



GEORGETOWN UNIVERSITY

Lombardi
COMPREHENSIVE
CANCER CENTER

How looking for a needle in a Haystack taught me to love the Isoform

Anne Deslattes Mays

Mentor: Anton Wellstein, MD, PhD

Trainer: Marcel Schmidt, PhD

Collaborator: Elizabeth Tseng, PhD Pacific Biosciences

8/20/2014

Wellstein/Riegel Laboratory,
Lombardi Cancer Center,
Washington DC 20007



PACIFIC
BIOSCIENCES™

Slide 1

1

I love the title! excellent!

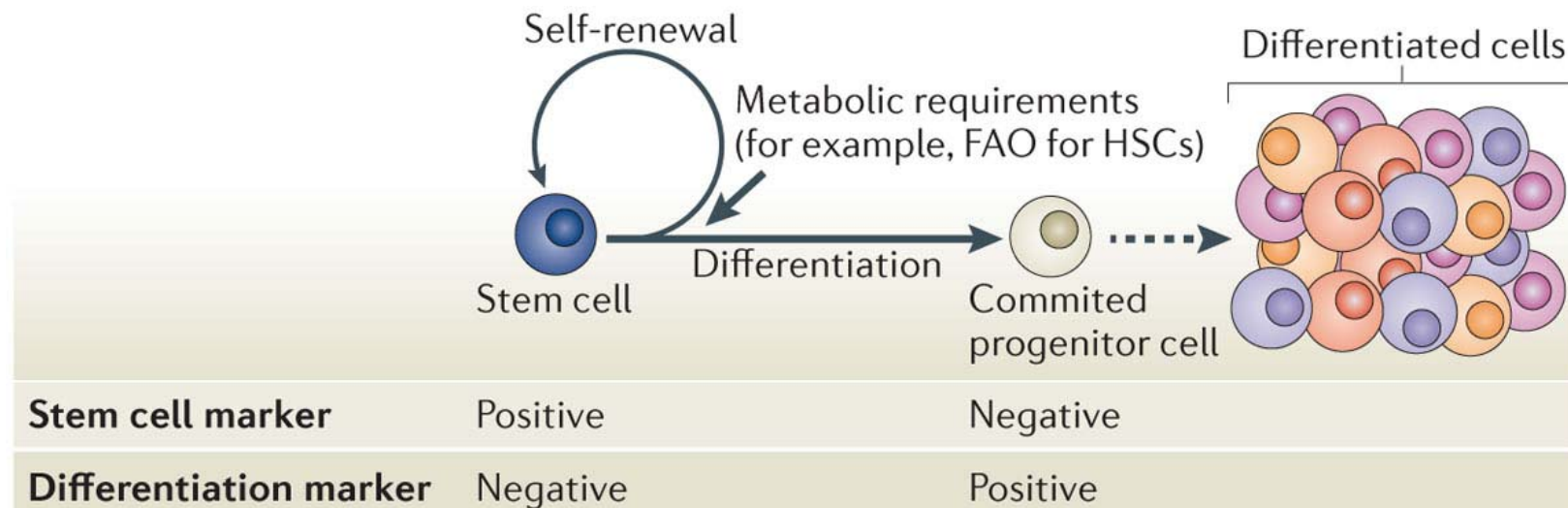
lachesis, 6/12/2014



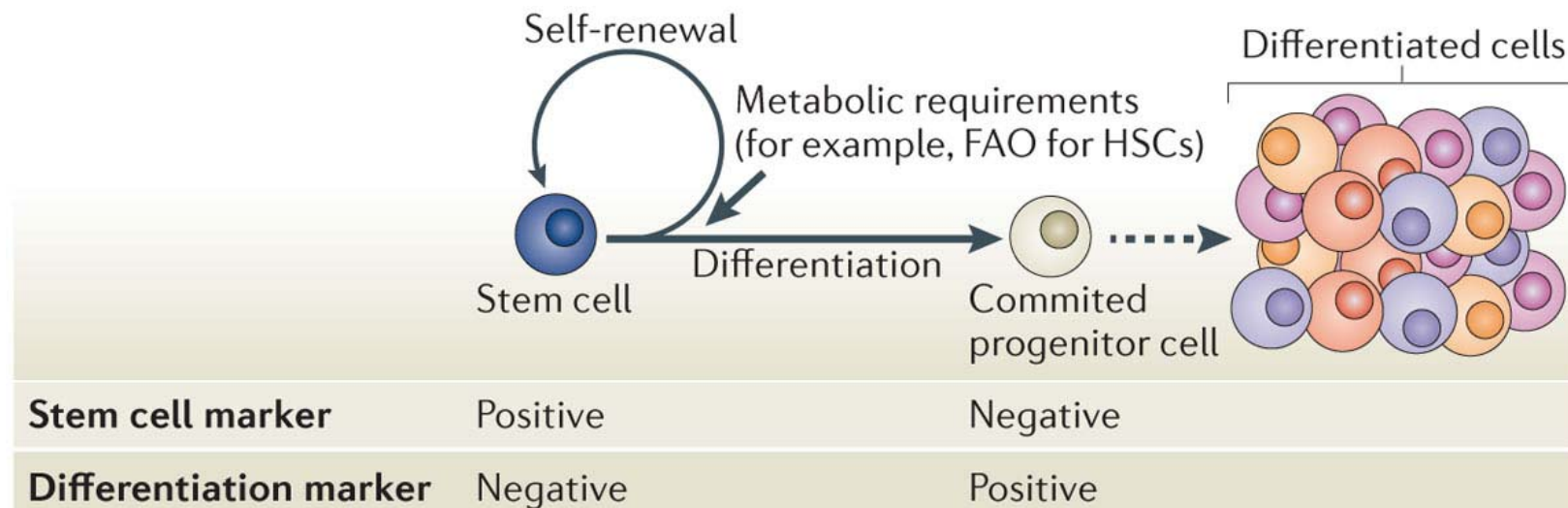
Outline

- Motivation
- PacBio Sequencing
- Methods
 - Short Read Workflow
 - Long Read Workflow
- Results
 - Population-specific isoform distributions
 - Information gain from PacBio
- Conclusion

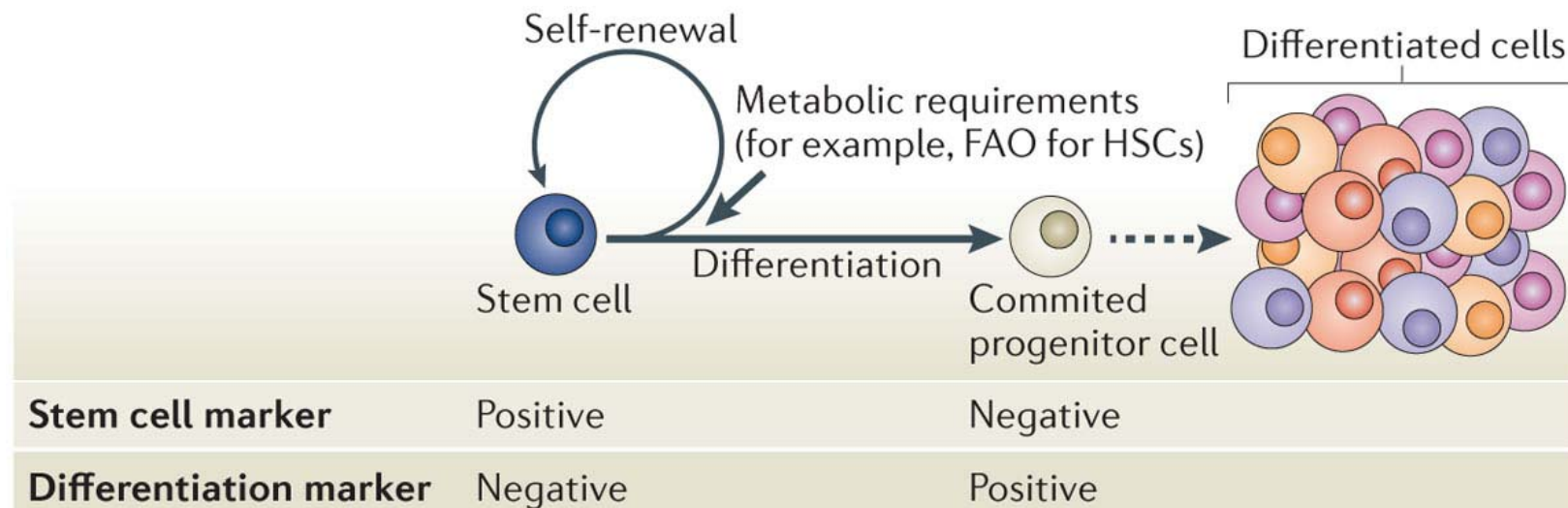
Bone Marrow is a rich heterogeneous mixture containing platelets, white blood cells, red blood cells (committed progenitor cells) and their uncommitted precursors or stem cells



Nature Reviews | [Molecular Cell Biology](#)



The Wellstein/Riegel lab focuses on studying and discovering what drives this differentiation with the hope to use this information to help discover both markers and potential therapeutic directions for the treatment of cancer

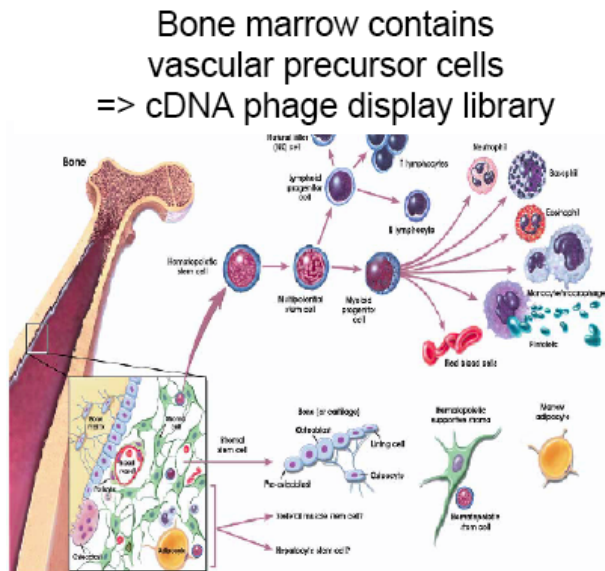


One particular experiment begun nearly a dozen years ago created a model system with which to study these drivers and potentially discover new ones

Using bone-marrow derived monocytes – a random human cDNA phage display library was created.

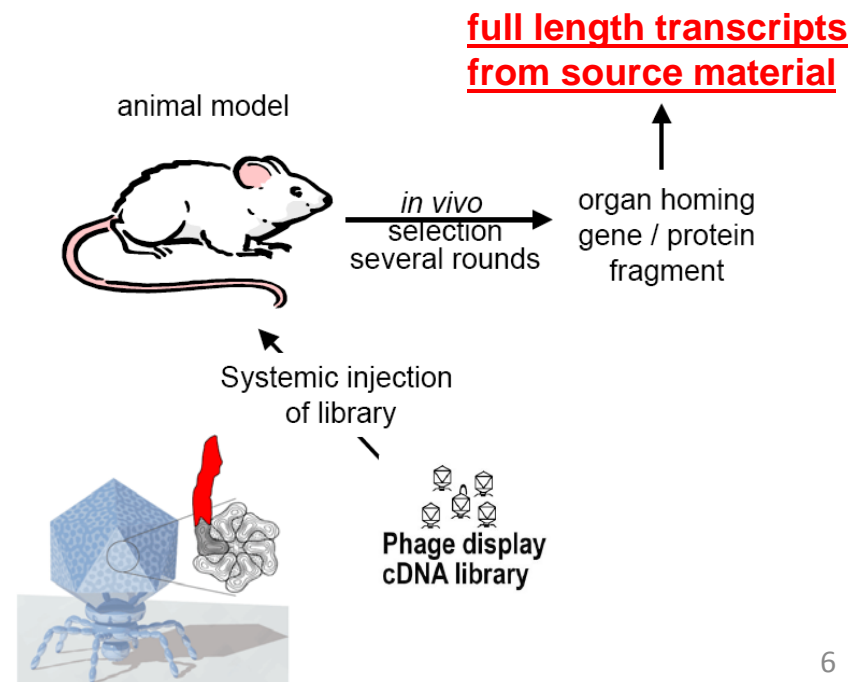
Questions:

1. Which proteins drive organ homing of hematopoietic cells?
1. Are there distinct homing proteins for diseased organs (cancer, wound healing, ischemia, infection)?



BM & derived cells

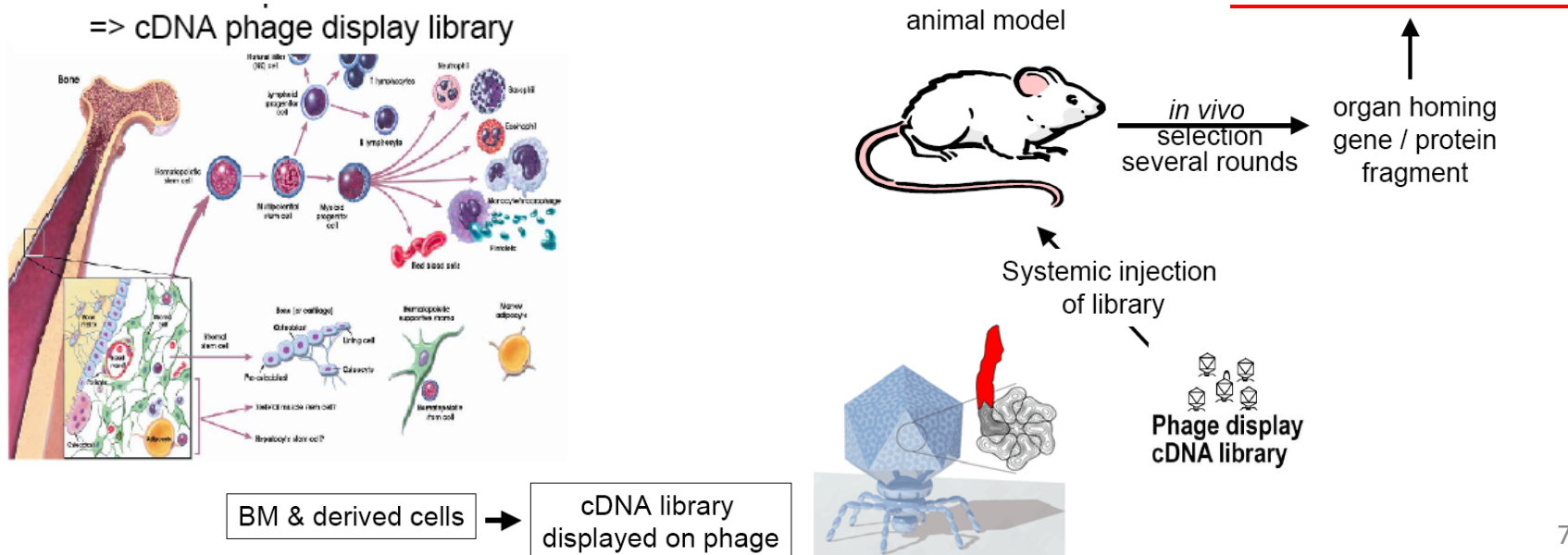
cDNA library
displayed on phage



The result was a set of 40 small (300-900 base pair) gene/protein fragments that were highly expressed in the lineage negative cell population of total bone marrow.

But they were fragments – no full length transcripts – and without the full length – further study on these apparently early homing genes hit a stand still

Enter high throughput RNA-Seq – and the effort to discovery the full length entered another stage





Find a needle in the haystack

Initial Project Goal:

Identify full-length transcripts using 2nd and 3rd generation sequencing in bone marrow cell populations



Find a needle in the haystack

Sample Extraction

- From freshly harvested, viable human bone marrow tissues

Cell Subpopulation Selection

- Negatively select (lin-) progenitor cells pulling out differentiated cells using hematopoietic lineage (lin+) antibodies to cell surface markers magnetic bead sorting

Next-gen Sequencing

- SOLiD (35bp, 50bp)
- Illumina HiSeq (100bp paired-end)
- PacBio (Iso-Seq, 1 – 6 kb)

Discarded harvesting equipment collected from MedStar Georgetown University Hospital Cell Processing Unit

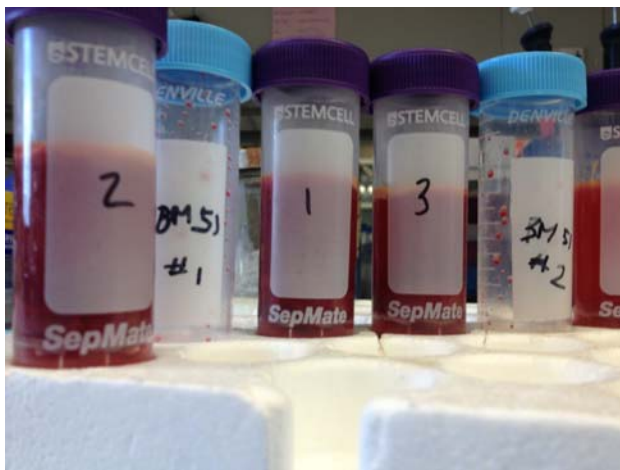
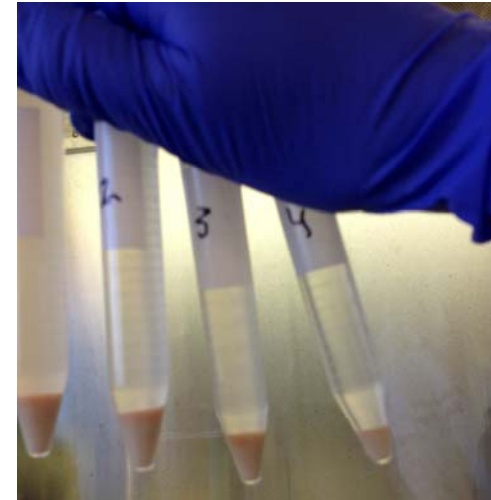
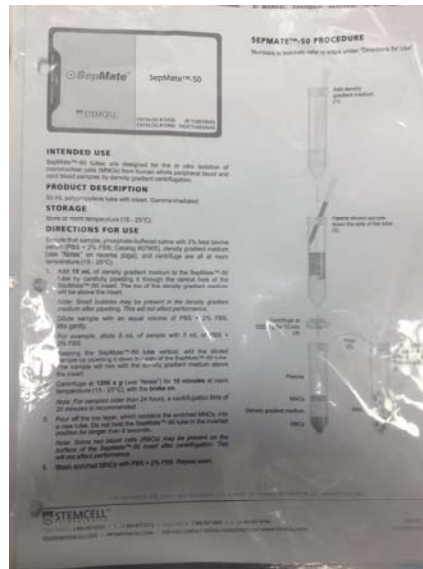


Cell Processing and Cellular Engineering Laboratory
HPC Waste Distribution Tracking Log

Page 1

Distributed Date	For Research Only		BIO	Product (Phase 1)		Batches	Comments
	Distributed by (Last Name)	Phone #		Batch Name	Batch Lot		
2/15/12	Alice Belletier M.D.	240.228.2205	7056	✓		1	2/15/12
2/15/12	3rd floor, 125 above	3rd floor, 125 above	7070	✓		1	2/15/12
2/15/12	3rd floor, 125 above	3rd floor, 125 above	7071	✓		1	2/15/12

Negatively select using antibodies hematopoietic differentiated surface markers with magnetic bead sorting



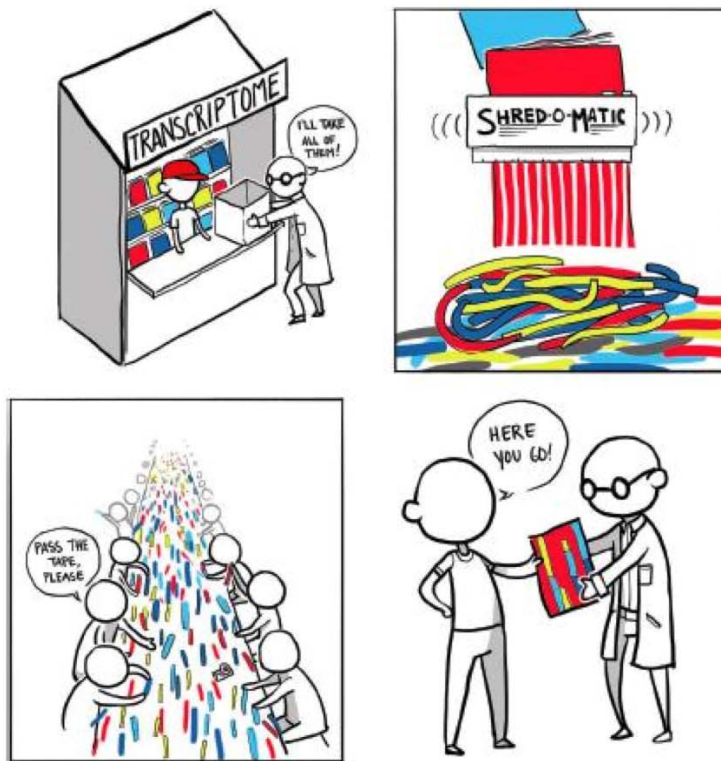
The result is three cell subpopulations:

- Total bone marrow
- Hematopoietic progenitor positive (lineage negative)
- Hematopoietic progenitor negative or differentiated cells (lineage positive)

Summarizing Short Read Results

- Short reads assembled using common assembly pipeline
 - Genome-guided: gsnap gapped alignment + transcript assembly with cufflinks
 - *De novo*: Trinity
 - Transcriptome Alignment: GMAP
- Short read assembly results show that
 - Strand information unclear (went to stranded)
 - Isoform structures remain unclear (scattered hits but not complete resolution)
 - Targeted selection showed we could make custom libraries to sequence – however after such careful library construction – what did we do?

Current State of Transcript Assembly



“The way we do RNA-seq now is... you take the transcriptome, you **blow it up into pieces** and then you try to figure out **how they all go back together again**... If you think about it, it's kind of a **crazy way to do things**”

Michael Synder
Professor and Chair of Genetics
Stanford University

Tal Nawy, End to end RNA Sequencing, *Nature Methods*, v10, n10, Dec . 2013, p1144–1145

Figure 1 | Transcriptome reconstruction—akin to reassembling magazine articles after they have been through a paper shredder.

[Ian Korf \(2013\) Genomics: the state of the art in RNA-seq analysis, *Nature Methods*, Nov 26;10\(12\):1165-6. doi: 10.1038/nmeth.2735.](#)

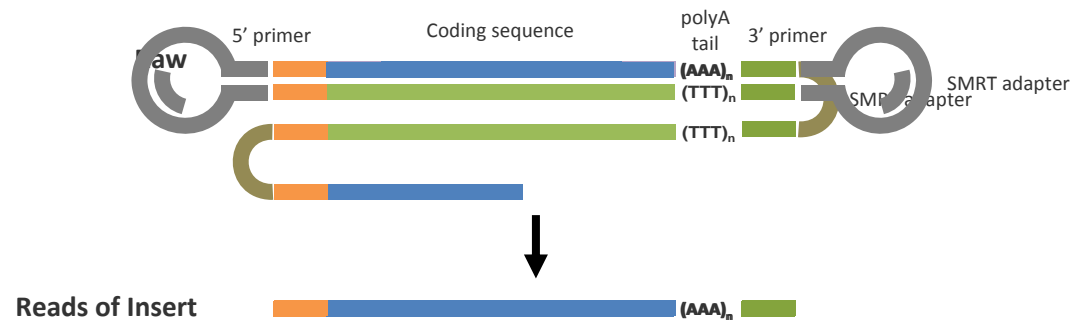
Source: Iso-seq webinar by Liz Tseng, Pacific Biosystems
https://github.com/PacificBiosciences/cDNA_primer/wiki/Understanding-PacBio-transcriptome-data

PacBio's Iso-Seq™ Method for High-quality, Full-length Transcripts

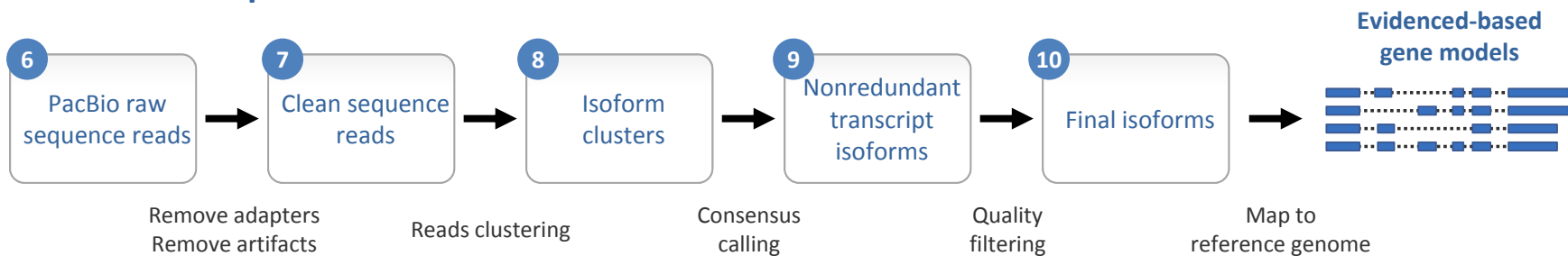
Experimental Pipeline



[SampleNet: Iso-Seq Method with Clontech cDNA Synthesis Kit](#)



Informatics Pipeline



[DevNet: Iso-Seq wiki page](#)



Final Workflow

Short Read Workflow

- Genome-guided assembly¹
- *de novo* assembly²
- Optimized short read alignment³

Long Read Workflow

- SMRTAnalysis RS_IsoSeq to get full-length transcripts⁴

Analysis

- Quantitation using Sailfish⁵
- Compare isoform distributions within and across three cell populations⁶

¹Cufflinks - http://cufflinks.cbc.umd.edu/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz

²Trinity - http://sourceforge.net/projects/trinityrnaseq/files/trinityrnaseq_r20140413p1.tar.gz/download

³GSNAP- <http://research-pub.gene.com/gmap/src/gmap-gsnap-2014-03-28.v2.tar.gz> (newer release available!)

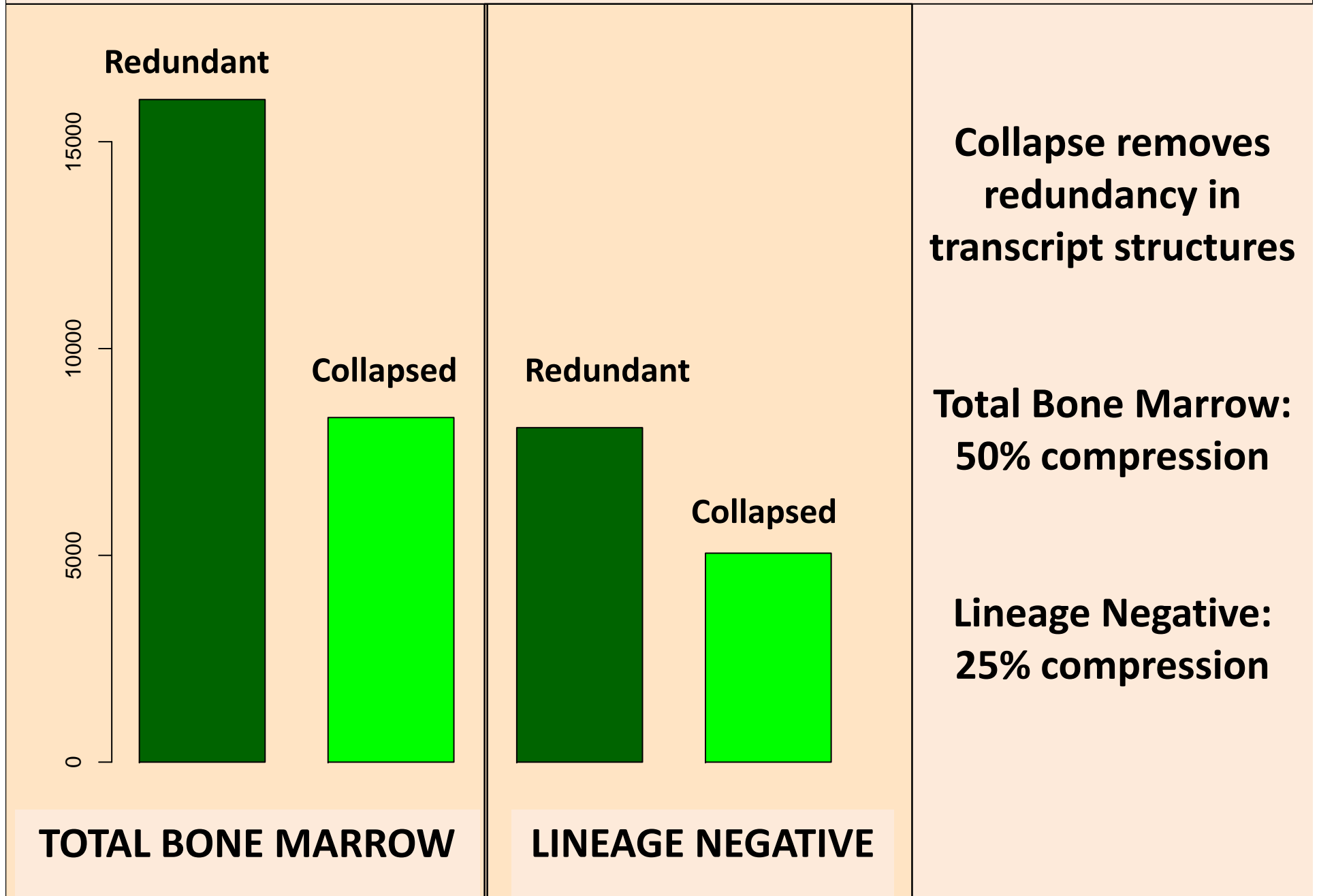
⁴ICE, Elizabeth Tseng, part of the SMRTAnalysis 2.2

⁵Sailfish

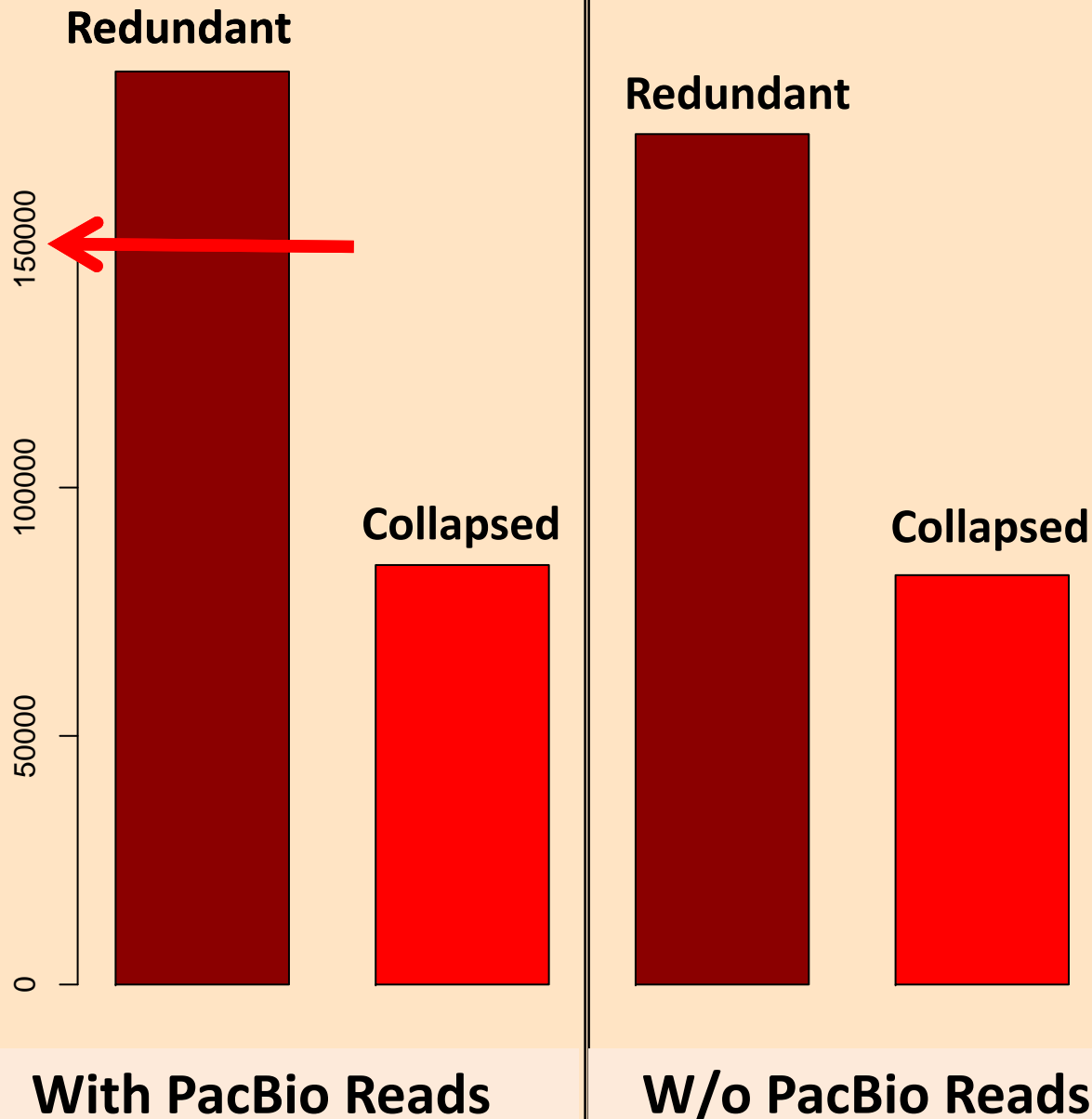
https://github.com/kingsfordgroup/sailfish/releases/download/v0.6.3/Sailfish-0.6.3-Linux_x86-64.tar.gz

⁶Custom R Script

Collapsing Redundant Transcripts: PacBio



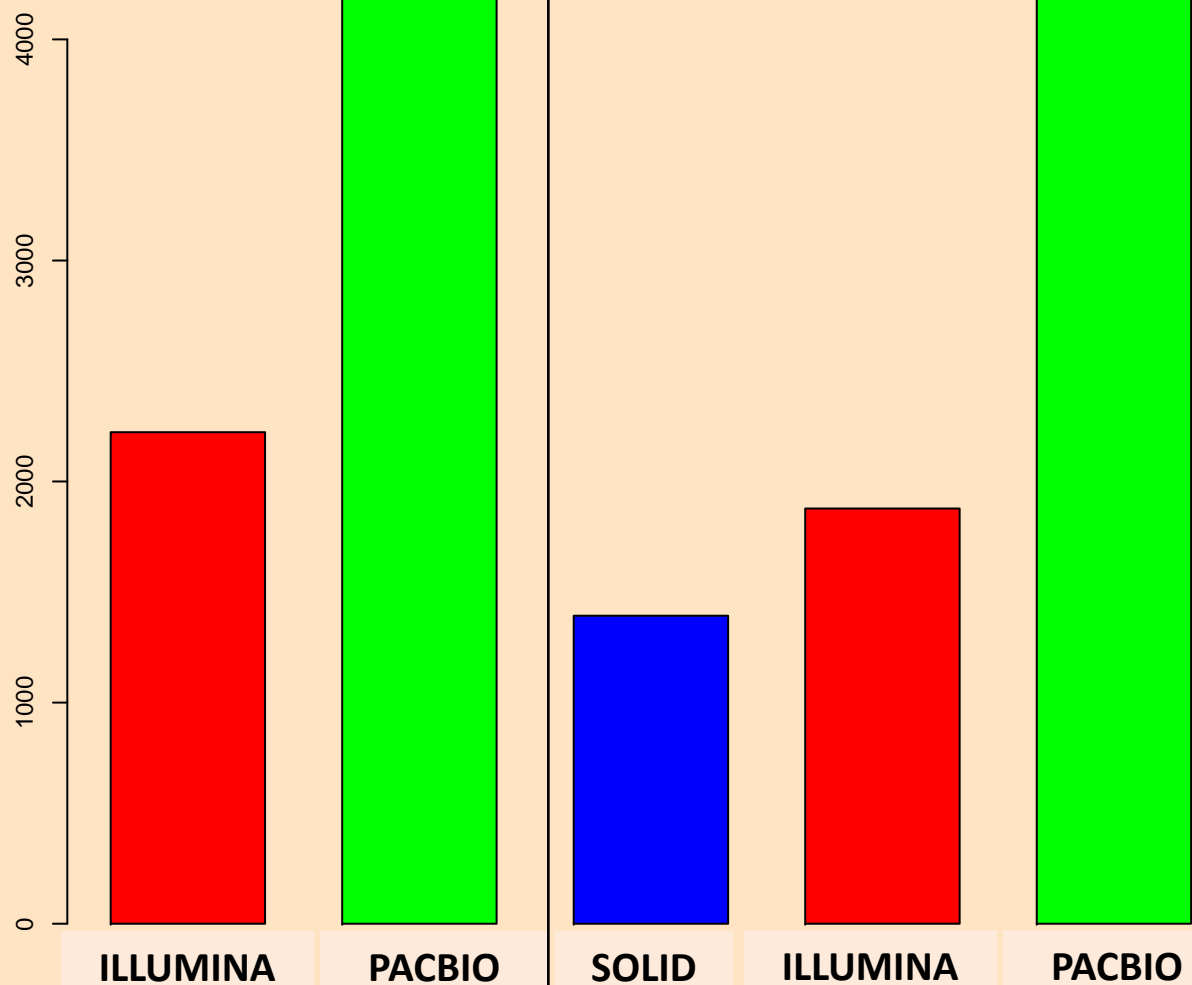
Collapsing Redundant Transcripts: Short Read Assembled



**Collapse removes
redundancy in
transcript structures**

**w/ PacBio Reads:
50% compression**

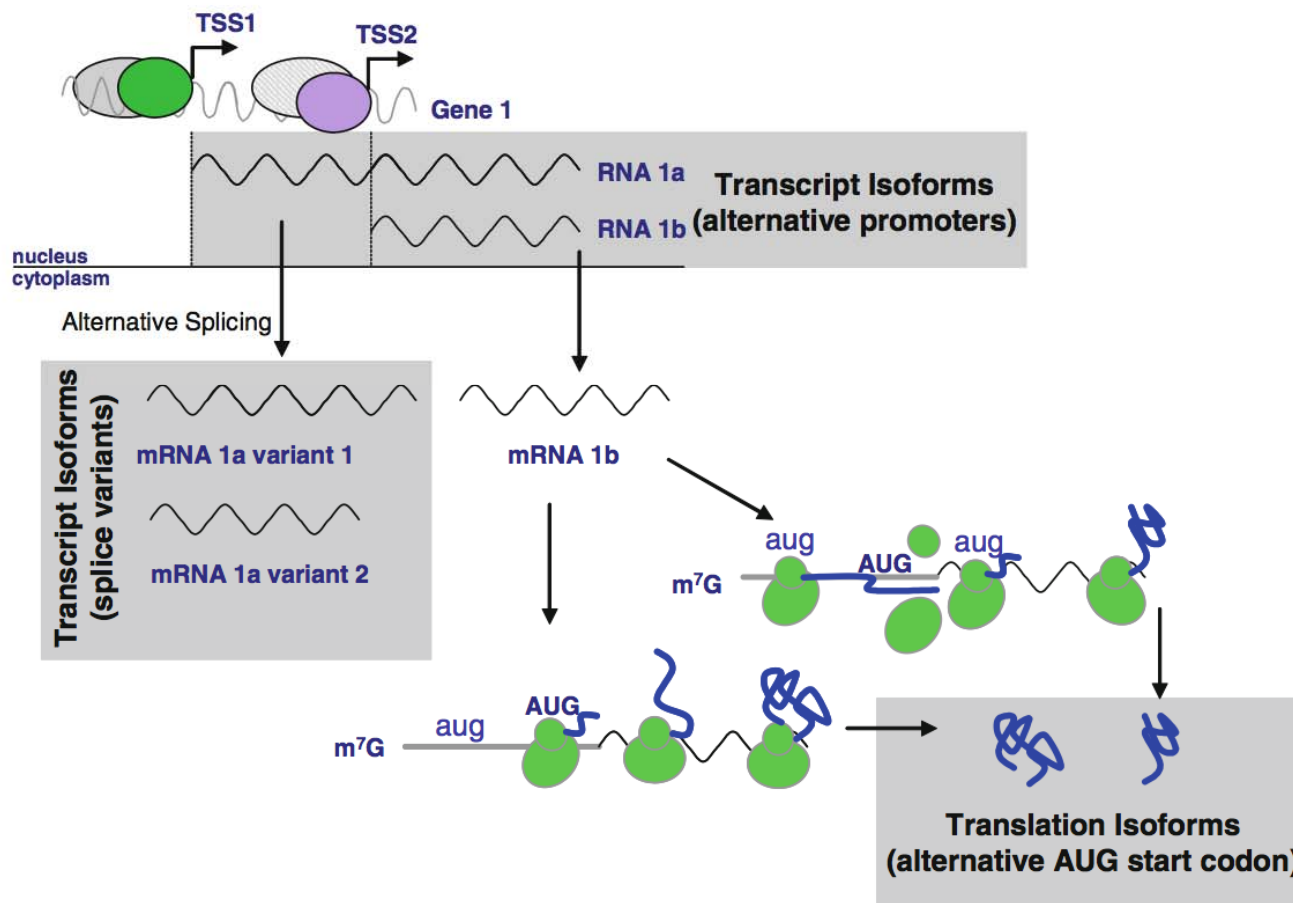
Unique Genes per platform



50% gain in gene information with PacBio long reads

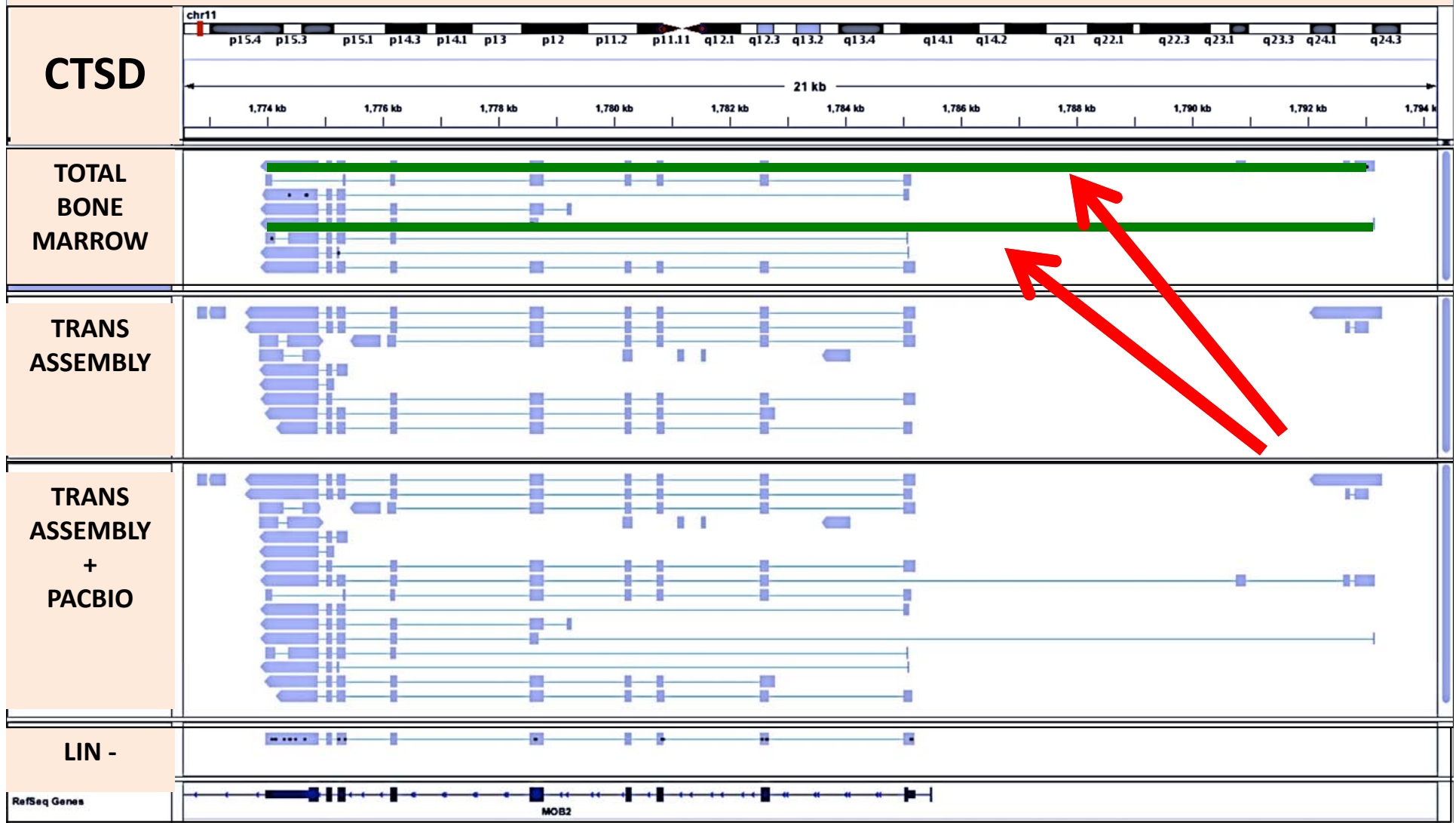
Isoforms in Hematopoiesis

Isoforms exist in a mixture specific to the sub-cell population

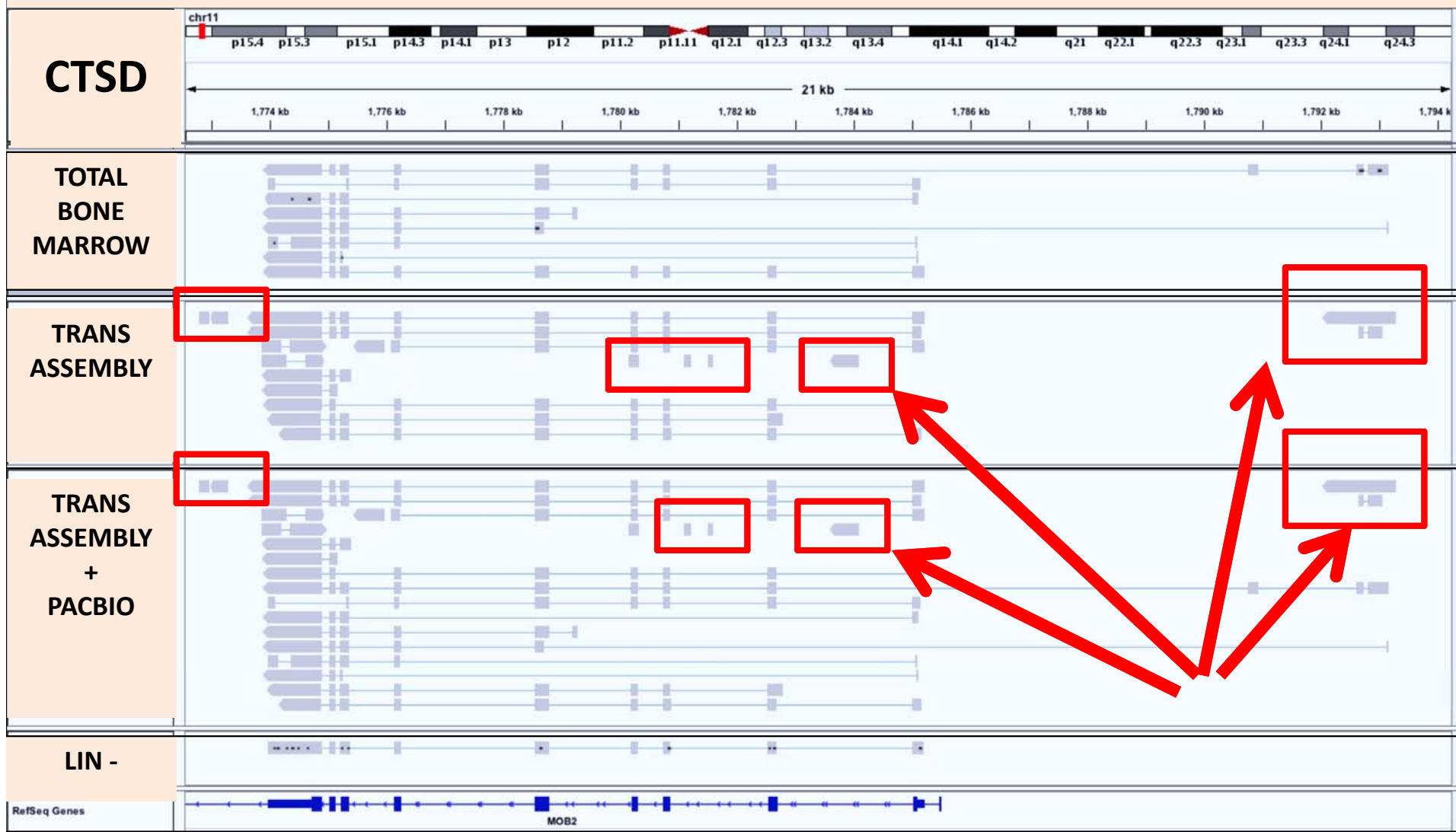


Grech, Godfrey, et al. "Expression of different functional isoforms in haematopoiesis." *International journal of hematology* 99.1 (2014): 4-11.























PacBio contributed additional isoforms at locus



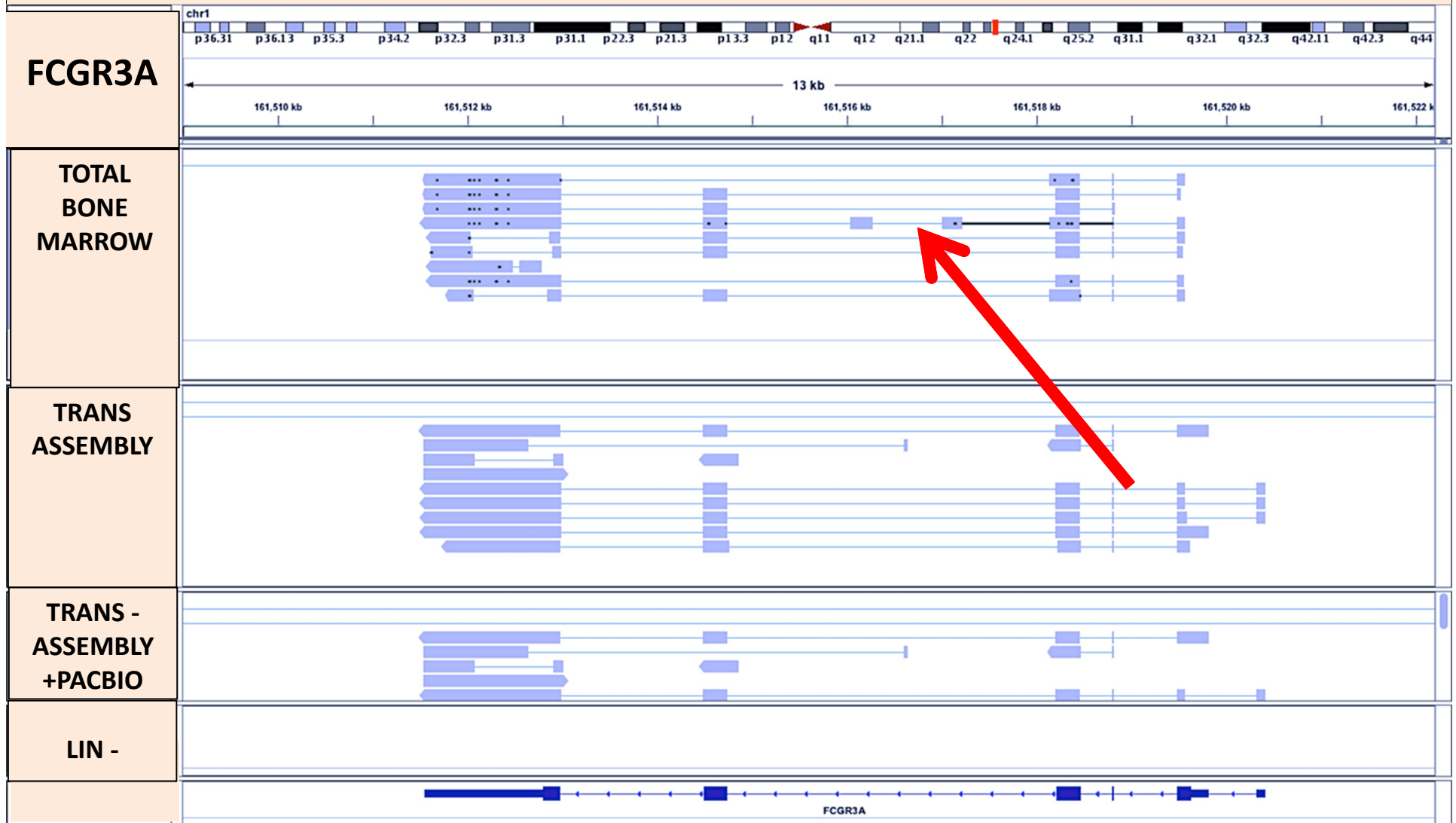
Short Reads have unassembled fragments at the locus



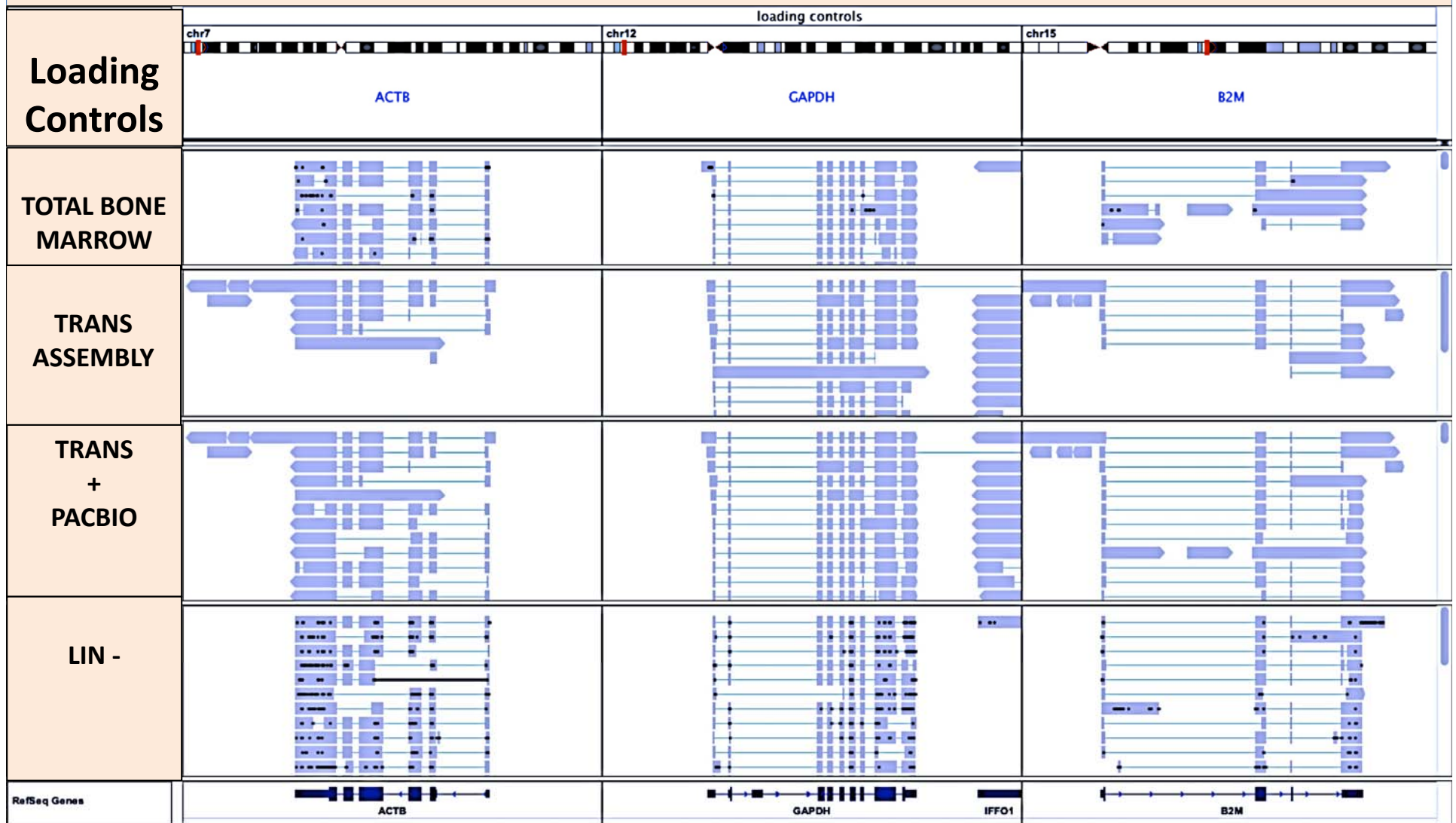
Negative Selection Markers (Differentiated Cells)

	negative controls										
	chr16	chr16	chr5	chr1	chr16	chr1	chrY	chr19	chr11	chr21	chr4
LIN+											
ITGAM											
ITGAX											
CD14											
FCGR3A											
CD19											
CD2											
CD24											
CD3EAP											
NCAM1											
NCAM2											
GYPA											
TOTAL BONE MARROW	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm
TRANS ASSEMBLY	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm
TRANS ASSEMBLY + PACBIO	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm
LIN -	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm	in to see alignm
RefSeq Genes											
ITGAM											
ITGAX											
CD14											
FCGR3A											
CD19											
CD2											
TTY14											
R13L											
ERCC1											
NCAM1											
NCAM2											
GYPA											




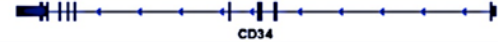
FCGR3A with PacBio added isoforms



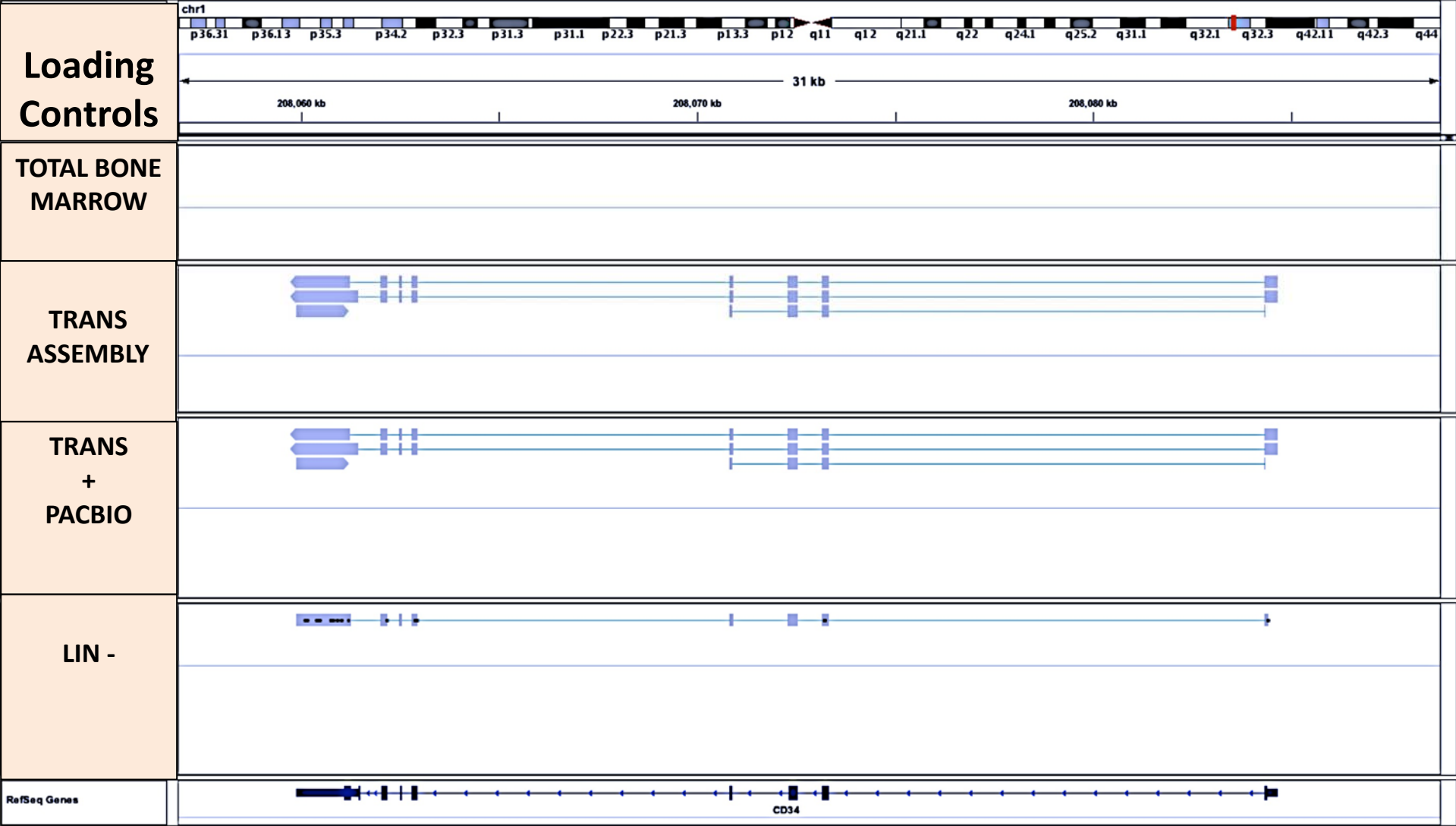
Loading Controls – different isoforms per lineage



Positive Controls

Loading Controls	positive controls		
	chr4 BST1	chr11 THY1	chr1 CD34
TOTAL BONE MARROW	Zoom in to see alignments.		Zoom in to see alignments.
TRANS ASSEMBLY	Zoom in to see alignments.		Zoom in to see alignments.
TRANS + PACBIO	Zoom in to see alignments.		Zoom in to see alignments.
LIN -	Zoom in to see alignments.		Zoom in to see alignments.
RefSeq Genes	 BST1	 USP2-AS1	 CD34

CD34 – no additional isoforms/no presence in TOTAL



So far, just a discussion of isoforms, not abundance

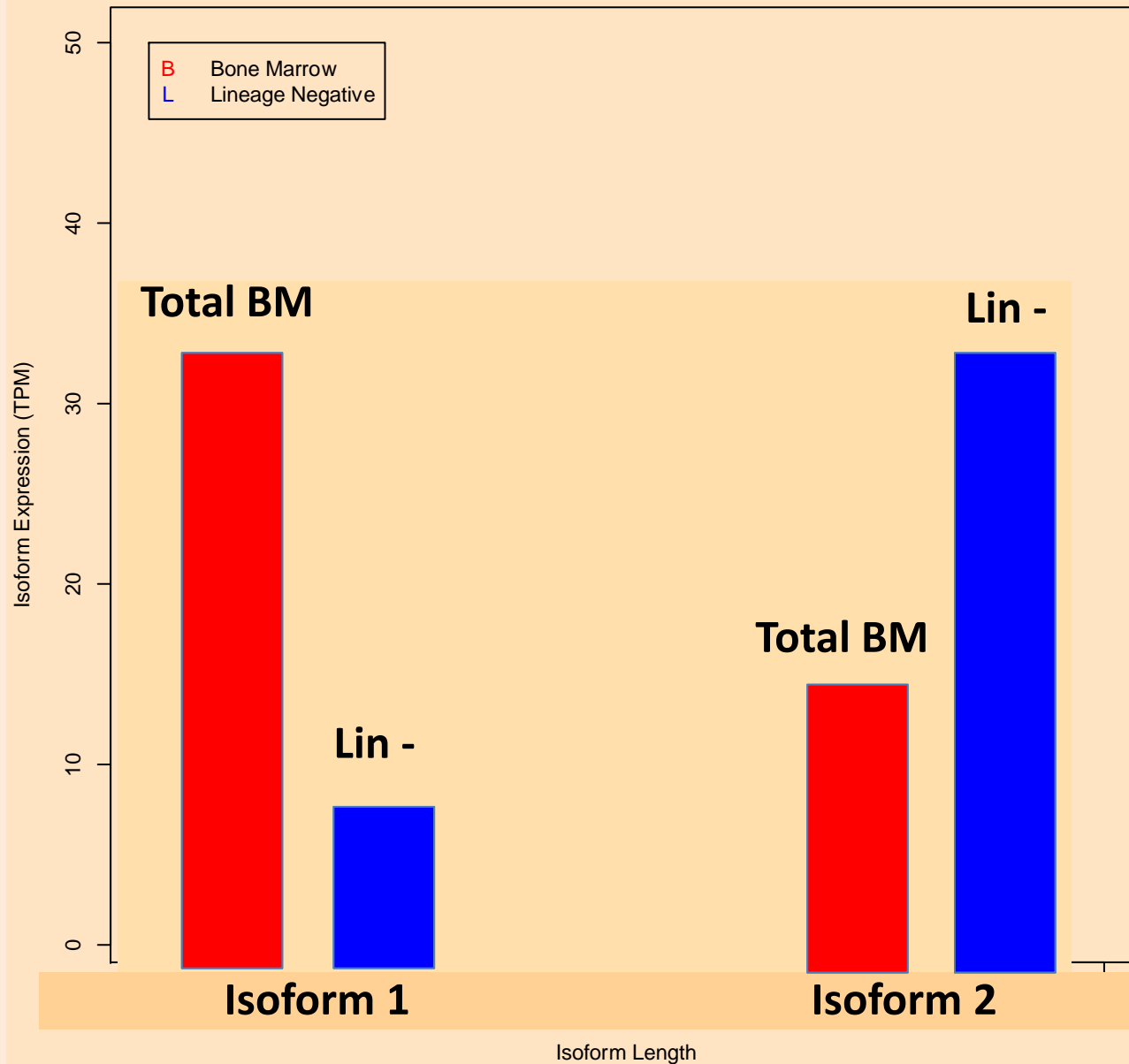
Isoforms exist in a mixture specific to the sub-cell population

The isoform distribution per sub-cell population is critical to the cell fate decision program

This involves both specific isoform and abundance

Illustrative Example – Runx1

RUNX1 in PacBio Only Lineage Negative and Total Bone Marrow Cell Populations



Isoforms exist
in a mixture
specific to the
sub-cell
population

CONCLUSIONS

- Short reads are not useless
 - But they become even more valuable with FL information!
- Got to the biology quicker
- Long read targeted sequencing allows for sequencing for specific regions of interest, without destroying the hard earned cDNA library
- Using the information from the reads, older experiments may be leveraged with the enhanced transcriptome
- Cell populations are maintained by specific isoform distributions regulated in tight balance
- For isoform detail, and when you do not know what you are looking for – PacBio offers a rapid solution for getting to the bottom of the haystack

Next Steps

- Characterize the distribution of isoforms within the specific transcriptomes show connectivity to the regulatory program involved in cell fate determination
- Perform long read sequencing on a targeted enriched cell population to identify full length transcripts of the phage

Acknowledgements

Wellstein Lab, Georgetown U

Anton Wellstein, Ph.D., M.D.

Anna Riegel, Ph.D.

Elena, Marcel, Virginie, Ghada, Ivana, Eveline, Khalid, Khaled, Eric, Nitya

Lombardi Cancer Center, Georgetown U

Yuri Gusev, Ph.D.

Dr. Anatoly Dritschilo

Michael Johnson, Ph.D.

Christopher Loffredo, Ph.D.

Habtom Ressim, Ph.D.

Terry Ryan, Ph.D.

Pacific Biosciences

Elizabeth Tseng, Ph.D.

Primo Baybayna, Ph.D.

Mike Hunkapillar, Ph.D.

Mount Sinai

Robert Sebra, Ph.D.

Eric Schadt, Ph.D.

Software Authors Correspondence

Brian Haas, Author of Trinity Software

Rob Patro, Stephen M. Mount,

and Carl Kingsford, authors of Sailfish