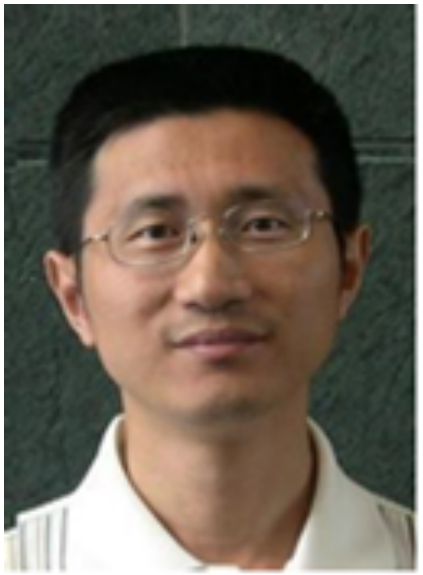
Optimizing your use of RNA-Seq tech & data analysis – 101

Shanrong Zhao, Pfizer

Alexander Dobin, CSHL

Baohong Zhang, Pfizer

Shanrong Zhao



shanrong.zhao@pfizer.com

Director of Computational Biology, Pfizer Inc

A recognized pioneer in next generation sequencing, big data analysis and cloud computing.

- 20 years of experience in computational biology and bioinformatics
- 14 peer-reviewed publications on NGS in the past 3 years
- 12 invited talks at NGS, cloud computing and precision medicine meetings
- Scientific reviewers for multiple journals

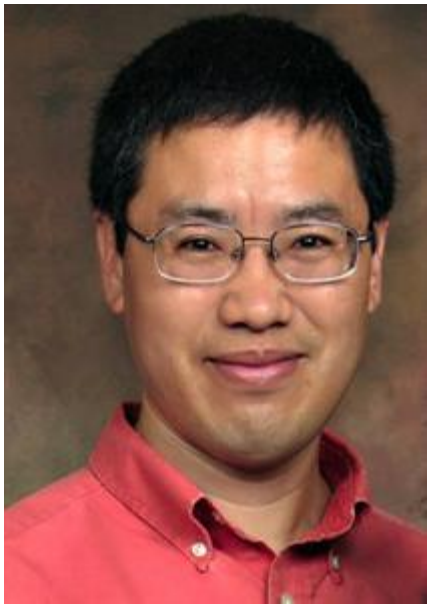
Alexander Dobin

dobin@cshl.edu

Computational Science Manager,
Cold Spring Harbor Laboratory

- Pioneer in RNA-seq data analysis
- 62 peer-reviewed publications on NGS, genomics
- Author of STAR, the most popular read mapper used in RNA-seq
- Invited speakers on RNA-seq, NGS conferences

Baohong Zhang



baohong.zhang@pfizer.com

Director of Clinical Bioinformatics, Pfizer Inc.

An industrial leader in next generation sequencing, bioinformatics and big data analytics and visualization.

- 20+ years of experience in bioinformatics, genomics and genetics
- 43 peer-reviewed publications and patents on protein structure modeling, SNP genotyping, exome-seq, RNA-seq, microRNA-seq, omics
- Invited speaker on next generation sequencing, big data visualization (https://baohongz.github.io/DataVizAlive/#0)

Outline of workshop

- **RNA-seq data analysis**
  1. **Part #A – Mapping**
  2. **Part #B – Quantification and normalization**
  3. **Part #C – Visualization**
- **Open questions**

Part A
Optimizing alignment of the RNA-seq data

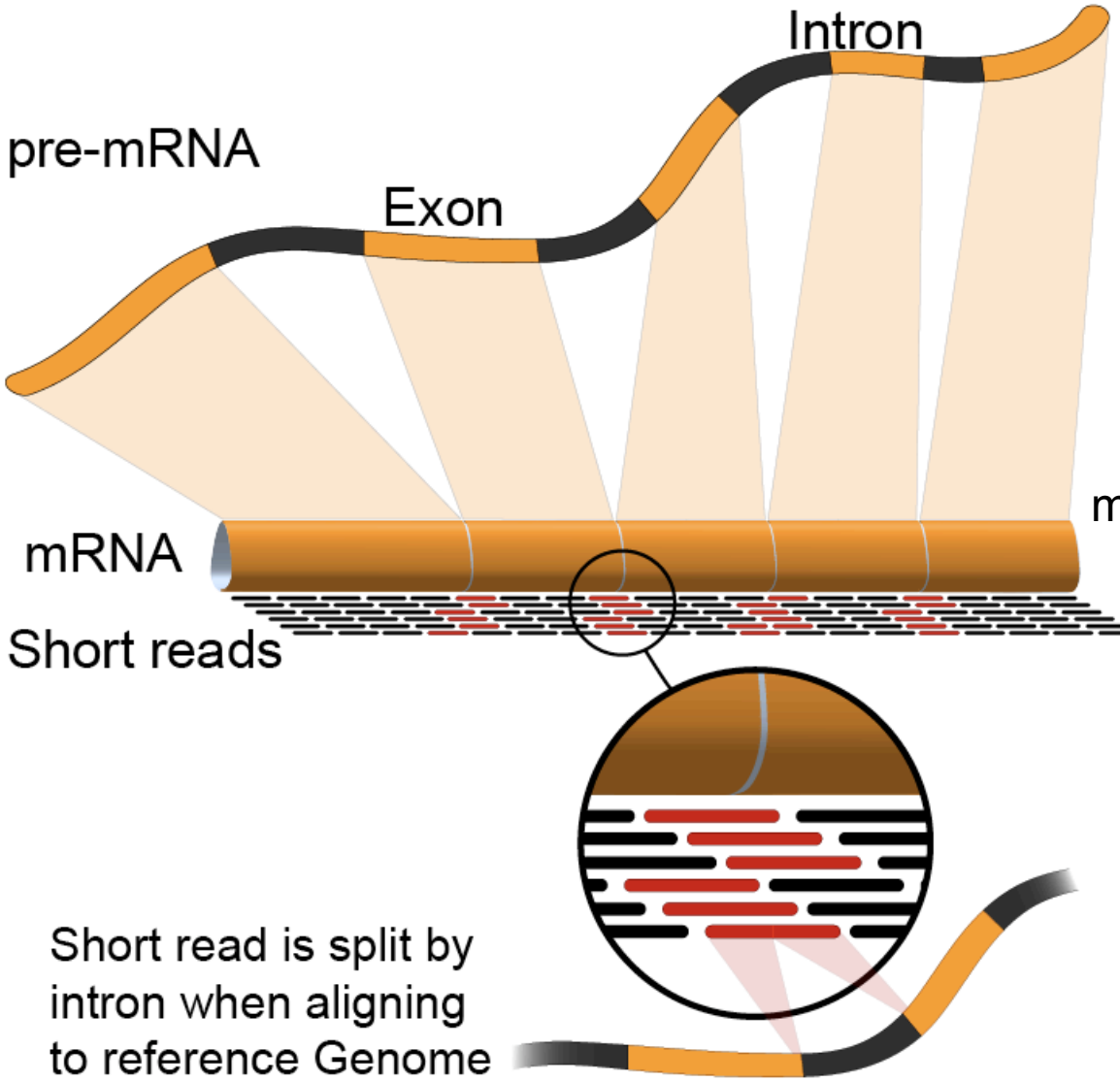
Alexander Dobin
CSHL

Outline

- Introduction: RNA-seq technology, analyses, pipelines
- Optimizing mapping of RNA-seq reads to the genome
- STARtools: post-mapping analyses at no extra cost

Introduction: RNA-seq technology, analyses, pipelines

RNA-seq

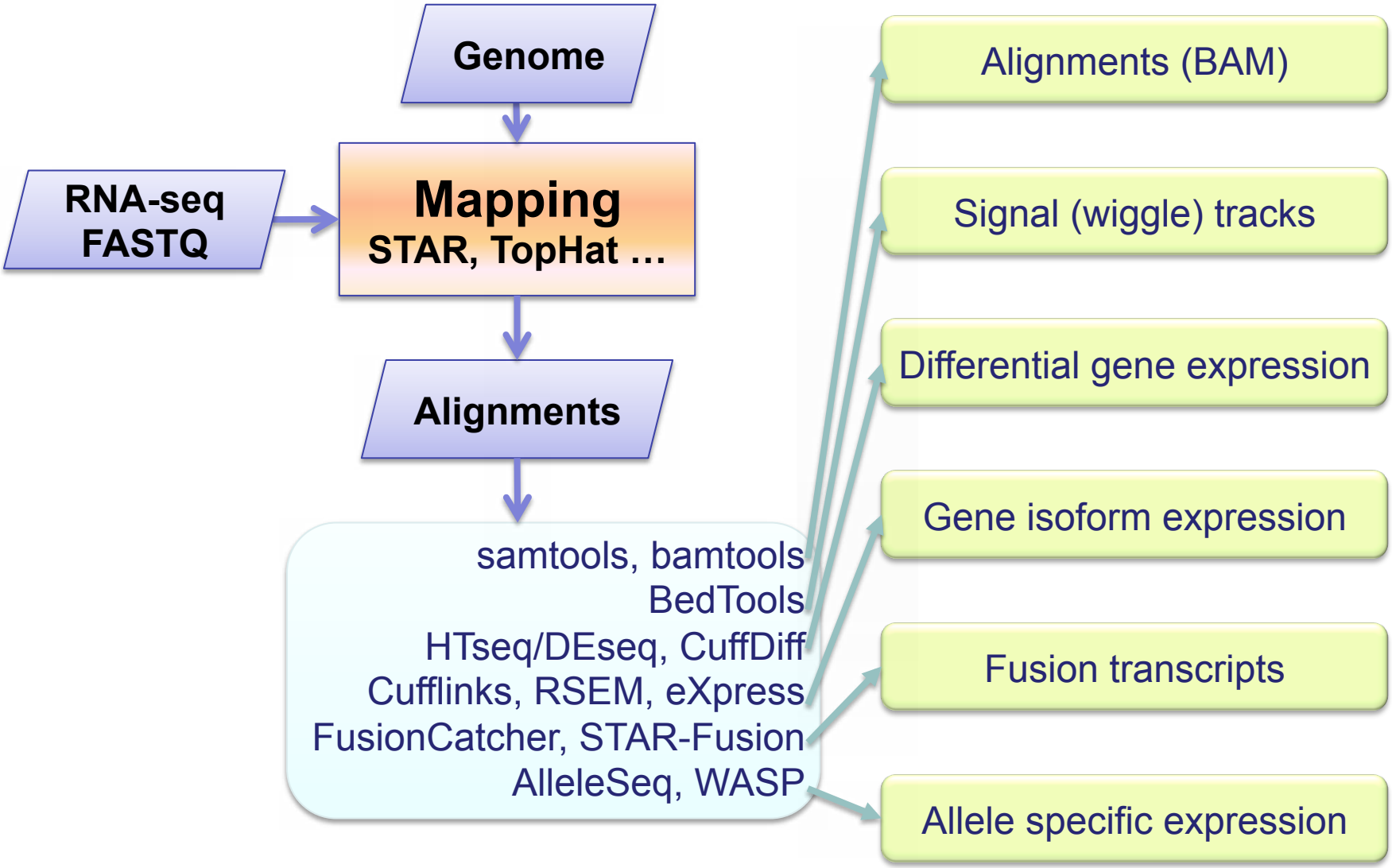


mRNA:0.3-10kb, median=1.4kb

short reads: 50-300b

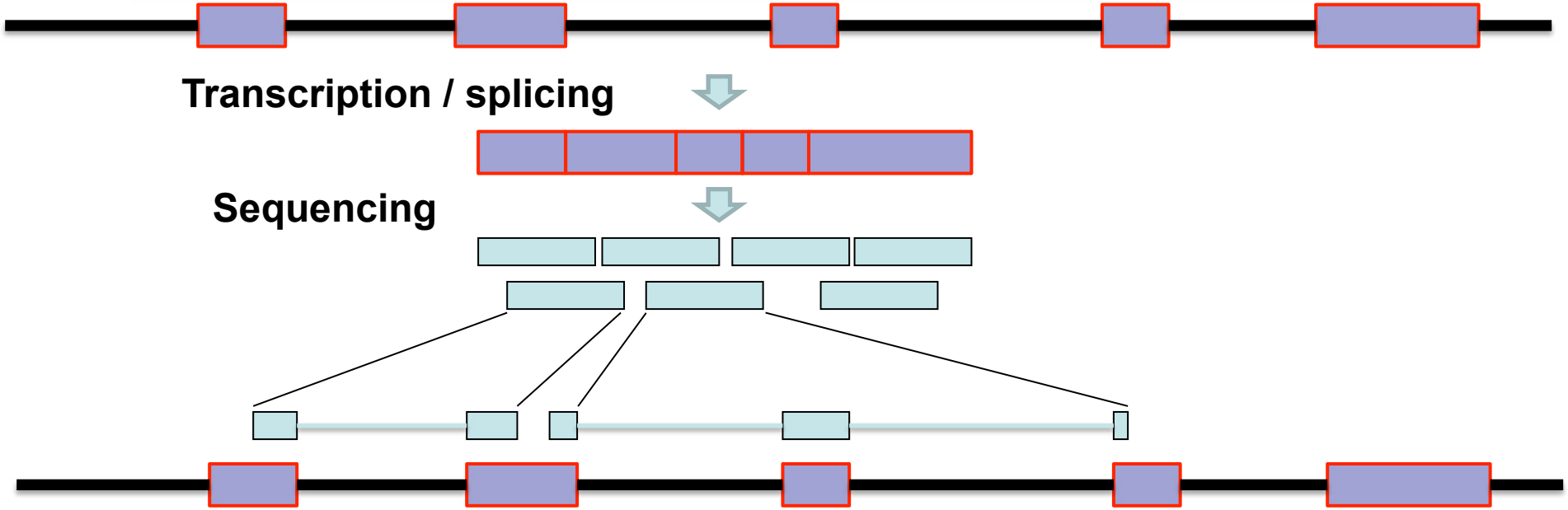
https://en.wikipedia.org/wiki/RNA-Seq#/media/File:RNA-Seq-alignment.png

RNA-seq pipeline



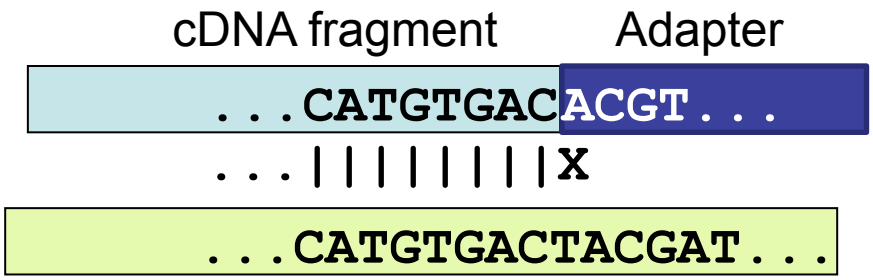
Optimizing mapping of RNA-seq reads to the genome

Challenges of RNA-seq mapping

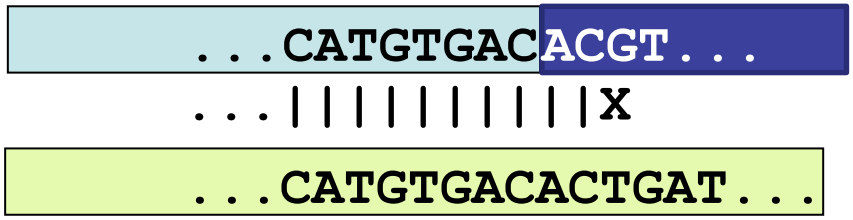


- Most long RNAs are spliced
- Short reads map non-contiguously, may contain >1 splice junction
- Large introns: ~0.1-1,000 kb in mammals
- Multi-mappers are important (expression of repeats, paralogs, pseudogenes)
- Highly expressed loci create mapping artifacts
- Genomic variations: SNPs, indels, SVs
- RNA editing
- Sequencing errors, short inserts,

Adapter trimming



Locus 1:
only the cDNA sequence
maps to the genome



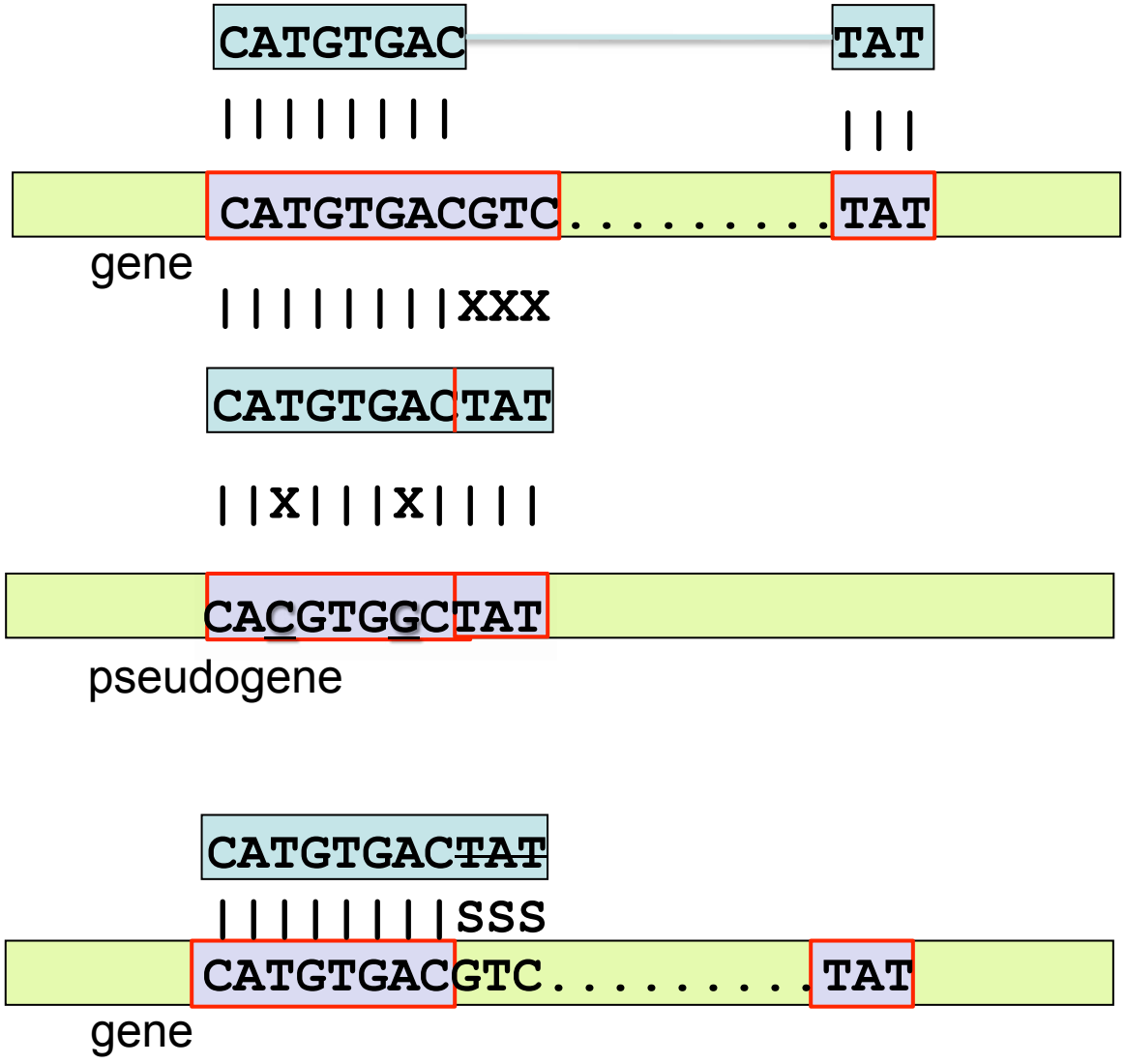
Locus 2:
cDNA sequence + 2 adapter bases
map to the genome

- Adapter at 3' of the read sequence if
  fragment ("insert") length < sequence length
- Untrimmed adapter can turn multi-mappers into unique mappers

- Trimming software: Cutadapt, Trimmomatic, FASTX, etc.
  - take care not to mess up the read order for paired-end reads
- Basic aggressive 3' adapter trimming in STAR with

```
--clip3pAdapterSeq <sequence>  --clip3pAdapterMMp 0.1
```

Soft-clipping



True alignment: spliced

Hard to find because of short overhang: 3b will randomly occur every 64b in the genome

End-to-end alignment: 3MM

May be filtered out

End-to-End alignment to the pseudogene: 2MM

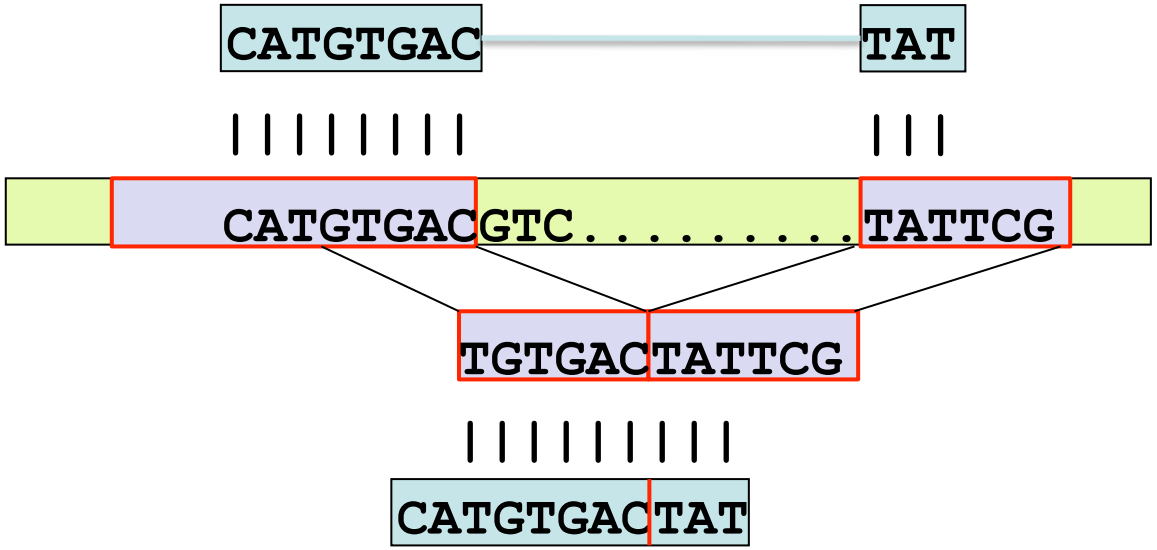
False alignment!

Soft-clip unmatched bases

Soft-clipping penalty < MM penalty

Also helps to catch polyA-tails, end modifications, adapters, poor quality tails

Mapping short splice overhangs



True alignment: spliced

Hard to find because of short overhang: 3b will randomly occur every 64b in the genome

Solution:

use annotated junctions

STAR concatenates exon sequences and adds them to the search space

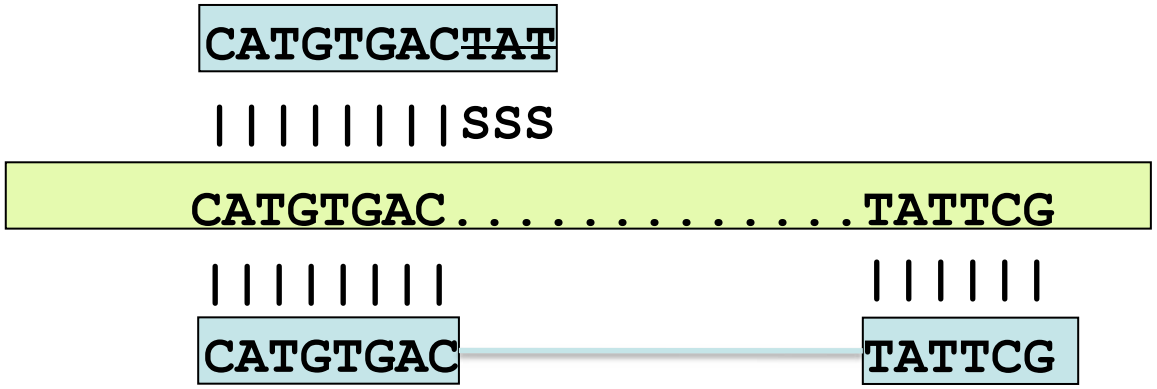
%	BLAT	Tophat	STAR	STAR + Annot
base FPR	2.9	5.4	2.0	0.1
base FPR: wrong loci	2.7	5.3	1.9	0.1
base FPR: wrong locus, missed splice	2.2	4.5	1.8	0.1
<u>% of all reads mapping to processed pseudogenes</u>				
all	0.7	1.6	0.8	0.1
False Positive, wrong locus, missed splice	81.8	82.3	71.1	26.0

~80% of false positive alignments arise from alignments missing a splice junctions and mapping to a wrong locus

~30% of false positive alignments map to processed pseudogenes

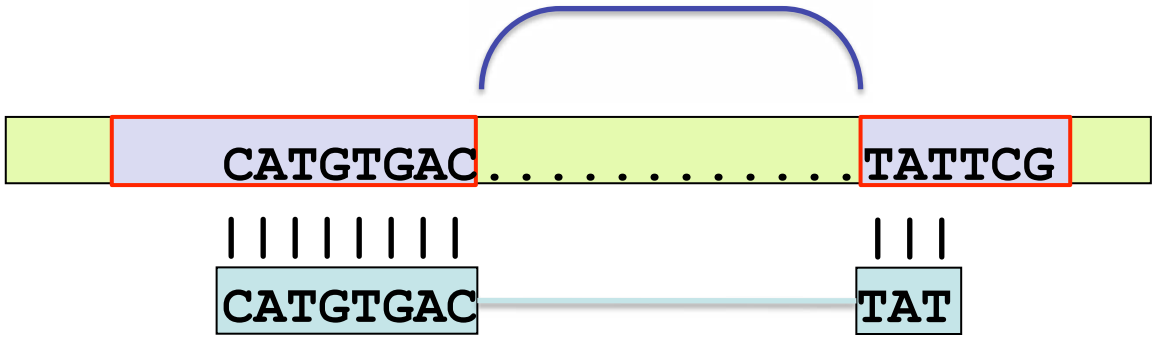
~80% of pseudogene alignments are false positive

2-pass mapping



1st pass:
Reads with short overhangs are soft-clipped

Read with long overhangs identify novel junctions



2nd pass:
Junctions from the 1st pass are added to the search space

Read with short overhangs map spliced to novel junctions

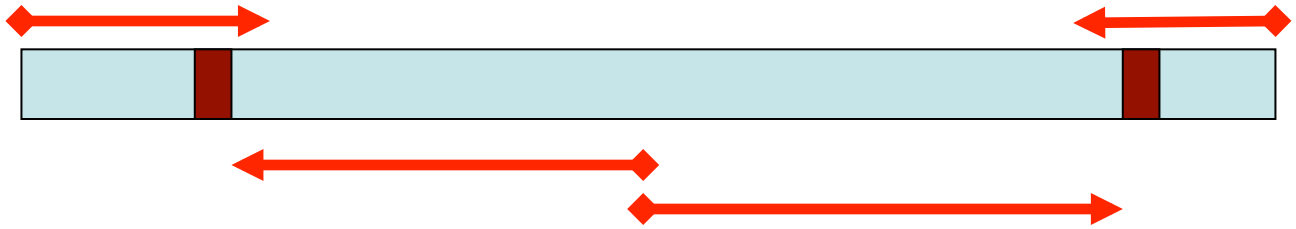
Increasing search sensitivity

- One mismatch near one of the ends

  Seed is too short – max exact search does not stop at the mismatch and maps to a wrong locus

  Solution: search backwards from the other end

- Two mismatches near the ends



  Solution: start search from the middle of the read

- **`--seedSearchStartLmax <N>`**

  user defined parameter to start search as often as needed

  – Reducing N will increase sensitivity, but reduce mapping speed

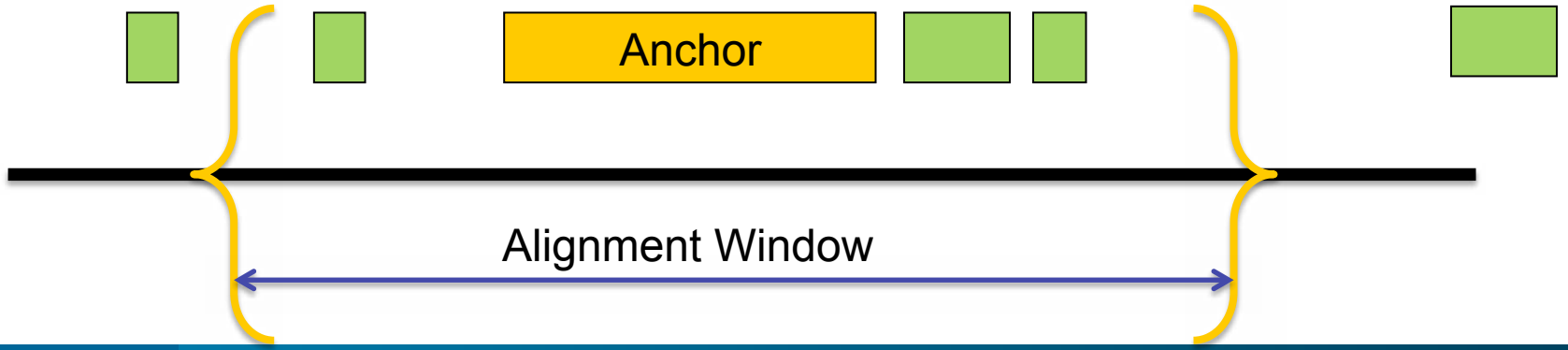
  – Default N= 50b, works well for Illumina reads

Anchor seeds and windows

- **`--seedMultimapNmax <N>`**
  All seeds that map less than N times are recorded: =10,000 by default
  10-mers map 10,000 on average in human genome

- **`--winAnchorMultimapNmax <N>`**
  "Anchors": seeds that map less than N times: =50 by default

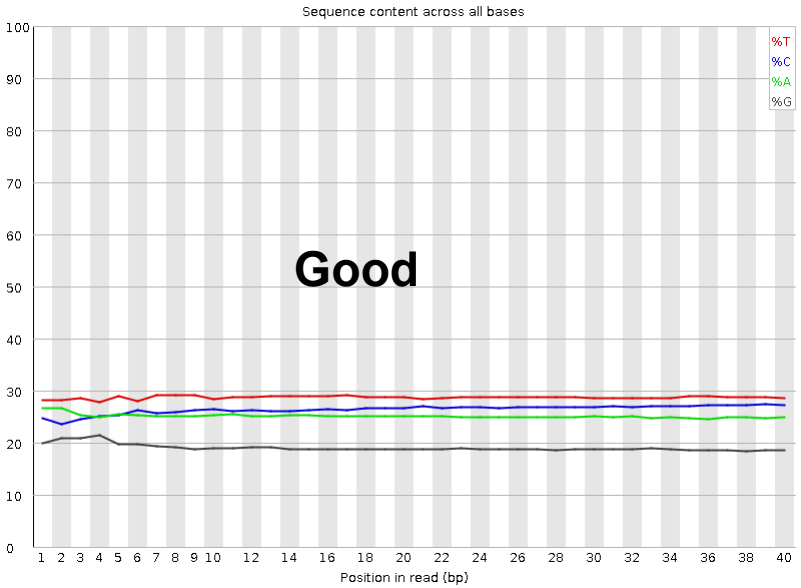
- "Alignment windows": genome regions around anchors
  All seeds inside alignment windows are stitched together
  Size of the window ~ maximum intron size, ~1Mb for human
  `--alignIntronMax <N>, --alignMatesGapMax <N>`

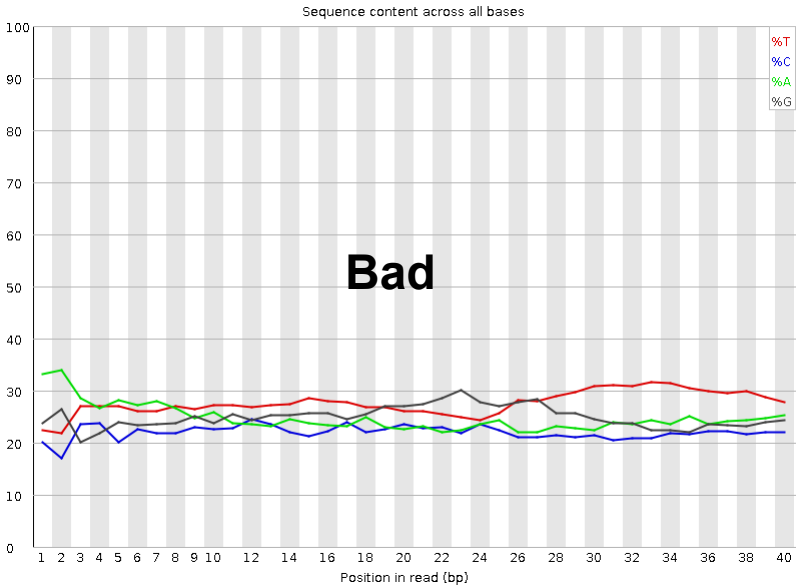


Pre-mapping QC with FASTQC

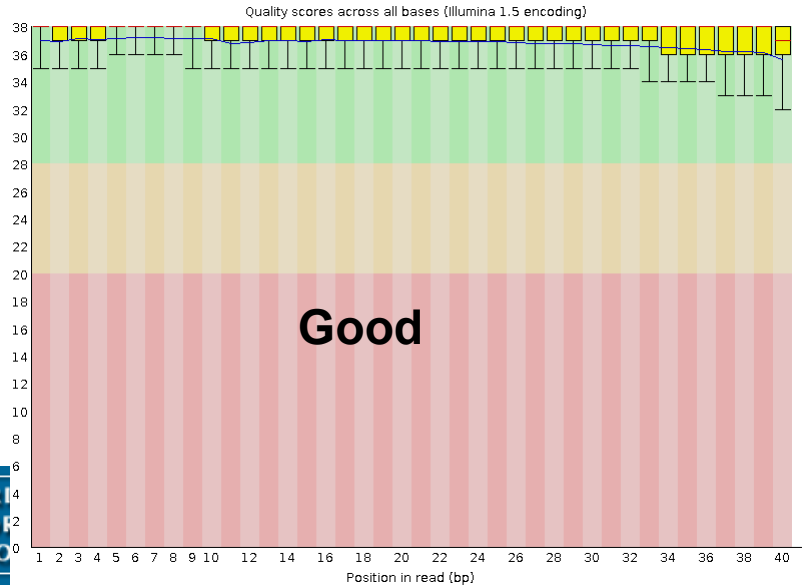
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

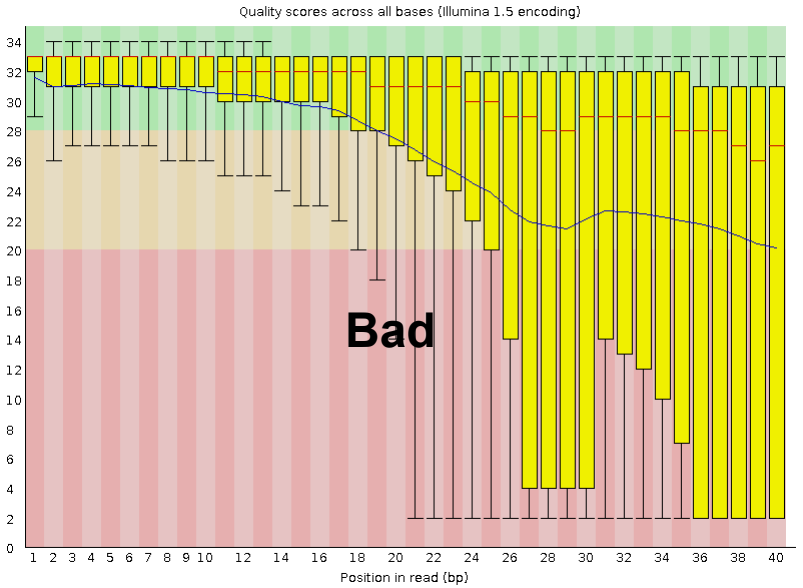
Per-base sequence content





Per-base sequence quality





COLD SPRING HARBOR LABORATORY

Understanding mapping results

STAR's Log.final.out file:

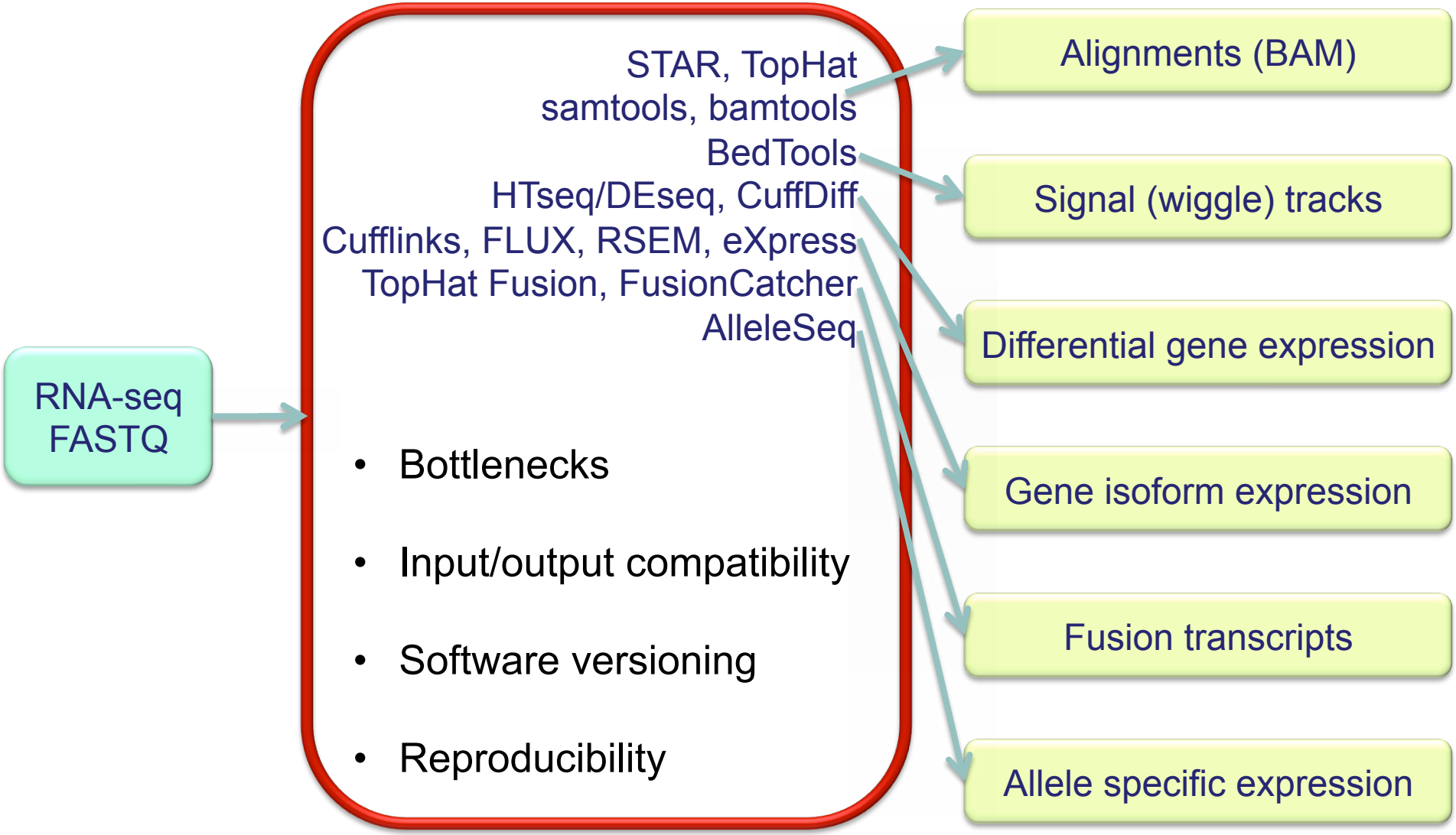
```
                          Average input read length |         202
                                       UNIQUE READS:
                            Uniquely mapped reads % |         90.08%
                              Average mapped length |         201.98
                           Mismatch rate per base, % |         0.30%
                             Deletion rate per base |         0.02%
                            Insertion rate per base |         0.01%
                                  MULTI-MAPPING READS:
               % of reads mapped to multiple loci |         3.55%
               % of reads mapped to too many loci | (>10)  0.02%
                                  UNMAPPED READS:
        % of reads unmapped: too many mismatches |         2.82%
                  % of reads unmapped: too short |         3.44%
                      % of reads unmapped: other |         0.08%
```

Why my mapping rate is low?

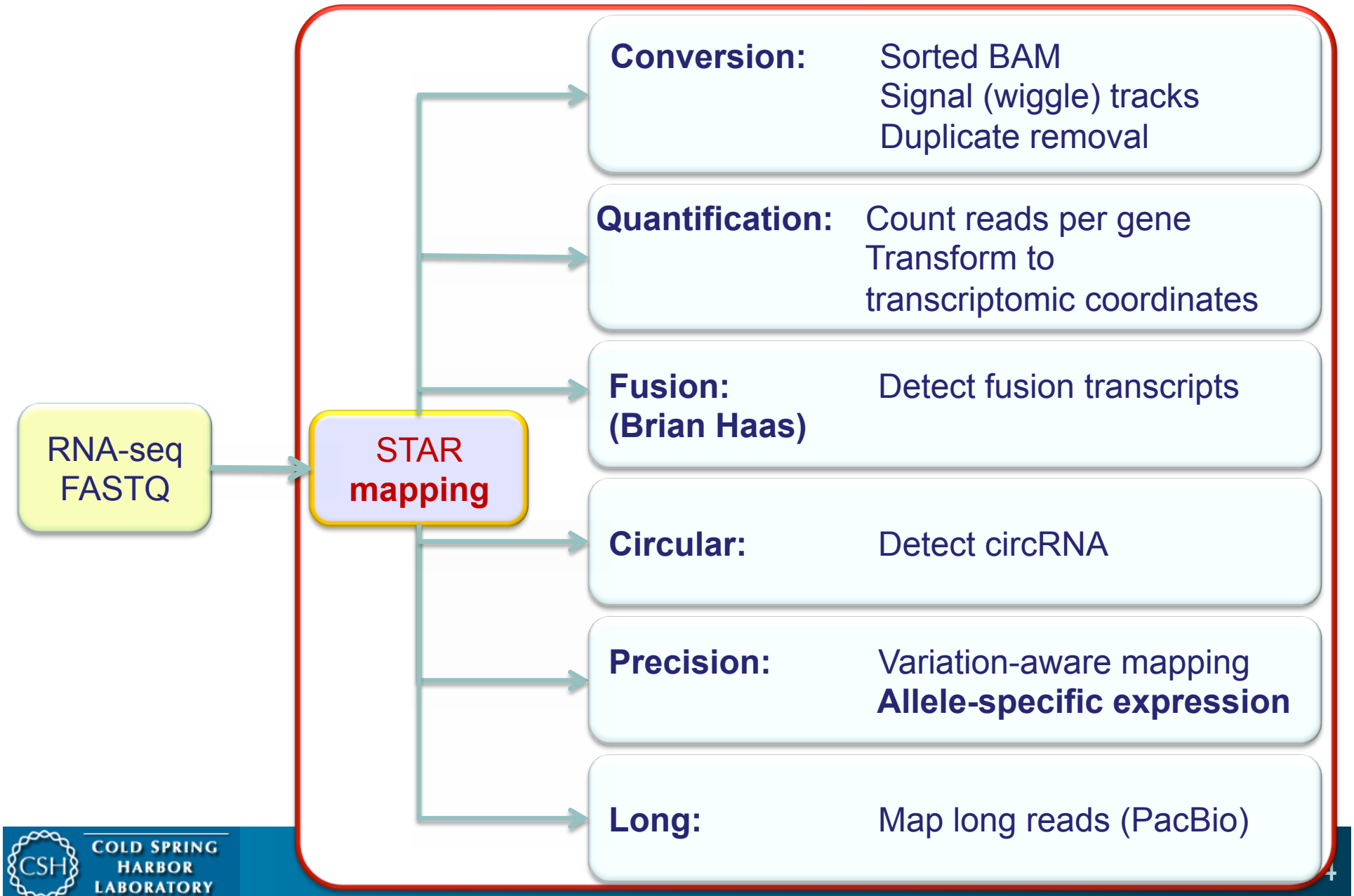
Possible problem	Checks/Solutions
• File formatting mix-up:	Ensure the same order of read1/2
– read1/read2 order broken	Map read1/2 separately
• Poor quality of sequencing	Plot quality scores vs read length
• Tails	
– Poor quality	Trim by quality
– Adapter - short insert	Trim adapter
• rRNA insufficient depletion	Include rRNA sequences in the reference
• Contamination with other species	BLAST unmapped reads

STARtools: post-mapping analyses at no extra cost

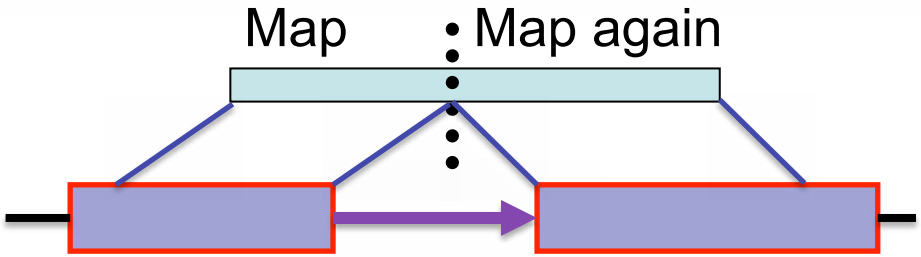
RNA-seq pipeline



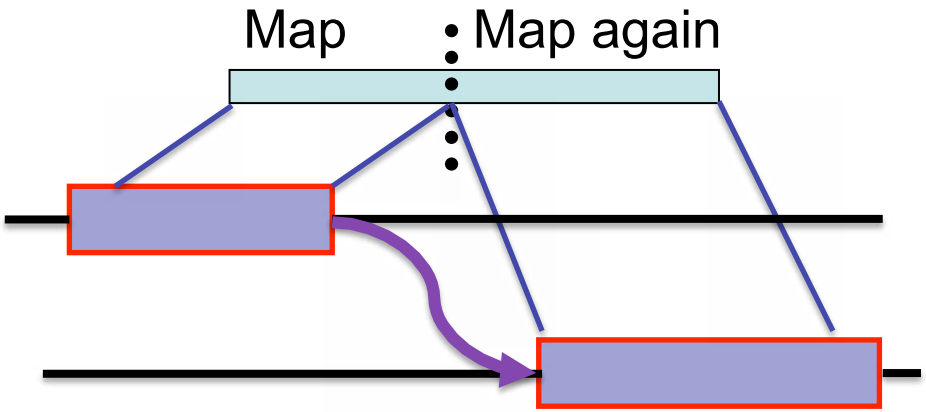
STARtools


RNA-seq
FASTQ
STAR
mapping
Conversion:
Sorted BAM
Signal (wiggle) tracks
Duplicate removal
Quantification:
Count reads per gene
Transform to
transcriptomic coordinates
Fusion:
(Brian Haas)
Detect fusion transcripts
Circular:
Detect circRNA
Precision:
Variation-aware mapping
Allele-specific expression
Long:
Map long reads (PacBio)

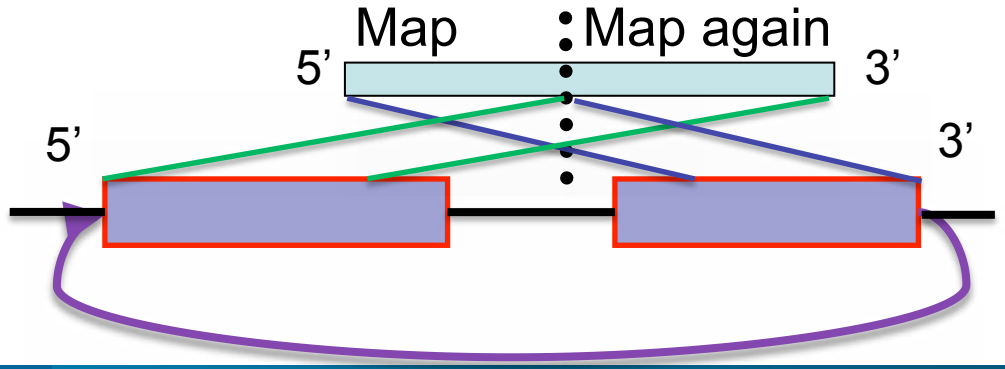
Chimeric and circular junctions



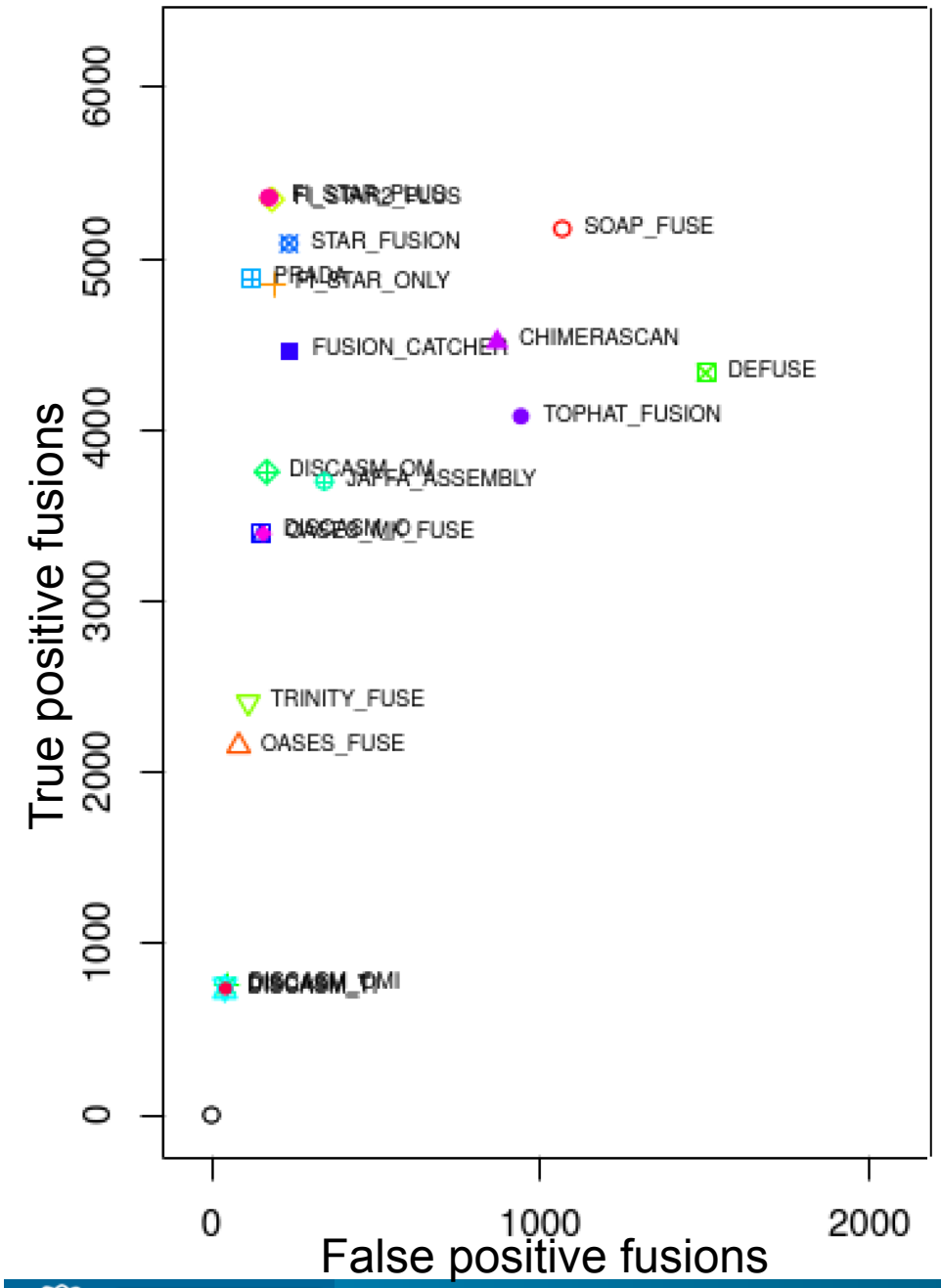
Linear junction



Chimeric junction



Circular junction



STAR-Fusion

STAR-Fusion
FusionInspector

developed by **Brian Haas** (Broad Institute)

Analyzes STAR chimeric alignments to detect fusion transcripts in RNA-seq data

https://github.com/STAR-Fusion/STAR-Fusion

Part B
Quantification and normalization

Shanrong Zhao & Baohong Zhang

Pfizer

B#1: A systematic investigation of the impact of gene annotations on RNA-seq data analysis

Experimental design and analysis

16 human tissues from Human Body Map 2.0

a. Adipose
b. Adrenal
c. Brain
d. Breast
e. Colon
f. Heart
g. Kidney
h. Leukocyte
i. Liver
j. lung
k. Lymphnode
l. Ovary
m. Prostate
n. Skeletal muscle
o. Testis
p. Thyroid

X

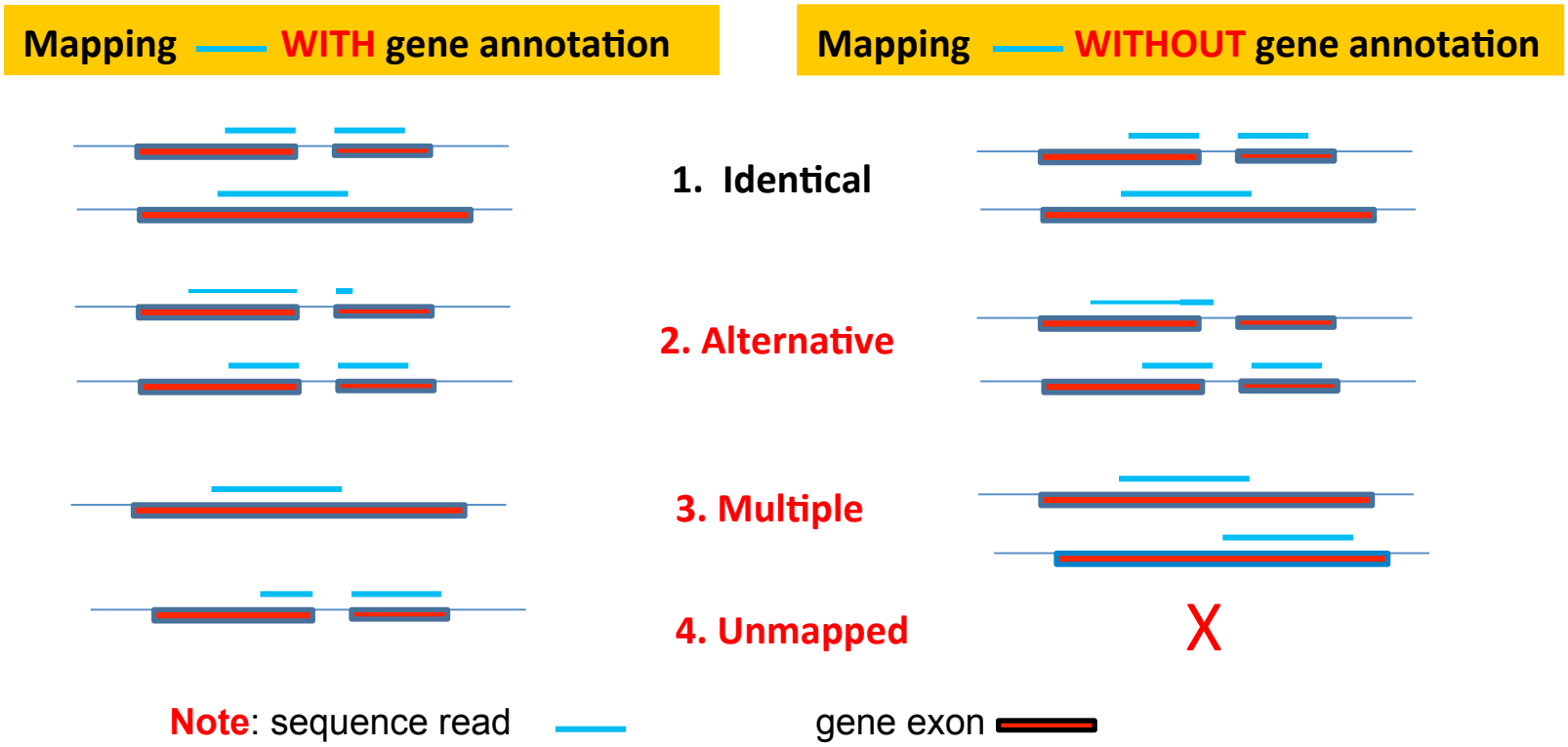
Gene Annotations

1. **RefGene (RefSeq Gene)**
2. **Ensembl**
3. **UCSC gene**
4. **None (no gene annotation)**

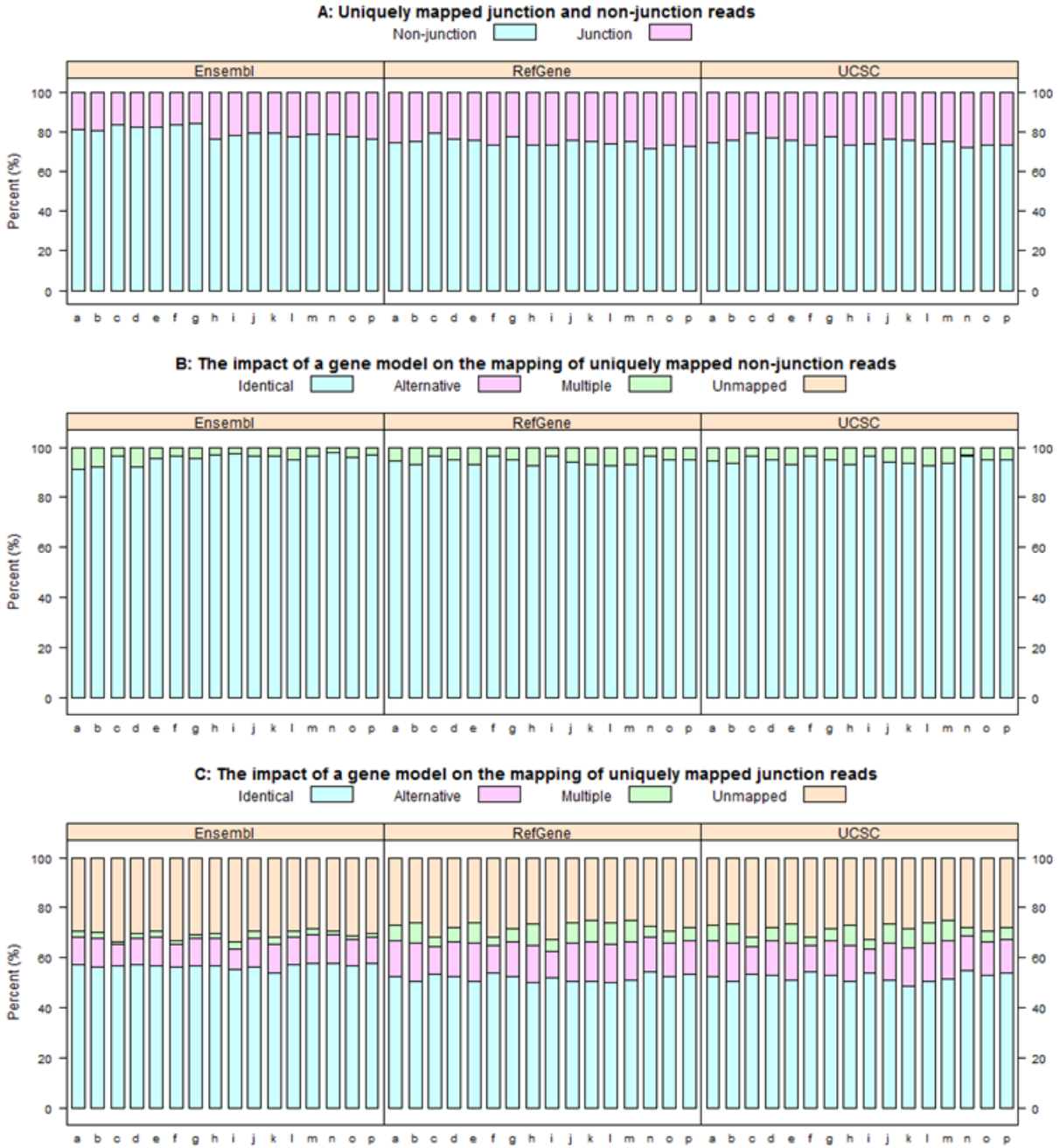
Analysis protocol

- All combinations of tissues and gene annotation were analyszed
  - **64** (=16X4) runs
- **Compare:**
  1. RefGene, Ensembl, UCSC versus None
  2. RefGene versus Ensembl

Classification of uniquely mapped reads based upon their mappings with and without gene annotation



1. **Identical**: the same alignment regardless of the use of a gene model;
2. **Alternative**: still mapped but mapped differently;
3. **Multiple**: a uniquely mapped read by gene annotation mapped to multiple genomic locations;
4. **Unmapped**: i.e., a read could not be mapped to anywhere in the genome without the assistance of a gene model.



The impact of gene annotation on RNA-Seq read mapping (read length:75 bp).

(A) Breakdown of reads

- Junction reads: **~23%**
- Non-junction reads: **~77%**

(B) Non-junction reads

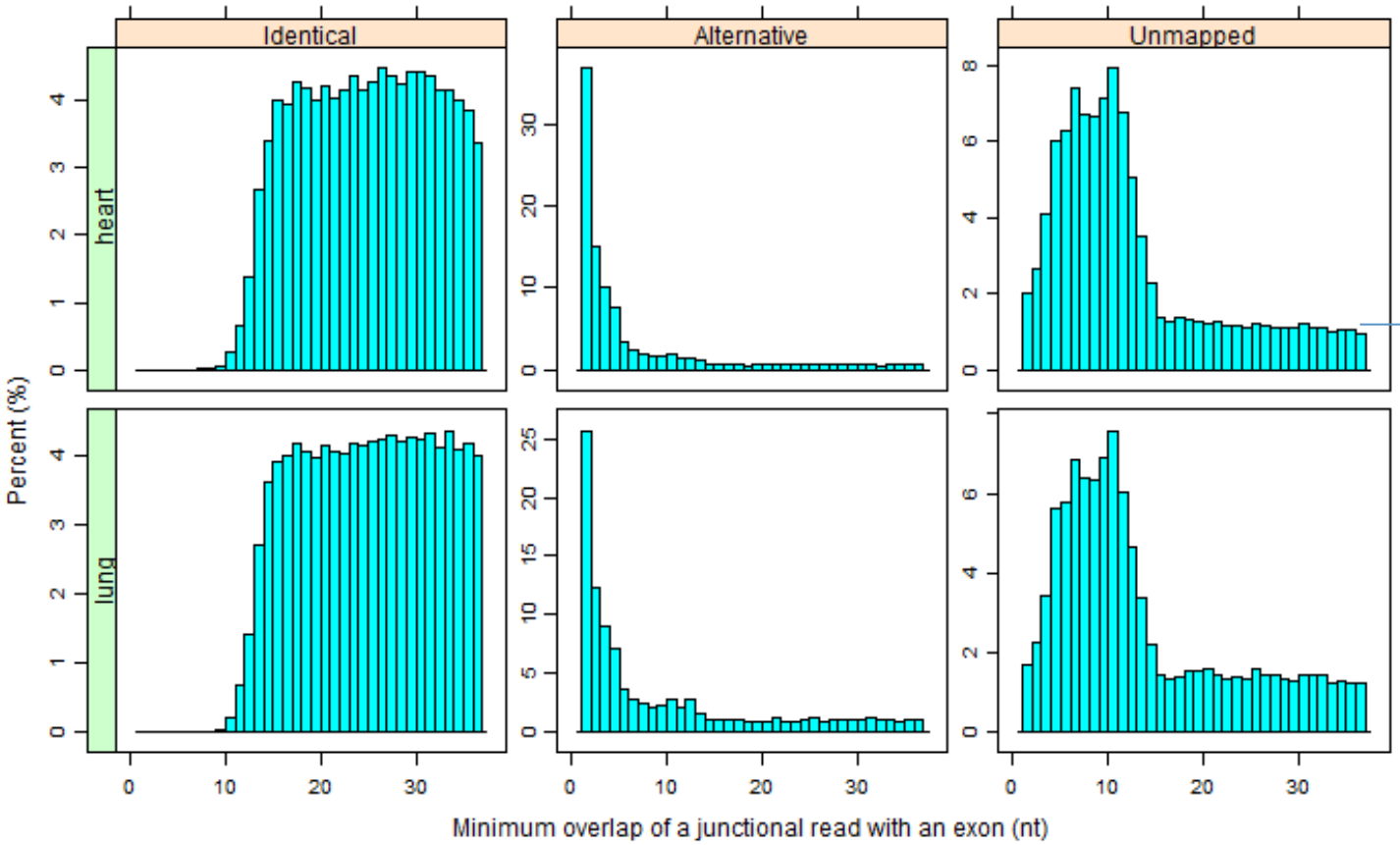
- **~95%** remain intact
- **3–9%** multiple mapped

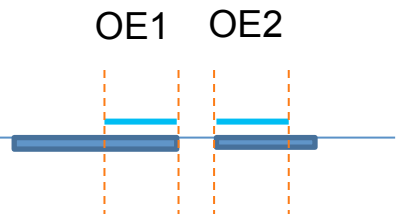
(C) **Junction reads**

- **~53%** of reads remain intact
- **3–9%** multiple mapped
- **10–15%** mapped alternatively
- **~30%** fail to be mapped

(Note: the 16 tissue sample names are denoted as follows: **a**: adipose; **b**: adrenal, **c**: brain; **d**: breast; **e**: colon; **f**: heart; **g**: kidney; **h**: leukocyte; **i**: liver; **j**: lung; **k**: lymph node; **l**: ovary; **m**: prostate; **n**: skeletal muscle; **o**: testis; and **p**: thyroid).

The splicing patterns for *"Identical"*, *"Alternative"* and *"Unmapped"* junction reads





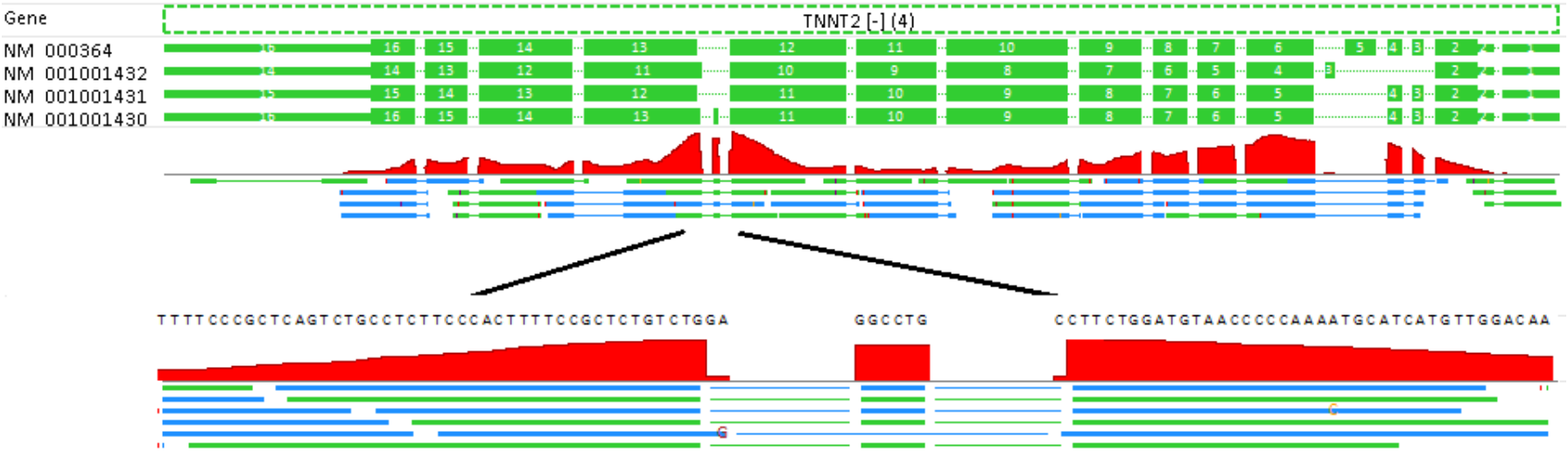
MOE=min(OE1,OE2)

MOE is calculated only for those junctional reads that span two exons

Since the read length is 75 bp long, the MOE ranges from 1 to 37 for any junction read.

- For "I**dentical**" junction reads, the typical MOE ranges from **15 to 37**, and the frequency drops to nearly 0 when MOE<10.
- For "**Alternative**" junction reads, the most dominant MOE is 1, representing an average 1/3 of cases. For those junction reads with MOE of **1,2** and **3**, it is virtually impossible to map them 'correctly' without the prior knowledge on transcripts.
- The MOE for "**Unmapped**" reads has a much broader range with peaks from **4** to 12.

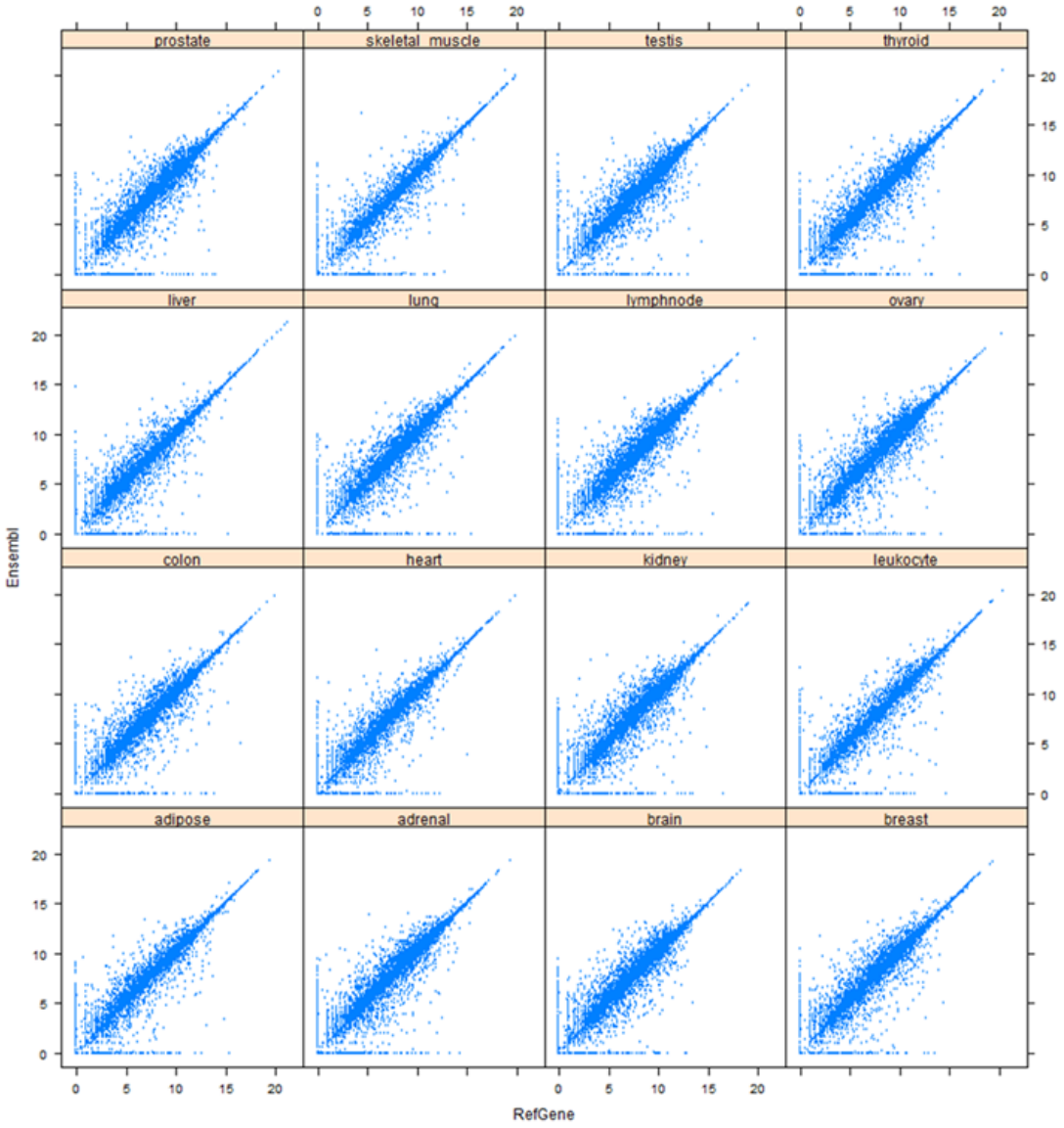
The proper annotation is critical for accurate mapping of spicing reads



- Gene TNNT2 has 4 transcripts, and **Exon #12** in transcript ***NM_001001430*** is only **6 bp** long.
- Without the prior knowledge of transcript structure on ***NM_001001430***, it is virtually impossible to correctly map those splicing reads spanning this exon to the reference genome

Note: Paired-end sequencing; a read is colored in **blue** if mapped to "**+**" strand, and **green** if mapped to "**-**" strand.

The correlation of gene quantification results between RefGene (x-axis) and Ensembl (y-axis)



- Both x and y-axes represent Log2(count +1).
- The majority of genes have highly consistent or nearly identical expression levels
- However, there are many genes whose expressions are dramatically affected by the choice of a gene model.

Summary of difference in gene quantifications between RefGene and Ensembl annotation

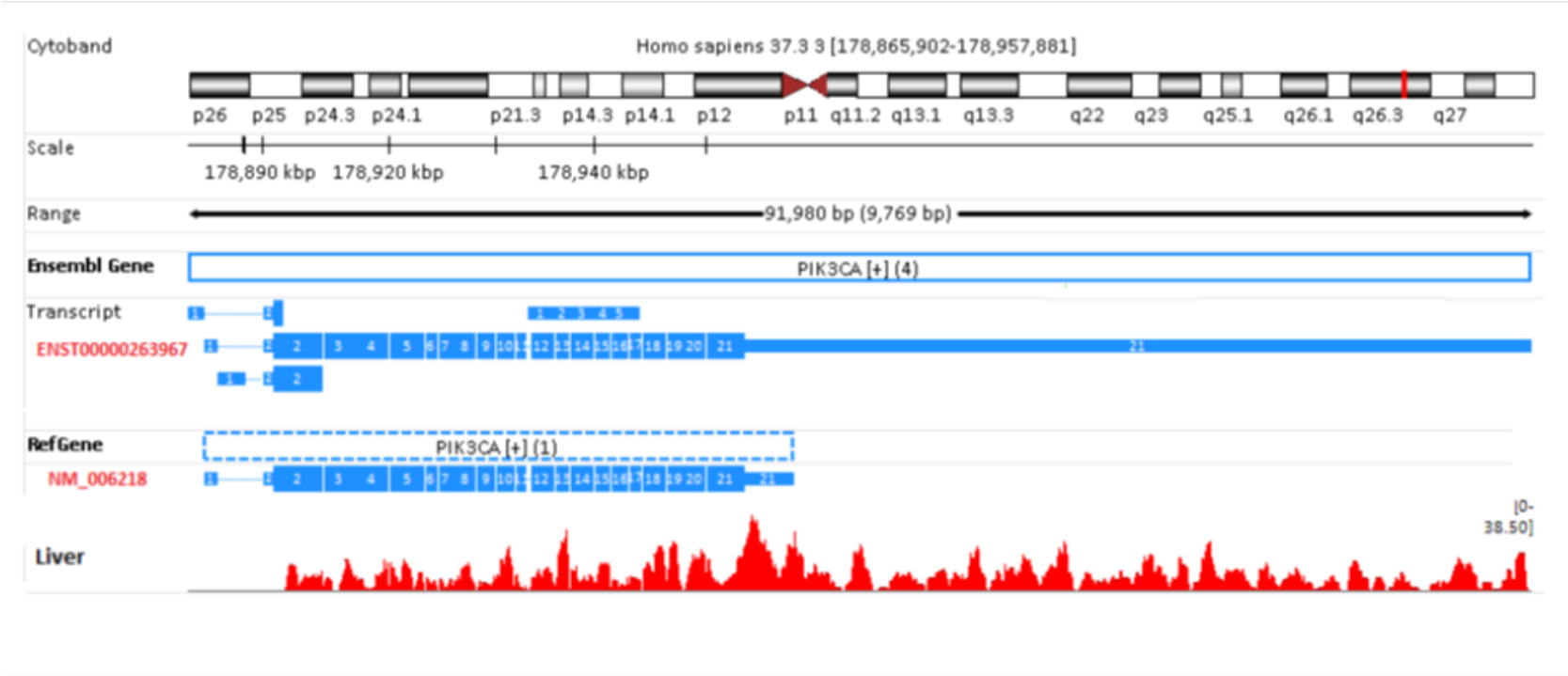
The number of common genes: **21958**

Difference	# of genes
No expression	20%
No difference	**16.3%**
>5%	**28.1%**
>50%	**9.3%** (**2038** genes)

Note:

- The results in the table represent the average of all 16 samples
- The choice of a gene model **has a large impact** on gene quantification.
- The discrepancy mainly results from inconsistent gene definitions

PIK3CA: gene definitions in RefGene and Ensembl



- The different gene definitions for PIK3CA give rise to differences in expressions
- PIK3CA in the Ensembl annotation is much longer than its definition in RefGene, explaining why **1094 reads** are obtained in Ensembl, while only **492 reads** in RefGene.

http://dx.doi.org/10.5772/61197

Chapter 16

Impact of Gene Annotation on RNA-seq Data Analysis

Shanrong Zhao and Baohong Zhang

Additional information is available at the end of the cha

http://dx.doi.org/10.5772/61197

OPEN ACCESS Freely available online

PLOS ONE

Assessment of the Impact of Using a Reference Transcriptome in Mapping Short RNA-Seq Reads

Shanrong Zhao*¤

Systems Pharmacology and Biomarkers, Janssen Research & Development, LLC, San Diego, California, United States of America

Zhao and Zhang *BMC Genomics* (2015) 16:97
DOI 10.1186/s12864-015-1308-8

BMC Genomics

RESEARCH ARTICLE

BMC Genomics 2015, 16:97

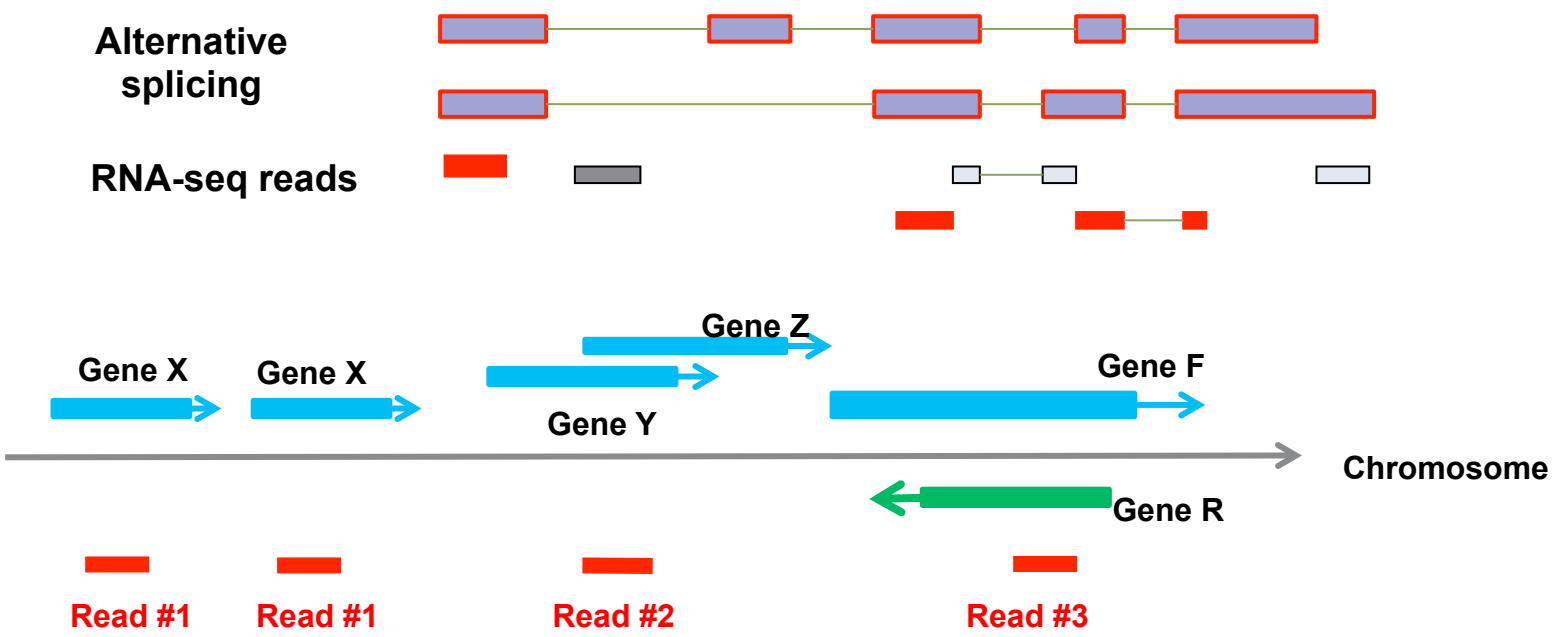
Open Access

A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification

Shanrong Zhao* and Baohong Zhang

B#2: Gene quantification: “Read Ambiguity” is true guilt for all kinds of complexities

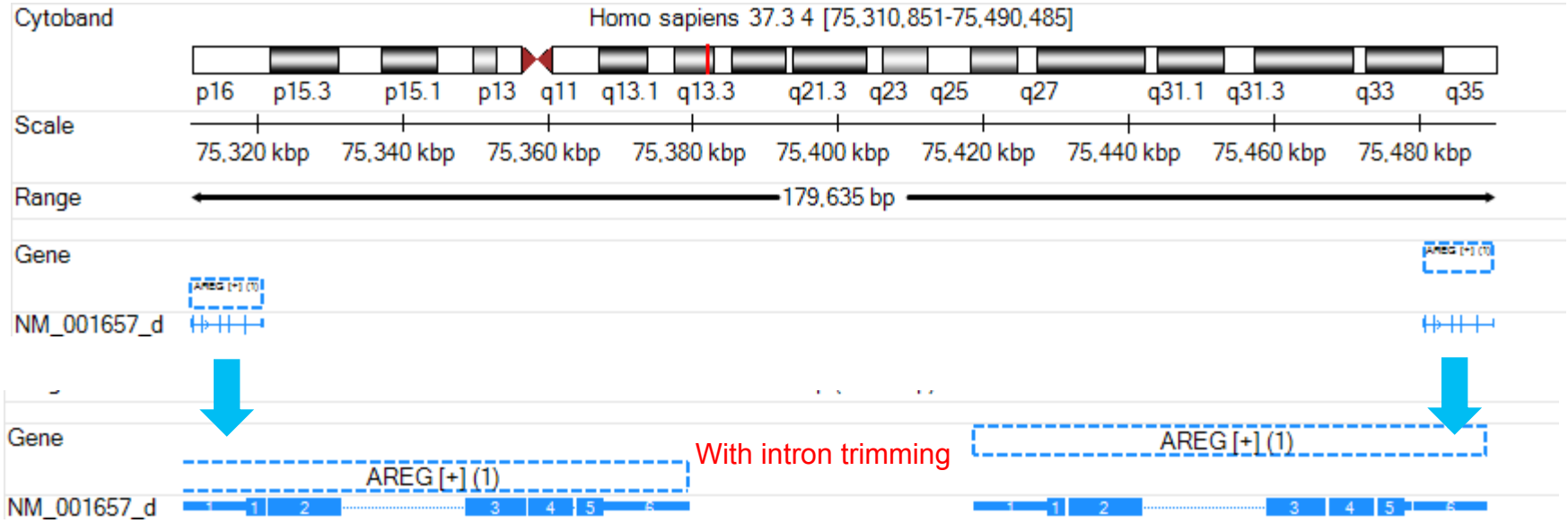
Quantification: count reads per gene/isoform



Reads Ambiguity

1. reads mapped to shared exons of different isoforms of the same gene
2. multiple mapped reads, such as Read #1
3. reads mapped to (same-strand) overlapping gene regions, such as Read #2
4. reads mapped to overlapping genes from **opposite strands**, such as Read #3

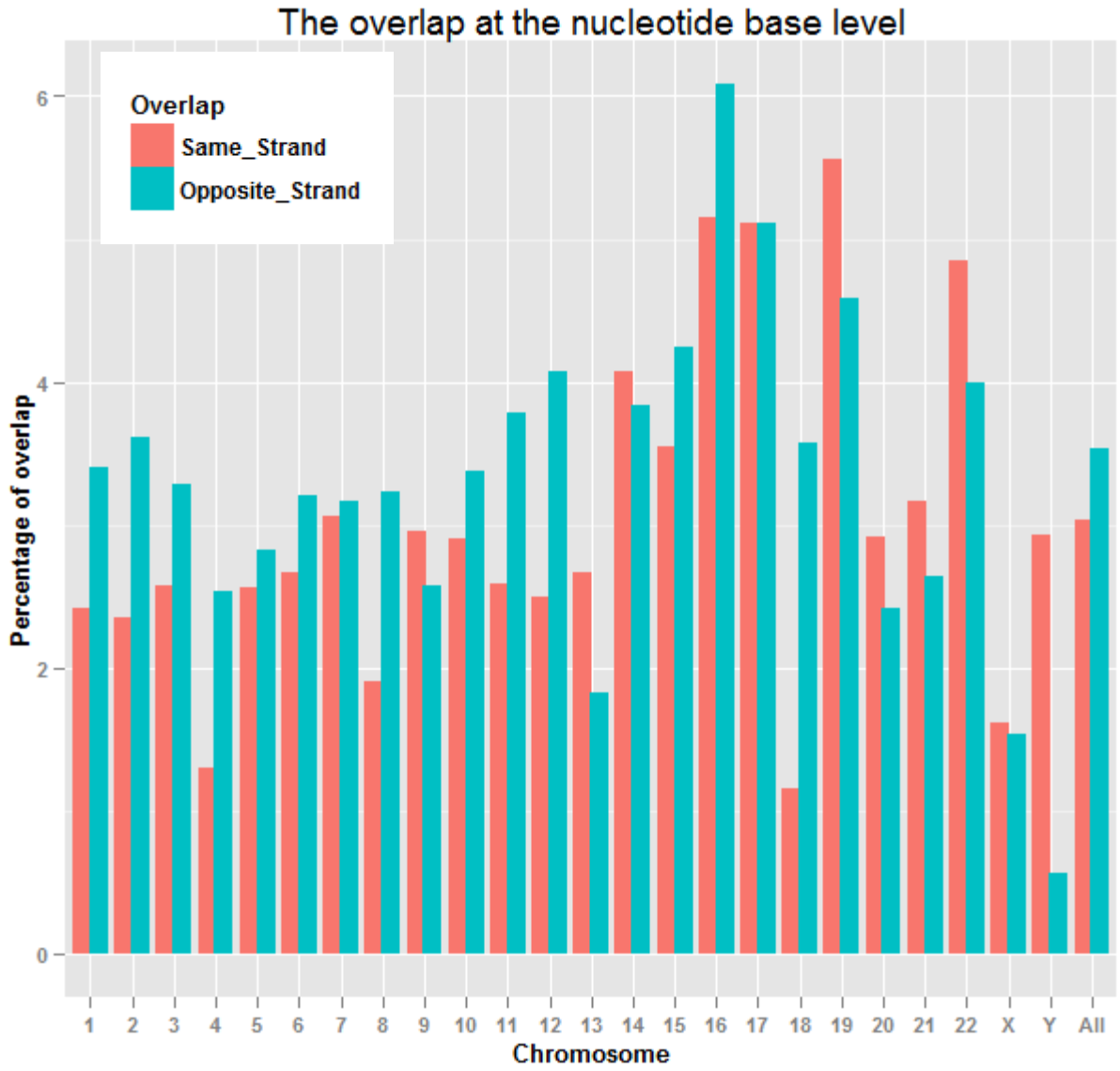
To exclude multiple mapped reads from counting is problematic for duplicated genes such as AREG



- Gene AREG has two locations in human genome, according to RefGene.
- Reads mapped to AREG are ALWAYS multiple-mapped ones.

Many counting algorithms ignore multiple mapped reads, and this practice is fine for most genes, but is problematic for those genes with duplications and homologous ones.

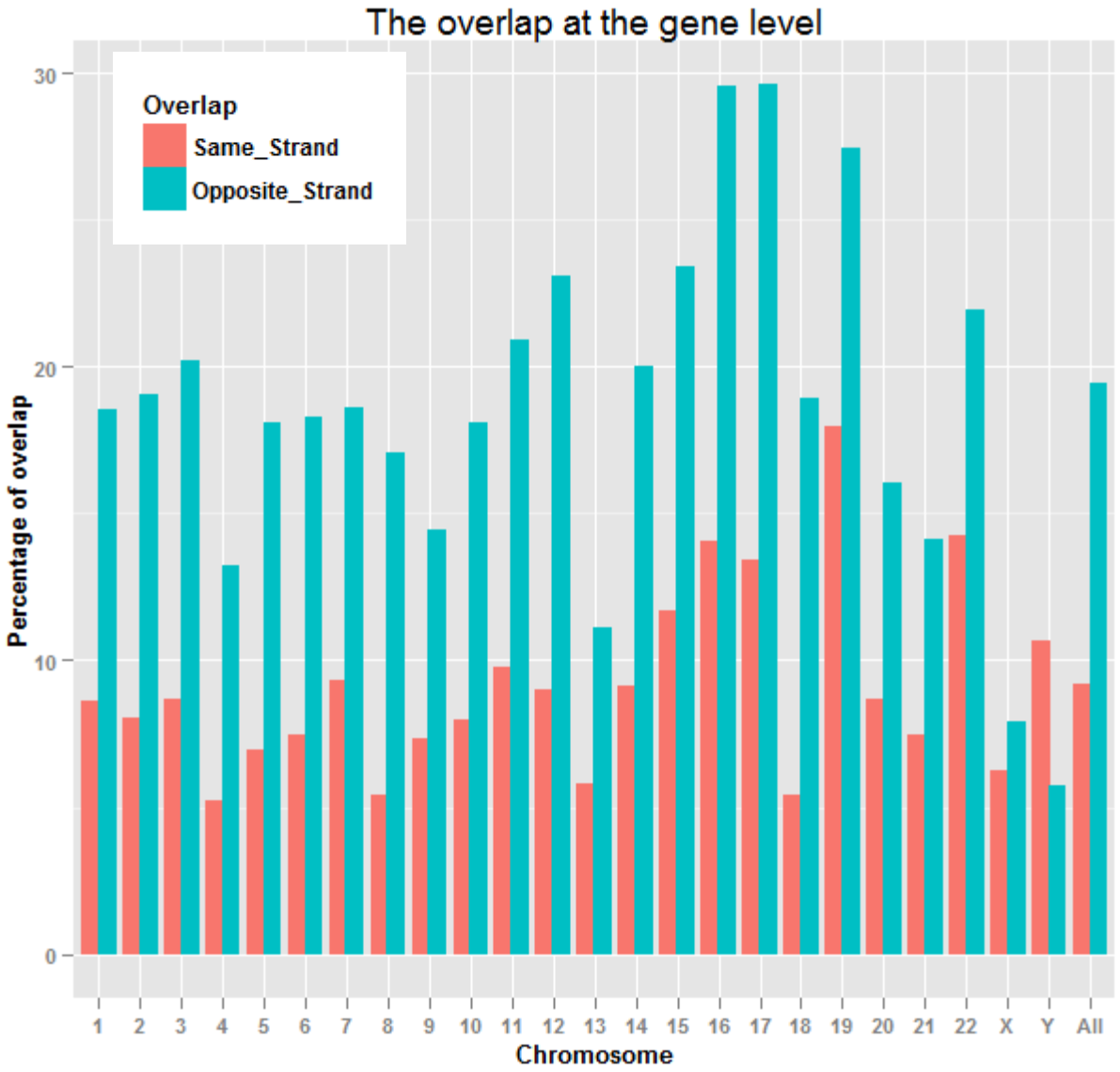
Theoretical overlap at the nucleotide base level in Gencode Human Release 19



On average:

- **Same strand: 3.03%**
- **Opposite strand: 3.54%**
- **Any strand: 6.57%**

Theoretical estimation of gene overlaps in Genecode Human Release 19: *Gene overlaps is not uncommon*

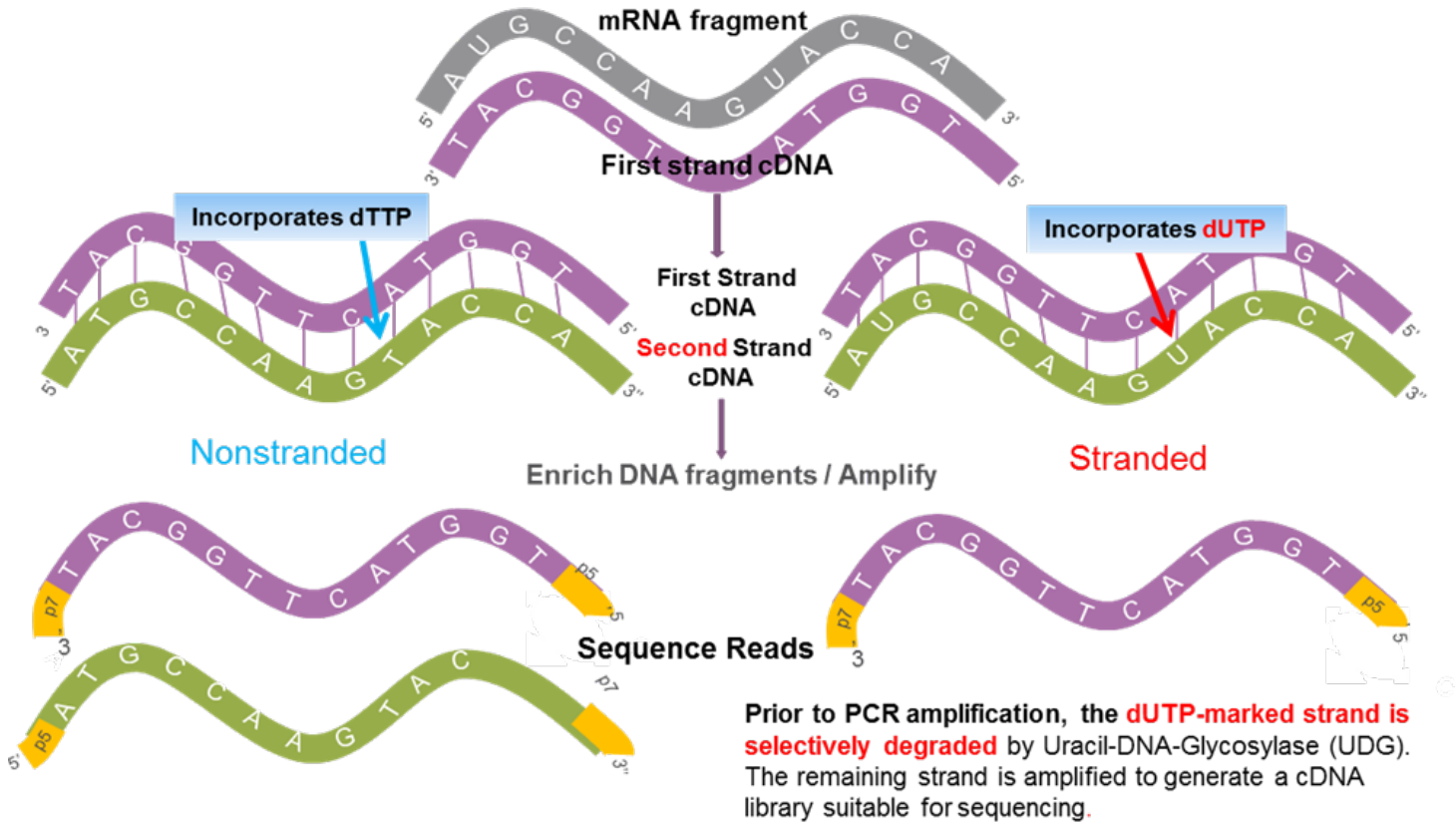


The percentage of genes that overlap with others at least 1bp

On average:

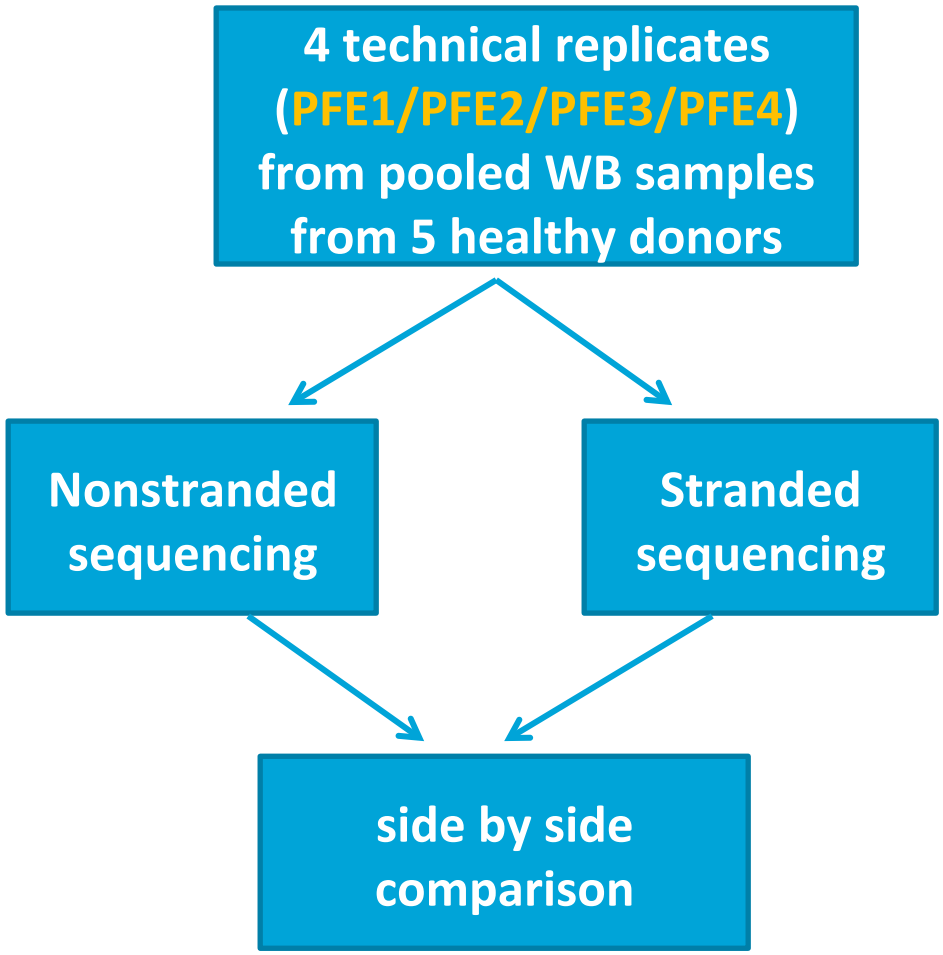
- **Same strand: 9.18%**
- **Opposite strand: 19.45%**
- **Any strand: 25.99%**

Stranded RNA-seq to resolve ambiguous reads mapped to overlapping genes from opposite strands

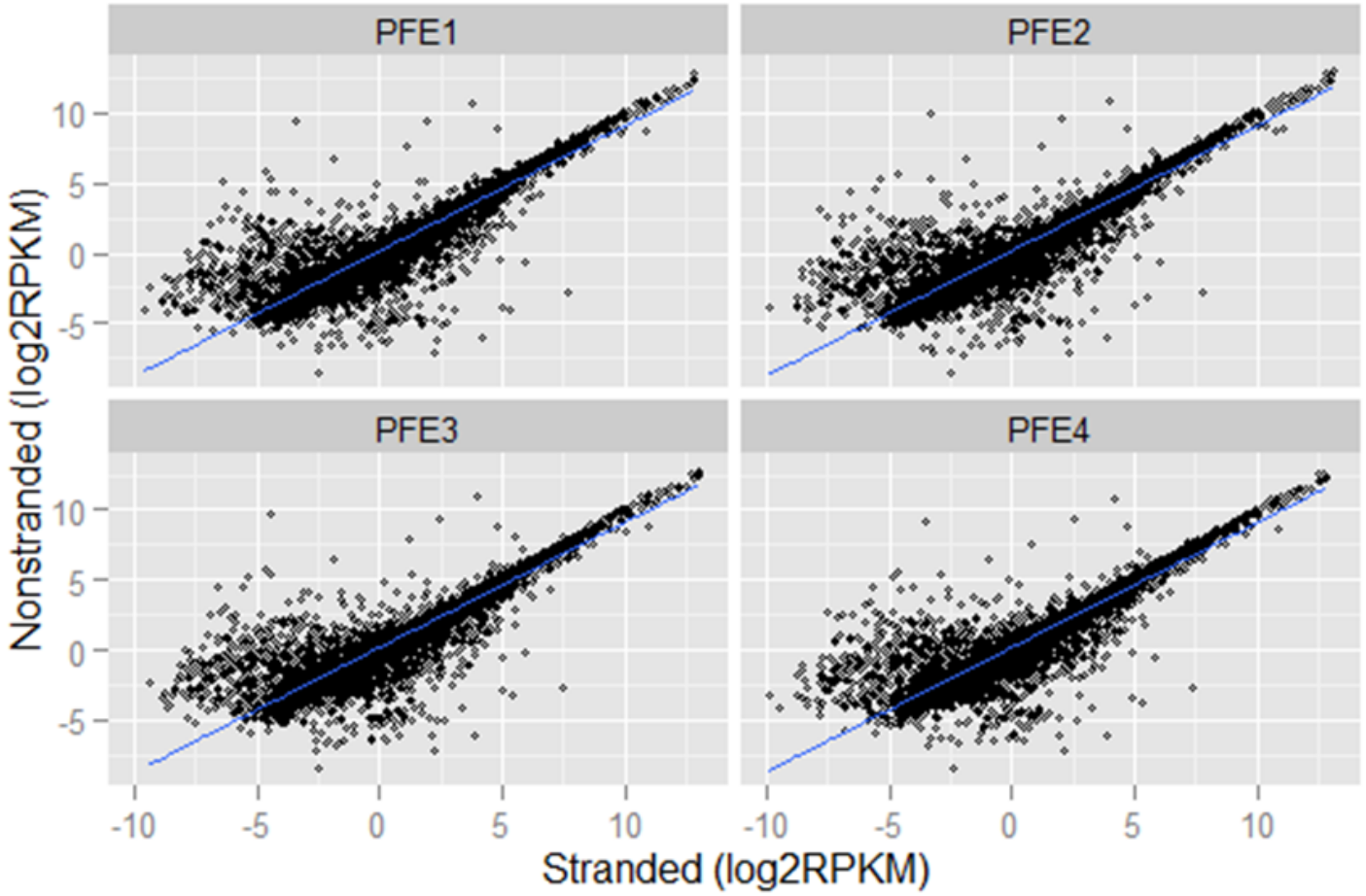


For dUTP based stranded sequencing protocol, the sequence reads are **reversely complementary** to is **originating mRNA transcript**.

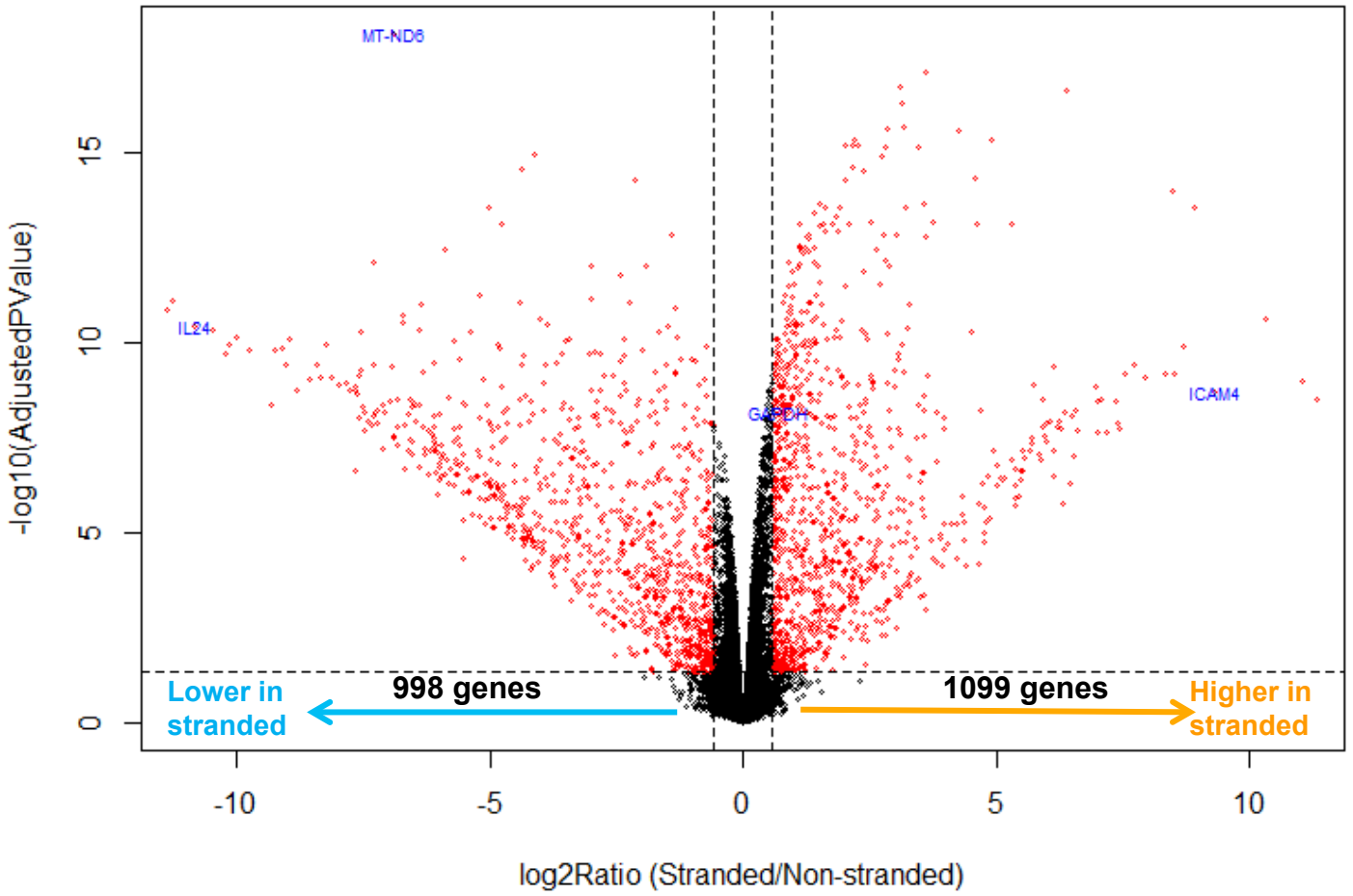
Comparison of stranded versus non-stranded



Scatter plot of gene expression: non-stranded versus stranded

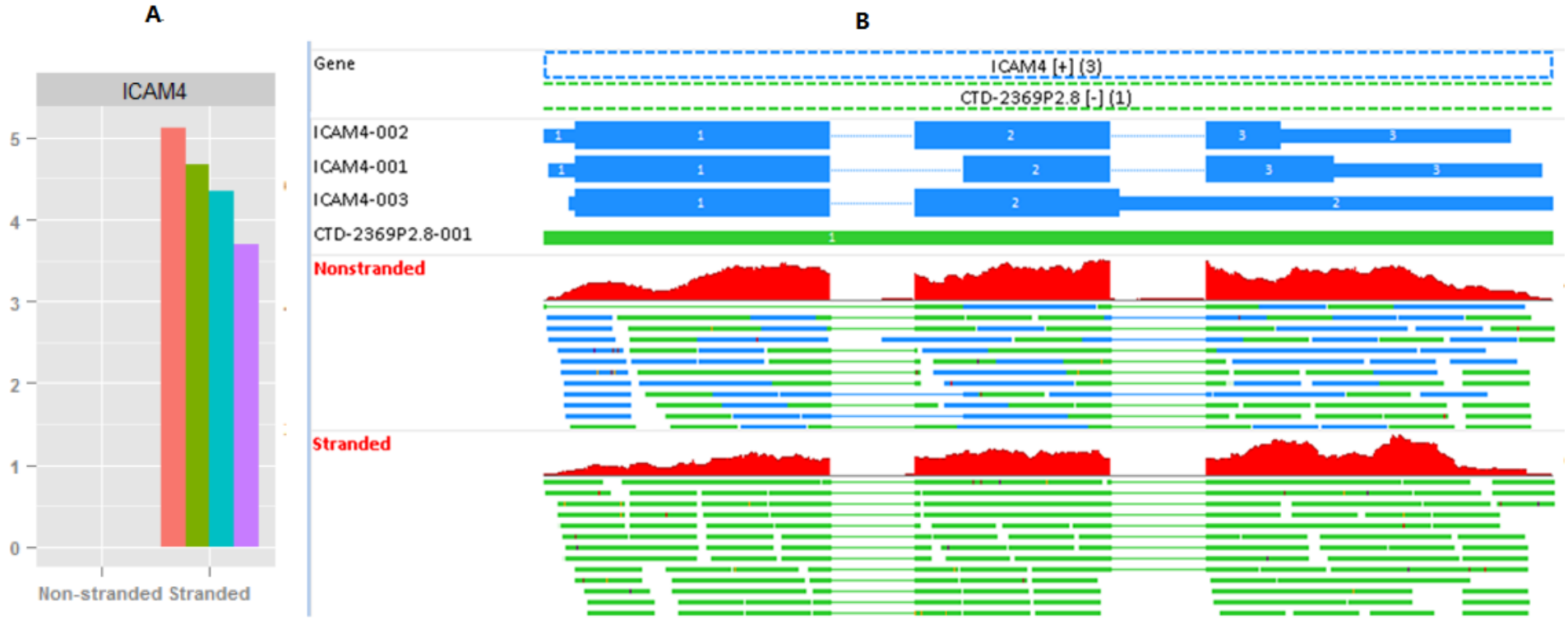


Volcano plot: summary of differential analysis



Note: each dot represents a gene, and all data points colored in **red** are significant genes.
The criteria for significance: (1) adjusted pvalue **< 0.05**; and (2) fold change **> 1.5**

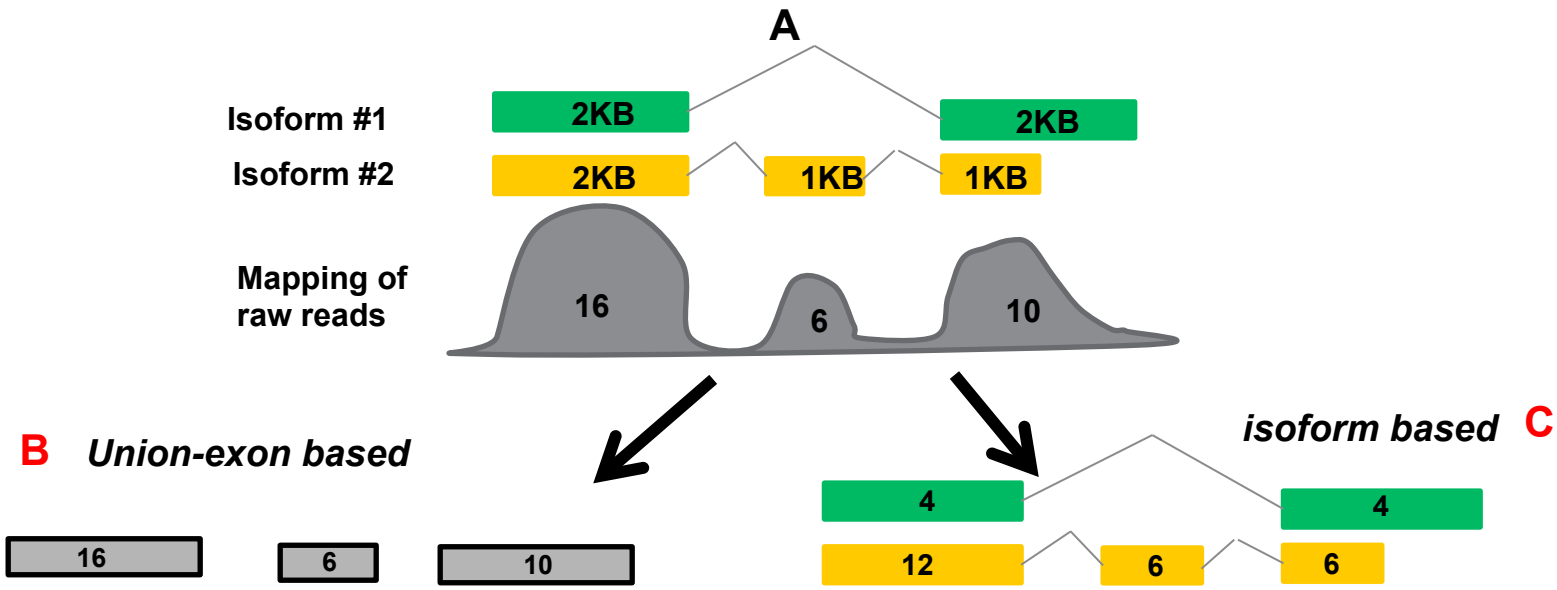
Stranded RNA-seq tells the truth on ICAM4



Note: forward and reverse strand reads are colored blue and green, respectively.

- ICAM4 is known to have medium level expression in whole blood
- ICAM4 has 3 isoforms, but all are 100% contained within gene CTD-2369P2.8
- In nonstranded reports, all reads mapped to overlapping gene are excluded from counting. In stranded RNA-seq, all the reads are reversely complementary to ICAM4, and they are counted towards to ICAM4 only.

“Union-exon” versus isoform based approaches for gene quantification

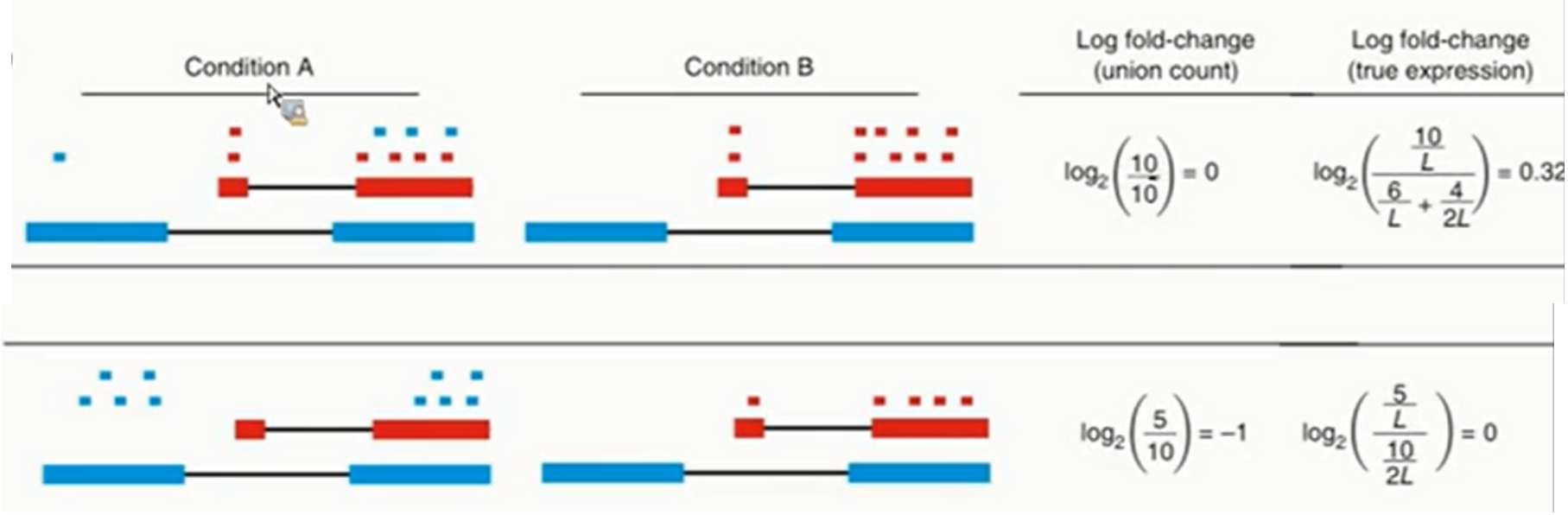


Gene length after exon flattening: **5KB**
Total reads **32**
RPKM for gene: **6.4 (=32/5)**

Relative isoform abundance (#1/#2): **25% / 75%**
RPKM for isoforms #1 and #2: **2 and 6**
RPKM for gene (=sum of isoforms): **8 (=2+6)**

Union-exon based approach is simpler, but can be misleading when the lengths of isoform differ significantly. Unfortunately, it’s notoriously challenging to counts read towards individual isoforms due to read ambiguilty

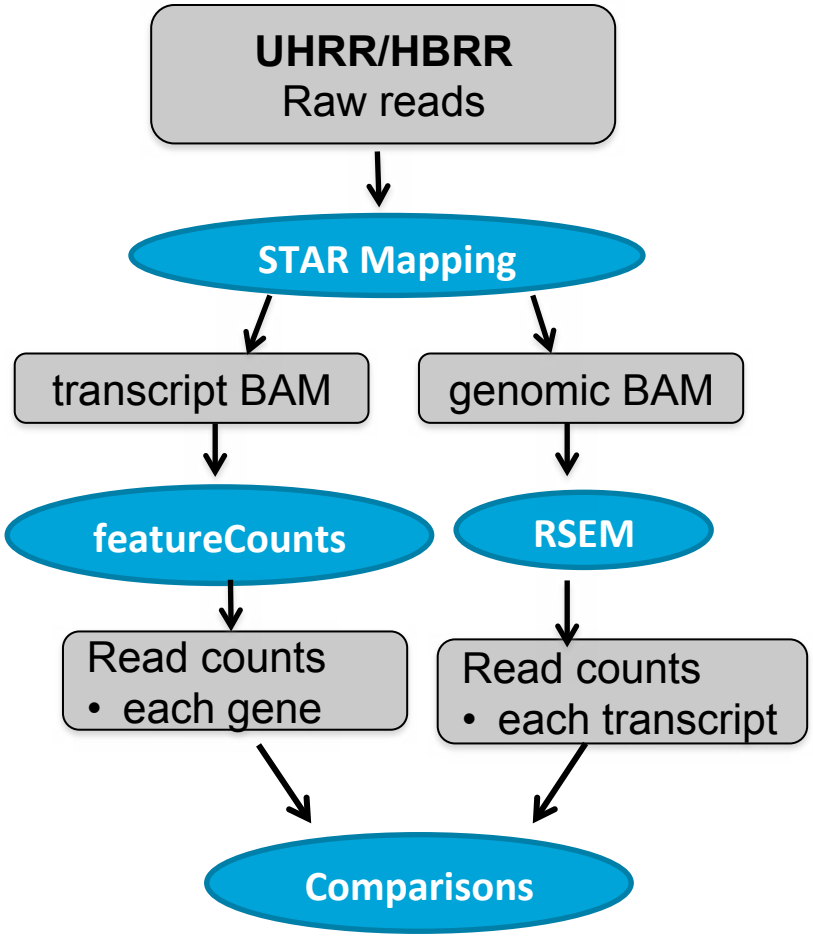
Illustration of union-exon based versus isoform based approaches in differential analysis



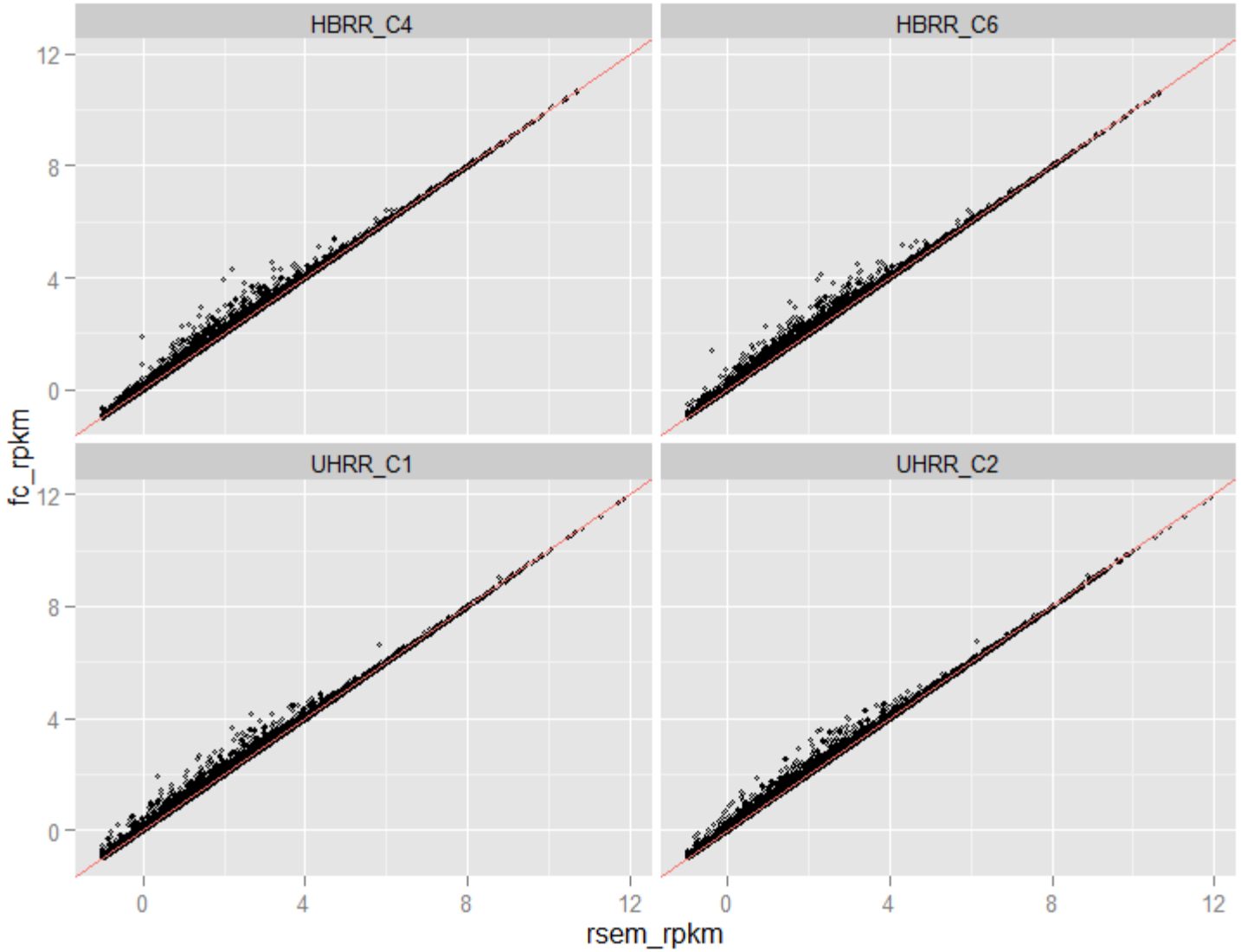
Adapted from Trapnell et al. Nat Biotech, 2013, 31:46–53

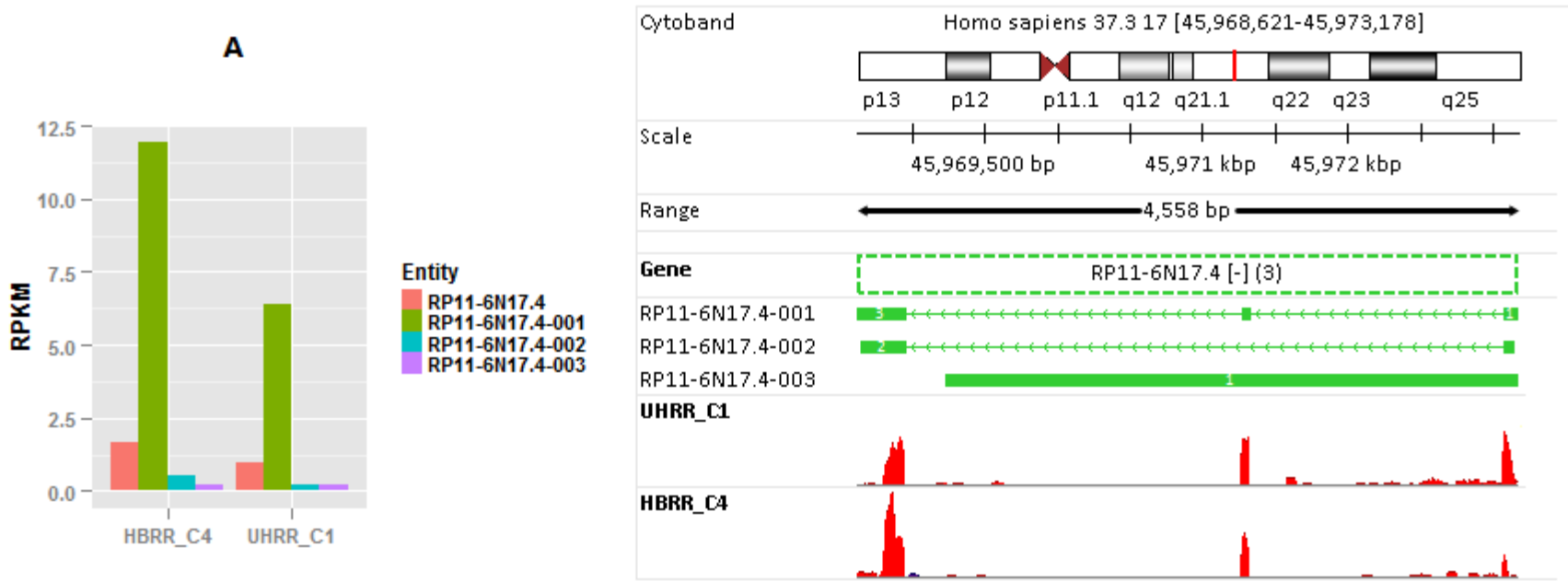
The expression changes in isoforms might be masked by gene level differential analysis, and this is the downside of union exon based approach.

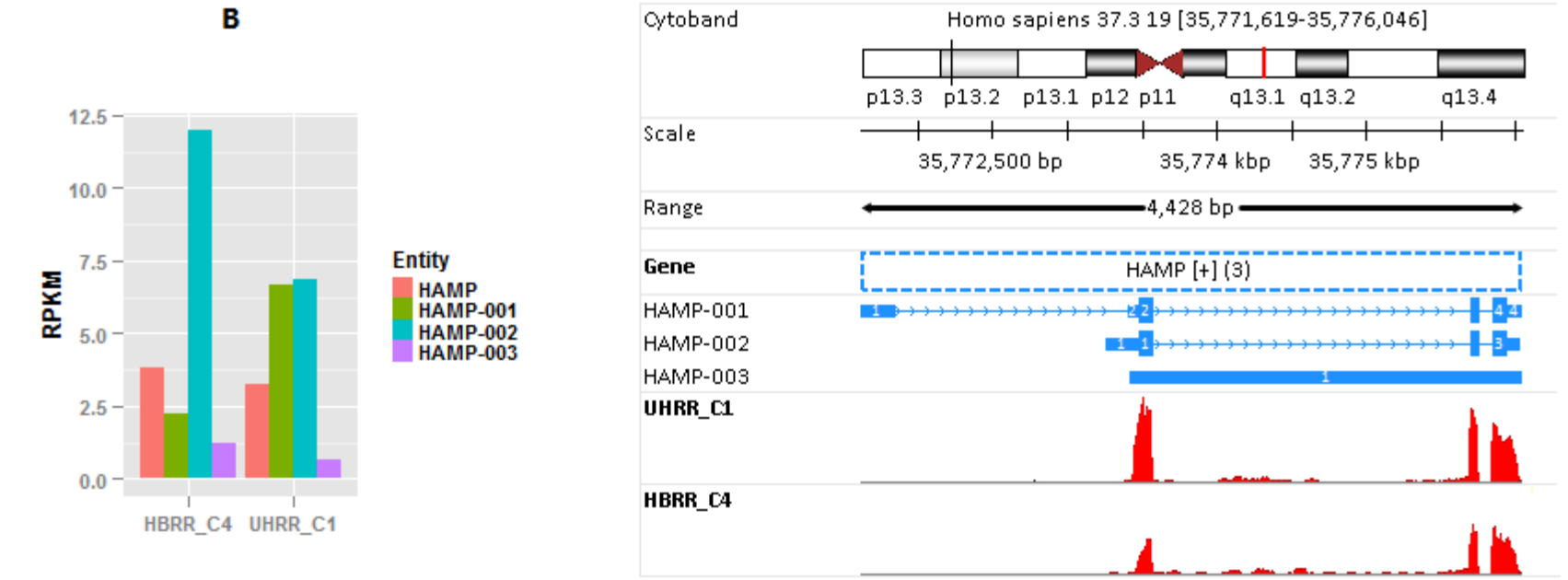
Evaluation of “union-exon” based versus isoform based approaches



Scatter plot: “union-exon” based (fc_rpkm) versus isoform based (rsem_rpkm)

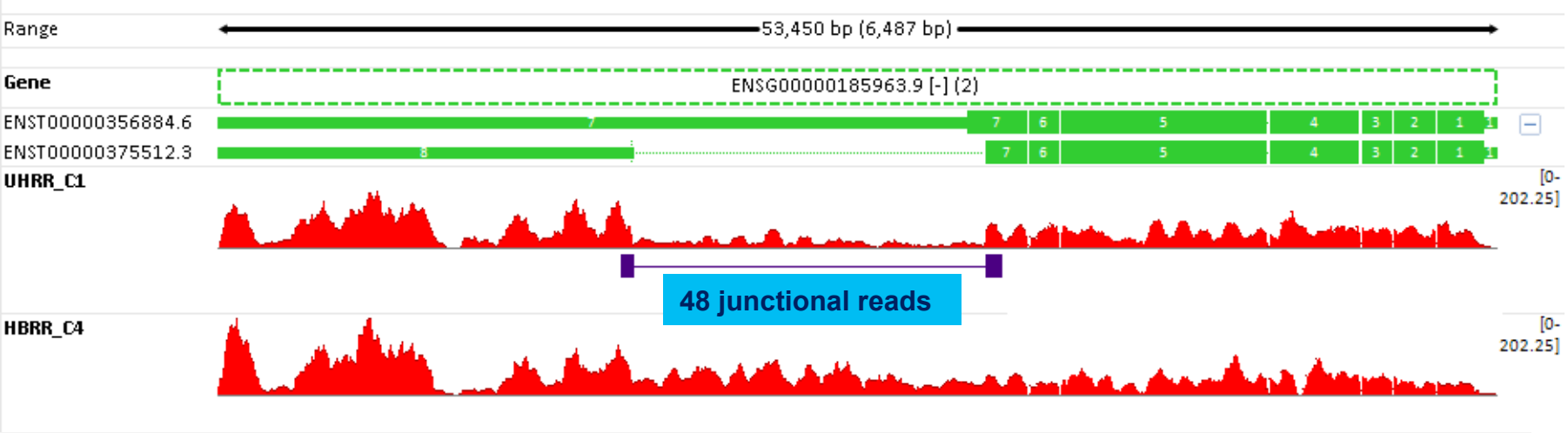



A
RPKM
12.5
10.0
7.5
5.0
2.5
0.0
HBRR_C4
UHRR_C1
Entity
RP11-6N17.4
RP11-6N17.4-001
RP11-6N17.4-002
RP11-6N17.4-003
Cytoband
Homo sapiens 37.3 17 [45,968,621-45,973,178]
p13
p12
p11.1
q12
q21.1
q22
q23
q25
Scale
45,969,500 bp
45,971 kbp
45,972 kbp
Range
4,558 bp
Gene
RP11-6N17.4 [-] (3)
RP11-6N17.4-001
RP11-6N17.4-002
RP11-6N17.4-003
UHRR_C1
HBRR_C4


B
RPKM
12.5
10.0
7.5
5.0
2.5
0.0
HBRR_C4
UHRR_C1
Entity
HAMP
HAMP-001
HAMP-002
HAMP-003
Cytoband
Homo sapiens 37.3 19 [35,771,619-35,776,046]
p13.3
p13.2
p13.1
p12
p11
q13.1
q13.2
q13.4
Scale
35,772,500 bp
35,774 kbp
35,775 kbp
Range
4,428 bp
Gene
HAMP [+] (3)
HAMP-001
HAMP-002
HAMP-003
UHRR_C1
HBRR_C4

Isoform changes masked by gene-level differential analysis

Type	Ensembl ID	log2Ratio	FDR	HBRR_C4	UHRR_C1
gene	ENSG00000185963.9	-0.434	1.8896E-05	4438	3496
Transcript	ENST00000375512.3 (**short**)	**1.509**	0.0088750	**643**	2170
Transcript	ENST00000356884.6 (**long**)	**-1.685**	0.0106644	**3795**	1326



Two issues:

1. This gene has two isoforms, In **HBRR**, only the **long** isoform is expressed. However, **14%** (643/(643+3795) of reads are counted towards short isoform by RSEM(**?**). In UHRR, the short isoform is also expressed evidenced by those 48 junctional reads
2. The large difference in the short isoform between **HBRR** and **UHRR** IS **masked** if we perform differential analysis only at the gene level.

PLOS ONE

RESEARCH ARTICLE

Union Exon Based Approach for RNA-Seq Gene Quantification: To Be or Not to Be?

Shanrong Zhao*, Li Xi, Baohong Zhang

Clinical Genetics and Bioinformatics, Pfizer Worldwide Research & Development, Cambridge, Massachusetts, 02139, United States of America

Zhao *et al. BMC Genomics* (2015) 16:675
DOI 10.1186/s12864-015-1876-7

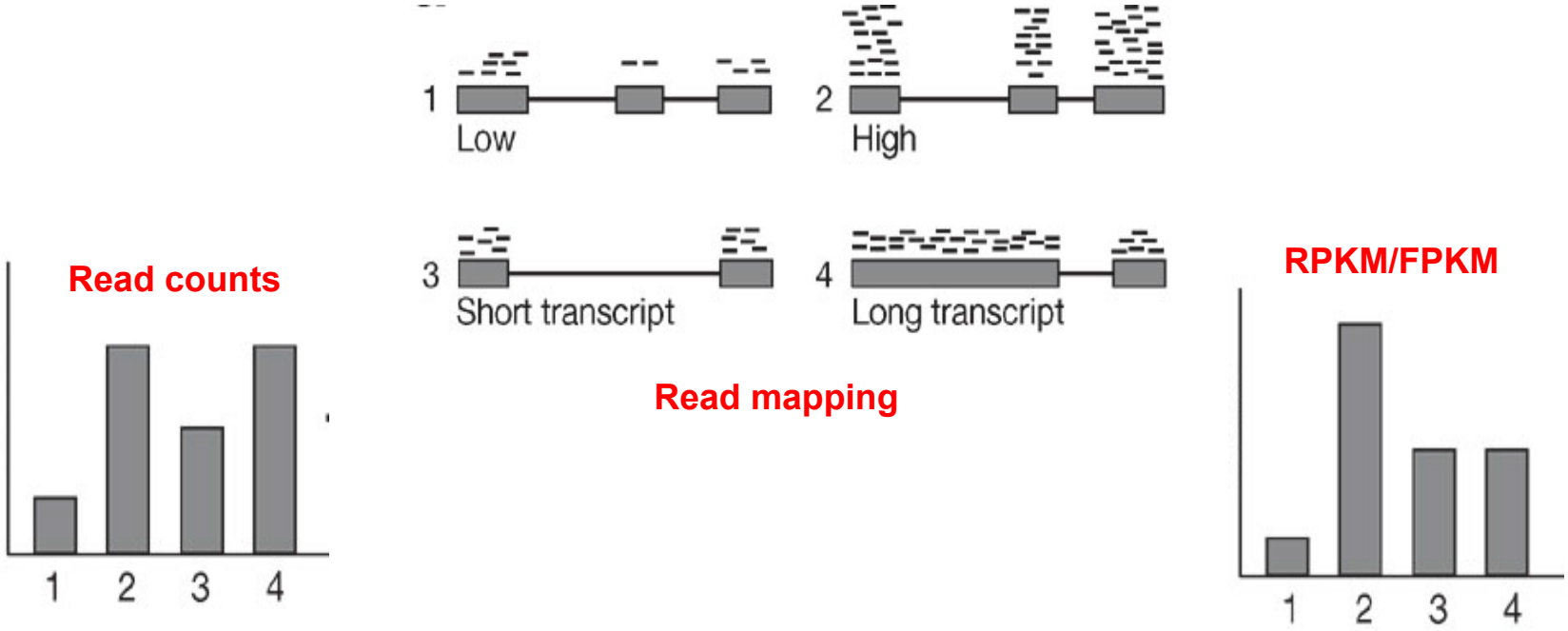
RESEARCH ARTICLE Open Access

Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap

Shanrong Zhao[1*], Ying Zhang[2], William Gordon[2], Jie Quan[3], Hualin Xi[3], Sarah Du[2], David von Schack[2*] and Baohong Zhang[1*]

B#3: RPKM (TPM, CPM): do we use them appropriately, or mis-use or abuse them?

RPKM to measure gene expression level

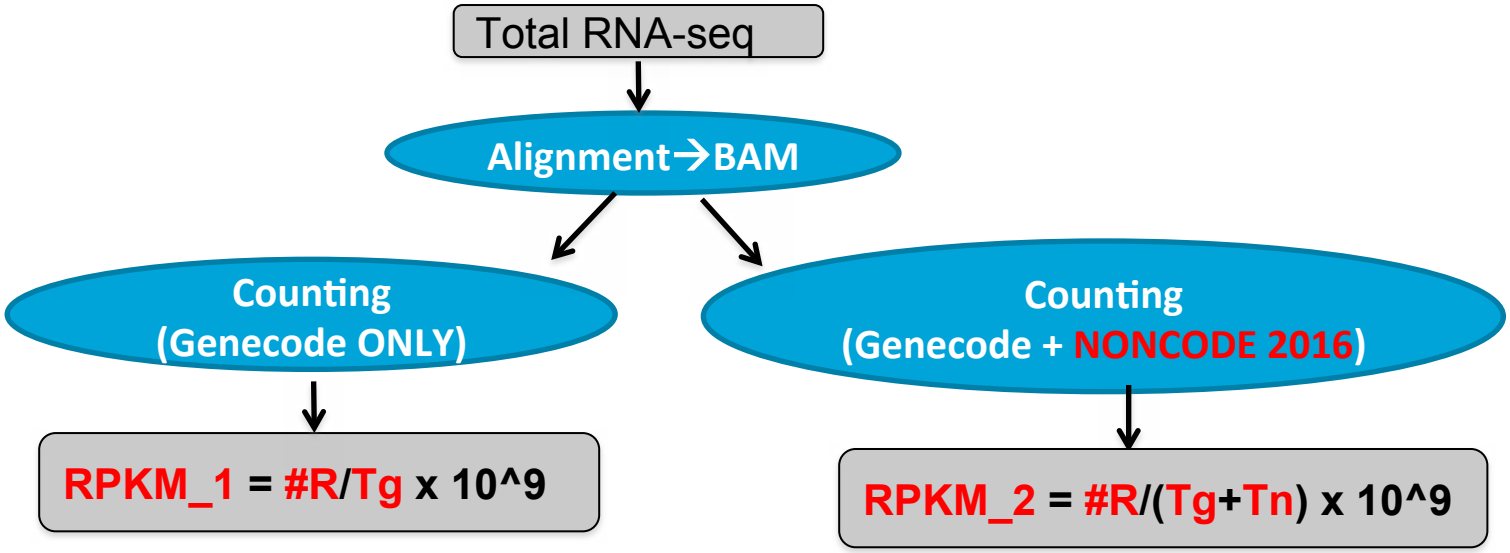


FPKM/RPKM (Reads per **K**ilobase of exon per **M**illion reads **M**apped) (**right**).

$$RPKM = 10^9 \times \frac{\#Reads}{Total\ Reads \times Gene\_Length}$$

In practice, RPKM for a given gene is not only dependent upon the number of reads mapped to that gene, but also the **total number of reads**.

RPKM for a given gene is dependent upon the total number of reads (1)



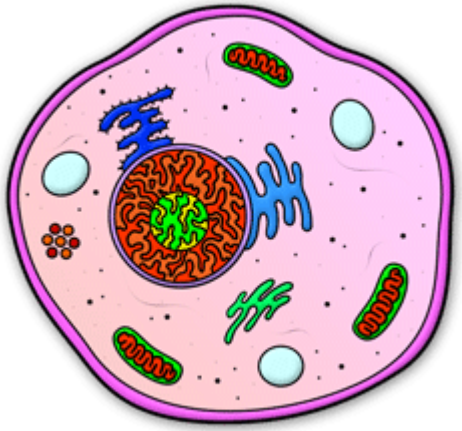
Note:

- Additional genes from NONCODE 2016 do not overlap with any gene in Genecode
- **#R:** the number of remad mapped to a gene; **Tg:** the total number of reads mapped to Gencode; **Tn:** the total number of reads mapped to NONCODE 2016
- For any gene, **RPKM_1 < RPKM_2**, ALWAYS.

It's important to interpret RPKM with respect to the gene set used in RPKM calculation

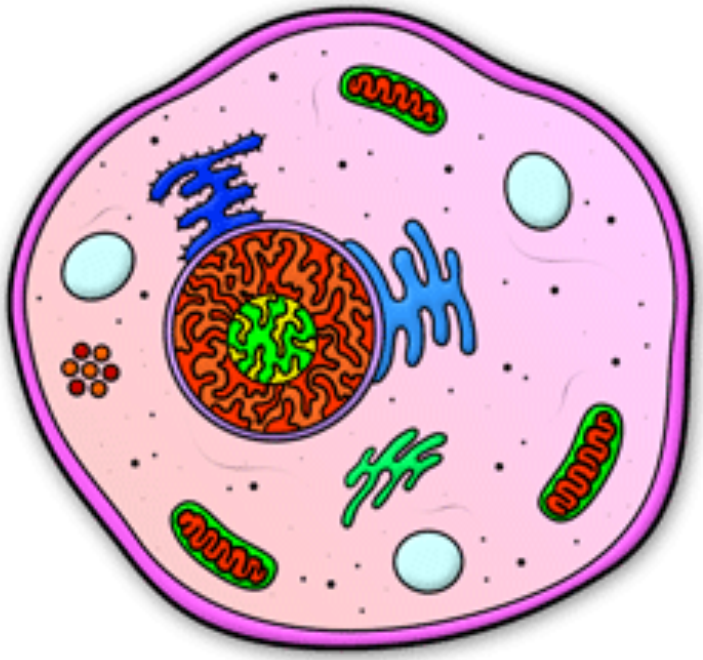
Comparison of RPKM across samples (2)

- RPKM represents the **relative** expression of genes in a sample(cell)
- Across-sample comparison makes a sense **if and only if** **(1)** the **total amount of mRNA** (mass? or mRNA copies?) are roughly the SAME; and (2) the compositions of mRNAs are comparable.



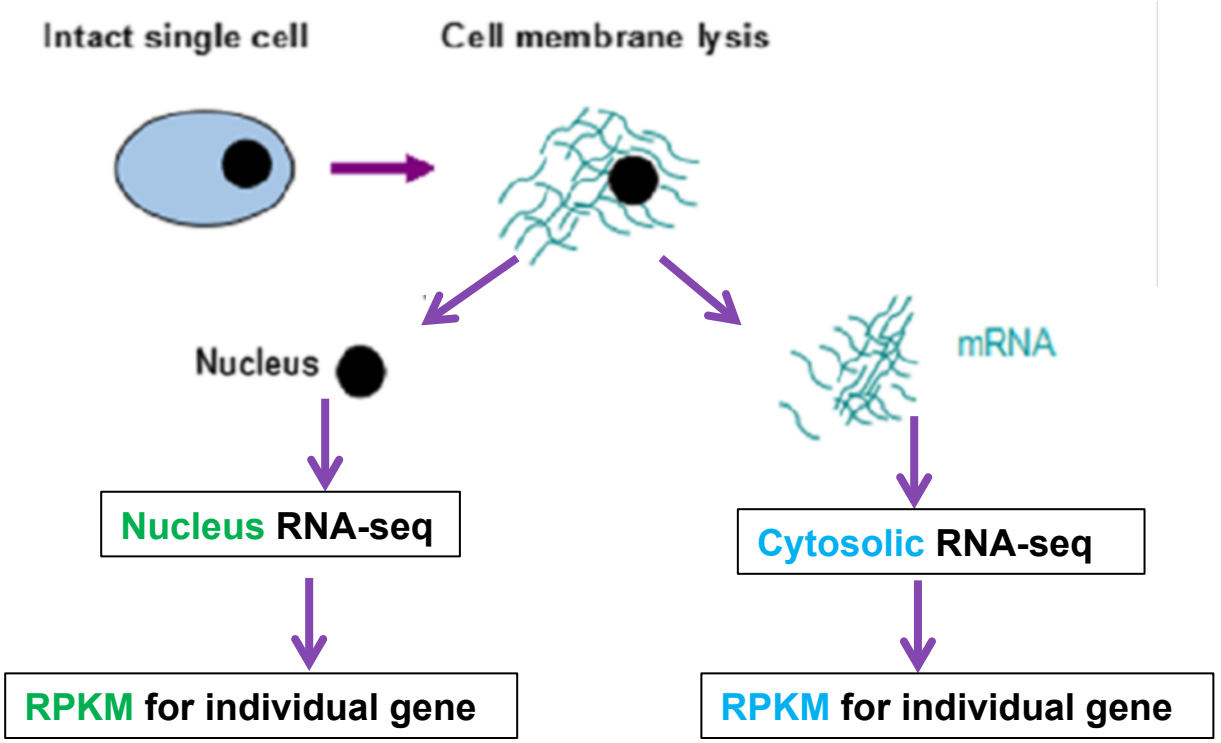
A

- Let's say the cell at the right (B) is **twice** as large as the cell on the left (A).
- Assume the number of mRNA copy in (B) is also twice that in A
- All genes will have the **SAME RPKMs** between A and B
- Always keep in mind RPKM **CANNOT tell the absolute** mRNA copy numbers in a sample(cell)



B

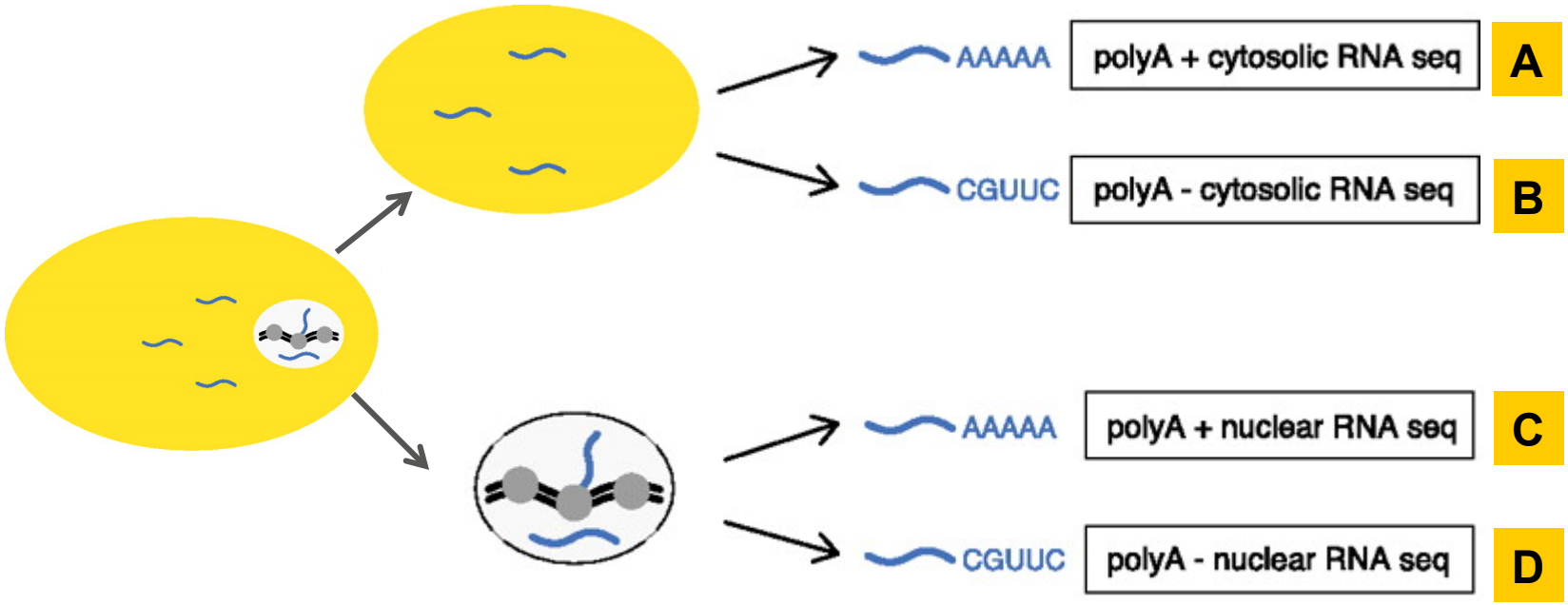
Comparison of RPKM across samples (3): nucleus versus cytosol



Q: Can we compare RPKM (in nucleus) with RPKM (in cytosol)?

Across-sample comparison is **problematic** when the composition of mRNA in the samples differ significantly

Comparison of RPKM across samples (4): polyA+ versus polyA- RNA-seq



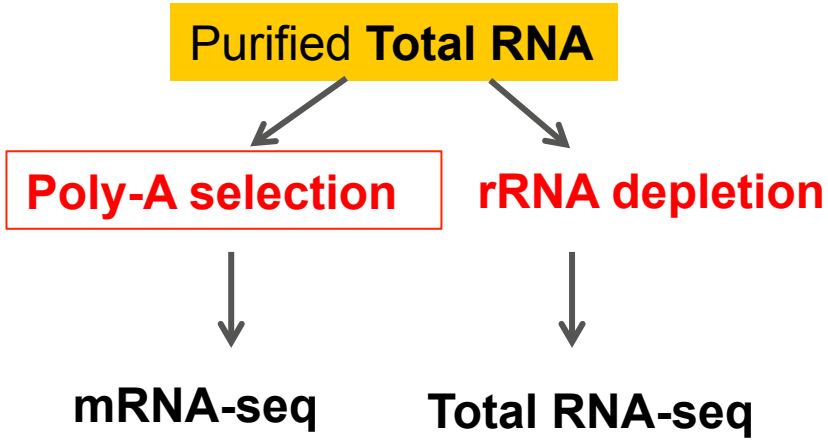
Research

Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs

Hagen Tilgner,[1,3] David G. Knowles,[1] Rory Johnson,[1] Carrie A. Davis,[2] Sudipto Chakrabortty,[2] Sarah Djebali,[1] João Curado,[1] Michael Snyder,[3] Thomas R. Gingeras,[2] and Roderic Guigó[1,4]

http://genome.cshlp.org/content/22/9/1616.full.pdf+html

Comparison of RPKM across samples (5): total RNA-seq versus mRNA-seq



Q: Is it fair to compare a gene's RPKM in total RNA-seq with its counterpart in mRNA-seq?

RPKM vs CPM or TPM

- **CPM** is similar to RPKM except for the number of reads NOT normalized by **the length of gene**

- **TPM** (transcripts per million) is an accurate measure of **relative molar RNA concentration.** TPM is very similar to RPKM and FPKM. The only difference is the order of operations. Here's how you calculate TPM:

    1. Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).

    2. Count up all the RPK values in a sample and divide this number by 1,000,000. This is your "per million" scaling factor.

    3. Divide the RPK values by the "per million" scaling factor. This gives you TPM.

- When you use TPM, the sum of all TPMs in each sample are the same. This makes it easier to compare the proportion of reads that mapped to a gene in each sample. In contrast, with RPKM and FPKM, the sum of the normalized reads in each sample may be different, and this makes it harder to compare samples directly.

Additional comments on RPKM

1. RPKM can represent the **relative abundance** of genes in a sample, but is not a 'true' measure of **relative molar RNA concentration**

2. Be cautious when comparing RPKMs across
   - **Gene set**
   - **Cellular fraction**
   - **Sequencing protocols**
   - **Biological conditions**

3. RPKM is often ***mis-used*** **unintentionally.**

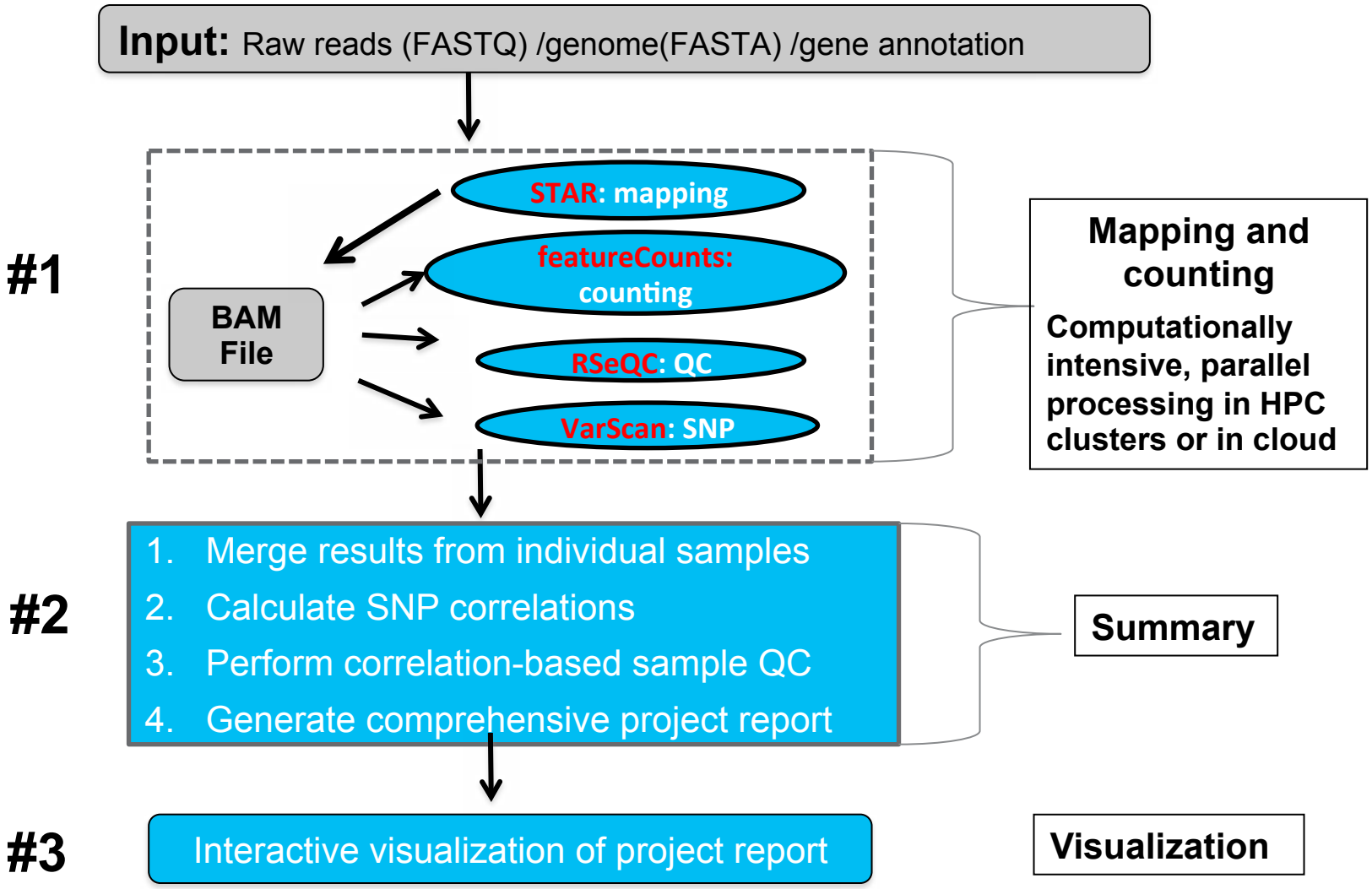
Key takeaways from Part #B

- Gene/isoform quantification is heavily dependent upon your choice of a **gene model**
- **Union-exon based approach** for gene quantification is simpler, but more insights are gained from isoform quantification
- More accurate quantification is obtained from **stranded RNA-seq**, and **"read ambiguity"** (multiple-mapping and gene overlapping) is responsible for almost all kinds of complexities in quantification.
- Comparison of RPKMs across samples should be done **with caution**.

Part C
Pipeline and Visualization

Baohong Zhang
Pfizer

Overview of QuickRNASeq Pipeline



2015's Most Influential Articles from BMC Genomics

BioMed Central
The Open Access Publisher

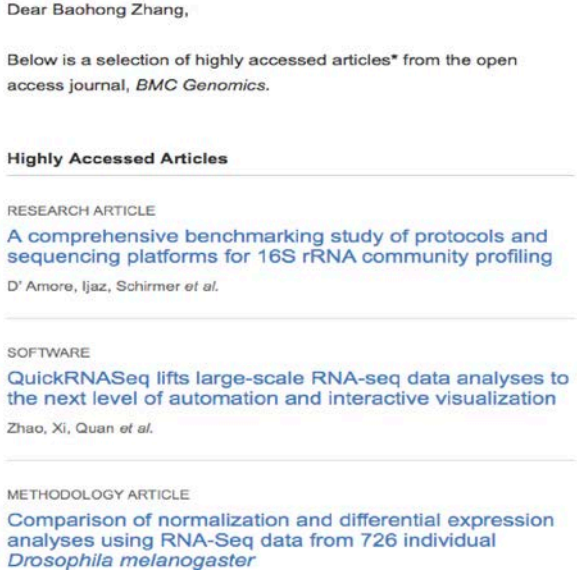
BMC Genomics

Dear Colleague,

We would like to share with you our most influential articles of 2015, according to Altmetric.com.

Influential Articles of 2015

- A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers
- Crowdsourced direct-to-consumer genomic analysis of a family quartet
- Gene expression during zombie ant biting behavior reflects the complexity underlying fungal parasitic behavioral manipulation
- QuickRNASeq lifts large-scale RNA-seq data analyses to the next level of automation and interactive visualization
- Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa
- Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community
- Analysis of expressed sequence tags from Actinidia: applications of a cross species EST database for gene discovery in the areas of flavor, health, color and ripening



Dear Baohong Zhang,

Below is a selection of highly accessed articles* from the open access journal, *BMC Genomics*.

Highly Accessed Articles

RESEARCH ARTICLE

A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling

D' Amore, Ijaz, Schirmer *et al.*

SOFTWARE

QuickRNASeq lifts large-scale RNA-seq data analyses to the next level of automation and interactive visualization

Zhao, Xi, Quan *et al.*

METHODOLOGY ARTICLE

Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*

Shanrong Zhao was invited to present a talk entitled "*QuickRNASeq lifts large-scale RNA-seq data analysis to next level of automation and visualization*" at ***Bio-IT World Conference and Expo*** in Boston on **April 5-7, 2016**.

Baohong Zhang, Shanrong Zhao are invited to hold workshop "**Optimizing your use of RNA-seq tech & data analysis – 101**" at Boston Festival of Genomics on June 27-29, 2016 and San Diego Festival of Genomics.

Sample Annotation File of QuickRNASeq

sample_id	subject_id	Tissue	Sex
SRR607214	GTEX-N7MS	Blood	M
SRR615261	GTEX-N7MS	Blood Vessel	M
SRR603068	GTEX-N7MS	Brain	M
SRR821282	GTEX-N7MS	Esophagus	M
SRR608096	GTEX-N7MS	Heart	M
SRR612839	GTEX-N7MS	Muscle	M
SRR816609	GTEX-N7MS	Pituitary	M
SRR821518	GTEX-N7MS	Testis	M
SRR607679	GTEX-N7MS	Thyroid	M
SRR809283	GTEX-N7MT	Blood	F
SRR808044	GTEX-N7MT	Blood Vessel	F
SRR598671	GTEX-N7MT	Brain	F
SRR598509	GTEX-N7MT	Heart	F
SRR600784	GTEX-N7MT	Lung	F
SRR813208	GTEX-N7MT	Pancreas	F
SRR821573	GTEX-N7MT	Pituitary	F
SRR810945	GTEX-NFK9	Blood	M
SRR811819	GTEX-NFK9	Blood Vessel	M
SRR820689	GTEX-NFK9	Esophagus	M
SRR602106	GTEX-NFK9	Heart	M
SRR607166	GTEX-NFK9	Lung	M
SRR598044	GTEX-NFK9	Muscle	M
SRR614287	GTEX-NFK9	Nerve	M
SRR811029	GTEX-NFK9	Pancreas	M
SRR815280	GTEX-NFK9	Prostate	M
SRR820839	GTEX-NFK9	Testis	M
SRR603834	GTEX-NFK9	Thyroid	M

sample_id	subject_id	Tissue	Sex
SRR808836	GTEX-NPJ8	Blood Vessel	M
SRR598124	GTEX-NPJ8	Brain	M
SRR817306	GTEX-NPJ8	Esophagus	M
SRR598148	GTEX-NPJ8	Heart	M
SRR603750	GTEX-NPJ8	Lung	M
SRR601695	GTEX-NPJ8	Muscle	M
SRR615790	GTEX-NPJ8	Nerve	M
SRR819771	GTEX-NPJ8	Pancreas	M
SRR807949	GTEX-NPJ8	Pituitary	M
SRR820234	GTEX-NPJ8	Prostate	M
SRR810899	GTEX-NPJ8	Testis	M
SRR602951	GTEX-NPJ8	Thyroid	M
SRR815494	GTEX-O5YT	Blood	M
SRR809785	GTEX-O5YT	Blood Vessel	M
SRR814003	GTEX-O5YT	Esophagus	M
SRR820316	GTEX-O5YT	Heart	M
SRR821525	GTEX-O5YT	Lung	M
SRR815044	GTEX-O5YT	Muscle	M
SRR812080	GTEX-O5YT	Nerve	M
SRR810761	GTEX-O5YT	Pancreas	M
SRR818850	GTEX-O5YT	Testis	M

- **48 samples** from **5 subjects**. All are downloaded from Genotype-Tissue Expression (GTEx) project
- The samples are from different tissues

Commands for Primary RNAseq Data Analysis

#ENVIRONMENT

export QuickRNASeq=/hpc/grid/shared/ngsapp/QuickRNASeq

export PATH=$QuickRNASeq:$PATH

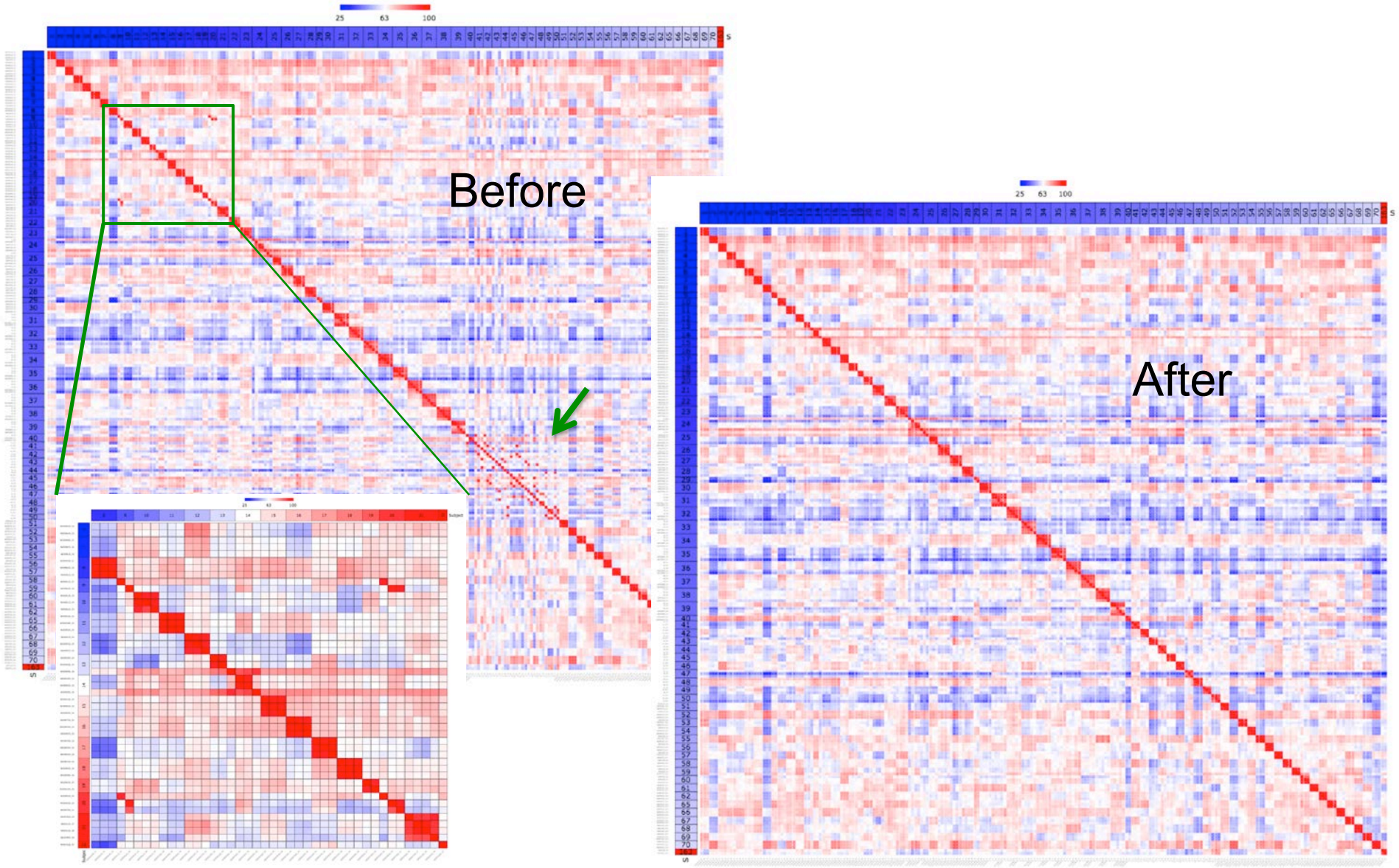
star-fc-qc.sh allIDs.txt run.config

#Summarization – after all jobs submitted by the first command finishes

export GENOME_ANNOTATION=/hpc/grid/shared/ngsdb/annotation/gencode/hg19.gencode.v19.annot

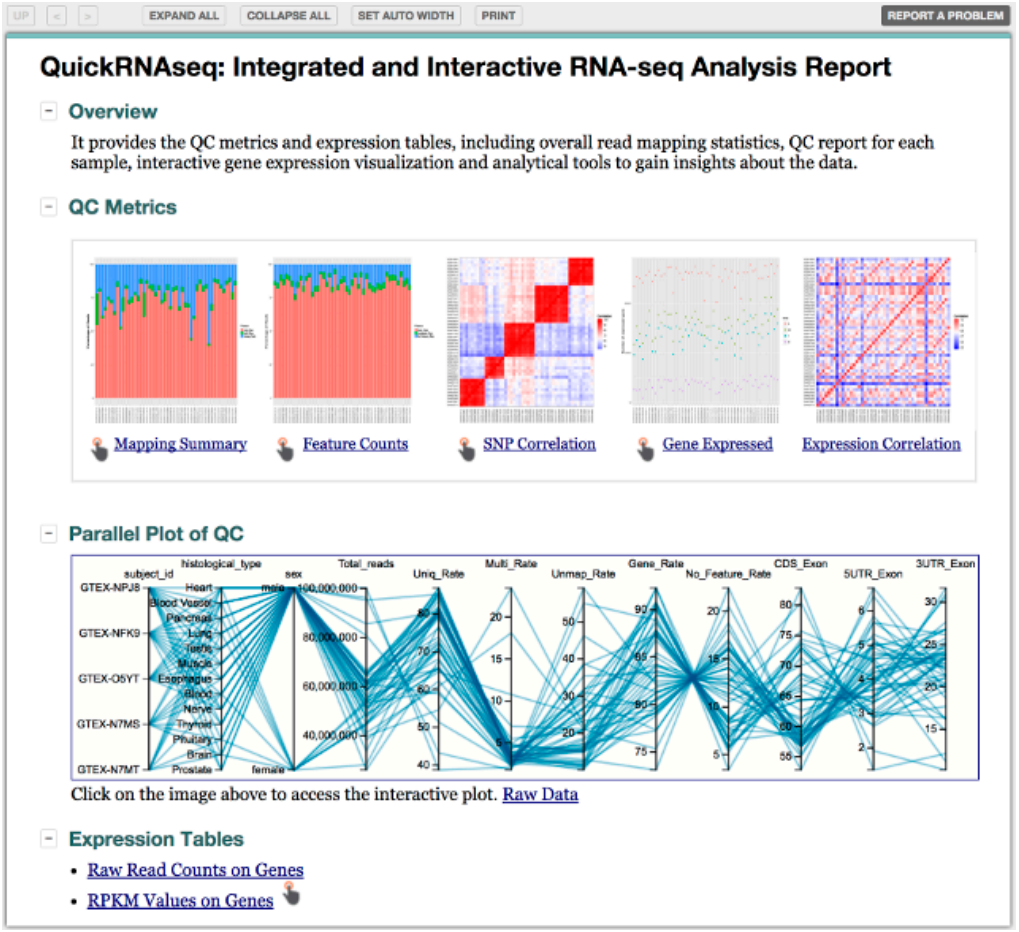
star-fc-qc.summary.sh allIDs.txt sample.annotation.txt

Identifying Sample Mix-up by SNP Concordance



Demo of QuickRNASeq Tool

http://baohongz.github.io/QuickRNASeq



User Guide: http://baohongz.github.io/QuickRNASeq/guide.html

Summary of QuickRNASeq

- QuickRNASeq was implemented by combing the best open source tool sets and the most advanced web 2.0 technologies.
- It significantly reduces the efforts involved in primary RNA-seq data analyses and generates an integrated project report.
- The dynamic visualization features enable end users to explore and digest RNA-seq data analyses results intuitively and interactively.
- The configuration file contains project, species, and software related parameters, and improves the reproducibility in RNA-seq data analyses.
- QuickRNASeq has been applied to in house large scale RNA-seq projects, and is mature for public deployment.
- No internet needed for exploring data on PC. The comprehensive report can be easily attached as supplementary material for publication.

QuickRNASeq resources

- QuickRNASeq website and user guide
  - http://quickrnaseq.sourceforge.net
  - http://baohongz.github.io/QuickRNASeq/guide.html
- Test run results
  - http://baohongz.github.io/QuickRNASeq/
- Publication
  - http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-2356-9
- Send your questions and comments to:
  - shanrong.zhao@pfizer.com
  - baohong.zhang@pfizer.com

Acknowledgements

Zhao *et al. BMC Genomics* (2016) 17:39
DOI 10.1186/s12864-015-2356-9

BMC Genomics

SOFTWARE Open Access

QuickRNASeq lifts large-scale RNA-seq data analyses to the next level of automation and interactive visualization

Shanrong Zhao[1*], Li Xi[1], Jie Quan[2], Hualin Xi[2], Ying Zhang[1], David von Schack[1], Michael Vincent[1] and Baohong Zhang[1*]

- Clinical Research: Karen Page, Mina Hassan-Zahraee
- HPC Team: Vassilios Pantazopoulos, Kirk Watrous