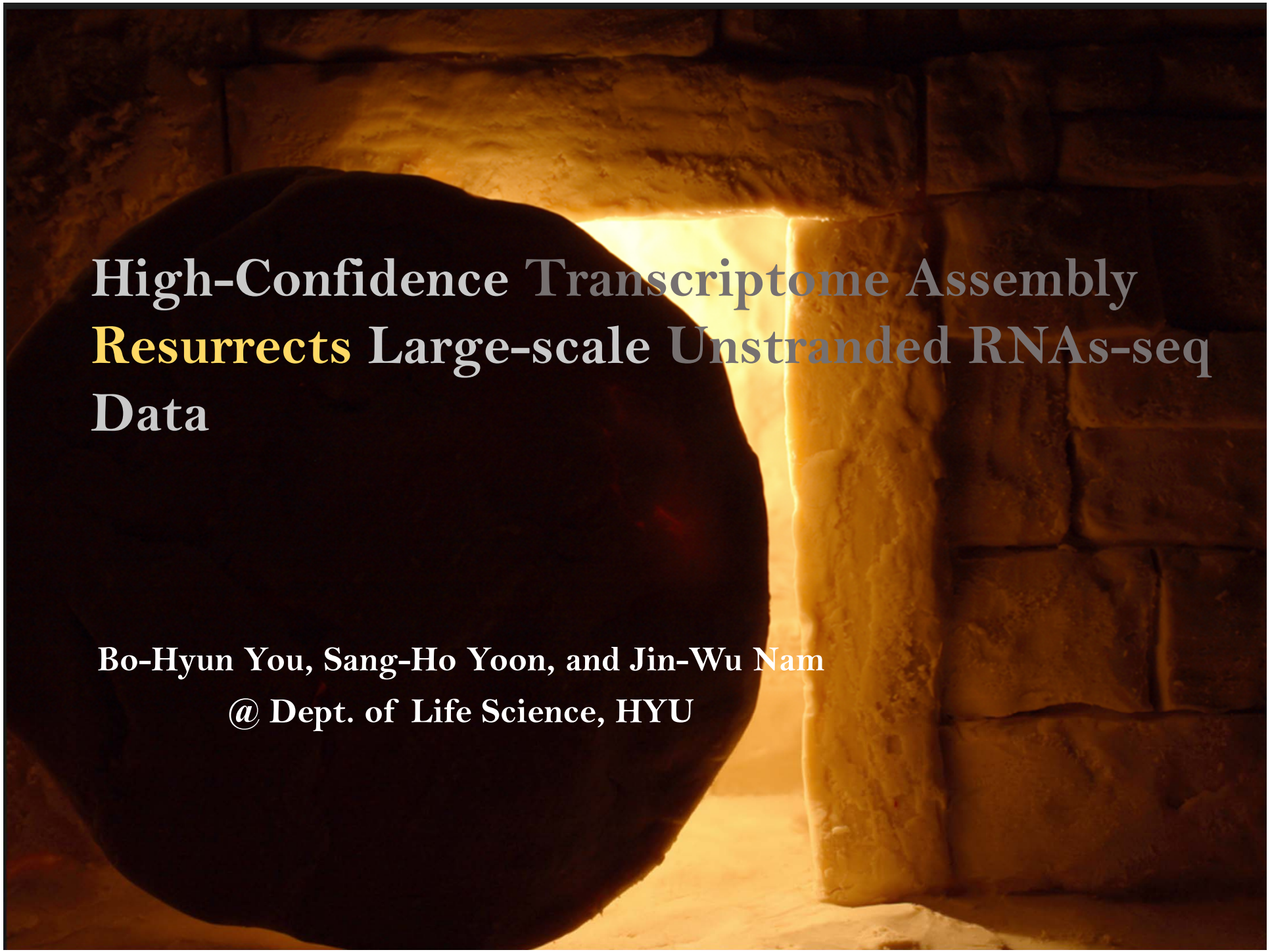
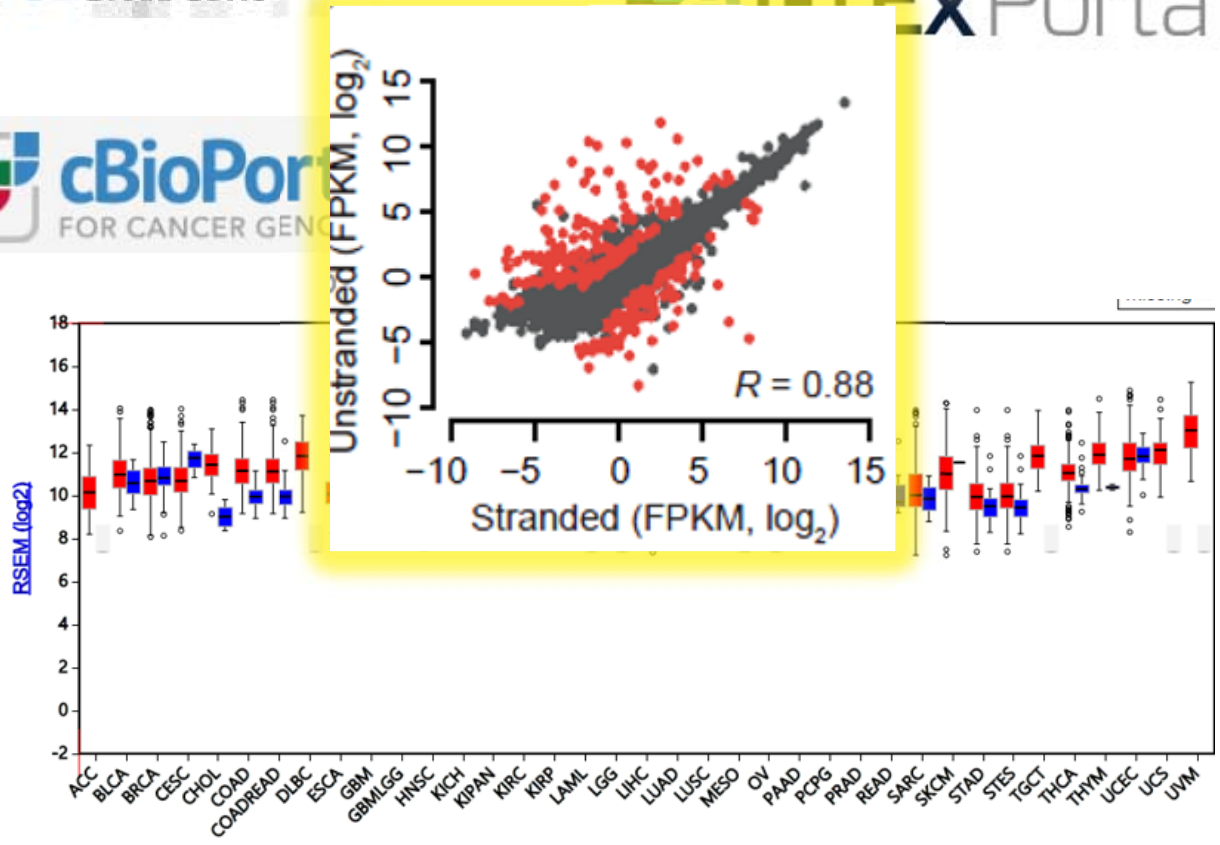

High-Confidence Transcriptome Assembly
Resurrects Large-scale Unstranded RNAs-seq
Data
Bo-Hyun You, Sang-Ho Yoon, and Jin-Wu Nam
@ Dept. of Life Science, HYU

Public data portal may include many errors due to analyzing unstranded RNA-seq data

FIREHOSE
Broad GDAC

GTEx Portal

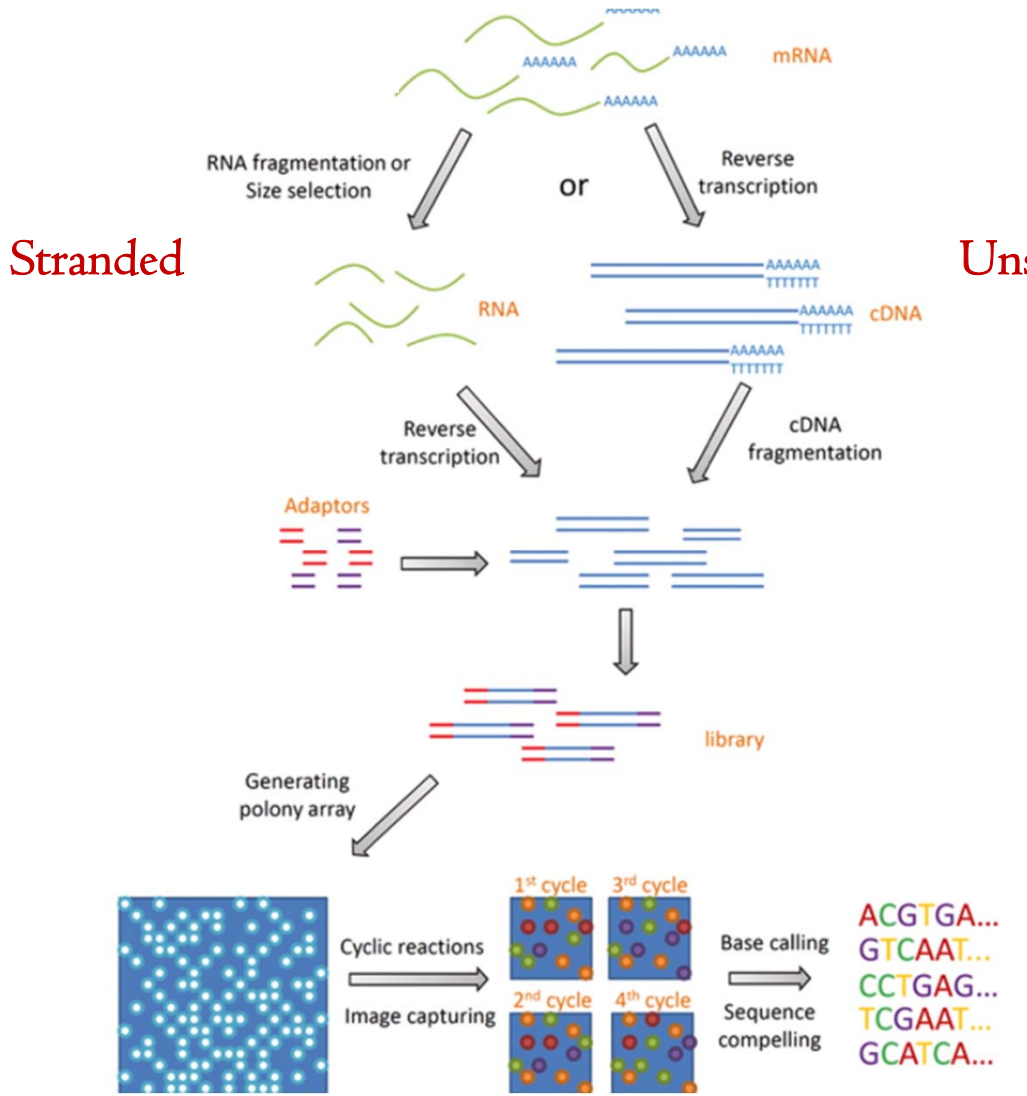
cBioPort
FOR CANCER GEN



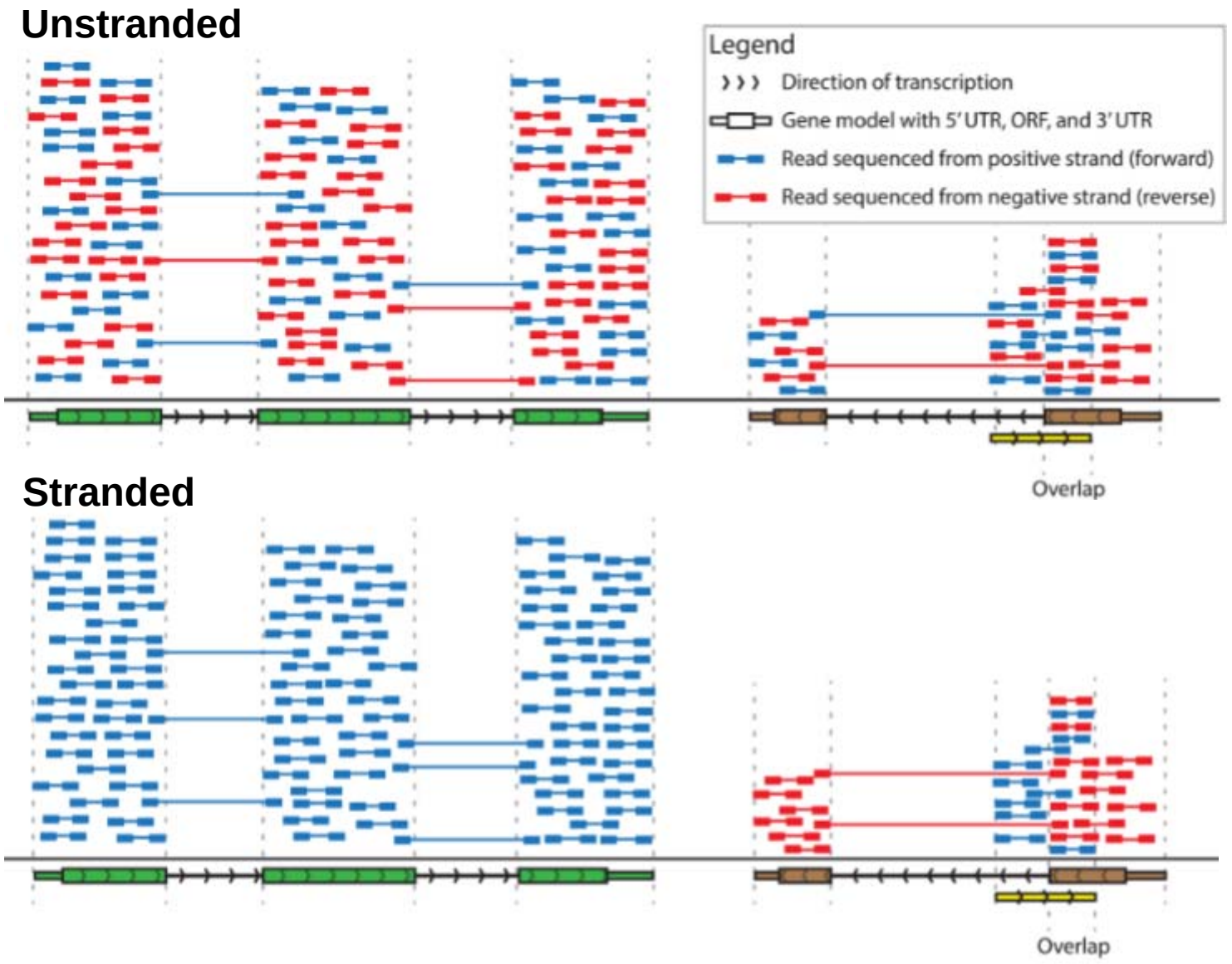
Unstranded vs stranded RNA-seq

Stranded

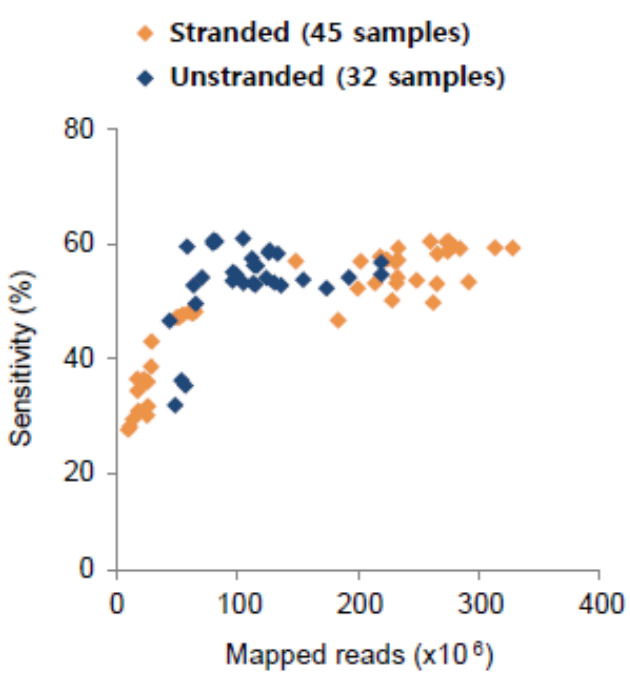
Unstranded

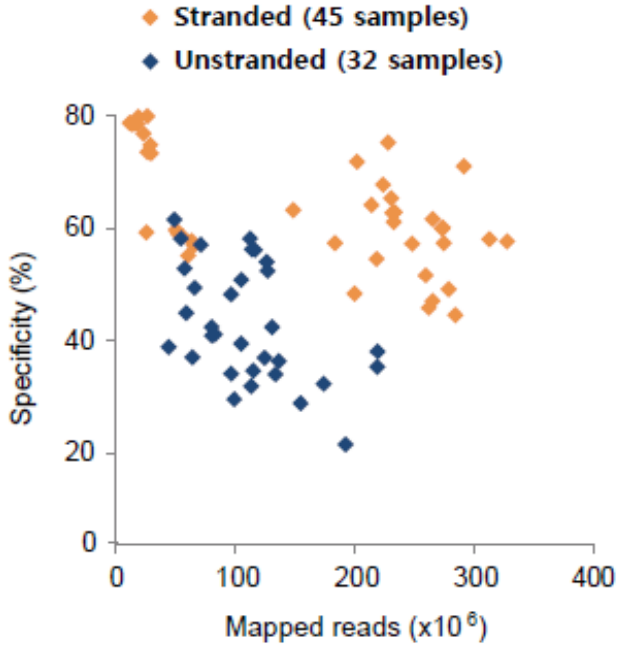


Unstranded RNA-seq reads



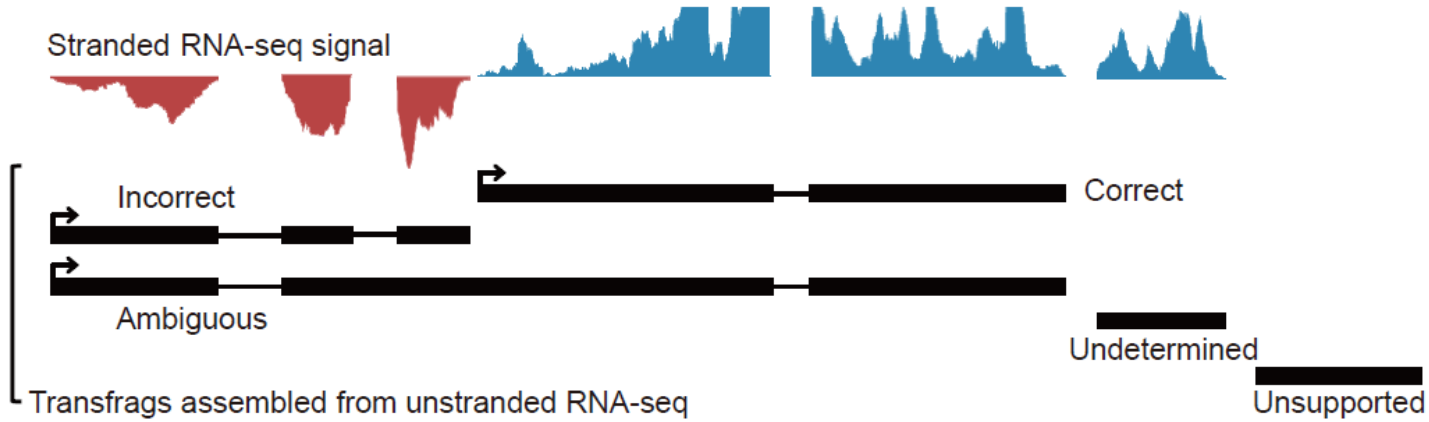
Problem 1. Unstranded RNA-seq causes error-prone assembly

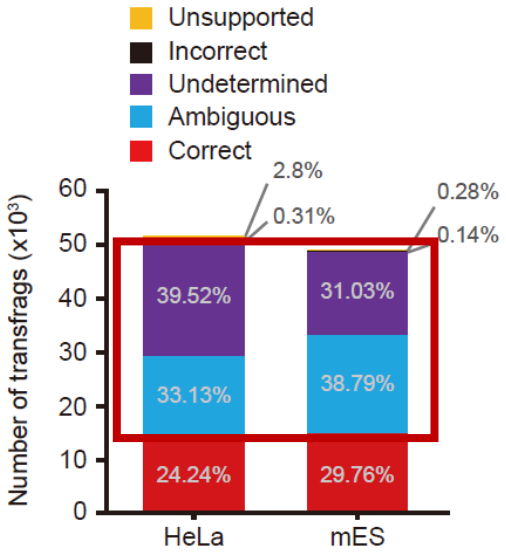




From ENCODE RNA-seq data

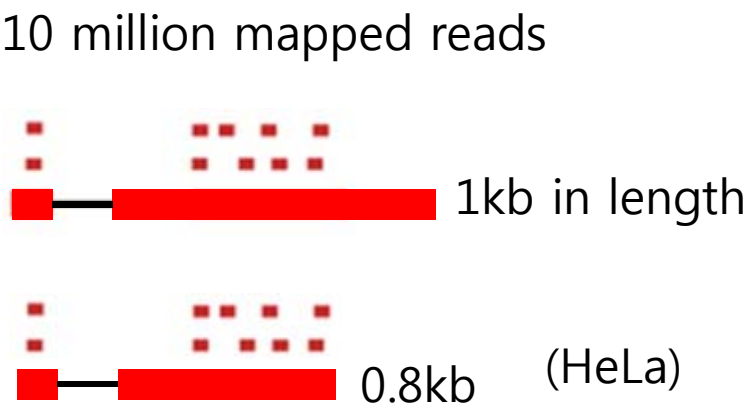
Problem 1. Unstranded RNA-seq causes error-prone assembly



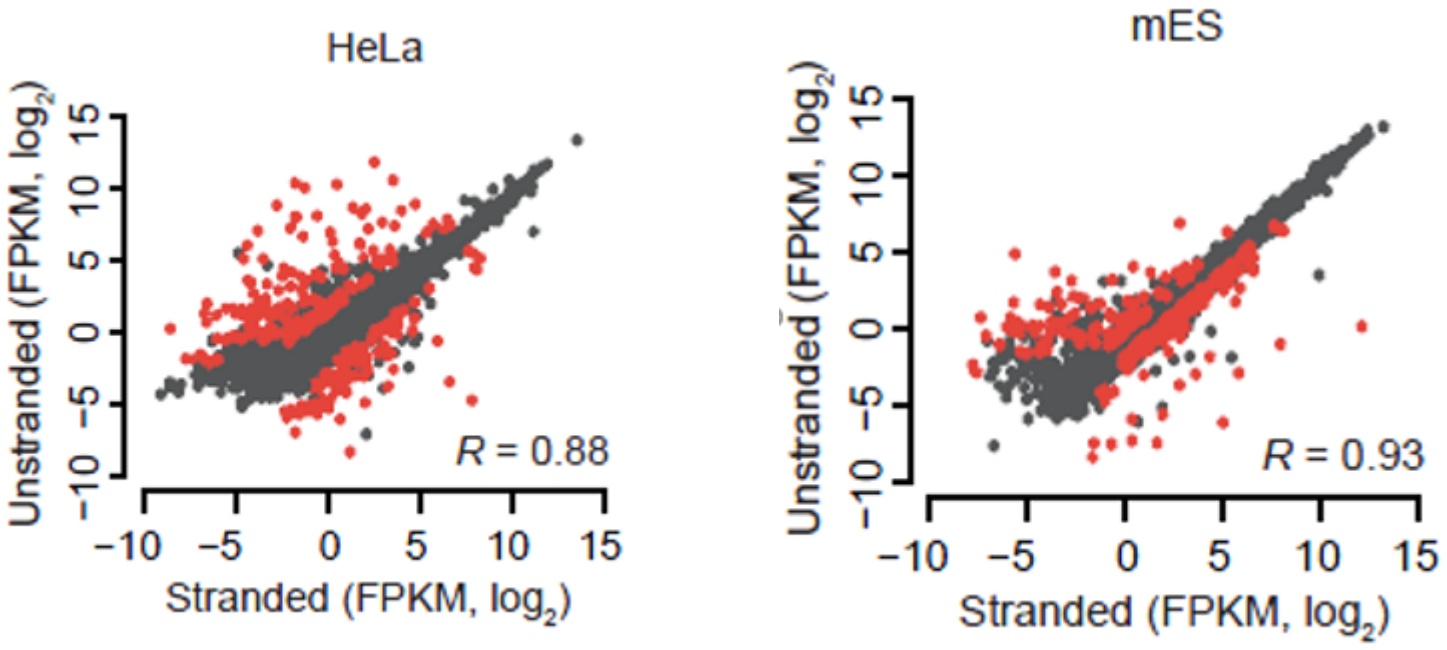


Problem 2. Mis-annotations causes quantification errors

- RPKM (~FPKM) = reads (fragment) per kilobases of exons per million mapped reads.
- 1 RPKM ~ 1 copy in a cell.



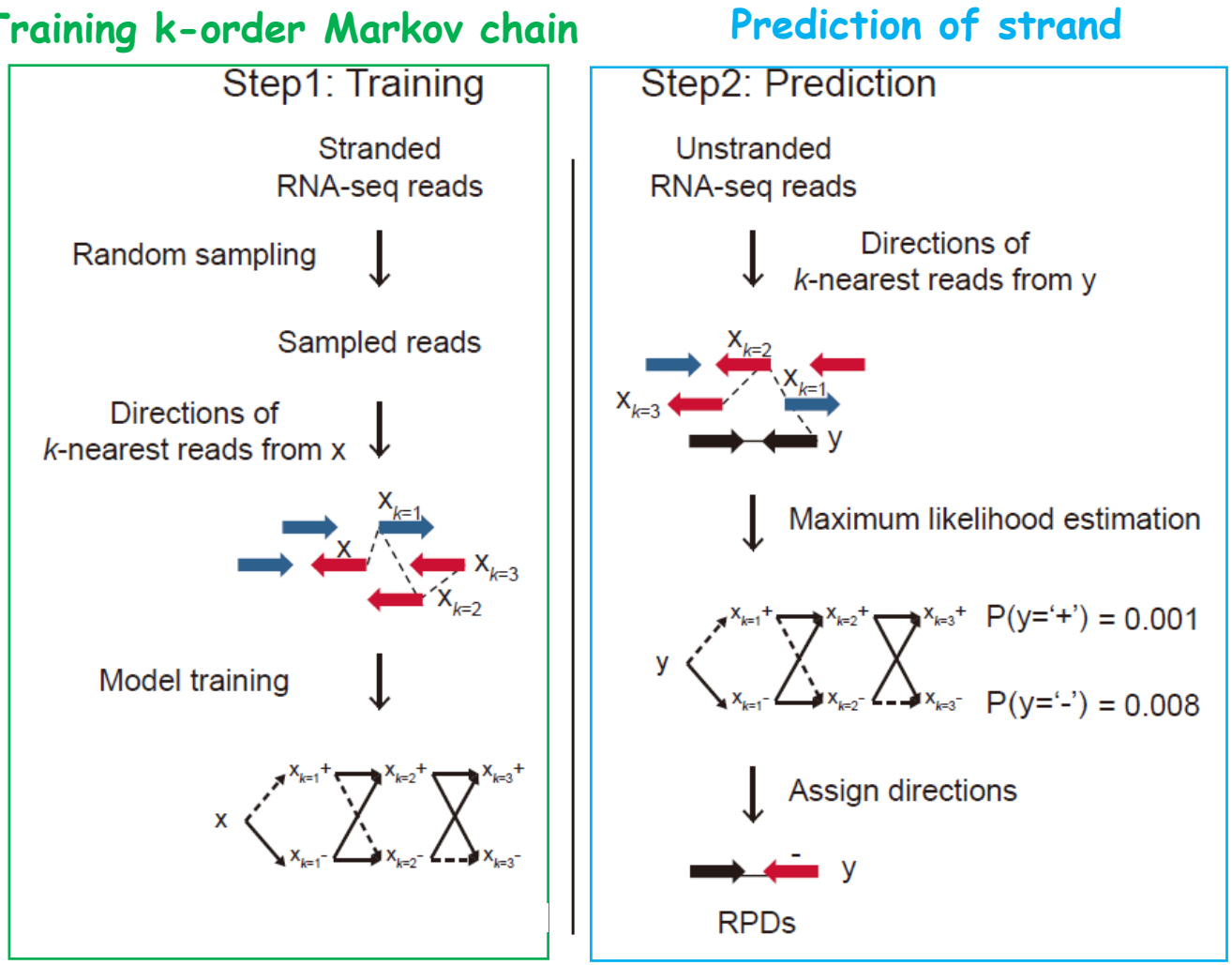
Problem 3. Unstranded RNA-seq causes quantification errors



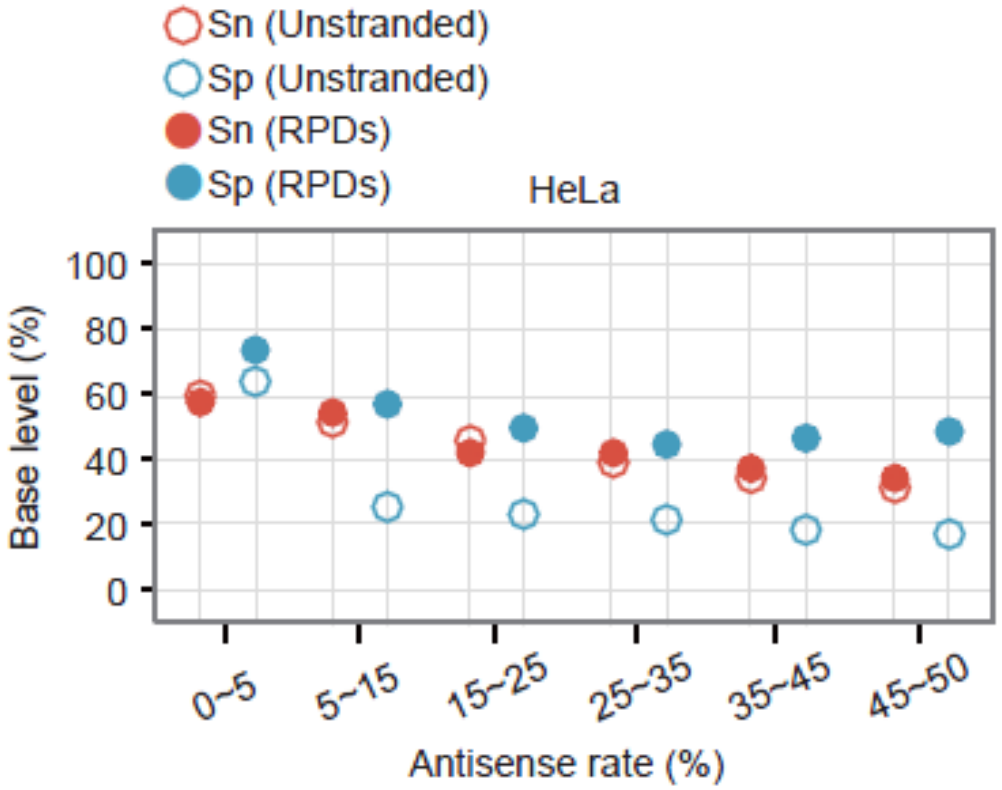
Gene annotation and personal transcriptome projects produced large-scale unstranded RNA-seq datasets

- ENCODE/GENCODE (Science 2012, Genome Res. 2012)
- modENCODE (Science 2010)
- Human Body Map project (Gene Dev. 2012)
- Zebrafish lncRNA annotation (Cell 2012)
- Worm lncRNA annotation (Genome Res. 2012)
- MiTranscriptome (Nat. Genetics 2015)
- GTEx (Nat Genet. 2013)
- TCGA (https://cancergenome.nih.gov/)
- ICGC (https://icgc.org/)
- The human protein atlas (http://www.proteinatlas.org/)

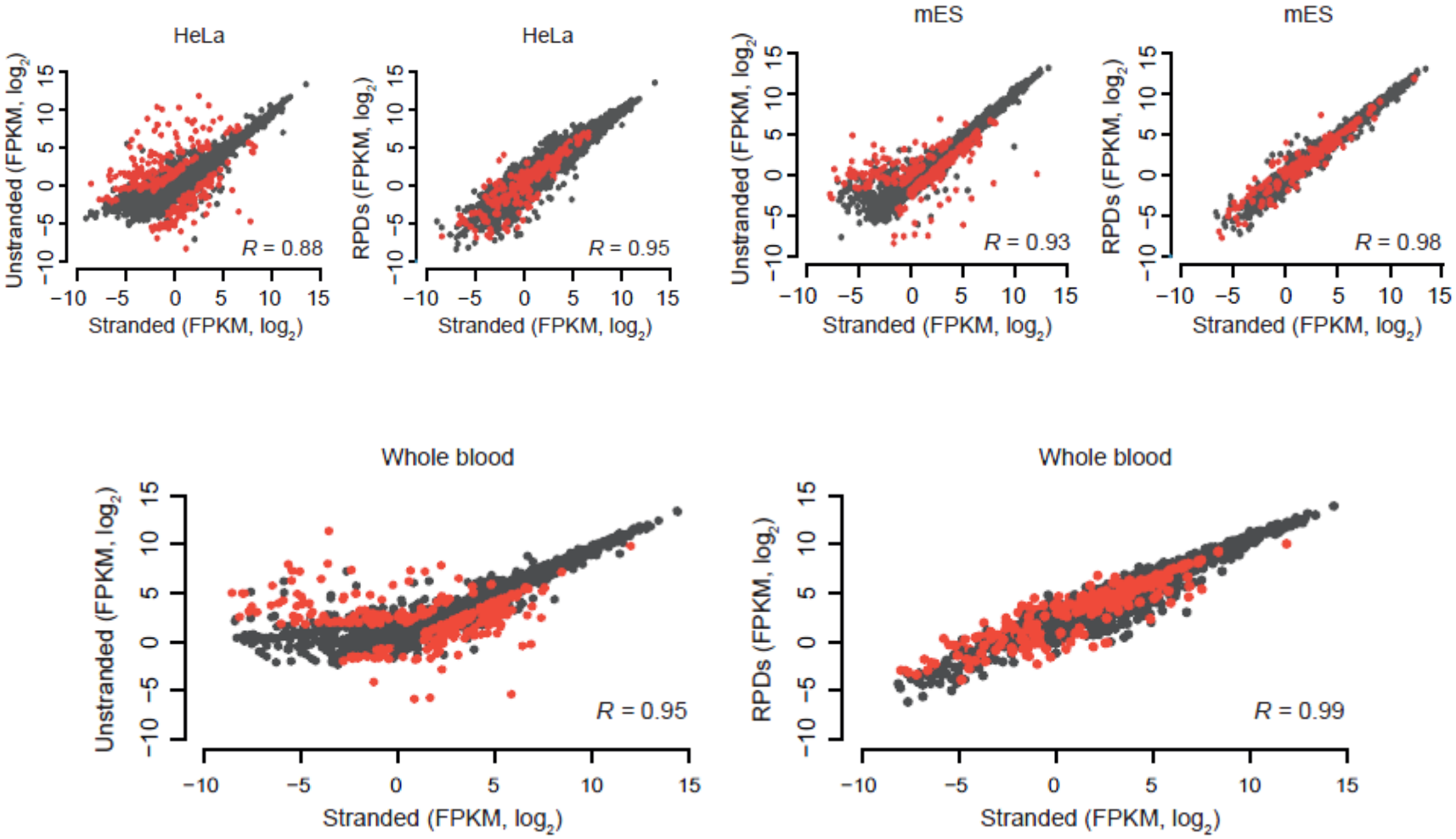
Probabilistic estimation of directions of unstranded reads



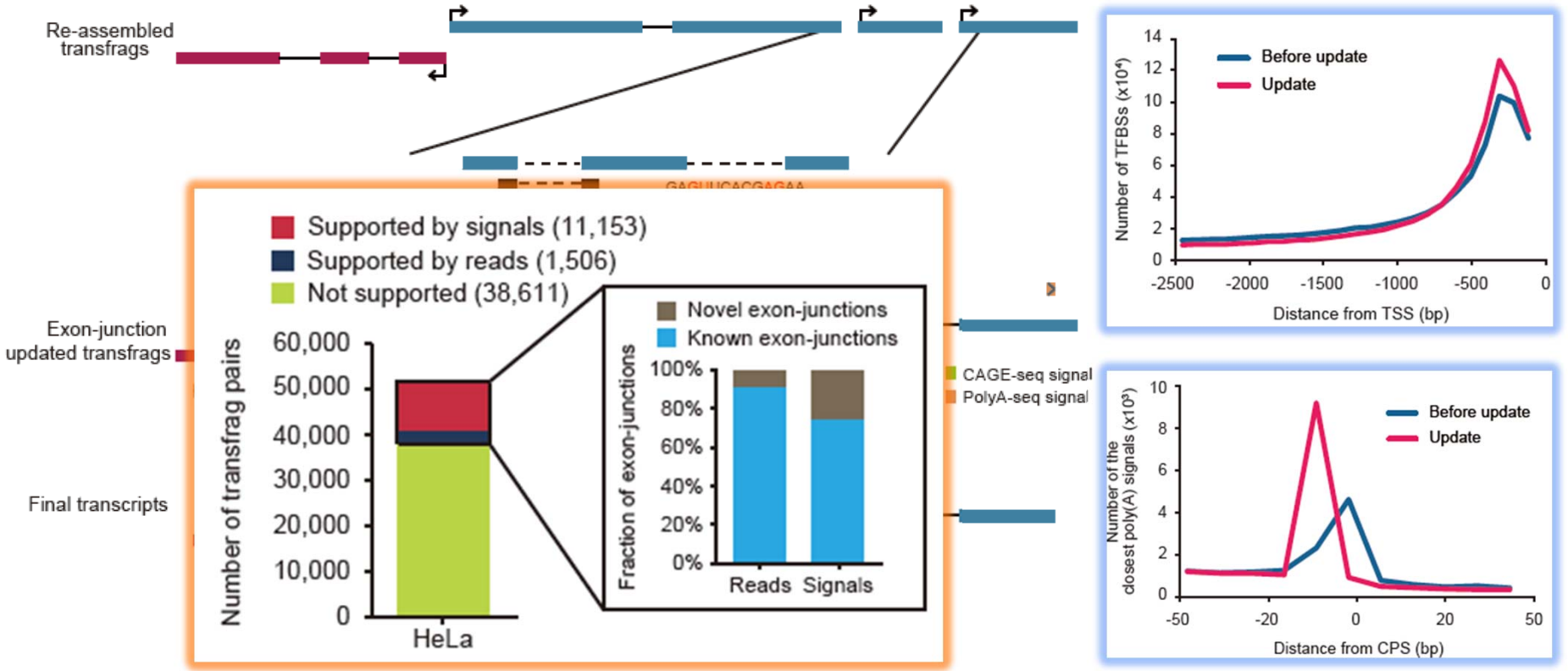
Antisense-overlapped reads benefited from strand prediction



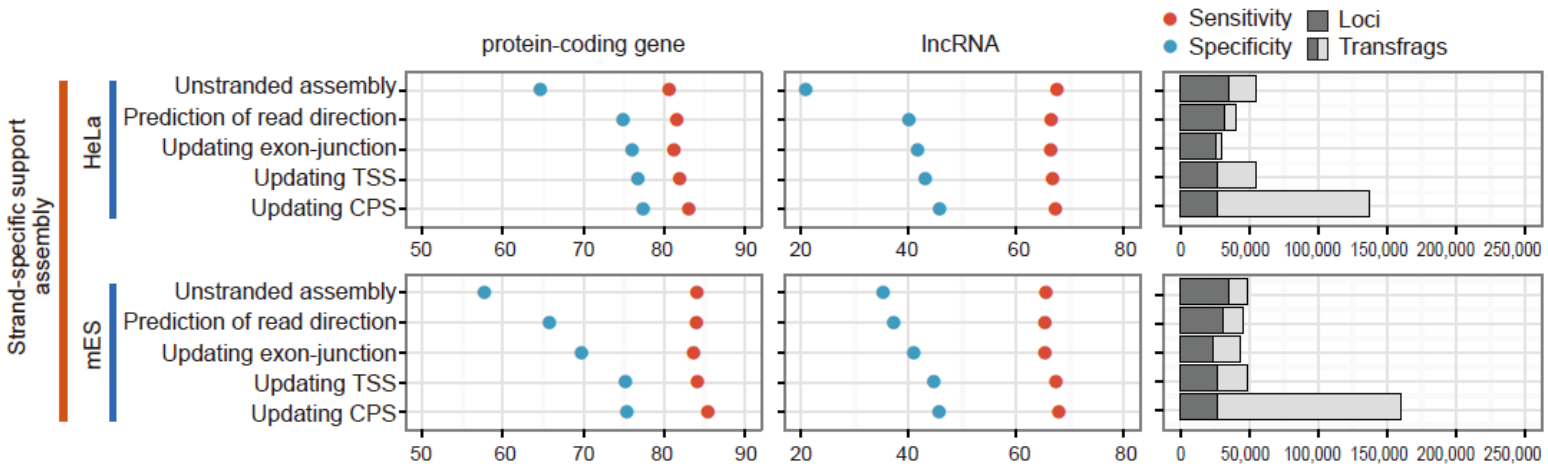
Strand prediction benefits the quantification



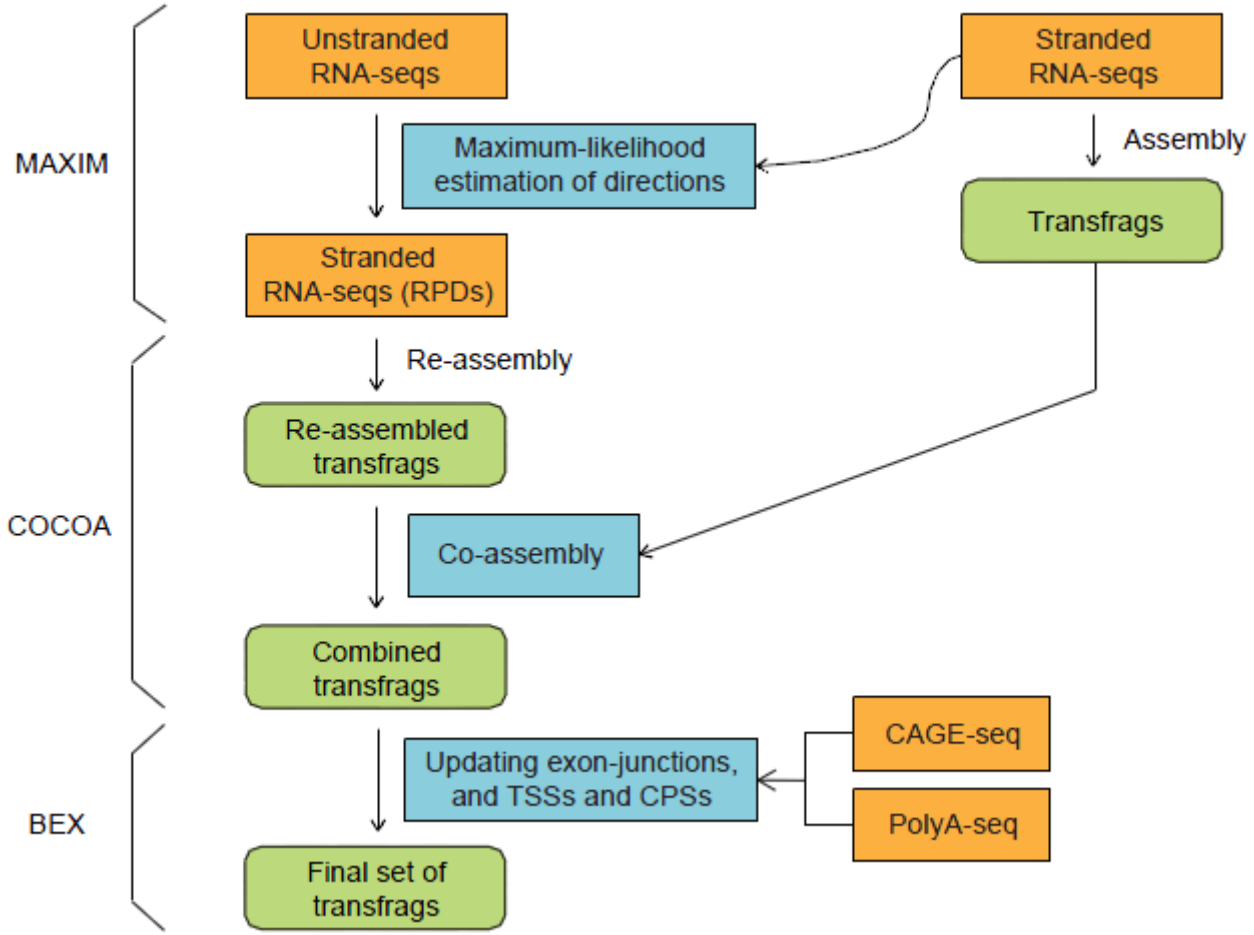
Exon-junction and boundary updates of transfrags



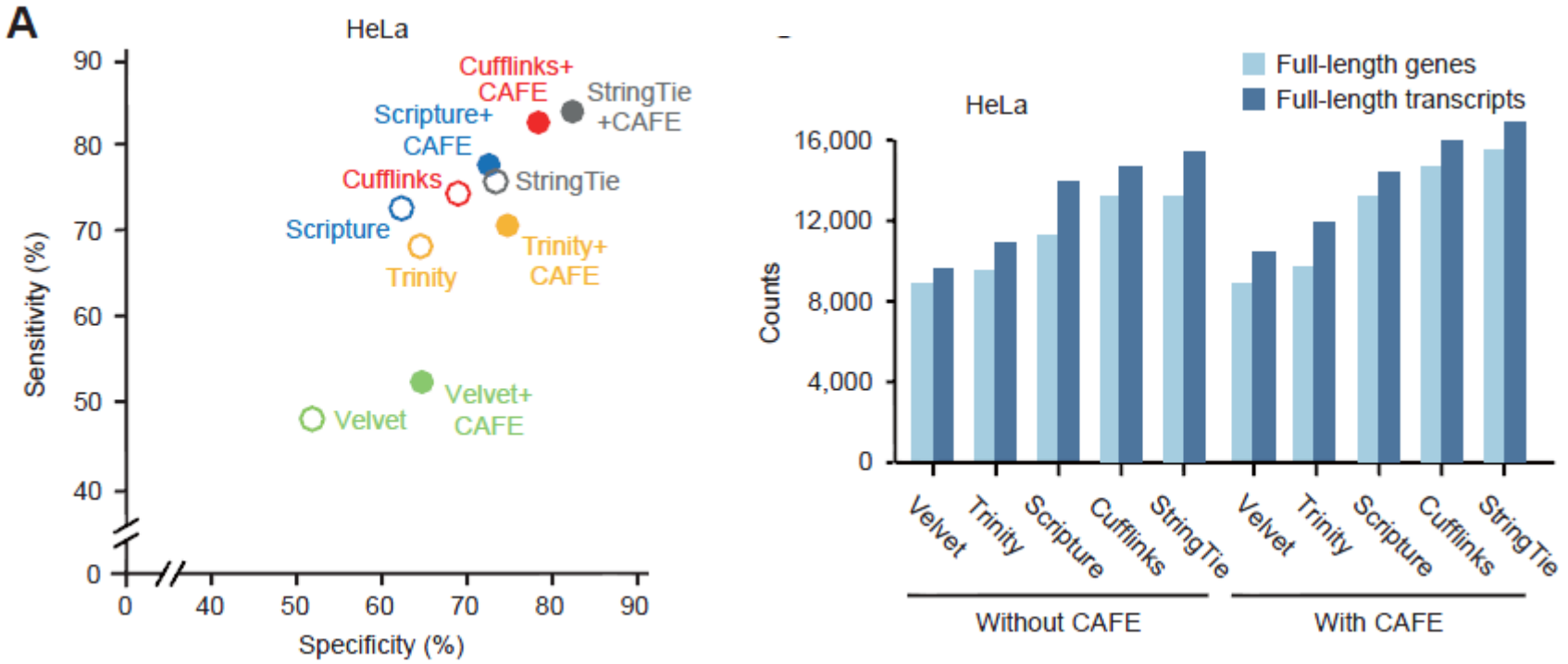
Co-assembly improves the quality of transcriptome maps



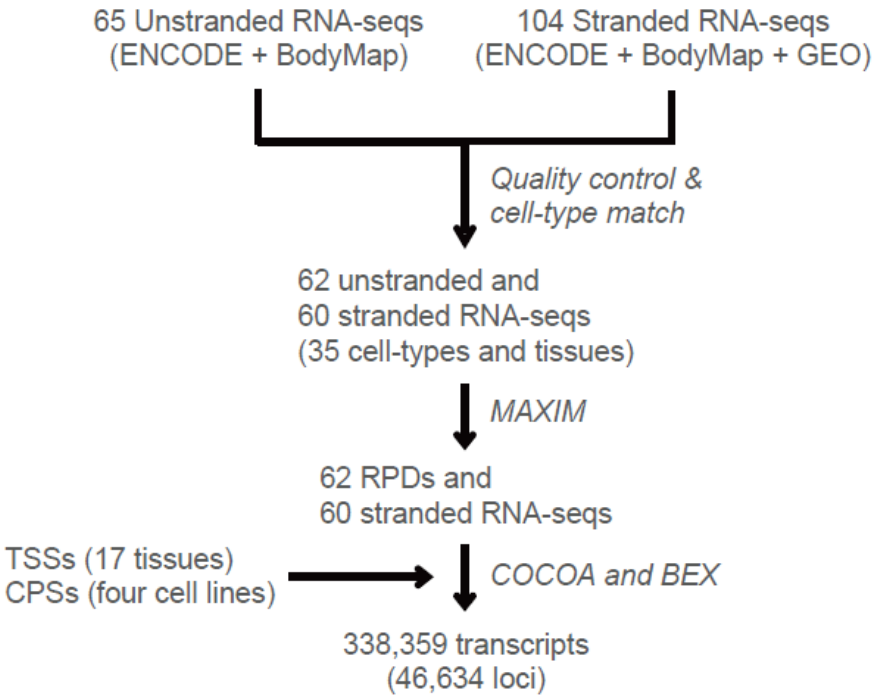
CAFE: Co-Assembly Followed by End-correction

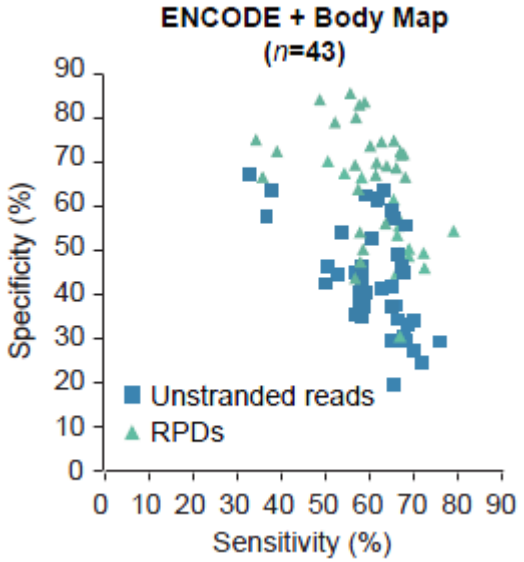


CAFE helps us reconstruct precise full-length transcriptomes



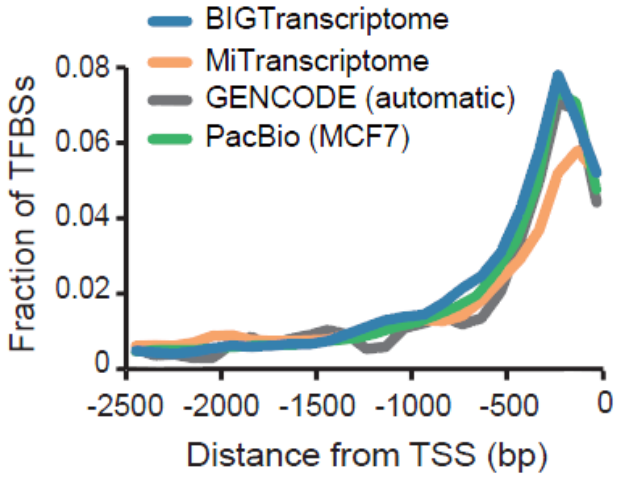
BIGTranscriptome from large-scale public RNA-seq data

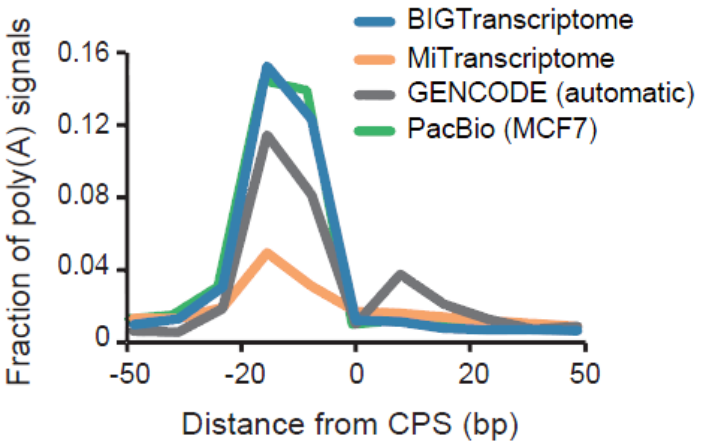




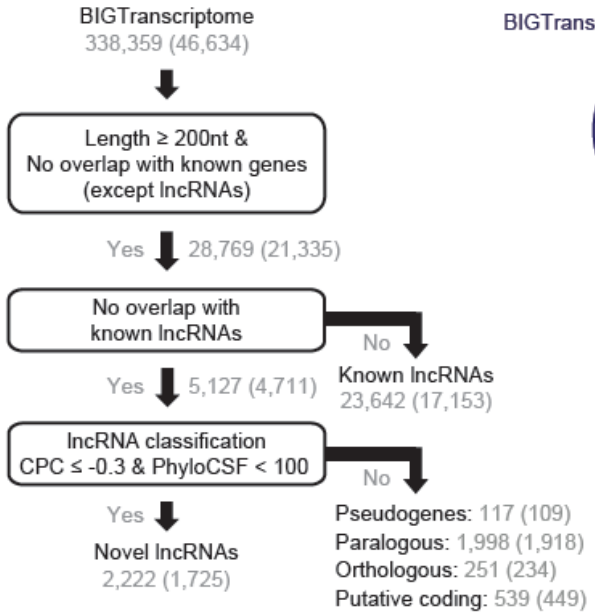
High-confidence BIGTranscriptome comparable to a long-read method

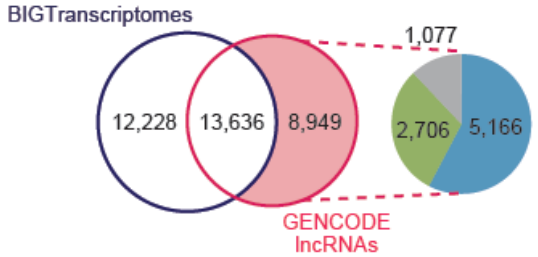
Annotation (%)	BIGTranscriptome				MiTranscriptome			
	Base level		Intron level		Base level		Intron level	
	SN	SP	SN	SP	SN	SP	SN	SP
RefSeq	91.4	48.3	98.4	55.1	94.3	33.6	93.2	37.7
GENCODE (manual)	86.6	66.4	99.7	88.7	77.8	29.7	74.3	47.7
GENCODE (automatic)	90.9	28.5	97.7	23.2	91.5	16.1	88.8	14.9
PacBio (MCF7)	85.6	50.2	92.1	52.9	80.2	30.1	86.6	46.0
EST	·	·	80.4	72.1	·	·	67.0	43.5
RefSeq + GENCODE + PacBio + EST	·	·	85.2	91.2	·	·	62.6	50.7



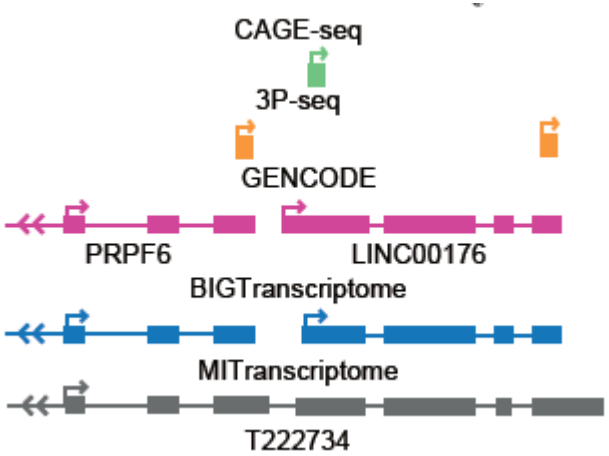


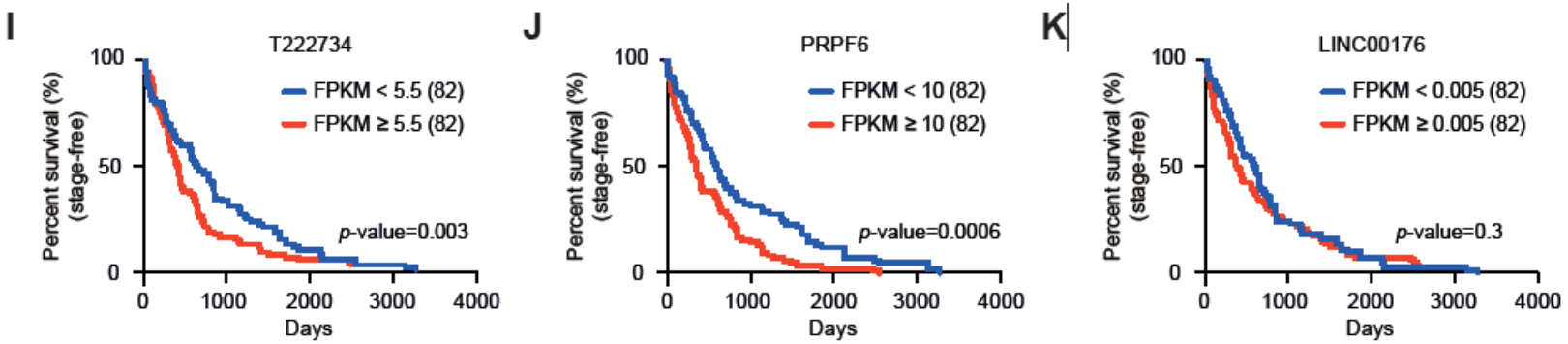
BIGTranscriptome includes known and novel lncRNAs



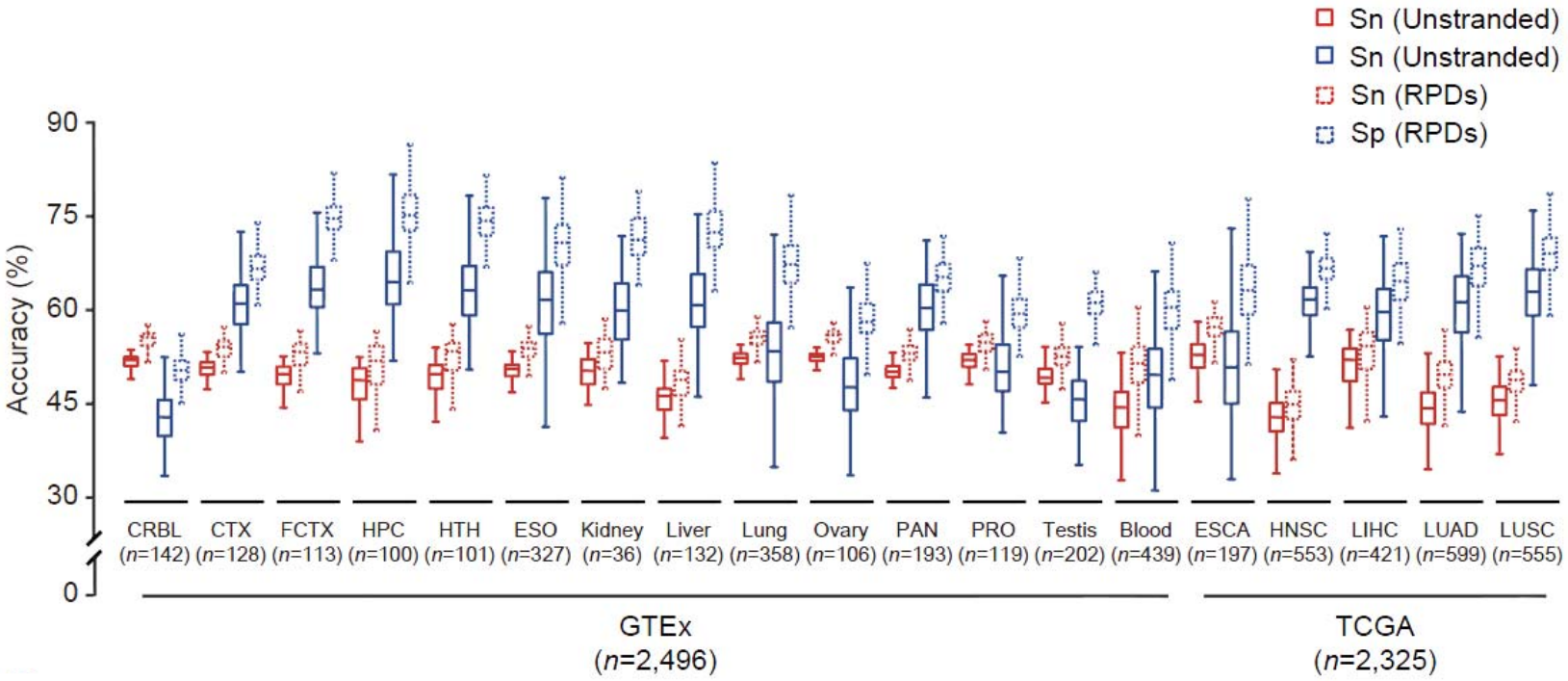


High-confidence noncoding transcriptome map leads to better downstream analyses





BIGTranscriptome-ST from large-scale personal RNA-seq data



High-confidence coding and noncoding transcriptome maps

Bo-Hyun You, Sang-Ho Yoon and Jin-Wu Nam

Genome Res. 2017 27: 1050-1062 originally published online April 10, 2017
Access the most recent version at doi:10.1101/gr.214288.116

Download

Home > Download

Program

Name	Description	Download
CAFE source code	Version 1.0.1	download
CAFE manual	User manual	download

Requirement

This program was developed on Linux environment (CentOS 6.8). Cufflinks, samtools, perl and python are required for running.

License

Annotations

Name	Description	Format	Download
BIGTranscriptome	GENCODE, Human BodyMap, and GEO (122 samples, 35 cell-types and tissues)	gtf	download
BIGTranscriptome-TS	GTEx and TCGA (4,821 samples, 19 tissues and tumors)	gtf	download
BIGTranscriptome lncRNA catalog	Known and novel lncRNAs annotated from BIGTranscriptome	gtf	download

RPDs

RPDs (the reads with predicted directions) files are available at the following link.

Download RPDs

Expression values

http://big.hanyang.ac.kr/CAFE

BIGLab

- http://big.hanyang.ac.kr

Funded by

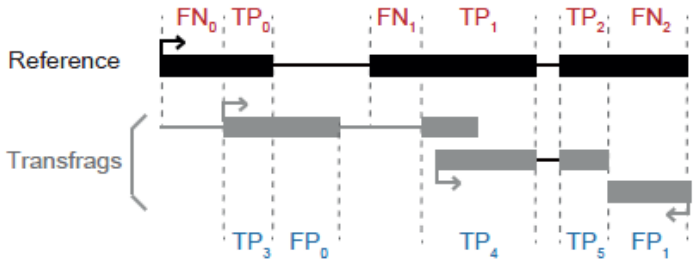
미래창조과학부
Ministry of Science, ICT and Future Planning

보건복지부

RDA 농촌진흥청

Strand-specific evaluation of transcriptome maps

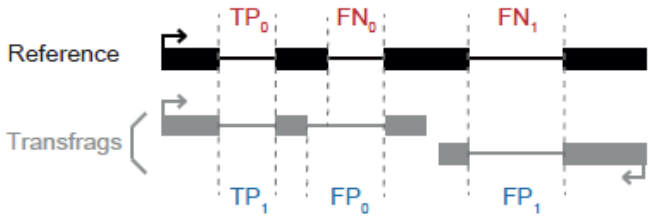
Base level



$$\text{Sensitivity} = \frac{(TP_0 + TP_1 + TP_2)}{(TP_0 + TP_1 + TP_2) + (FN_0 + FN_1 + FN_2)}$$

$$\text{Specificity} = \frac{(TP_3 + TP_4 + TP_5)}{(TP_3 + TP_4 + TP_5) + (FP_0 + FP_1)}$$

Intron level



$$\text{Sensitivity} = \frac{(TP_0)}{(TP_0) + (FN_0 + FN_1)}$$

$$\text{Specificity} = \frac{(TP_1)}{(TP_1) + (FP_0 + FP_1)}$$