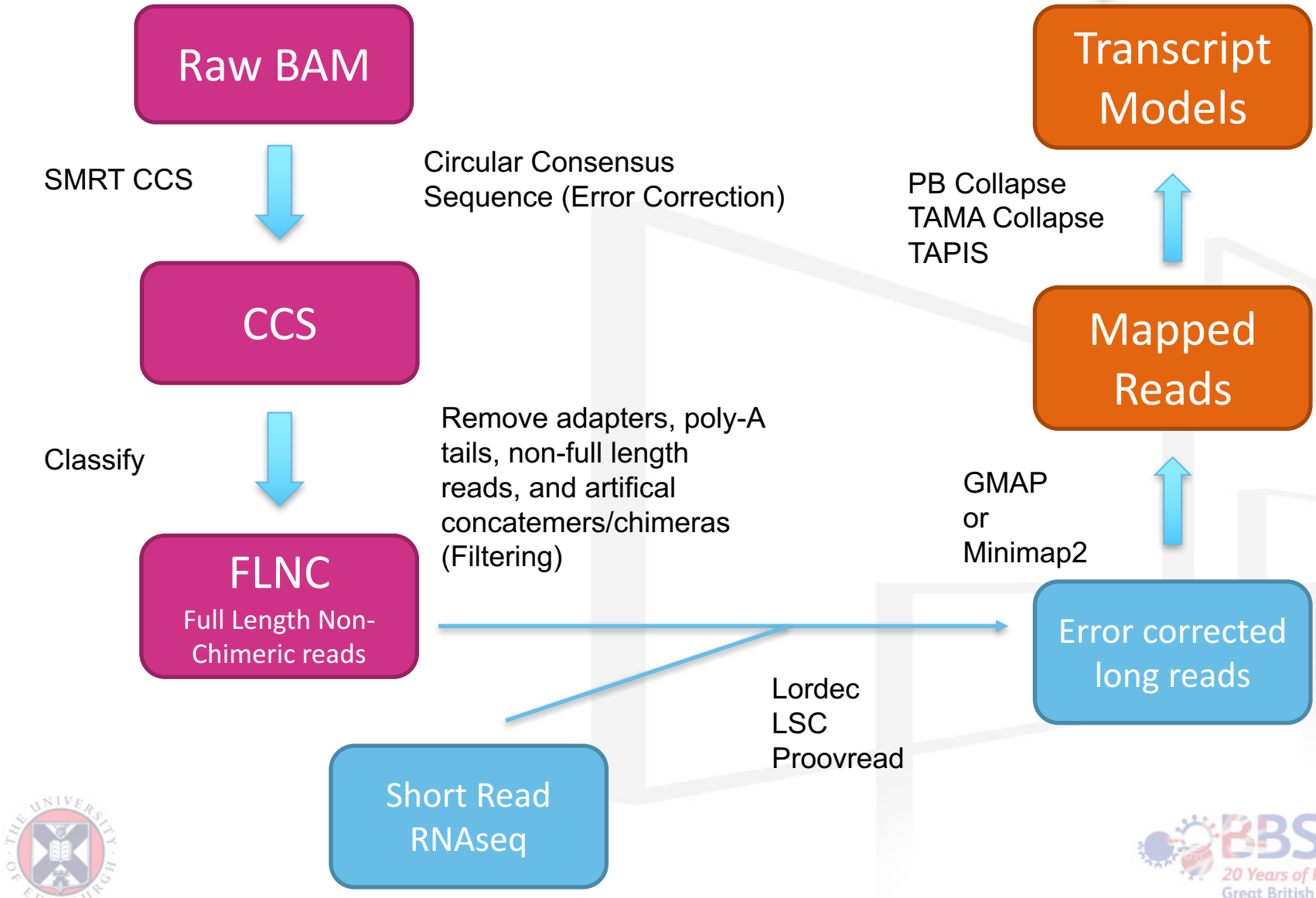
FIG. 1. Flow diagram depicting HAP-mediated separation of known and environmental viral nucleic acids.

From "Hydroxyapatite-Mediated Separation of Double-Stranded DNA, Single-Stranded DNA, and RNA Genomes from Natural Viral Assemblages"

Iso-Seq Analysis Pipeline

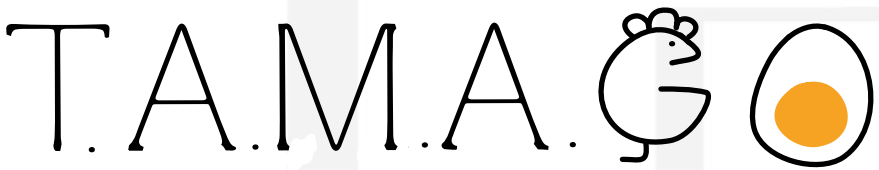


Iso-Seq pipeline w/ RNAseq



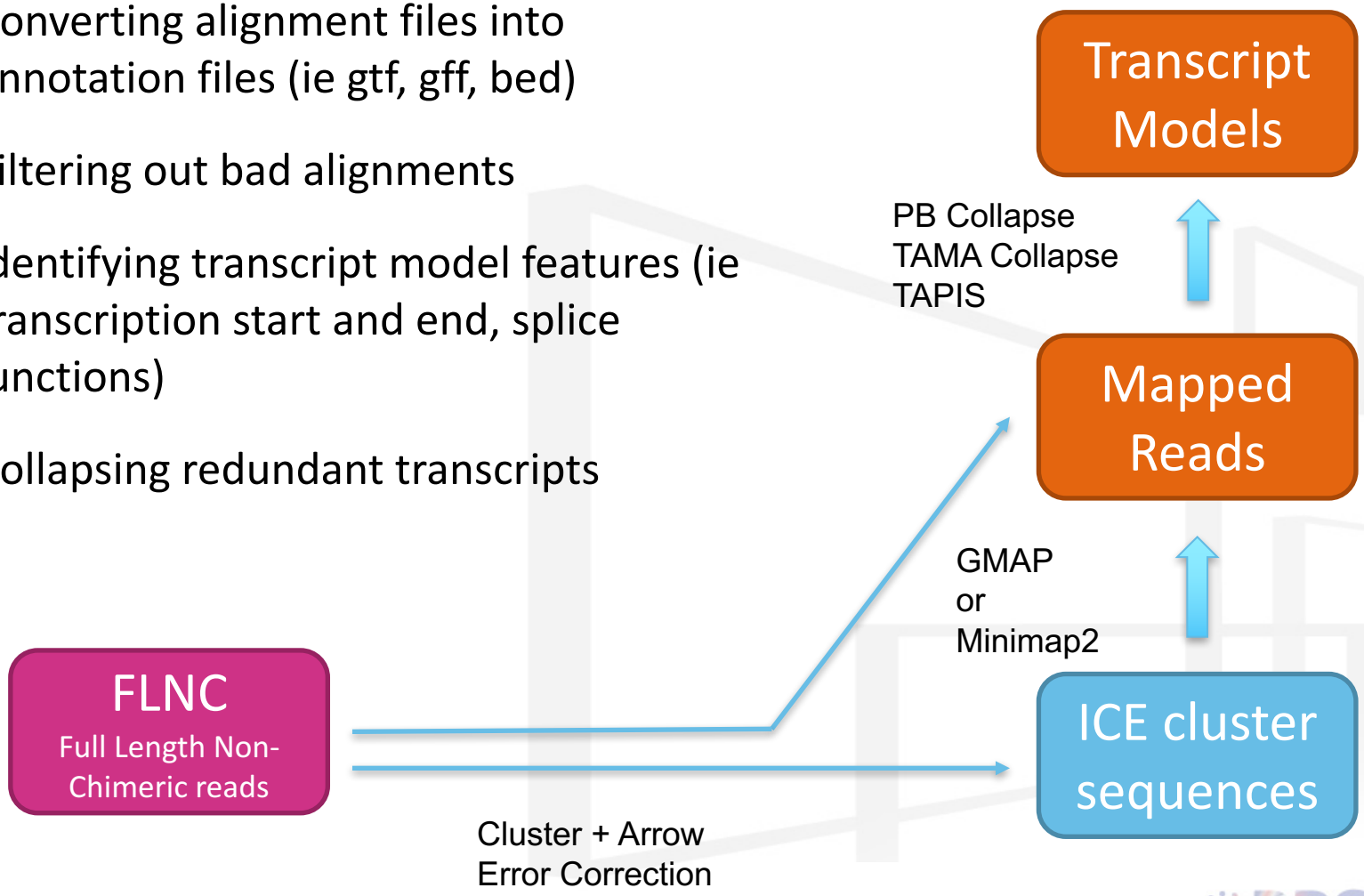
TAMA

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



Collapse/Annotation

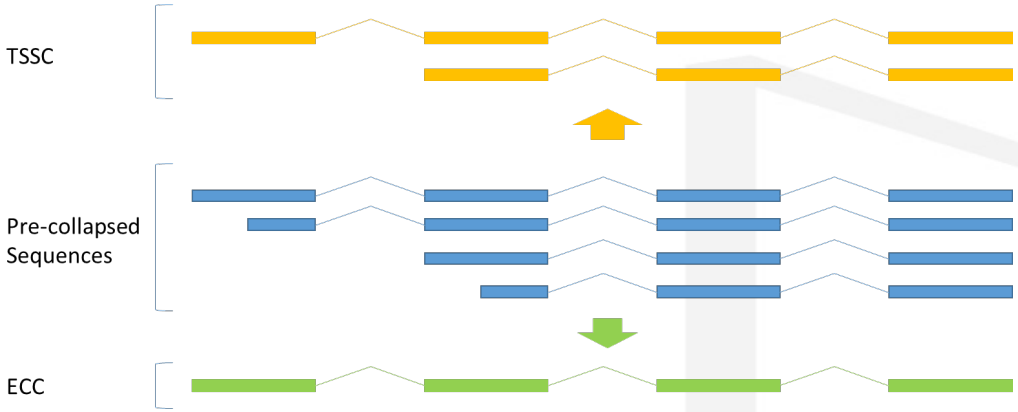
- Converting alignment files into annotation files (ie gtf, gff, bed)
- Filtering out bad alignments
- Identifying transcript model features (ie transcription start and end, splice junctions)
- Collapsing redundant transcripts



@GenomeRIK
#tamatools

PacBio Collapse

TSSC – Transcription Start Site Collapse
ECC – Exon Cascade Collapse



Transcript Models

PB Collapse
TAMA Collapse
TAPIS

Mapped Reads

GMAP
or
Minimap2

ICE cluster sequences

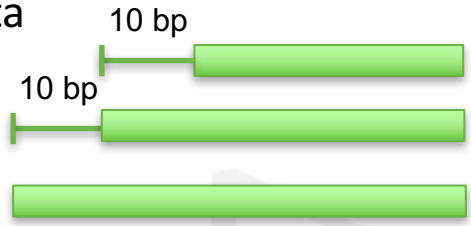
FLNC
Full Length Non-Chimeric reads

Cluster + Arrow Error Correction

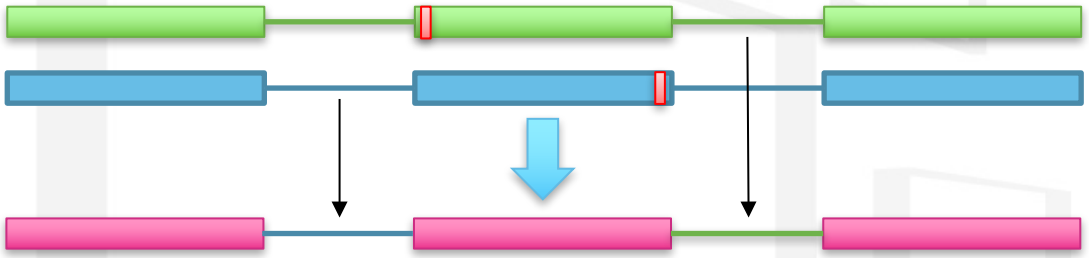
@GenomeRIK
#tamatools

TAMA Collapse

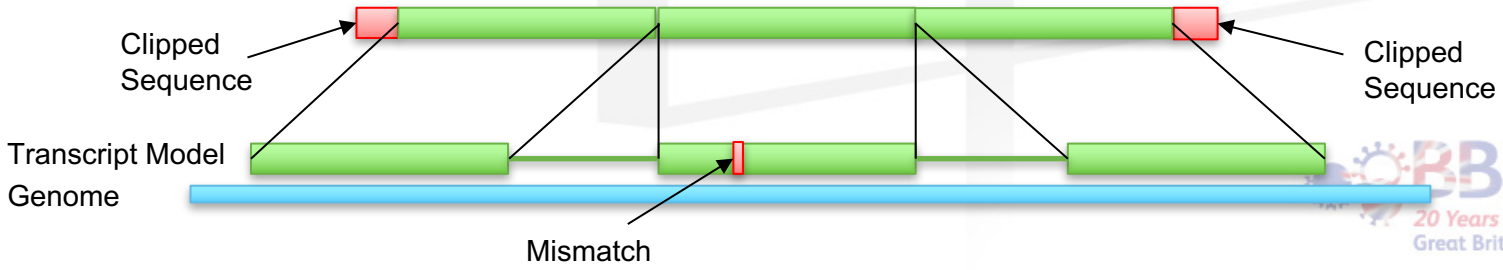
- Control over transcript collapsing
- Manages 5’ cap selected and non cap selected sequencing data
- Provides source information for all predicted events
  - Support for each final model
  - Support for each transcript feature (TSS/TTS, splice junctions)



- Flags uncertainties
  - Poly A truncation
  - Variation
  - Wobble



- Splice junction priority
  - Uses mapping mismatch information near splice junctions to choose best evidence



Using TAMA Collapse

TAMA Collapse trans_report

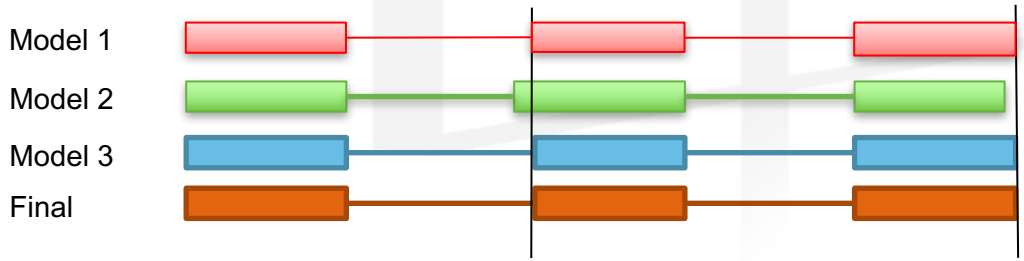
Line from trans_report.txt:

G1.6 47 100.0 99.3 99.62 93.33 52,0,0,0,0,0,0,4 0,0,0,0,0,0,0,20 0,0,0,0,0,0,0,1 0,0,0,0,0,0,2,0 0>0;0>0;0>0;0>0;0>0;0>0;10.G.A_1D_5M>0-10.G.A>0

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	52,0,0,0,0,0,0,4
end_wobble_list	0,0,0,0,0,0,0,20
collapse_sj_start_err	0,0,0,0,0,0,0,1
collapse_sj_end_err	0,0,0,0,0,0,2,0
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

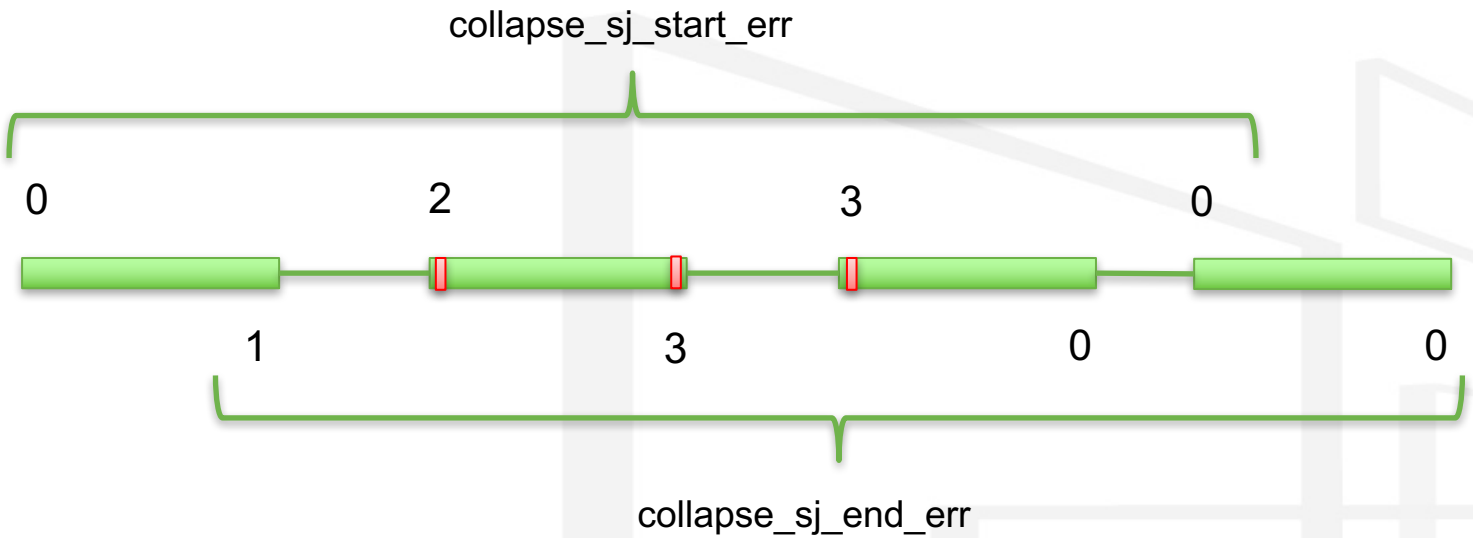
This is the interesting stuff!!



TAMA Collapse SJ Error

Column/Field identities

collapse_sj_start_err	0,2,3,0
collapse_sj_end_err	1,3,0,0

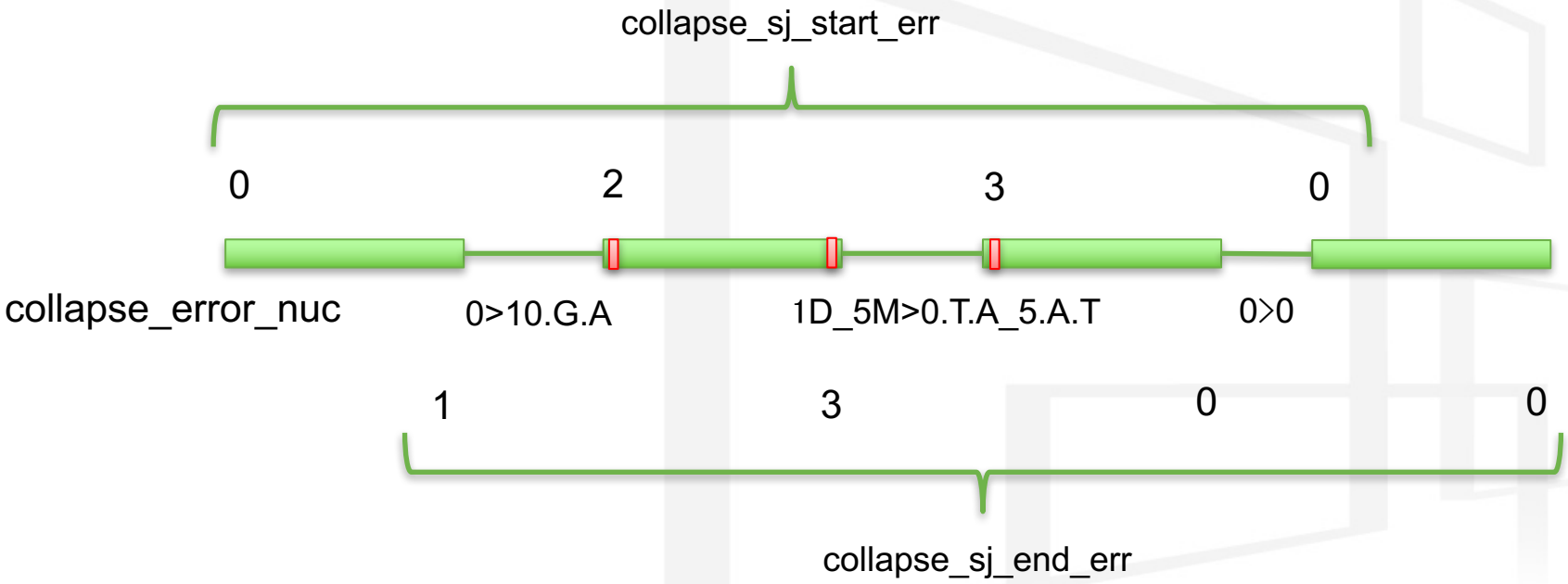


0	No mismatches on either side of the splice junction
1	One mismatch on the other side of the splice junction
2	One mismatch on the same side of the splice junction
3	There are mismatches on both sides of the splice junction

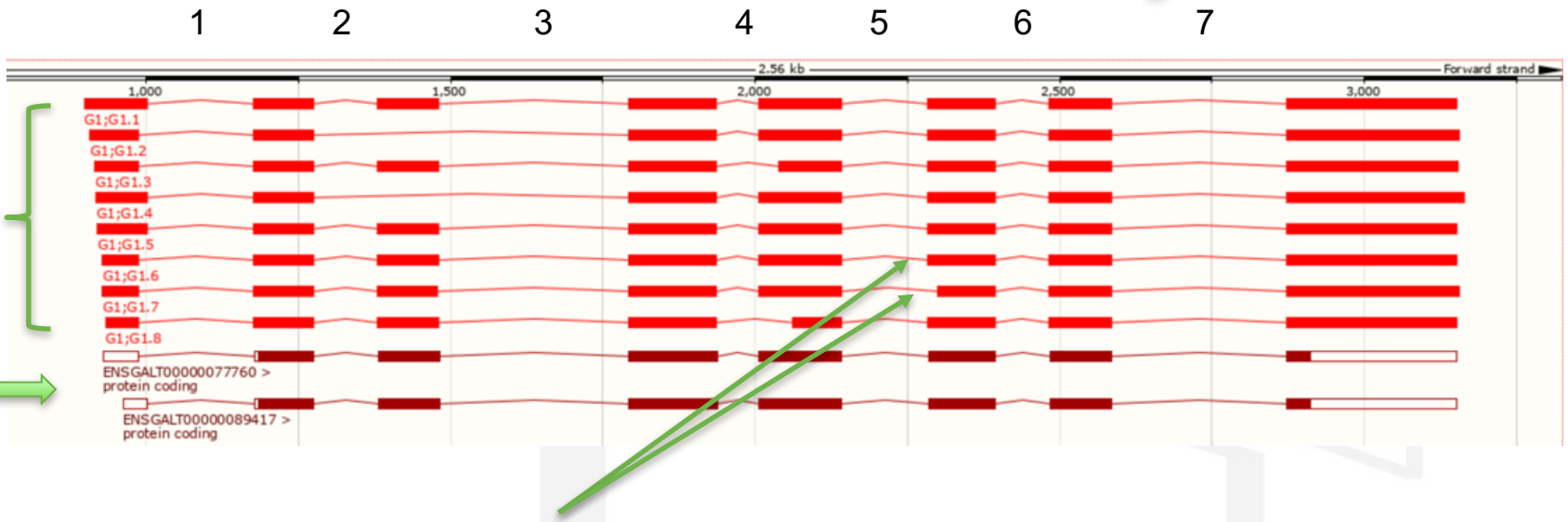
TAMA Collapse Error Nuc

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



transcript_id	clusters	high_cov	low_cov	high_quality_percent	low_quality_percent	start_wobble_list	end_wobble_list	collapse_sj_start_err	collapse_sj_end_err	collapse_error_nuc
G1.6	47	100.0	99.3	99.62	93.33	52,0,0,0,0,0,0,4	0,0,0,0,0,0,0,20	0,0,0,0,0,0,0,1	0,0,0,0,0,0,2,0	0>0;0>0;0>0;0>0;0>0;0>0;10.G.A_1D_5M>0-10.G.A>0
G1.7	1	100.0	100.0	93.98	93.98	0,0,0,0,0,0,0,0	0,0,0,0,0,0,0,0	0,3,0,2,3,3,1,3	3,0,1,3,3,2,3,0	3.G.C>9M_2I;0>0;0>0.T.A_5.A.T_6.T.A;1D_3M_1D_1M>0.T.C_2.C.T;7.A.C>9I_1M_2D;1D_6M>0;10.G.A>10M_1D

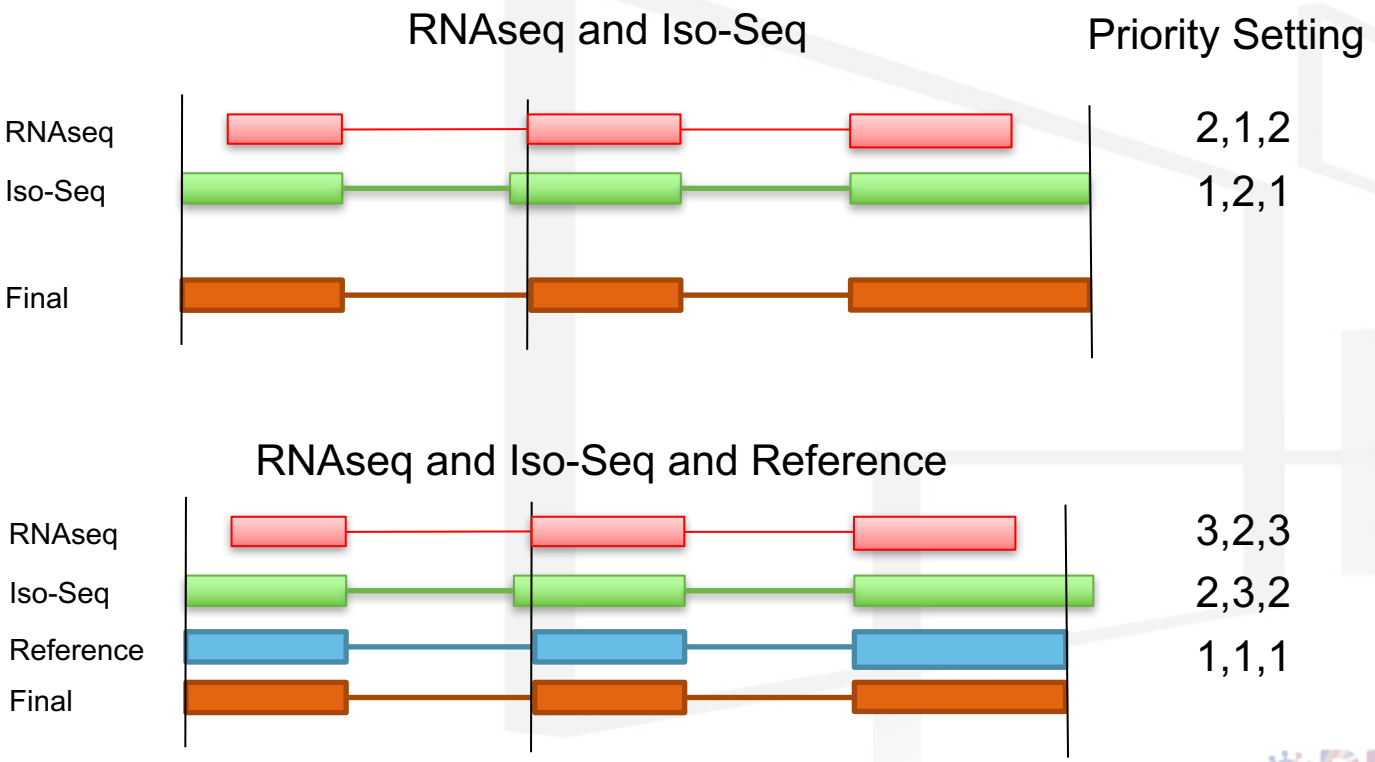
1	3.G.C>9M_2I
2	0>0
3	0>0.T.A_5.A.T_6.T.A
4	1D_3M_1D_1M>0.T.C_2.C.T
5	7.A.C>9I_1M_2D
6	1D_6M>0
7	10.G.A>10M_1D

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

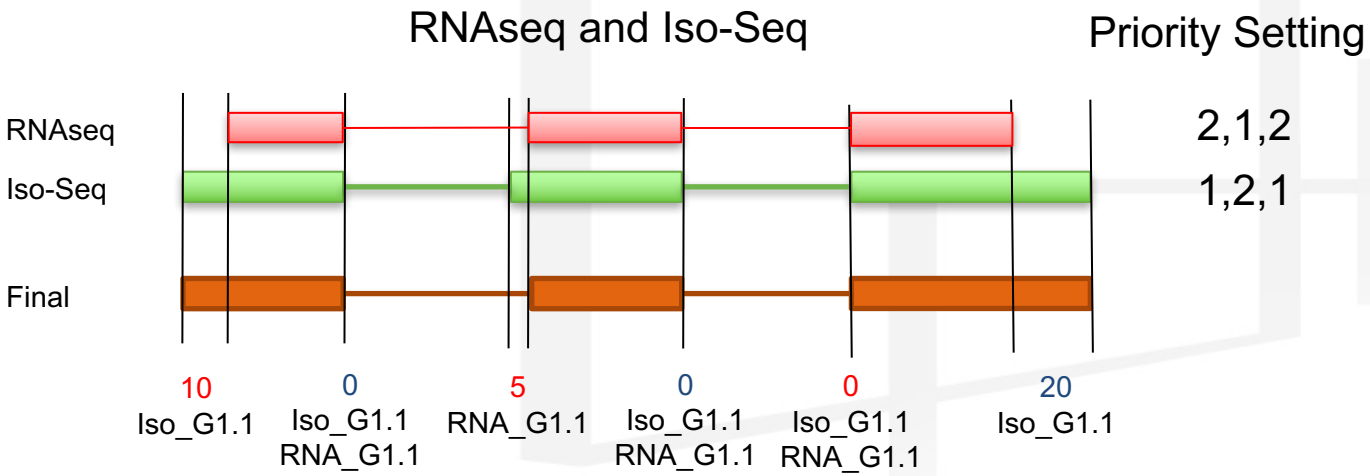
- Similar algorithm for merging transcripts as TAMA collapse
- Some nuanced (but important!) differences



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

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
5. 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

ROSLIN

ROSLIN

PACIFIC BIOSCIENCES®

Professor Dave Burt

Professor Alan Archibald

Jacqueline Smith

Katarzyna Miedzinska

Bob Paton

Lel Eory

Elizabeth Tseng

edinburgh genomics.

Karim Gharbi

Marian Thomson

EBRC Easter Bush Research Consortium

wellcome trust

BBSRC bioscience for the future

Contact

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