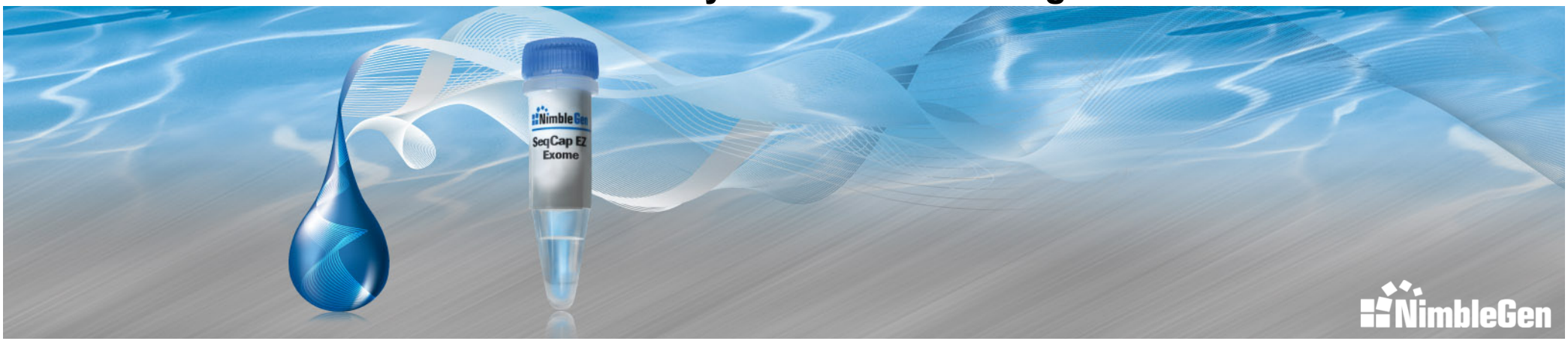
www.nimblegen.com

cDNA CaptureSeq

Reveals unappreciated diversity of the transcriptome

J.A. Jeddeloh, Ph.D.
Director, Business Development, Roche NimbleGen
jeffrey.jeddeloh@roche.com

For Life Science Research Only. Not for Use in Diagnostic Procedures.

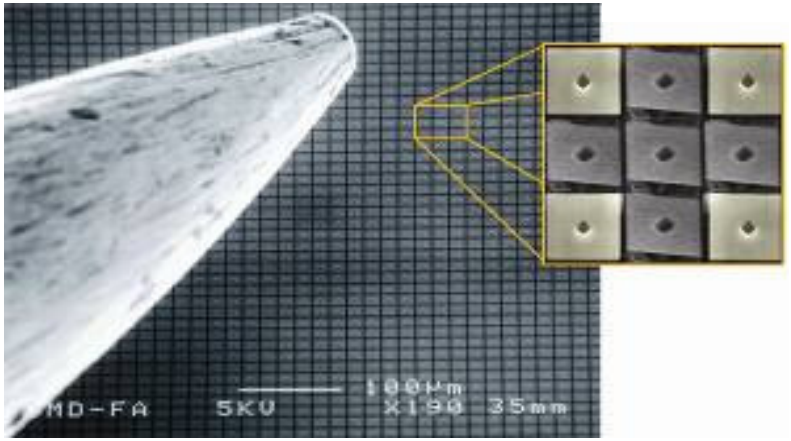


Core Technology: Maskless Array Synthesis (MAS)

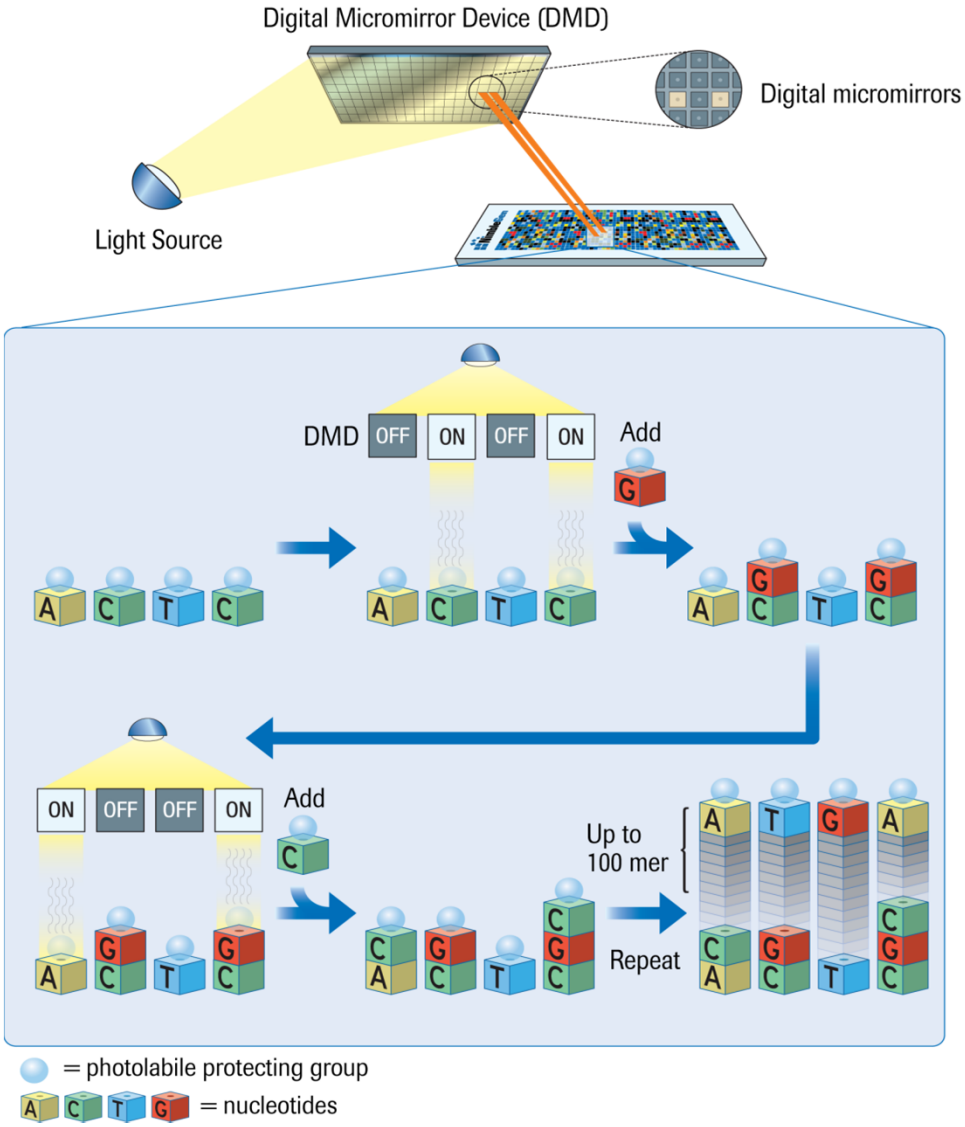
Differs from earlier approaches to light-directed synthesis in two critical respects:

1. Novel way to direct light
- High info content (density)
- Design speed and flexibility

2. Proprietary high-yield chemistry
- Highest synthesis yields
- Fast deprotection/coupling
- High-fidelity long oligo probes



NOW 4.2M FEATURES



Thank you to the following contributors

Roche

NimbleGen

Thomas J. Albert
Todd Richmond
Tracy Millard
Daniel Gerhardt
Mark D'Ascenzo

454 SEQUENCING

Faheem Niazi
Brian Dessany
Brian Godwin
G. Ferreri
Todd Arnold
Jason Affourtit
Thomas Jarvie

cDNA Capture/Splicing

John Rinn
Mitch Guttman
Manuel Garber
Joanna Crawford
Cole Trapnell

BROAD INSTITUTE

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Tim Mercer
Marcel Dinger
John Mattick

NimbleGen SeqCap EZ Choice Workflow

Roche

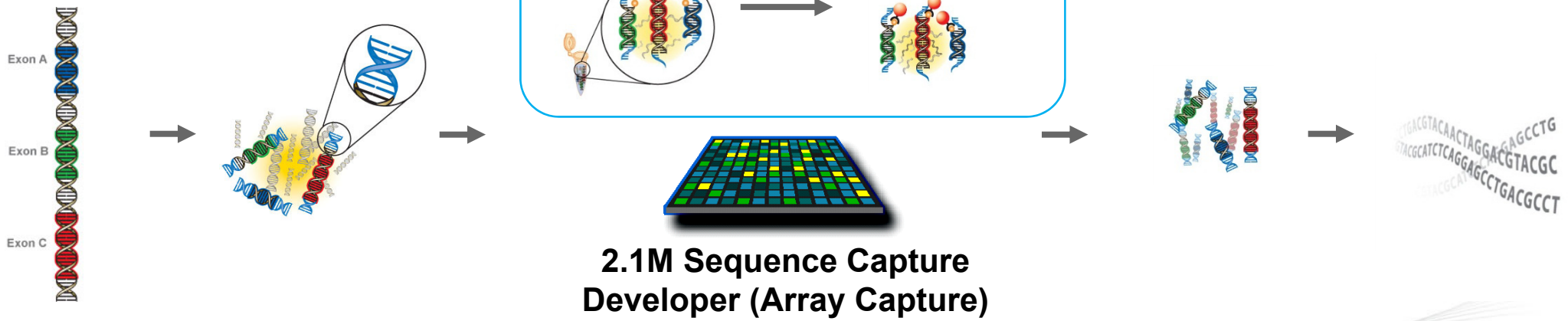
Target Regions

Prepare with Next-Gen Sequencing Adaptors

SeqCap EZ Choice (Solution Capture)

Amplify DNA and Enrichment QC

Sequence DNA on a Next-Gen Sequencer



2.1M Sequence Capture Developer (Array Capture)

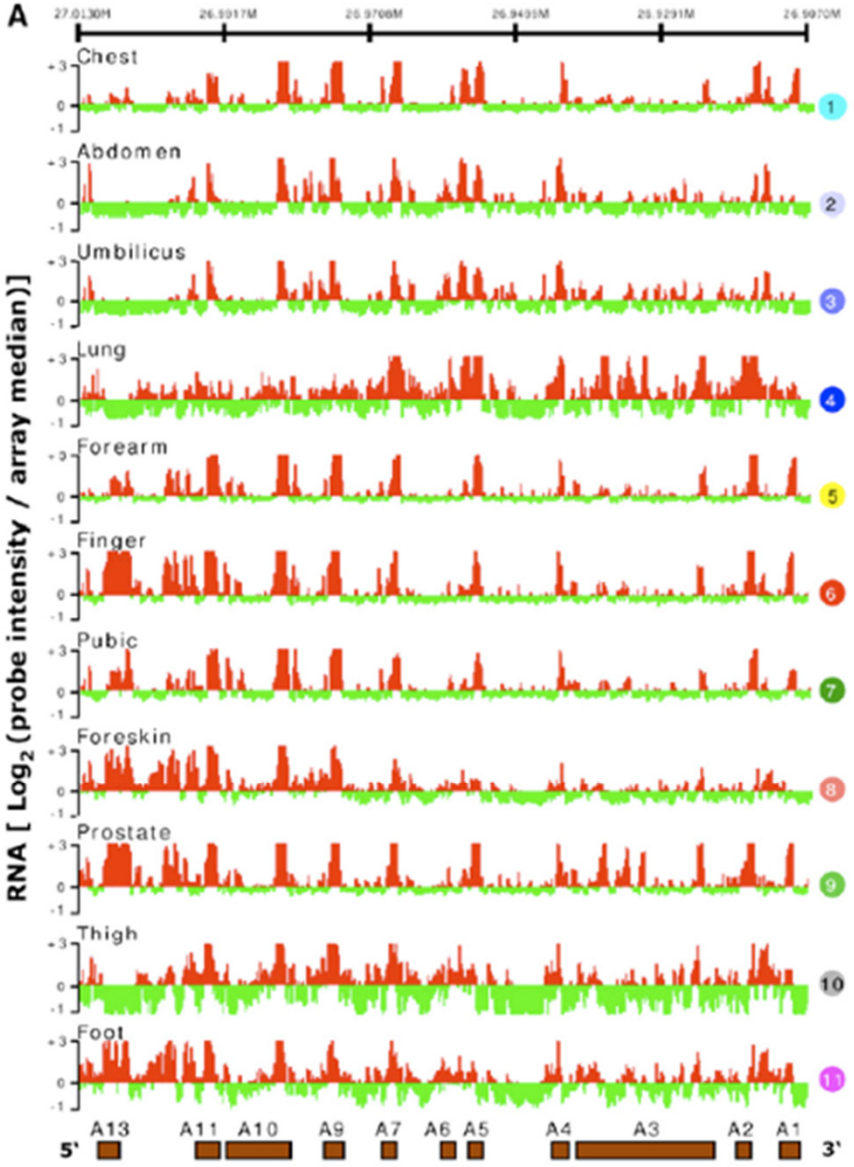
Genomic DNA	Library Preparation	Hybridization	Capture and Washing	Amplification and QC	Sequencing

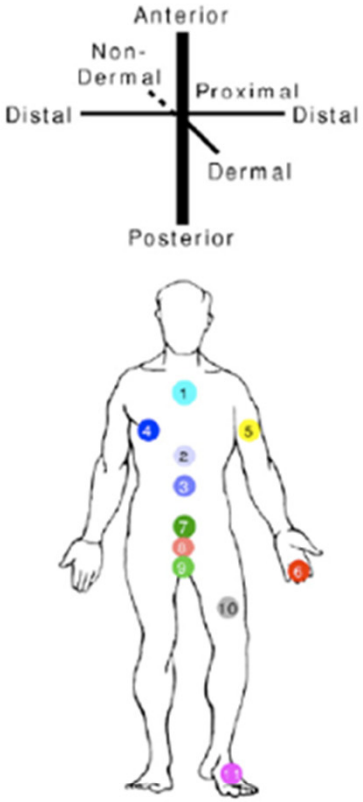
HOXA expression correlates with point of origin

J.L. Rinn et al 2007

In this study active chromatin was demonstrated to be demarcated by Noncoding RNA. Locus regulating HOX was discovered and named HOTAIR. PC2 partner

Discovered to be a class of ncRNA subsequently renamed lincRNA





lincRNA - Characterized by expression and tiling array

If we capture them can we better map their exons

- GOAL
  - Translate expression tiling information about lincRNA into transcript maps
    - RNAseq
      - Problem... about 20% of the reads go to most highly expressed mRNA. Some lincRNA are rare.
      - Splicing (even with PE) is imputed
    - Targeted capture of cDNA
      - Allow deep diving of molecular population
      - May complement RNAseq by measuring abundance of lowest expressed proportion
      - 454 AND Illumina -> Mixed read assembly?
        - » Long reads should help directly characterize diversity of splicing by focusing read depth against a defined target list from the RNA population
        - » Large “N” will help understand Quant
  - Open questions
    - Will it work? (need careful controls to guide interpretation)
    - Will it be quantitative?

Capture cDNA but use gDNA as a control

Establish probe response and evaluate expression normalization

- Array Capture Design (385K array)
  - ~2,000 exons or short regions targeted for capture
    - Putative Linc (Long Intergenic Non-Coding) RNA
    - Protein coding controls
  - 0.8Mb total
- 454 Titanium AND Illumina optimized capture library created and hybridized to capture array
  - Do RNASeq on Foot and Lung – compare to Capture of cDNA
    - 300bp insert libraries for PE sequencing
  - Make sequencing library from cDNA and capture from it
  - Samples
    - Foot fibroblast cell line (c45) cDNA and gDNA control
    - Lung fibroblast cell line (FL) cDNA and gDNA control
      - cDNA made with and without a size selecting gel slice
        - » "Cut" and standard
      - gDNA was a ~700bp average fragment Ti library

Mapping Statistics (Shallow 454 Coverage)

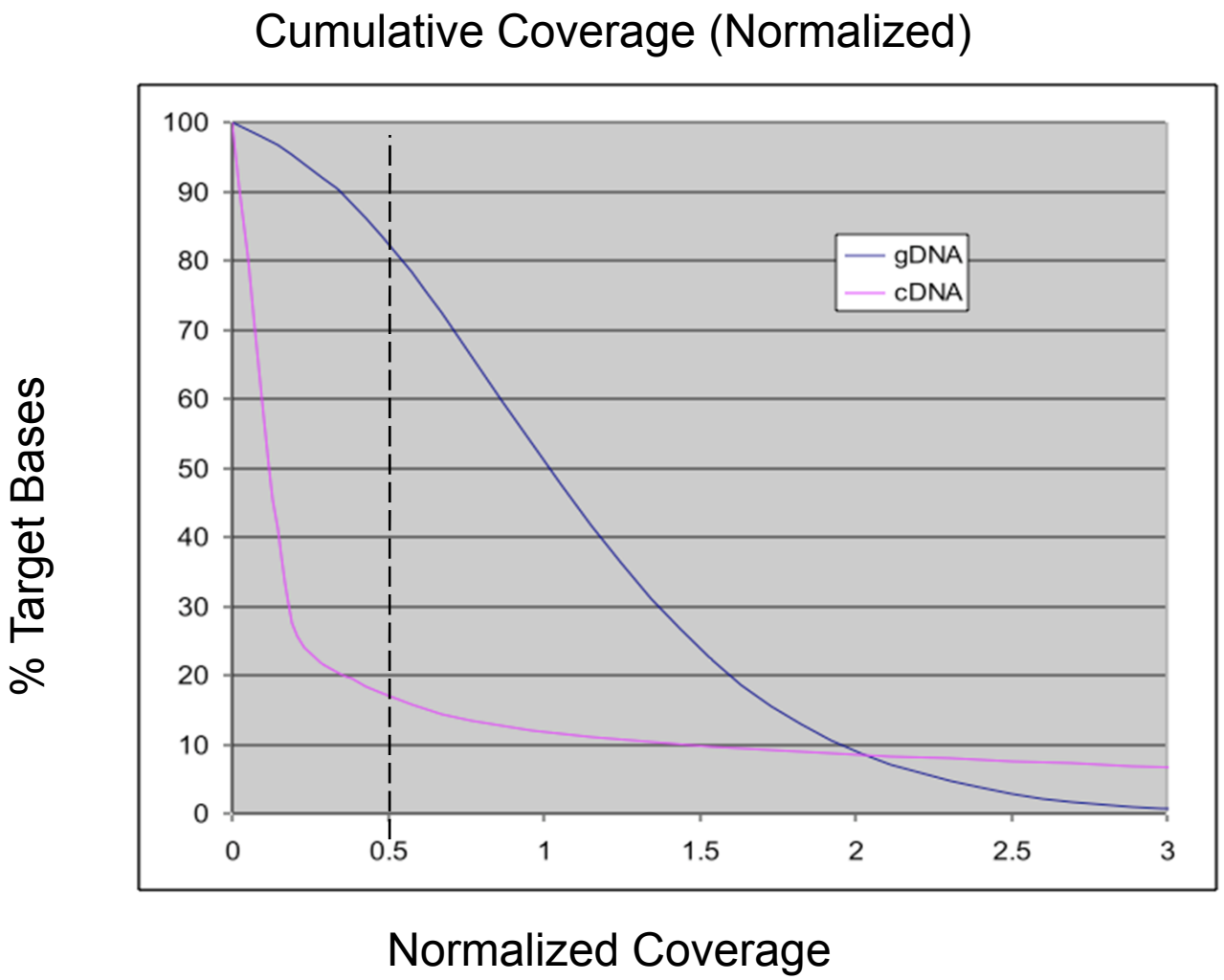
gDNA vs. cDNA

	cut gDNA 45	cut cDNA 45
Initial targets	2092	2092
Initial target bases	830202	830202
Final targets	1846	1846
Final target bases	811370	811370
Target bases covered	773185	223146
Percent target bases covered	95.3	27.5
Target bases not covered	38185	588224
Percent target bases not covered	4.7	72.5
Total reads	22564	28945
Number of reads in target regions	15259	18746
Percent of reads in target regions	67.6	64.8
Average coverage	5.2	5.8
Median coverage	5	0

Shallow coverage:

- ~10Mb per sample
- Capture highly specific for both gDNA and cDNA
- gDNA covered well
- cDNA covered less well
- Expected due to non-uniform representation of transcribed regions

Uniformity Statistics – gDNA vs. cDNA



Uniformity:

- gDNA normal coverage distribution
- >80% of target exonic bases have 50% of the expected coverage
- cDNA coverage very non-uniform, as expected

Mapping Statistics (Deeper 454 Coverage)

gDNA vs. cDNA

	Lung gDNA	Foot gDNA	Lung cDNA	Foot cDNA
TOTAL_READS	185437	186637	262103	268573
TOTAL_BASEPAIRS	56102185	55133762	73425525	81247544
TOTAL_MATCHES	2826924	3786172	946025	1140888
PERCENT_READS_NO_MATCH	9.80%	10.60%	2.80%	3.10%
NUM_BASEPAIRS_NO_MATCH	2788580	2993466	820234	986838
PERCENT_BASEPAIRS_NO_MATCH	5.00%	5.40%	1.10%	1.20%
NUM_READS_UNIQUELY_MAPPED	165993	165063	250832	256681
PERCENT_READS_UNIQUELY_MAPPED	89.50%	88.40%	95.70%	95.60%
NUM_BASEPAIRS_UNIQUELY_MAPPED	51915060	50543591	57643720	61507568
PERCENT_BASEPAIRS_UNIQUELY_MAPPED	92.50%	91.70%	78.50%	75.70%
NUM_BASEPAIRS_HSP_TRIMMED	1079817	1134159	14084613	17956647
PERCENT_BASEPAIRS_HSP_TRIMMED	1.90%	2.10%	19.20%	22.10%
Initial targets	2092	2092	2092	2092
Initial target bases	830202	830202	830202	830202
Final targets	1846	1846	1846	1846
Final target bases	811370	811370	811370	811370
Target bases covered	807557	806437	404035	334221
Percent target bases covered	99.5	99.4	49.8	41.2
Target bases not covered	3813	4933	407335	477149
Percent target bases not covered	0.5	0.6	50.2	58.8
Total reads	165993	165063	250832	256681
Number of reads in target regions	120481	109857	158006	154725
Percent of reads in target regions	72.6	66.6	63	60.3
Average coverage	36	32.5	40.4	42.3
Median coverage	35	32	0	0

cDNA maps better than gDNA (no repeats, no rRNA)

Exon junctions cause trimming

All targets from gDNA present
Less than 1/2 from cDNA

Highly specific capture

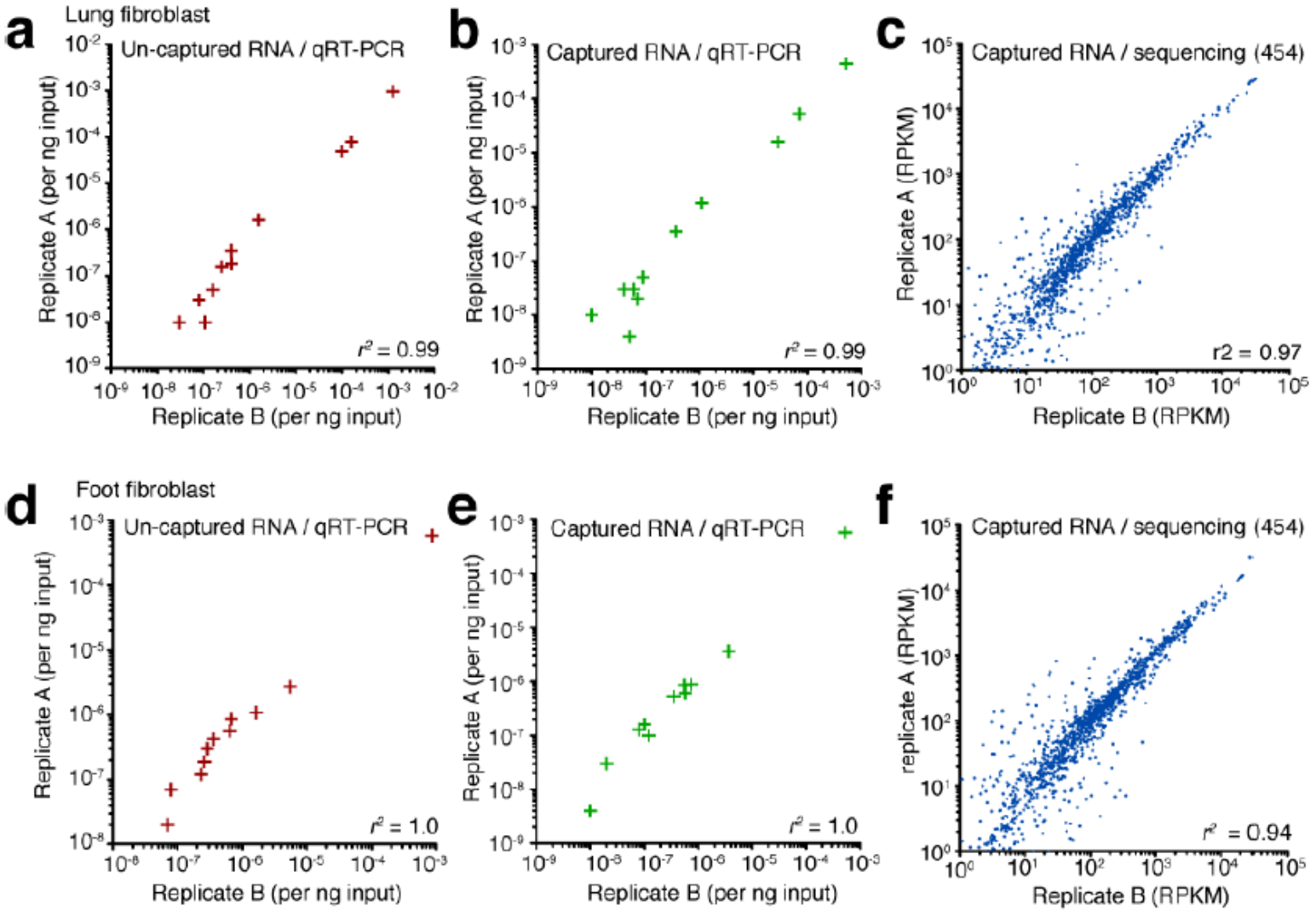
~73% reads from 0.024% of genome = >3000 Fold Enrichment

Correlation analysis confirms quantitative nature

qRT-PCR vs RNAseq (il) vs CaptureSeq vs Replicates (454)

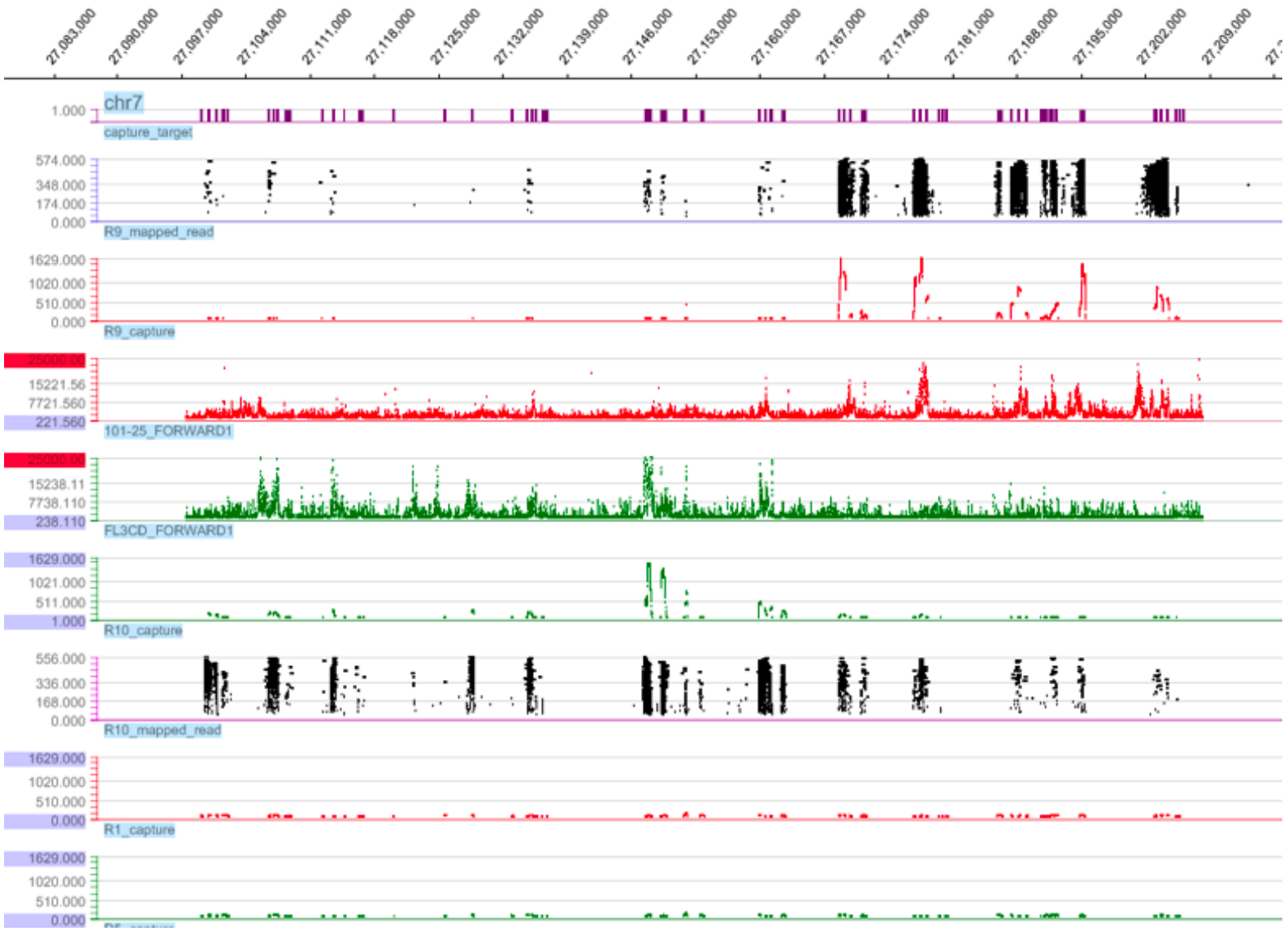
LUNG

FOOT



Capture cDNA provides unparalleled resolution

Confirmed previous expression tiling data



Capture of cDNA provides unparalleled resolution

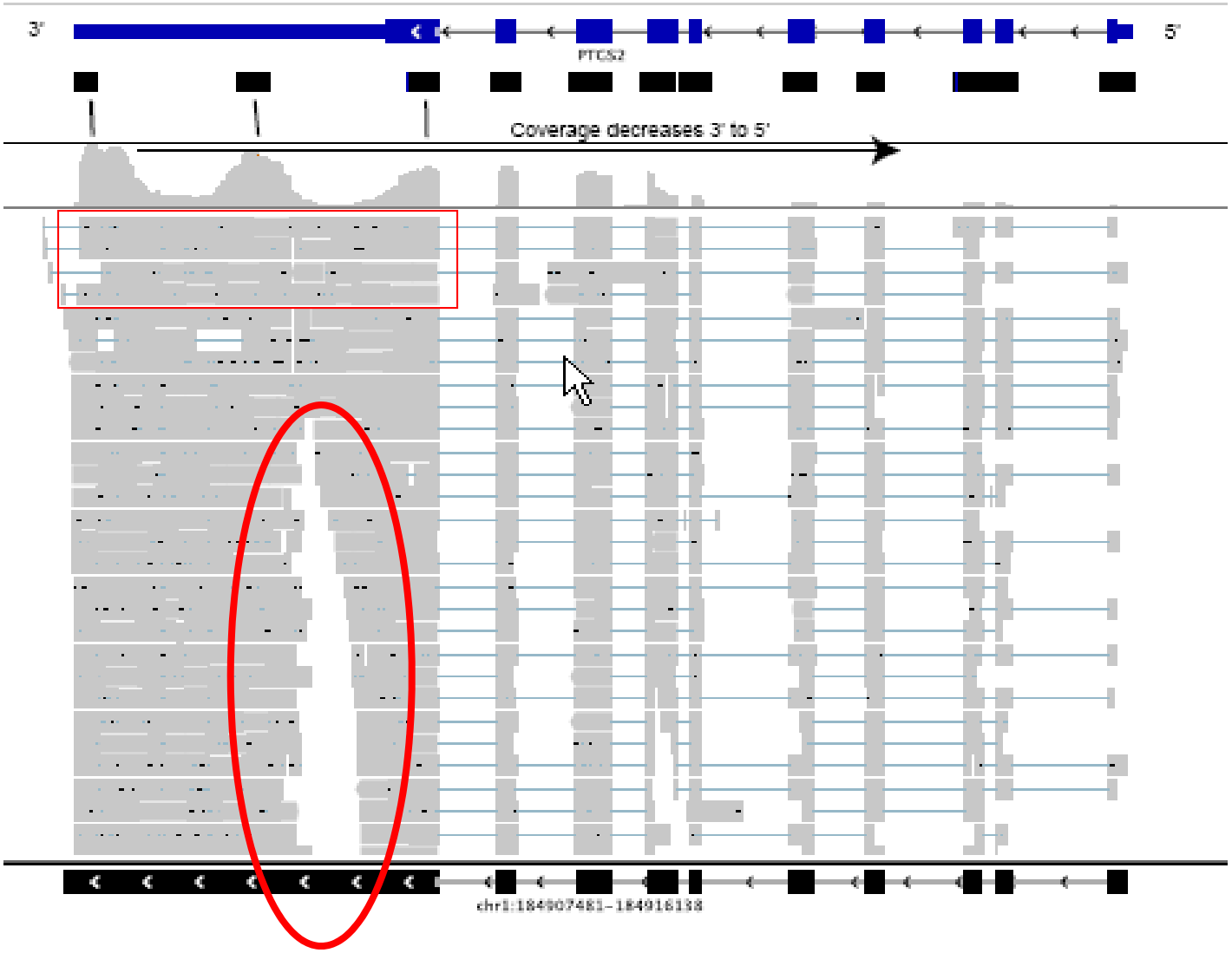
Splicing diversity amazing with long reads

HOTAIR

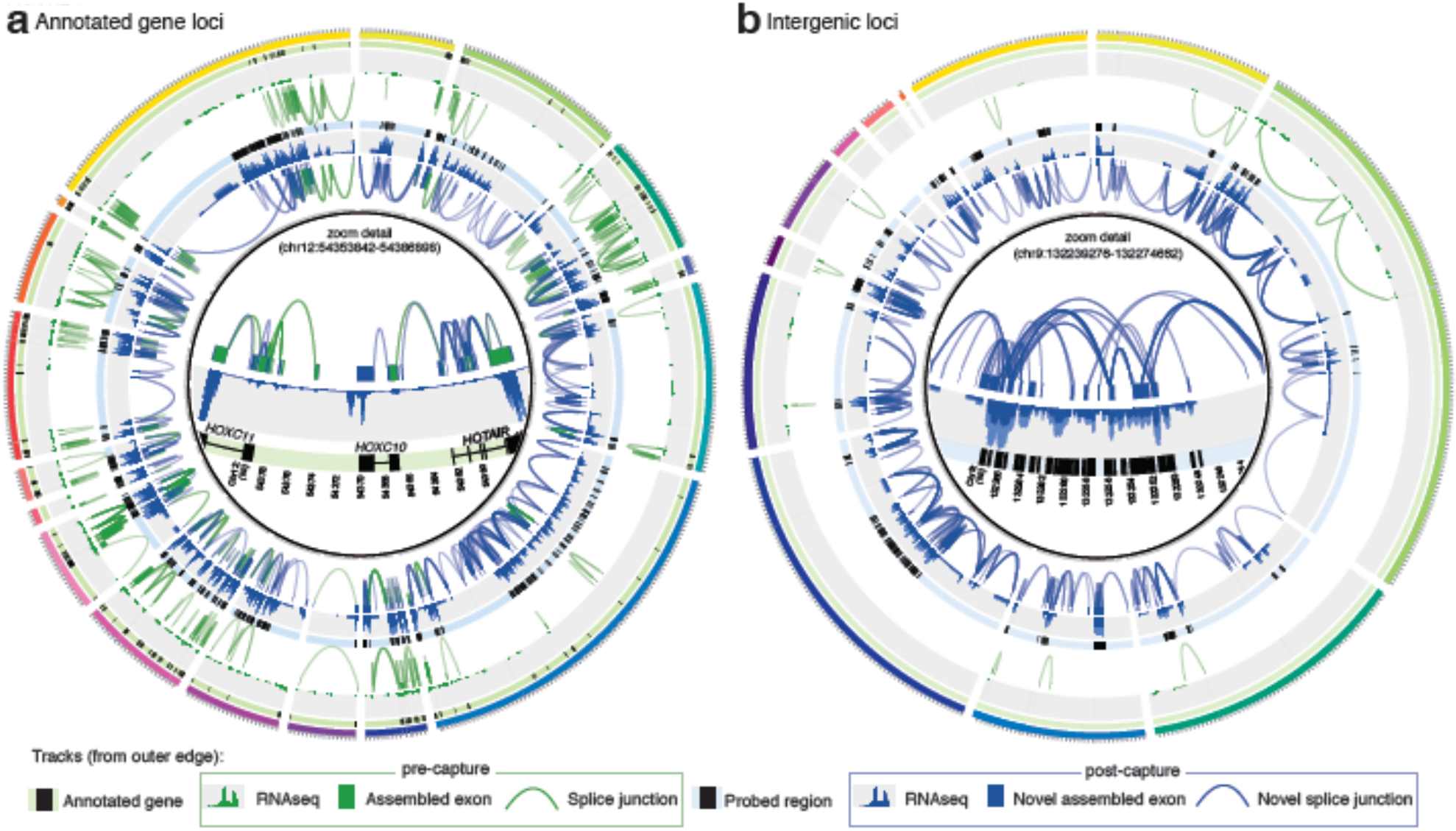
Capture of cDNA provides unparalleled resolution

Splicing diversity amazing with long reads

- The full length UTR in the reads suggests that despite where the probes are spaced the coverage along the UTR is appropriate.
- The missing bases in the center of the UTR have nothing to do with the probes
- ….And everything to do with processing of the UTR. ~15 different isoforms of this message's UTR

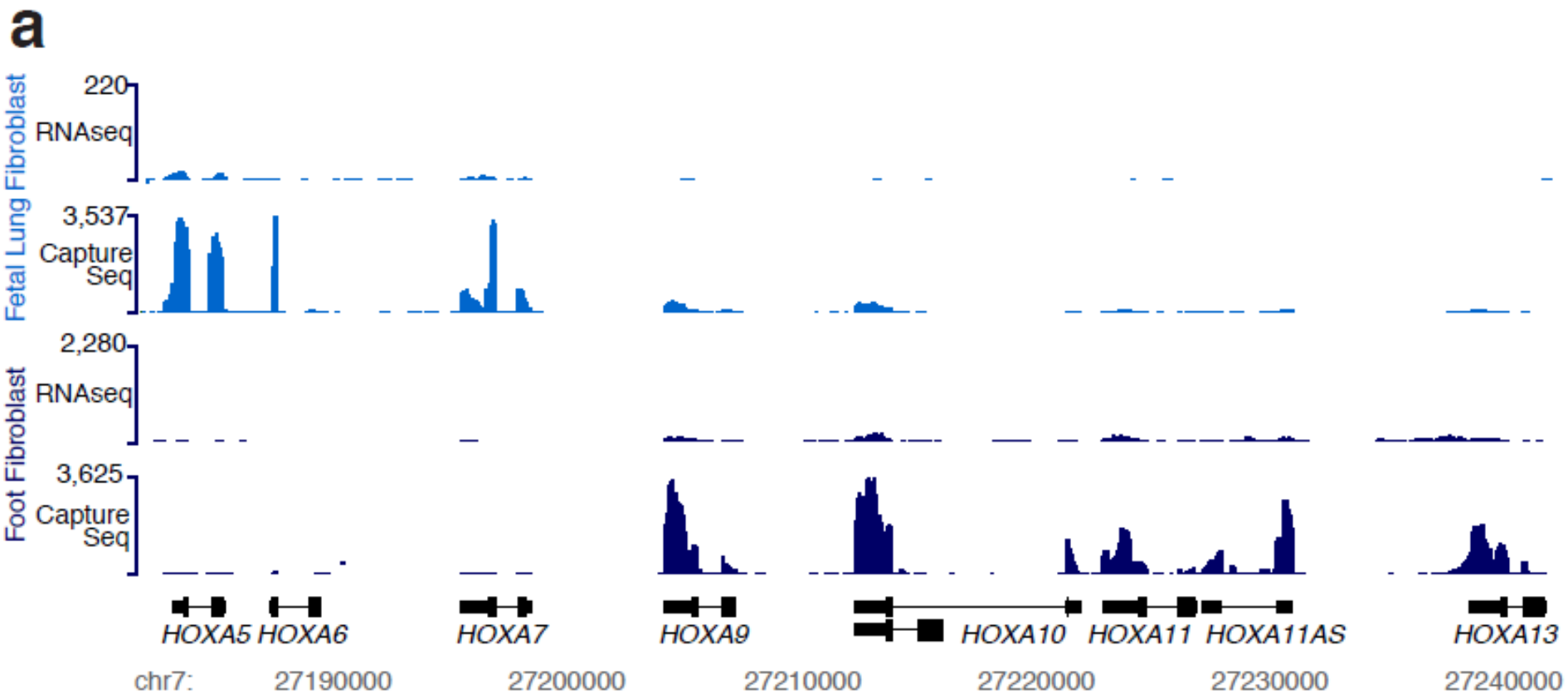


Mixed read length assembly splicing analysis



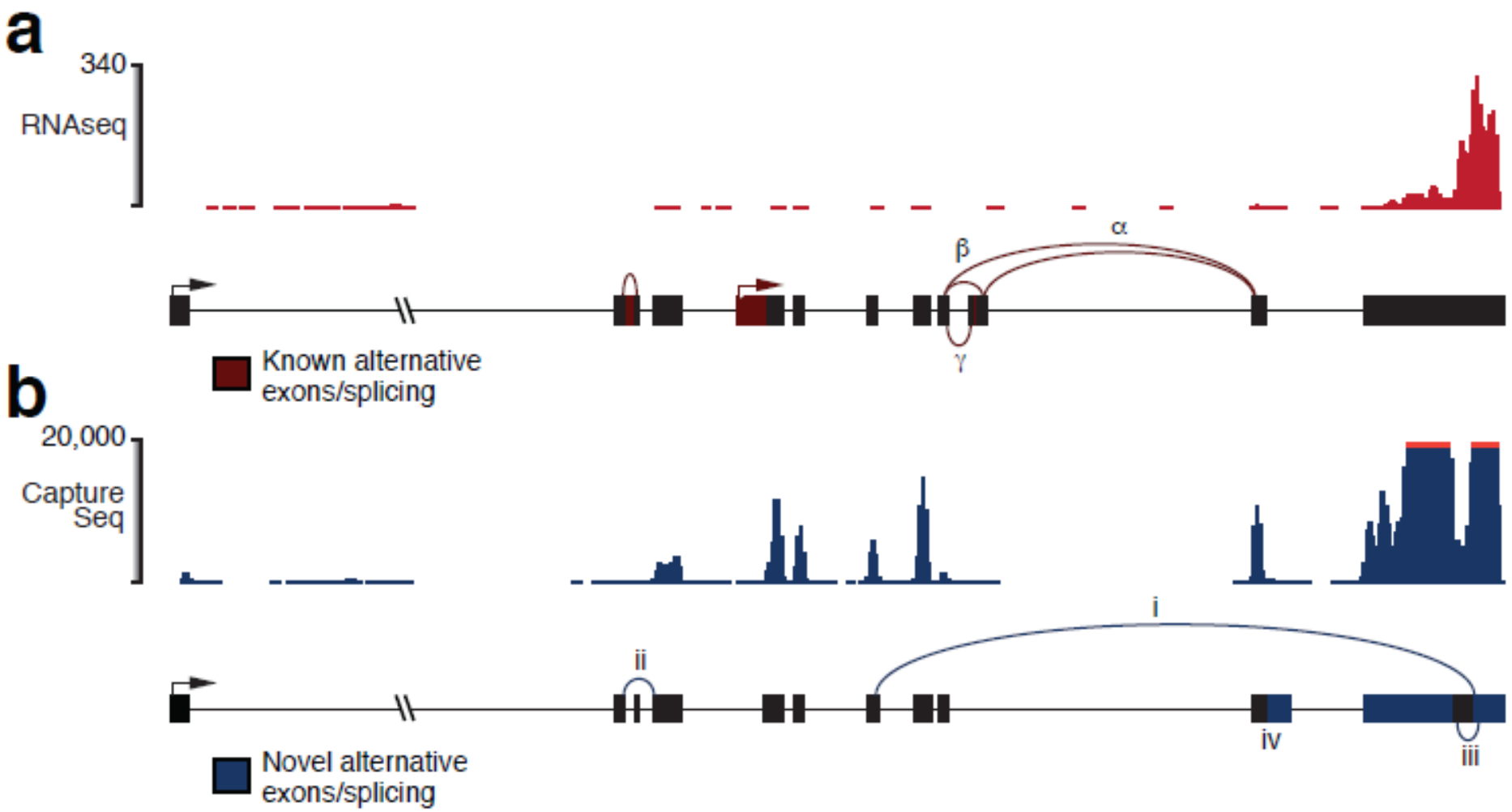
HOX-A differential expression compare

Illumina analysis

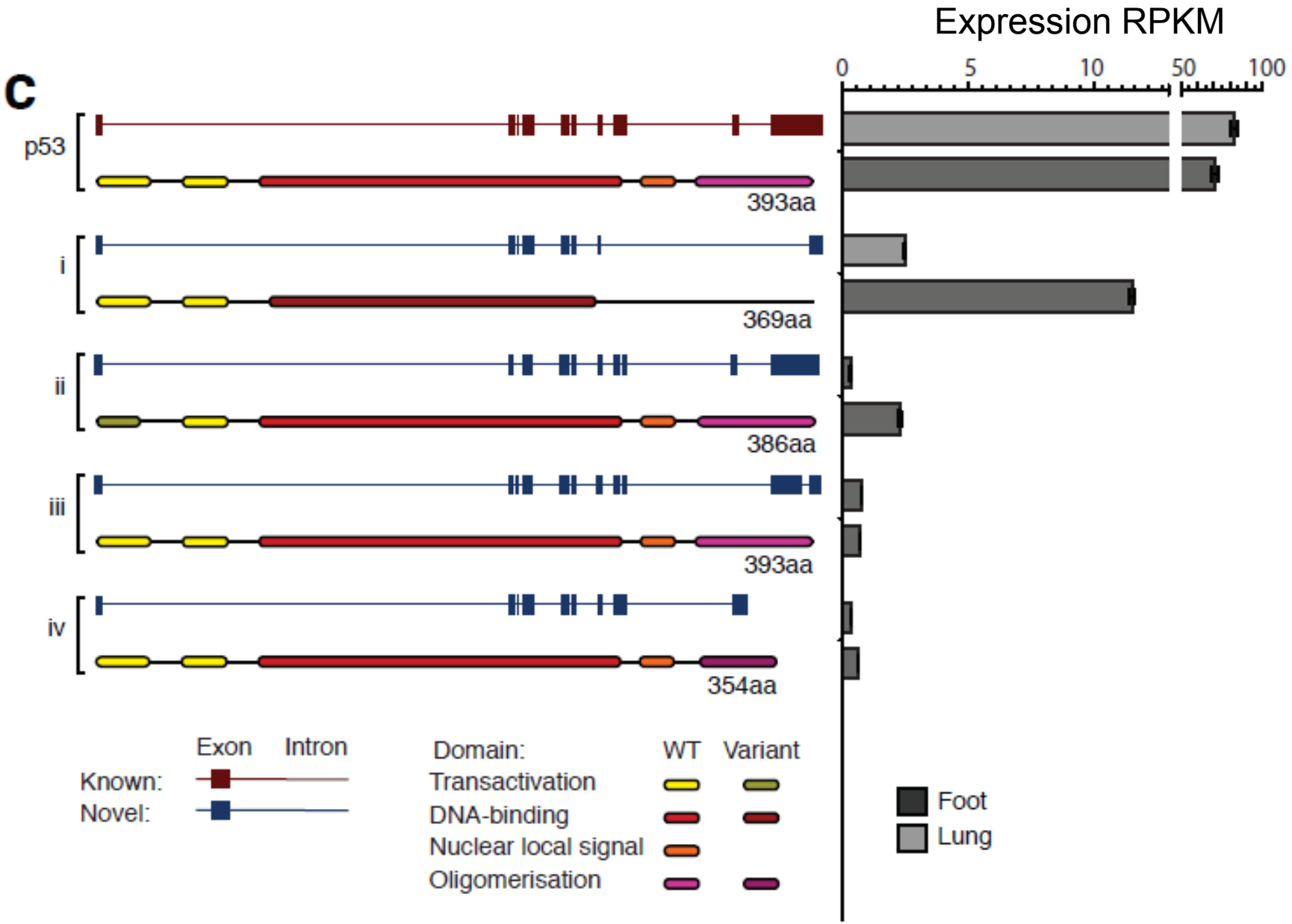


P53– Coverage compare Capture vs. RNAseq

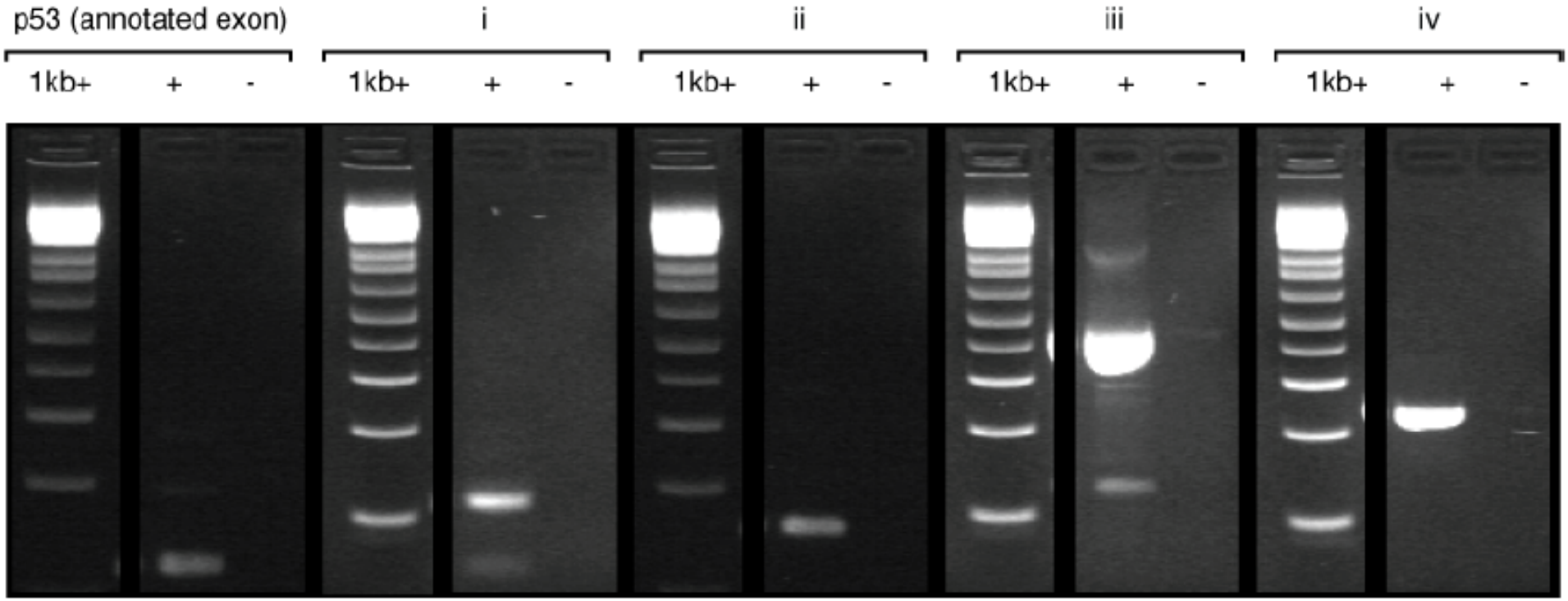
Illumina



P53 – novel alternative splicing and expression



P53- RT-PCR validation of 4/4 NOVEL EXONS



Conclusions

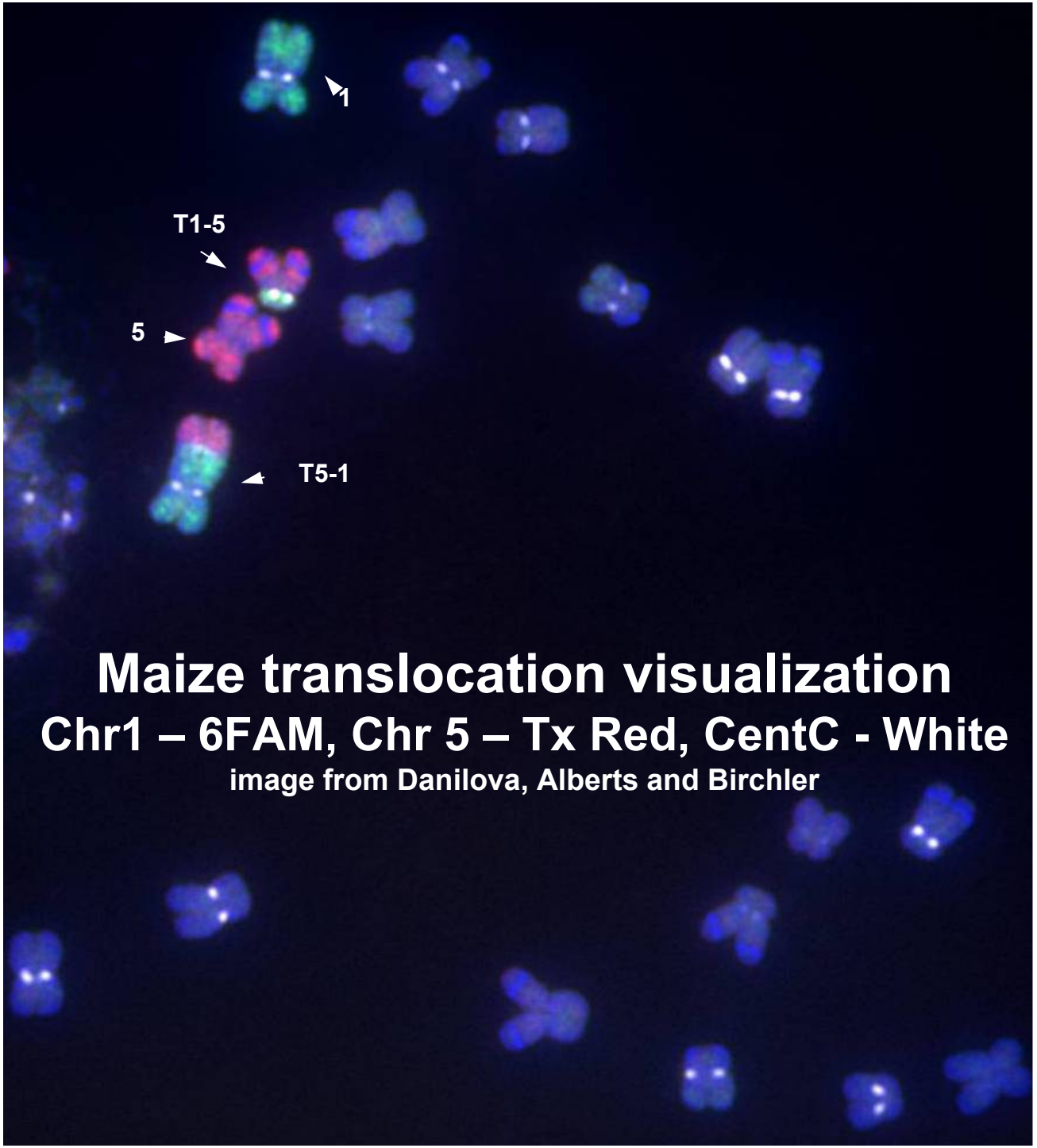
cDNA CaptureSeq

- Supplements information available from RNAseq by affording a targeted zoom on data analysis
- DOES NOT NORMALIZE expression but quantitatively samples
- Preserves known differential expression patterns
  - Reproducibly recapitulated known truth- Foot vs. Lung
- Mixed read analysis afforded expanded validation
  - Long-reads provided support despite only ~300K attempts

Conclusions

cDNA CaptureSeq

- Discovered 745 novel intergenic transcripts
  - Exp - subset of cells, perhaps less than 1 in 1000 cells
- Established the baseline of the transcriptome is far below the ability of a single HiSeq lane to resolve
  - >10 B attempts of RNAseq would be required to garner same coverage level as that achieved by CaptureSeq
- Enhanced zoom afforded discovery of four RARE transcripts of P53



Maize translocation visualization

Chr1 – 6FAM, Chr 5 – Tx Red, CentC - White

image from Danilova, Alberts and Birchler

Unless explicitly stated otherwise, all Roche NimbleGen and 454 products and services referenced in this presentation are intended for the following use:

For Life Science Research Only.

Not for Use in Diagnostic Procedures.

We Innovate Healthcare