ROSLIN

Comparing short and long read methods for transcriptome discovery

Richard Kuo

THE UNIVERSITY of EDINBURGH

BBSRC
20 Years of Pioneering
Great British Bioscience

Why Iso-Seq: Alt. Transcripts

Chicken Ensembl v85

Gene counts

Coding genes	15,508
Non coding genes	1,558
Small non coding genes	1,408
Misc non coding genes	150
Pseudogenes	42
Gene transcripts	17,954

Iso-Seq annotation

# of Transcripts	Biotype
43,738	Coding RNA
20,516	LncRNA
23	Short ncRNA
4,735	NMD transcript
12	Processed Pseudogene
27	Unprocessed Pseudogene
13,873	Antisense Exonic
2,139	Antisense Intronic

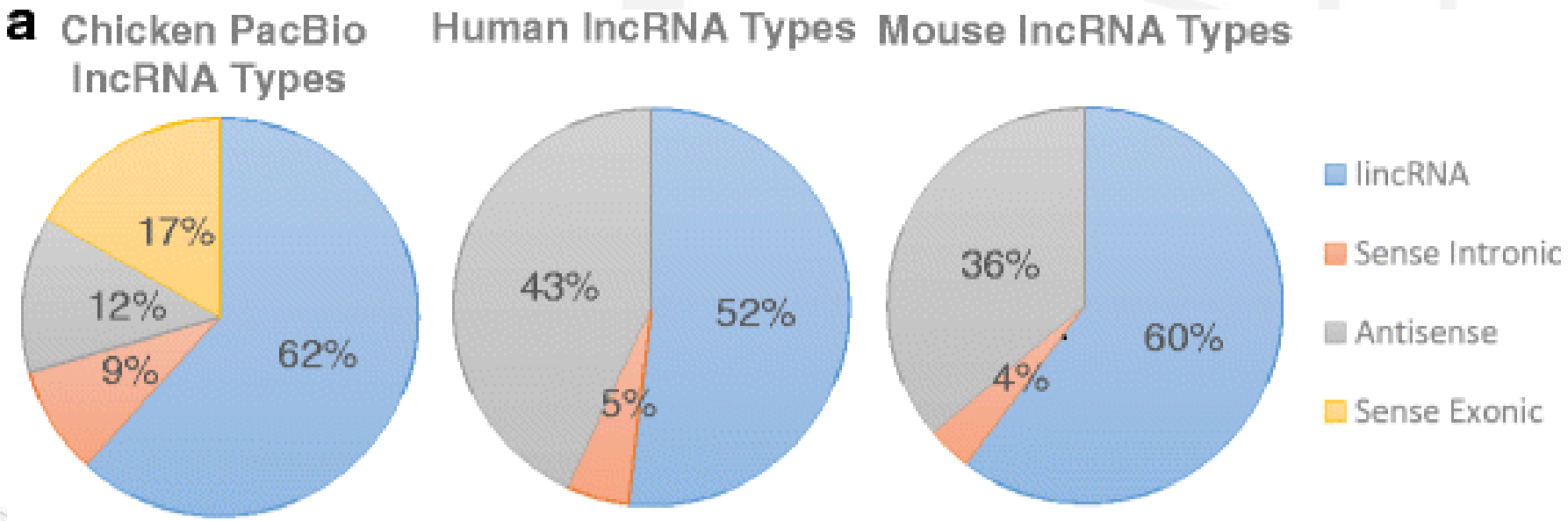
Coding genes	14,421
lncRNA genes	17,178
Transcripts	64,277

http://weknowyourdreams.com/image.php?pic=/images/chicken/chicken-01.jpg

Why Iso-Seq: lncRNA

Human	Ensembl 83
13473	lincRNA
977	sense_intronic
5	bidirectional_promoter_lncrna
1	macro_lncRNA
11186	antisense
343	sense_overlapping

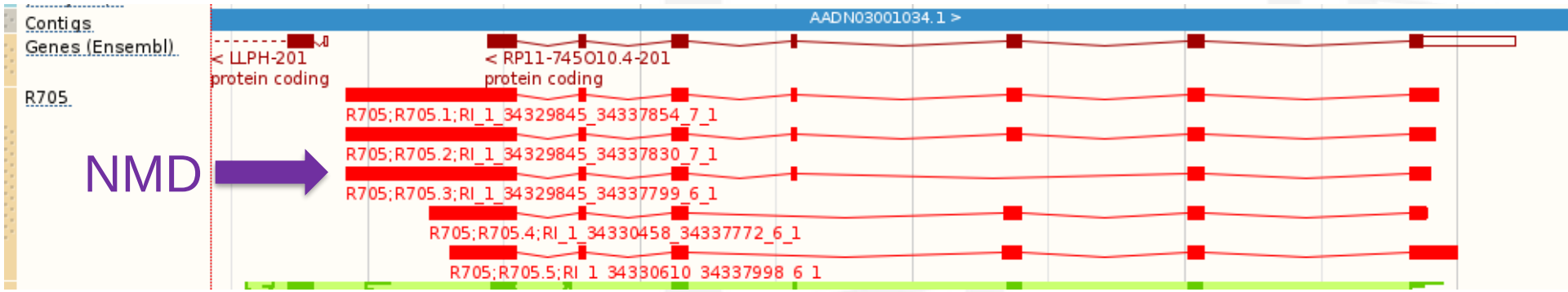
Chicken	Pacbio
8653	lincRNA
2329	sense intronic
4207	antisense exonic
2035	antisense intronic
7265	sense exonic



Why Iso-Seq: NMD

- Nonsense mediated decay transcripts
  - Looks a lot like a coding transcript
  - Premature stop codon
  - Does not make a protein
  - Implicated for fine tuning regulation

NMD Transcripts	Ensembl 83
Human	13401
Mouse	5229
Chicken	0
NMD Transcripts	Pacbio
Chicken	4735



Overview

- Goals for transcriptome discovery/annotation
- The project
- RNAseq Issues
  - TSS/TTS
  - Alternative Splicing/Exon Chaining
  - Gene merge
  - Annotation merging
- IsoSeq Issues
  - 5’ cap
  - Poly-A internal or genomic
  - RT Switch
  - Throughput

Transcriptome Annotation

- Full length transcript models
  - Accurate starts and ends
  - Accurate splice sites
  - Exon chaining
- Strand
- Expression profile
- Function
- Pathways

The project

- Incorporate long read and short read data to create a modern transcriptome/genome annotation
- Iso-Seq sequencing of Chicken tissues: brain, embryo, spleen, macrophage, ovary, testes, etc.
  - Comparing methods for library preparation: 5’ cap selection, normalization
- Short read sequencing from same samples: RNAseq, CAGEseq
- Presented comparison data:
  - Iso-Seq Sequel sequencing from chicken brain with 5’ cap selection (3 cells)
  - Iso-Seq RSII different chicken brain with normalization (25 cells)
  - RNAseq 150bp paired end stranded sequencing from same sample

http://weknowyourdreams.com/image.php?pic=/images/chicken/chicken-01.jpg

Chicken Brain Data

5’ cap selected, size selection free Iso-Seq (Sequel 1m cell)

	# Reads	% Reads ZMW	# CCS	# FLNC	% FLNC
Cell 1	84,208	8.42	65,565	53,614	63.67
Cell 2	358,268	35.83	266,640	194,539	54.30
Cell 3	314,211	31.42	234,102	174,010	55.38
Total	756,687	25.22	566,307	422,163	55.79

No cap selection, normalized, size selected Iso-Seq (RSII P4C2 150k cell)

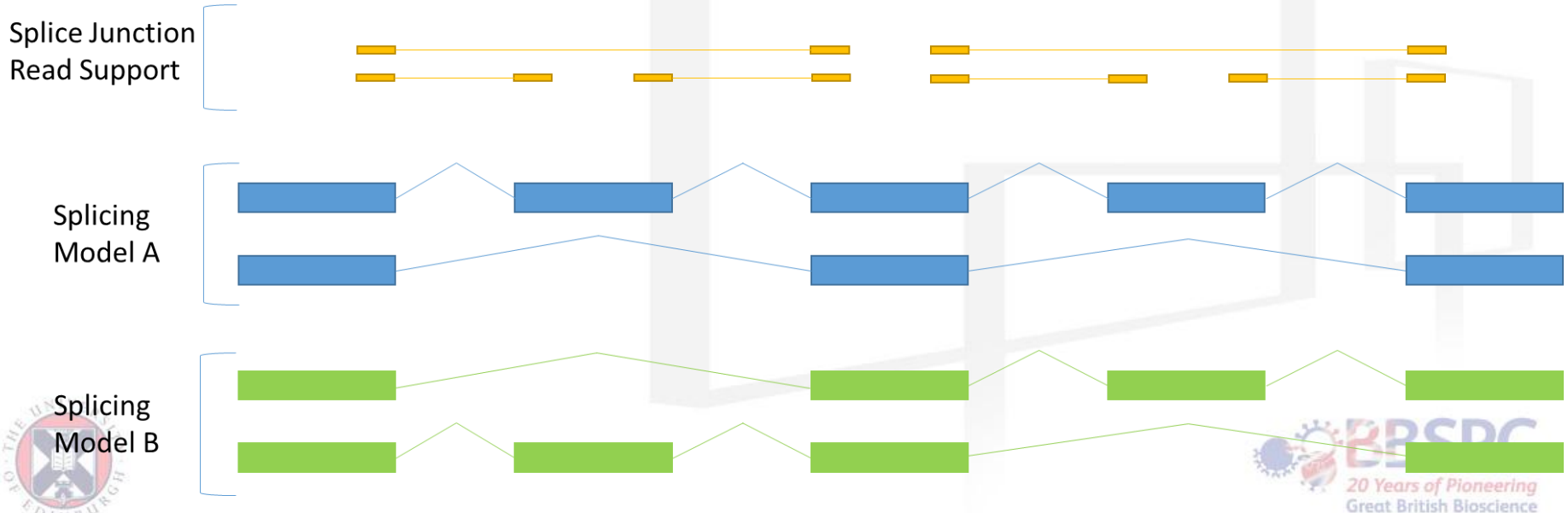
	# Reads	% Reads ZMW	# CCS	# FLNC	% FLNC
1kb	592,180	35.82	405,717	283,750	47.92
2kb	878,276	41.74	399,889	231,425	26.35
Total	1,470,456	39.14	805,606	515,175	35.04

381,444,876 PE reads from RNAseq

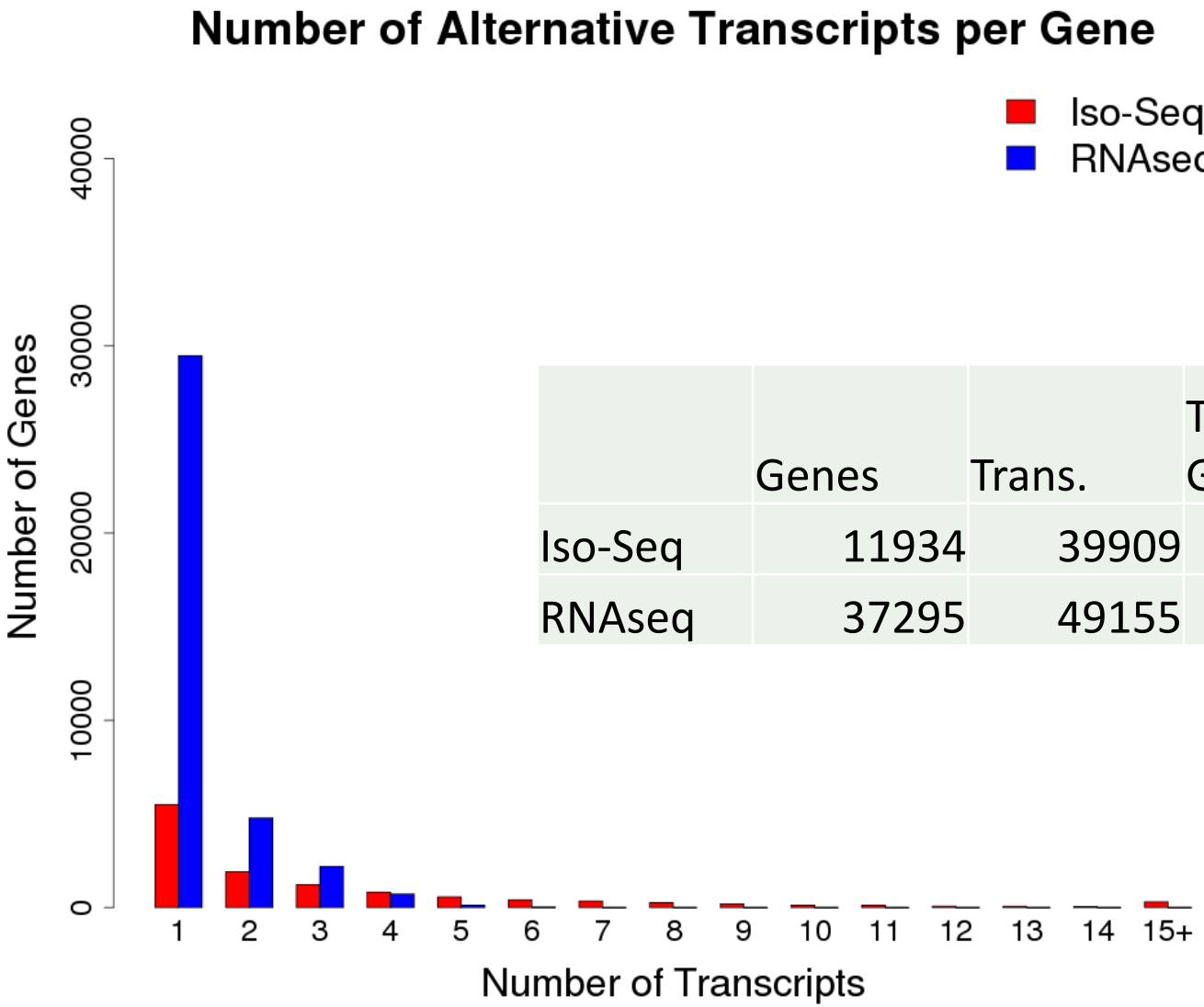
FLNC – Full length non-chimeric reads

Short read issues

- Impossible to be sure of transcript model assembly
- Alternative start and end sites hidden
  - Transcription Start Sites/transcription termination sites (TSS/TTS)
- Coding potential is difficult to ascertain
  - Exon chaining
- Transcriptional noise
- Gene merging

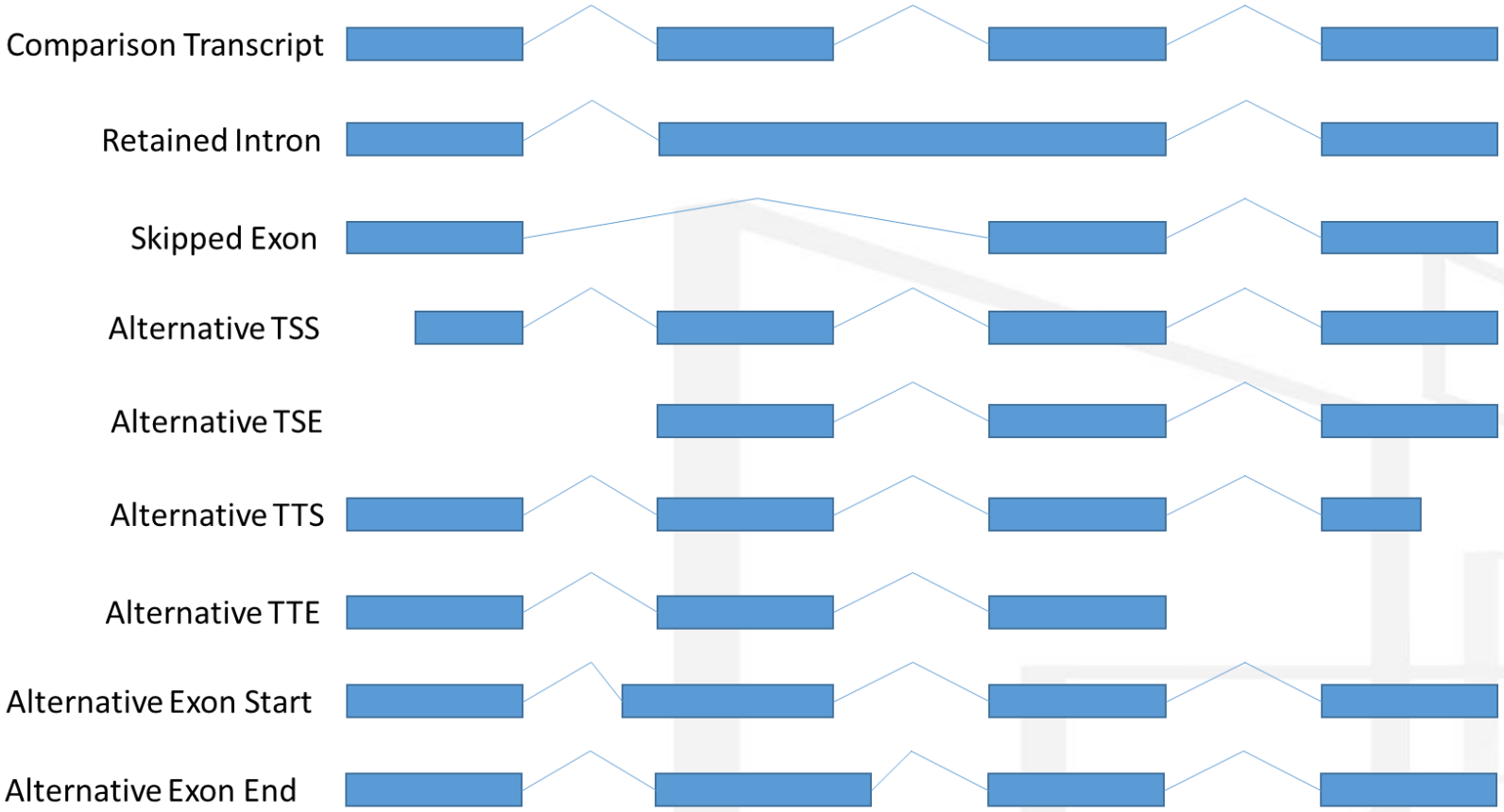


RNAseq: Alt. Transcripts

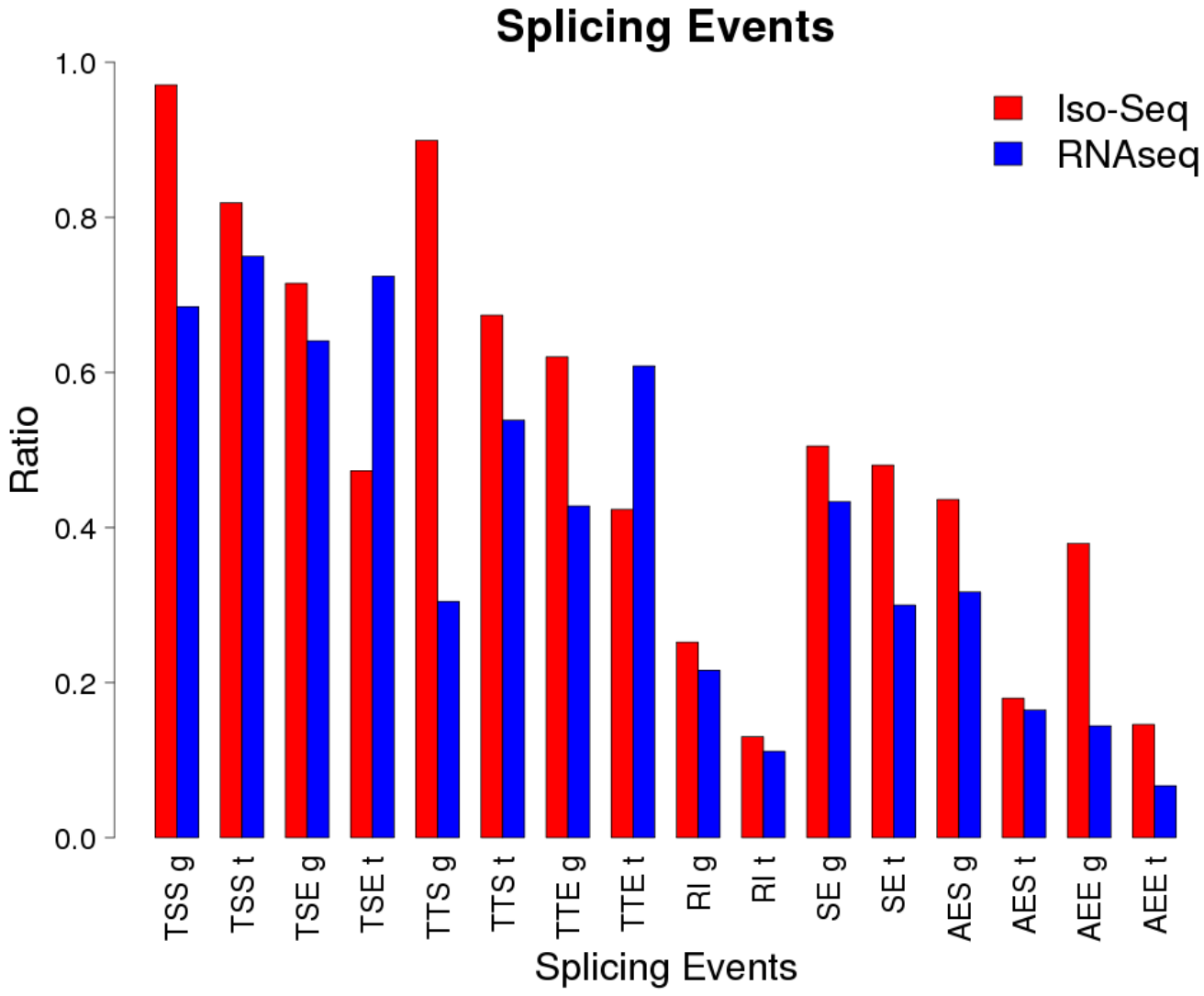


	Genes	Trans.	Trans. per Gene
Iso-Seq	11934	39909	3.34
RNAseq	37295	49155	1.31

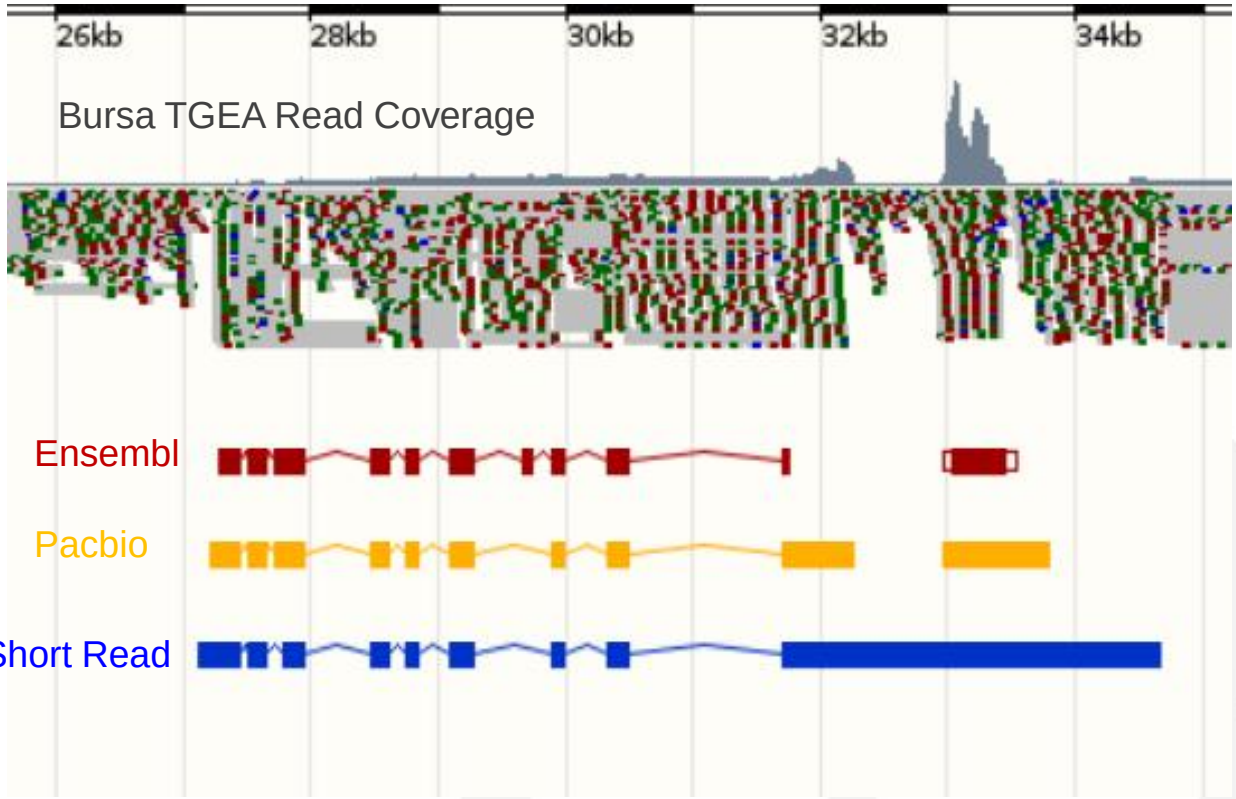
Alternative Transcript Types



RNAseq: Alt. Transcript Events

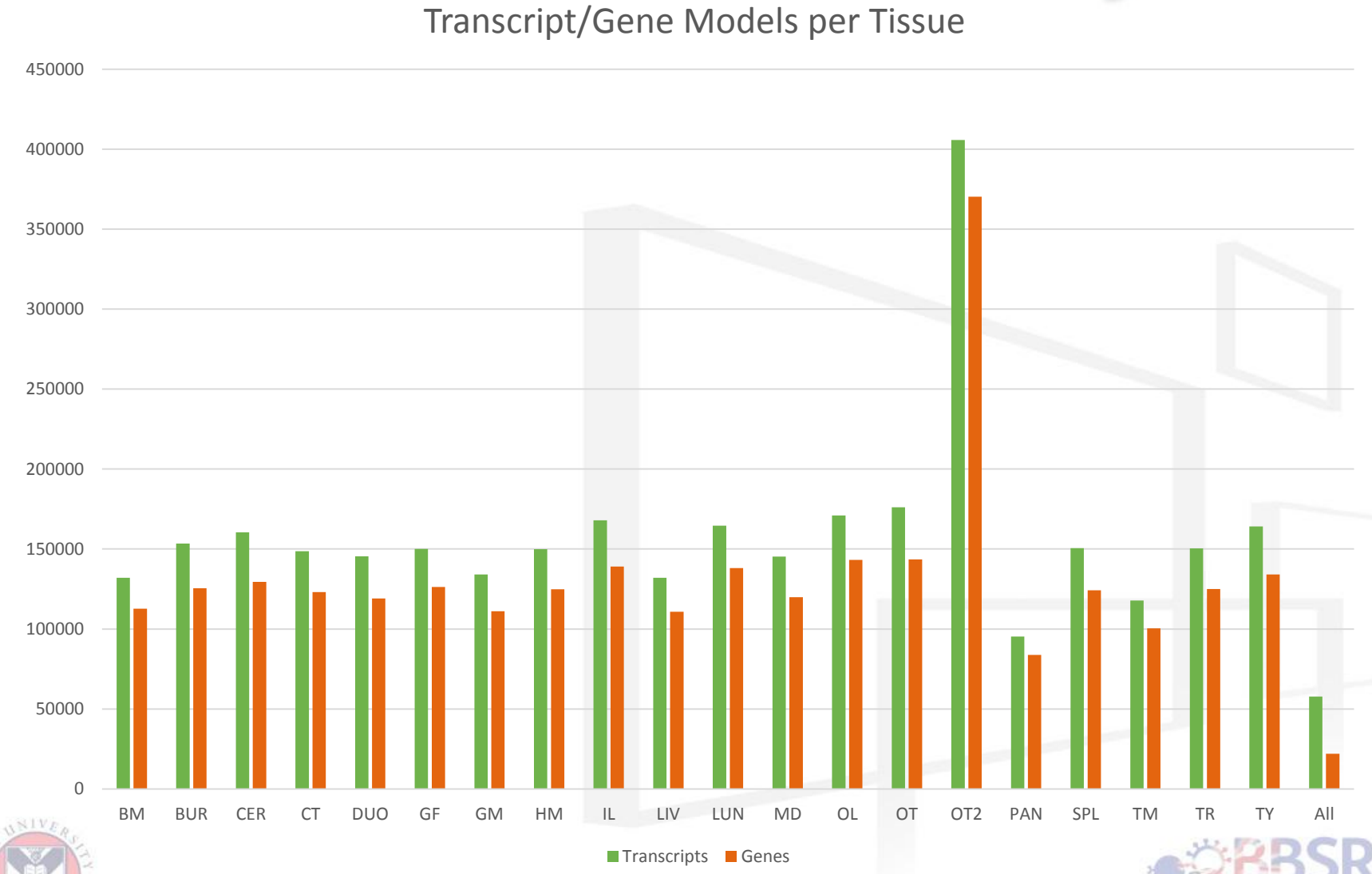


RNAseq: Gene Merge



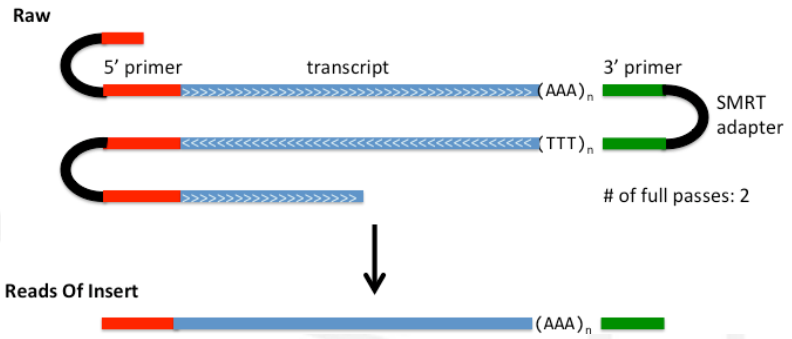
- 459 merged gene events involving 950 Iso-Seq genes
- 11,934 total Iso-Seq genes

RNAseq: Annotation Merging



Long read issues

- Throughput
- Transcript ends
  - 5’ degradation : Solved by 5’ cap selection
  - 3’ truncation from internal Poly A
- Splice Sites
  - RT Switch : Minimal occurrence
- Error Rate
  - Solved with multiple sub-read passes
- Genomic contamination
  - Solved by 5’ cap selection



https://github.com/PacificBiosciences/cDNA_primer/wiki/Understanding-PacBio-transcriptome-data

Iso-Seq: 5’ Cap

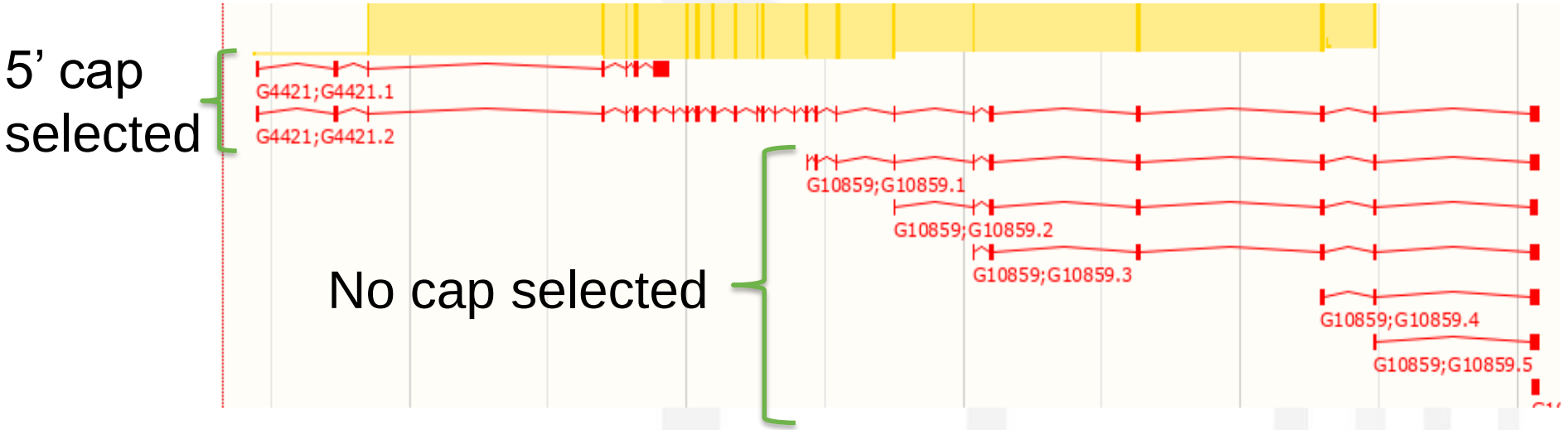
- Does 5’ cap selection make a difference?
  - Collapsed using Iso-Seq Tofu Collapse tool
  - Used both methods of collapsing to compare

TSSC – Transcription Start Site Collapse
ECC – Exon Cascade Collapse

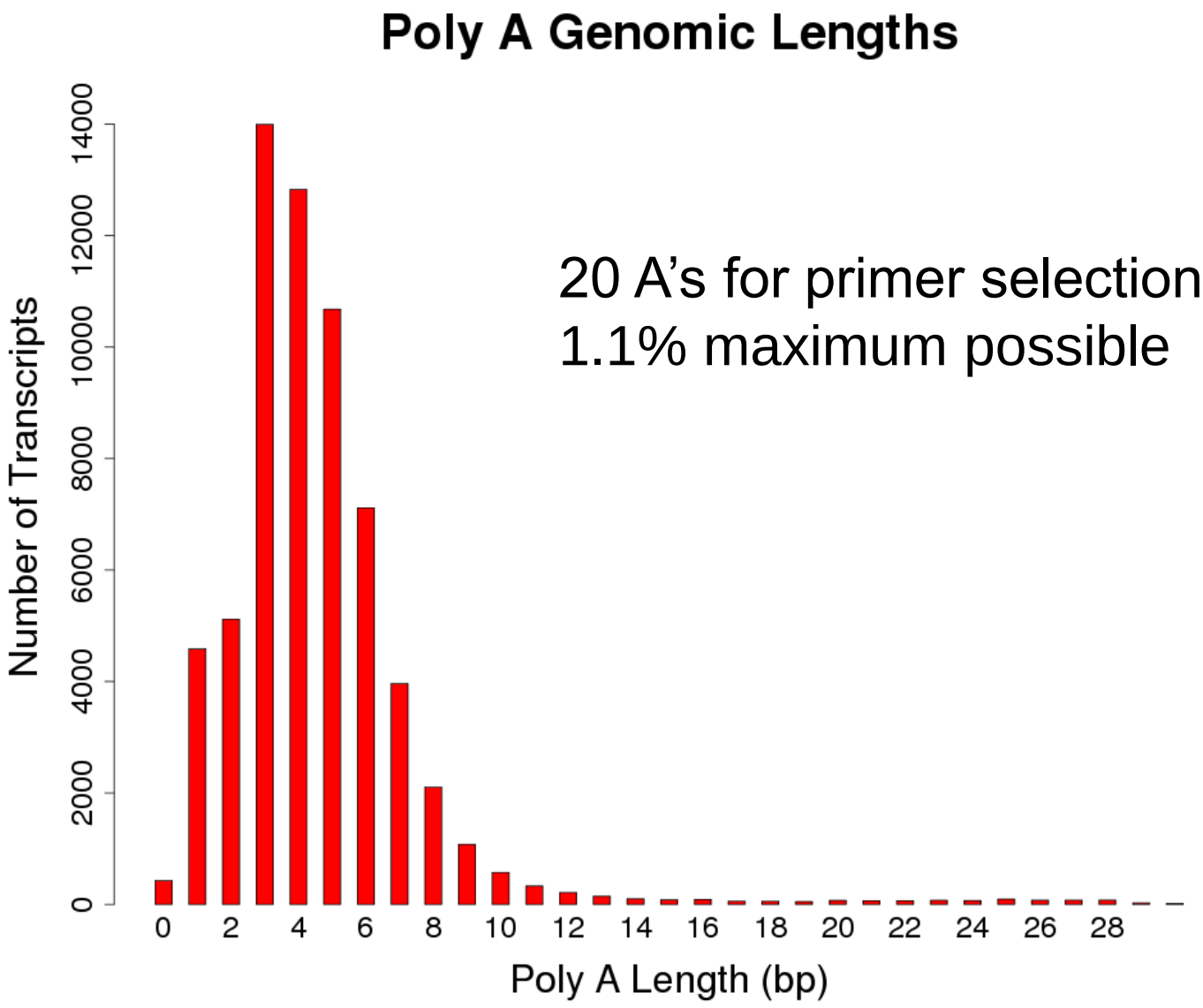
	Pre-collapsed	TSSC	ECC	TSSC % decrease	ECC % decrease
Brain	199,560	80,814	55,932	59.50%	72.00%
Embryo	11,881	9,368	8,468	21.20%	28.70%

http://cdn.analyteguru.com/uploads/2015/05/Campylobactor-jejuni-in-chicken-feces-2.jpg

Iso-Seq: 5’ Cap


5’ cap selected
G4421;G4421.1
G4421;G4421.2
No cap selected
G10859;G10859.1
G10859;G10859.2
G10859;G10859.3
G10859;G10859.4
G10859;G10859.5

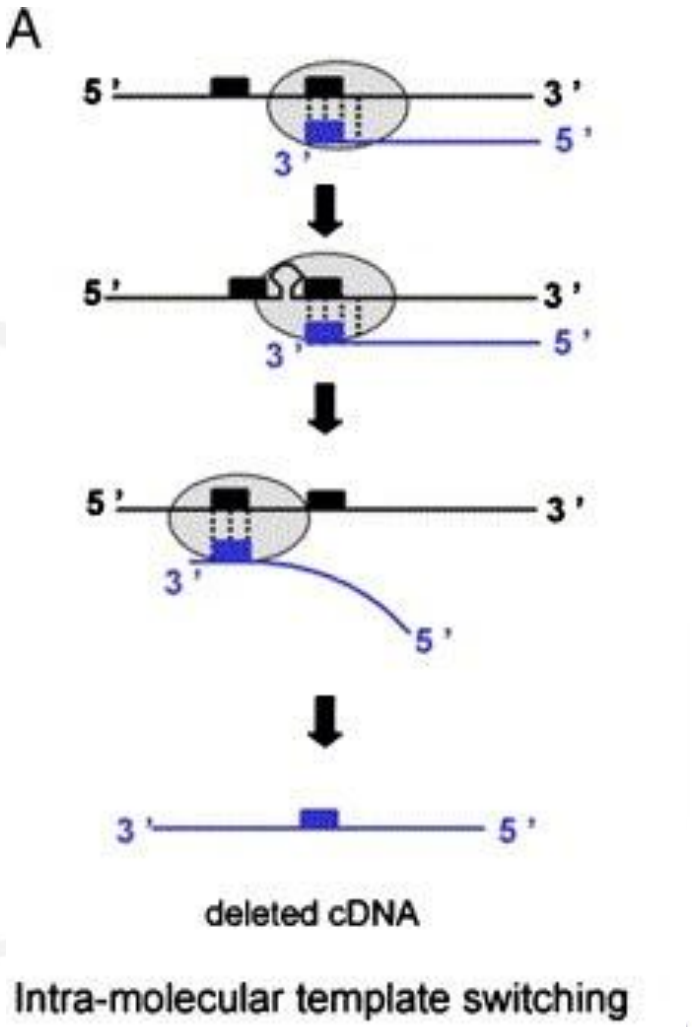
Iso-Seq: Poly-A



20 A's for primer selection
1.1% maximum possible

Iso-Seq: RT Switch

- Reverse transcriptase template switching
- Appears as novel splice junction
- Result of RNA structure and repeat regions
- SQANTI predicts these events
- <1% occurence



"Reverse transcriptase template switching and false alternative transcripts" by Cocquet et al

Iso-Seq: Throughput

Non-Normalized Iso-Seq (Sequel 1m cell)

	# Reads	% Reads ZMW	# CCS	# FLNC	% FLNC
Total	756,687	25.22	566,307	422,163	55.79

Normalized Iso-Seq (RSII P4C2 150k cell)

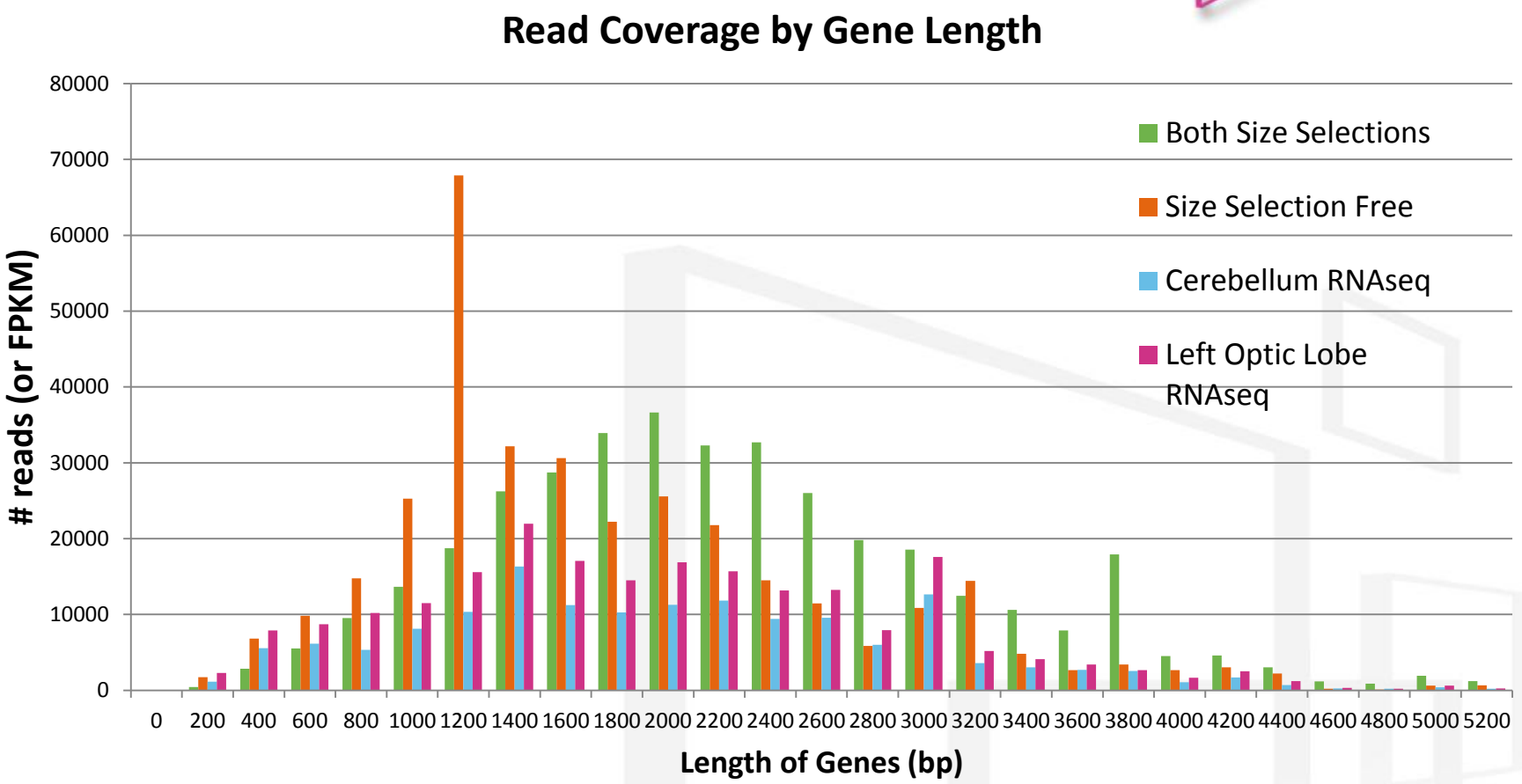
	# Reads	% Reads ZMW	# CCS	# FLNC	% FLNC
Total	1,470,456	39.14	805,606	515,175	35.04

	Genes	Transcripts	Transcripts per Gene
Non-Norm.	11934	39909	3.34
Normalized	30913	55932	1.81

Normalized library contained ~14k single exon lncRNA

FLNC – Full length non-chimeric reads

Iso-Seq: Throughput



- 2 genes in 1200 bin for Sequel run associated with 37,679 reads

Future/Ongoing Work

- Full characterization of methodology using:
  - 5’ cap selected non-normalized libraries
  - 5’ cap selected normalized libraries
  - RNAseq
  - CAGEseq
- Transcriptome Annotation Construction Software (TACoS)
- lncRNA functional predictions

Acknowledgement

ROSLIN

PACIFIC BIOSCIENCES®

Professor Dave Burt

Professor Alan Archibald

Katarzyna Miedzinska

Bob Paton

Lel Eory

Steve Picton

Elizabeth Tseng

edinburgh genomics.

Karim Gharbi

Marian Thomson

EBRC Easter Bush Research Consortium

wellcome trust

BBSRC bioscience for the future