Thesis for the Doctor of *Philosophy*

Reconstruction of High-confidence Transcriptome Maps and Pan-cancer Analysis of Long Noncoding RNAs

Bo-Hyun You

Graduate School of Hanyang University

February 2020

Thesis for the Doctor of *Philosophy*

Reconstruction of High-confidence Transcriptome Maps and Pan-cancer Analysis of Long Noncoding RNAs

Thesis Supervisor: *Jin-Wu Nam*

A Thesis Submitted to the graduate school of
Hanyang University in partial fulfillment of the requirements
for the degree of Doctor of *Philosophy*

Bo-Hyun You

February 2020

Department of Life Science
Graduate School of Hanyang University

This thesis, written by *Bo-Hyun You,*
has been approved as a thesis for the *Doctor of Philosophy.*

November 2019

Committee Chairman: Tae-Min Kim (Signature)

Committee member: Daehyun Baek (Signature)

Committee member: Jin-Wu Nam (Signature)

Committee member: Jiwon Shim (Signature)

Committee member: Junho Choe (Signature)

Graduate School of Hanyang University

Table of Contents

List of Figures

List of Tables

Abstract

Reconstruction of High-confidence Transcriptome Maps and Pan-cancer Analysis of Long Noncoding RNAs

Bo-Hyun You
Department of Life Science
Hanyang University

Ever since the burst of high-throughput RNA sequencing (RNA-seq), there has been a huge stream of newly annotated genes. These novel genes are comprised mostly of non-coding RNAs, especially long non-coding RNA (lncRNA) genes, which have been shown to play critical roles in myriad of biological processes. Even though numerous coding and non-coding transcriptome maps have been introduced in recent years and helped researchers to find missing information in diverse cellular processes, they are still incomplete partly because they were mostly reconstructed based on RNA-seq reads that lack strand information (known as unstranded reads) and accurate gene boundary information. To improve the accuracy of transcriptome maps, I developed a high-performing transcriptome assembly pipeline, CAFE. CAFE predicts the directions of unstranded reads using the maximum likelihood estimation and refines gene boundaries by integrating information about transcription start sites and cleavage and polyadenylation sites. Applying CAFE to the transcriptomic data from the ENCODE Project enabled us to construct high-confidence transcriptome map, named BIGTranscriptome, which is comparable to the manually curated map.

To identify novel cancer-driving lncRNAs, I applied CAFE pipeline to RNA-seq datasets of ESCC patients from multiple cohorts (Korean, Chinese, and GTEx+TCGA cohorts) and constructed a comprehensive ESCC transcriptome. As a results, I annotated 1,924 novel lncRNAs and identified 113 commonly dysregulated lncRNAs in ESCCs. Six of the dysregulated lncRNAs were significantly associated with the clinical outcomes of ESCC patients and defined four ESCC subclasses with different prognoses. Among the six lncRNAs, we found a novel lncRNA which named as HERES (highly expressed lncRNAs in esophageal squamous cell carcinoma), promotes cell proliferation, migration, invasion, and colony formation in ESCC cell lines and tumor growth in xenograft models. HERES appears to be a trans-acting factor that regulates CACNA2D3, SFRP2, and CXXC4 simultaneously to activate Wnt signaling pathways through an interaction with EZH2 via its G-quadruple structure-like motif.

Finally, applying CAFE pipeline to a collection of large-scale transcriptome data comprising more than 16,000 RNA-seq samples from the ENCODE Project, the Human BodyMap 2.0 Project, the CCLE Project, the TCGA Project, and the GTEx Project led to the creation of the most comprehensive and accurate tissue-/cancer-specific transcriptome maps. Our integrative transcriptome encompasses numerous novel lncRNAs, including thousands of antisense lncRNAs and hundreds of tissue-/cancer-specific lncRNAs.

As of conclusion, the CAFE pipeline and the resulting transcriptome maps will not only help to expand the universe of coding and non-coding genomes but also enable discovery of novel biomarkers and therapeutic target of value such as HERES in ESCC.

Chapter 1. Introduction

With the invention of shotgun sequencing, the era of omics data begun, and it led to parallel analysis of the vast number of identified genes and discovery of thousands of novel genes (*1-5*). Large-scale high-throughput RNA sequencing (RNA-seq) data from the ENCODE Project were used to characterize highly complex, overlapping transcription units on both strands, revealing that more than 60% of the human genome is reproducibly transcribed in at least two different cell types (*4, 6*). Intriguingly, a significant portion of these extensive transcription signals, mostly from intergenic regions, turned out to be unannotated. As comprehensive transcriptome maps essential for understanding of gene expression regulation in both coding and non-coding genomic regions (*1, 3*), the need to identify the unannotated transcriptome quality arose. Gene annotation projects, such as GENCODE (*4*), Human BodyMap 2.0 (*7*), MiTranscriptome (*8*), FANTOM CAT (*9*), and CHESS (*10*), have massively reconstructed whole transcriptomes by assembling large-scale RNA-seq data and have characterized transcriptome-wide non-coding RNAs (ncRNAs).

Unknown transcripts can be identified via assembly of RNA-seq data by two approaches: the genome-guided approach (known as reference-based assembly) (*11-16*) and the *de novo* approach (*16-22*). Because the *de novo* approach assembles RNA-seq reads without a guide genome, it generally requires RNA-seq data with strand information (called stranded RNA-seq data). However, for the reference-based approach, the stranded RNA-seq data had been regarded as dispensable because the sense-orientation of some reads spanning exon-junctions could be predicted based on the splicing signal. For that reason, most of the largest sources for transcriptome data such as the ENCODE Project (*5*), the Cancer Cell Line Encyclopedia (CCLE) Project (*23*), the Genotype-Tissue Expression (GTEx) Project (*24*), and The Cancer Genome Atlas (TCGA) (*25, 26*) Consortium produced large-scale unstranded RNA-seq data without strand information. Following genome-wide gene annotation projects thus have proceeded using these unstranded data. For an instance, MiTranscriptome was built from 6,810 publicly available unstranded RNA-seq data from ENCODE, TCGA, and other studies (*23*). Unfortunately, transcriptome assembly using unstranded RNA-seq data often results in erroneous transcript models, including chimeras, particularly when there are convergent, divergent, or antisense overlaps between two genes (*3, 27-30*). Nevertheless, the re-use of publicly available big unstranded data with the stranded data could not only enhance detection of new transcripts but also reduce the generation of erroneous transcript models (*29*).

RNA-seq-based transcriptome assembly is also challenged by the ambiguous ends of assembled transcripts (*29, 31*). Early methods roughly defined the transcription structures with the support of histone modification signals, such as H3K4me3 for activate promoters and H3K36me3 for active gene bodies (*11, 32*). Later, specialized RNA sequencing techniques, such as cap analysis gene expression by sequencing (CAGE-seq) (*33-35*) and poly(A) position profiling followed by sequencing (3P-seq) (*27, 32*), have been successfully applied to define the ends of transcripts at single base resolution. Determination of accurate gene boundaries through integrative analysis of the specialized RNA-seq data including CAGE-seq and 3P-seq would enable the appropriate functional studies of the novel genes, especially regulatory non-coding genes that are expected to play significant roles in diverse biological processes.

The Wnt signaling pathway is a well-known, evolutionarily conserved pathway that plays important roles in embryonic development; it has also been widely implicated in numerous tumor malignancies (*36-39*). Wnt signaling can activate both β-catenin-dependent (canonical) and -independent (non-canonical) signal transduction cascades (*38, 39*). Canonical Wnt signaling results in translocation of the transcriptional activator β-catenin into the nucleus during embryonic development and cell differentiation (*40*). Constitutive activation of this pathway by various causes leads to developmental diseases and carcinogenesis (*41*). In contrast, noncanonical Wnt pathways are known to be transduced by Wnt polarity, Wnt-Ca2+, and Wnt-atypical protein kinase signaling, independent of β-catenin transcriptional activity (*42*). These pathways have also been reported to be independently involved in cancer development as well as embryonic development. In particular, abnormal intracellular levels of the second messenger Ca2+ promote the Wnt signaling pathway, which in turn promotes the development and progression of many types of cancers (*43*).

Controlling Wnt signaling may be a useful strategy for curing cancers caused by aberrations in such signaling. The inhibition of either aberrant canonical or noncanonical Wnt signaling, however, has been shown to decrease progression in only a subset of cancers in a context-dependent manner (*44*). Because aberrations in Wnt signaling pathways result from various causes, such as mutations in different Wnt signaling-related genes, ligand overexpression, and dysregulation of regulators, targeting only the canonical Wnt signaling pathway might not be a universal therapeutic approach for cancers. Thus, the simultaneous inhibition of aberrant canonical and noncanonical Wnt signaling pathways could be a promising approach for cancer therapy.

Esophageal squamous cell carcinoma (ESCC), a major histological type of primary esophageal cancer in east Asia and other developing countries, is associated with a very poor survival rate that is only 5-15% at five years (*45, 46*), mainly due to delayed diagnosis, a high rate of metastasis, and a lack of effective treatment strategies (*45-47*). Moreover, the benefits of curative surgery for advanced stages of ESCC are still unclear (*46, 48*), and although cisplatin-based chemotherapy is commonly used, the effects are inconsistent among individuals (*46, 48*). Despite ongoing trials with combination therapy, efforts to identify appropriate targets to improve the therapy for ESCC have been largely unsuccessful (*49, 50*).

Long noncoding RNAs (lncRNAs), defined as transcripts longer than 200nt that do not code for functional proteins (*51, 52*). Further investigation of lncRNAs functions hinted that lncRNAs are a rather heterogeneous group of RNAs, with each individual exerting diverse roles in a wide range of biological process through different mode of actions (*51-54*). Because lncRNAs can modulate multiple targets at the transcriptional and posttranscriptional levels, lncRNAs tend to play functional roles in more than one biological pathway. Moreover, mounting evidence indicates that aberrant lncRNA expression, by modulating cancer-related pathways, can be responsible for cancer progression (*55, 56*). HOTAIR is a trans-acting lncRNA that promotes cancer progression via different pathways depending on the cancer type (*57, 58*). The most well-known mechanism of HOTAIR involved interactions with

PRC2 and LSD1 histone modification complexes to promote cancer metastasis through chromatin state reprogramming (*57*).

To date, hundreds of lncRNAs have been reported to be dysregulated in cancers and tens of them have been associated with cancer progression. With respect to ESCC development, the function of a few lncRNAs, including LUCAT1 and CASC9, have been investigated via a candidate-gene approach (*59, 60*). Recently, a Chinese group performed RNA-seq on tissue from 15 paired ESCC patients and normal individuals and identified lncRNAs dysregulated in ESCCs (*61*). Furthermore, they described a lncRNA that affects cell proliferation and invasion in ESCC cell lines but did not pursue further to elucidate the mechanism of action. Thus, the identification of novel ESCC-driving lncRNAs and an investigation of their cancer-driving mechanisms have not been simultaneously carried out.

Despite these findings that successfully describe the importance of lncRNAs, global investigation of lncRNAs via transcriptome reconstruction from large-scale RNA-seq data suffered from aforementioned difficulties, more so due to their low expression levels and condition-specific expression patterns. These particularities combined with the limited computational resource and large portion of unstranded data among public available datasets and hindered accurate annotation of lncRNAs. Numerous attempts to construct lncRNA maps from large collection of RNA-seq datasets, such as FANTOM CAT and CHESS annotations, relied heavily on reference annotations to avoid inaccurate transcript models rising from unstranded reads and overlooked novel lncRNA genes and isoforms.

In Section 2, I will proceed to explain a high-performing transcriptome assembly pipeline, CAFE, which is designed to tackle the challenges associated with construction by utilizing unstranded and stranded RNA-seq reads, and CAGE-seq and 3P-seq data. Section 3 describes a study of novel ESCC-driving lncRNA, HERES, from multi-cohort RNA-seq datasets, discovered using CAFE pipeline. Construction of high-confidence transcriptome maps and annotation of lncRNAs in pan-cancer from more than 16,000 RNA-seq samples comprising diverse types of tissue, cancer, and cell line is described in Section 4. Finally, the results are discussed in Section 5, followed by a list of references in Section 6 and supplementary information in the Supplementary Materials section.

Chapter 2. Computational approach for reconstruction of high-confidence transcriptome map

2.1. Unstranded RNA-seq causes error-prone assembly

2.1.1. Factors of transcriptome assembly quality

To investigate the factors that affect quality of transcriptome assembly, we reconstructed 45 stranded and 32 unstranded assemblies from public available RNA-seq data from the ENCODE Project using Cufflinks (*62*). The resulting assemblies were evaluated based on the GENCODE v19 annotation. The evaluation was done by counting false negative (FN), false positive (FP), and true positive (TP) bases upon agreement between the reference and the resulting assembly at the base level.

The recall (TP/(TP+FN)) of the resulting assemblies appeared to be simply correlated with the size of mapped reads up to about 100 million mapped reads but converged beyond that size (Fig. 1A), which suggests that many samples from the ENCODE Project still need more data to reach their maximum recall. On the other hand, the precision (TP/(TP+FP)) of the unstranded assemblies was much less than that of the stranded assemblies, regardless of the size of the mapped reads (Fig. 1B). This result indicates that stranded reads provide more accurate information for transcriptome assembly.

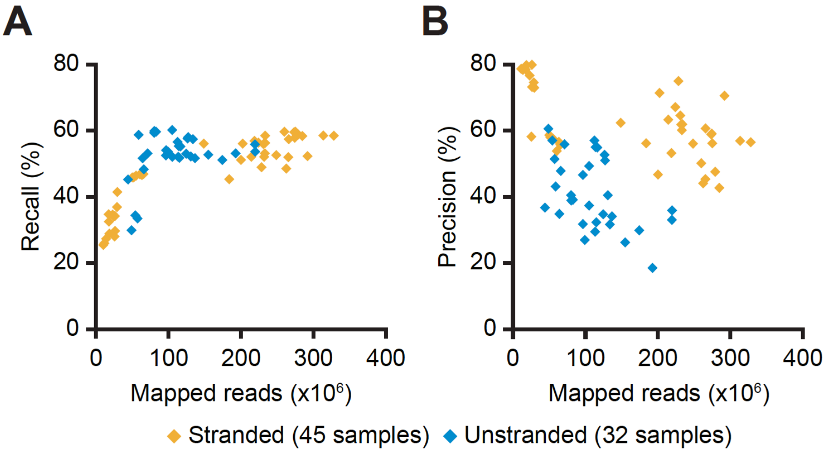


Figure 1. Important factors for quality of transcriptome assembly. (A-B) Recall (A) and precision (B) of stranded (orange diamond) and unstranded (blue diamond) assemblies constructed from ENCODE RNA-seq data are shown over the number of mapped reads.

2.1.2. Classification of erroneous transfrags

To examine the nature and cause of the errors in unstranded assembly, we next sequenced both stranded and unstranded RNA-seq libraries that were simultaneously prepared in mouse embryonic stem (mES) cells. We also obtained a pair of publicly available stranded and unstranded RNA-seq datasets of human HeLa cells from the NCBI gene expression omnibus (GEO). These reads were mapped to reference genomes (hg19 for human and mm9 for mouse) using TopHat (*63*), and unstranded reads (~40 million mapped reads for HeLa cells and ~68 million mapped reads for mESC) were assembled using Cufflinks. In total, 51,045 and 48,509 transcript fragments (transfrags) whose full lengths were not examined were assembled from HeLa and mES cells, respectively. The resulting transfrags were divided into five groups based on their directions validated by stranded RNA-seq signals: (1) correct, (2) incorrect (those with an RNA-seq signal on the opposite strand), (3) ambiguous (those with RNA-seq signals on both strands), (4) undetermined (those with no direction), and (5) unsupported (those with no stranded RNA-seq signals in either direction) (Fig. 2A). All transfrags in the correct group (24.24% for HeLa cells and 29.76% for mES cells) were multi-exonic (Figs. 2B and C); this high accuracy was the result of exon-junction reads that define the direction of the resulting intron with the splice-signal ‘GU-AG’ at the ends of the intron (Fig. 2C). The remainder were regarded as problematic transfrags (75.76% for HeLa cells and 70.24% for mES cells). They displayed low accuracies and were placed in the incorrect (0.31% and 0.14%), ambiguous (33.13% and 38.79%), undetermined (39.52% and 31.03%), and unsupported (2.8% and 0.28%) groups (Figs. 2B and C). They appeared to be severely defective in their structures and/or directions (Fig. 3), and the majority in the undetermined group were single-exonic transfrags (Fig. 2C). However, except for those in the unsupported group (Fig. 2A), the defective transfrags (72.96% for HeLa cells and 69.96% for mES cells) could be corrected using the guide of the matched, stranded RNA-seq data.

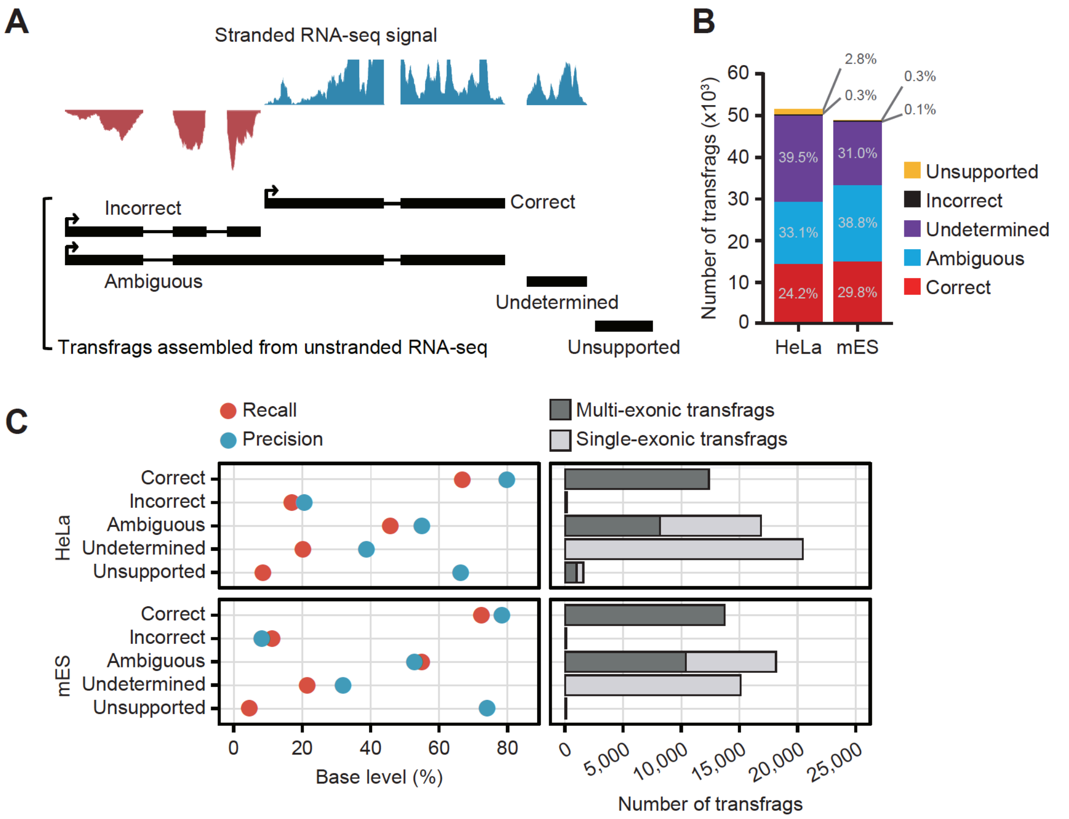


Figure 2. Five groups of transfrags assembled from unstranded RNA-seq. (A) Classification of transfrags assembled from unstranded RNA-seq data. Graphs on the top are signals from stranded RNA-seq data (blue is the signal in the forward direction and red is the signal in the reverse direction). (B) Shown are the percentages of transfrags belong to the five groups - correct (red), ambiguous (blue), undetermined (purple), incorrect (black), and unsupported (yellow) in HeLa and mES cells. (C) The precision (light blue) and recall (red) of the five groups compared to the reference protein-coding genes in HeLa (left top) and mES cells (left bottom). The number of multi-exonic (dark gray) and single-exonic (gray) transfrags are indicated in each group (right).

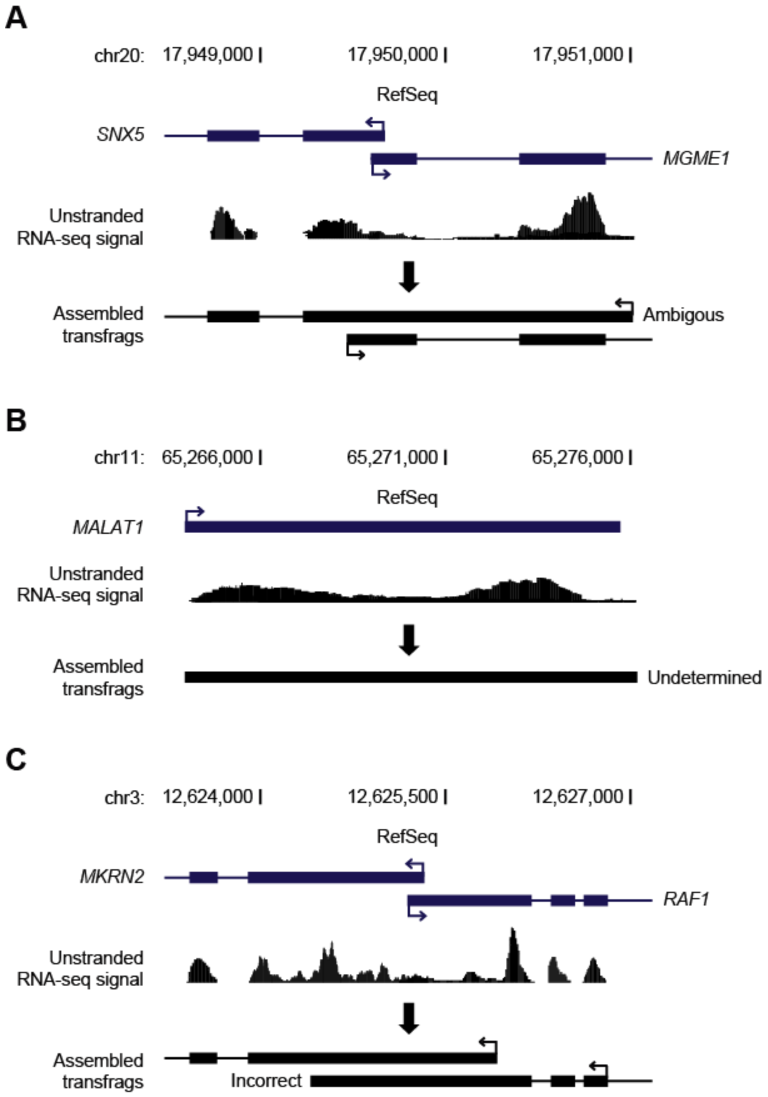


Figure 3. Problematic transfrags assembled from unstranded RNA-seq. (A-C) Examples of problematic transfrags belong to three groups - ambiguous (A), undetermined (B), and incorrect (C). Graphs are signals from unstranded RNA-seq data.

2.2. Prediction of the directions of unstranded reads

2.2.1. *K*-ordered Markov chain (*k*MC) model for read direction

To predict the direction of unstranded reads mapped to the genome, *k*-ordered Markov chain (*k*MC*)* models were trained with the directions of the *k*-nearest stranded reads relative to a target read. We built a training dataset including *S* base reads randomly selected from stranded reads mapped to genomes and their matched *k*-nearest reads. To acquire the *k*-nearest reads, we used a step-wise *k*-nearest method, in which the read $\mathrm{x}_{k=1}$ nearest to a query read $\mathrm{x}_{k=0}$ was first selected, then the read $\mathrm{x}_{k=2}$ nearest to the current read $\mathrm{x}_{k=1}$ was selected, then the read $\mathrm{x}_{k=3}$ nearest to the current read $\mathrm{x}_{k=2}$ was selected, and so on. To train unbiased models, we used 10 million as *S*, a large enough sampling number that is proportional to the *k* (also proportional to the number of states and edges to train). Practically, 2 X *K* matrix M_{+} *or* M_{-} for each emission value (+ and -) were constructed from the training data and each cell $\mathrm{m}_{+i,j}$ or $\mathrm{m}_{-i,j}$ in the matrix indicates the fraction of + or - direction of the *j*th-nearest read $\mathrm{x}_{k=j}$ when the emission value (direction) of the previous state is *i*.

2.2.2. Maximum likelihood estimation (MLE) of read direction

The direction of an unstranded read, *r*, was inferred from the trained *k*MC models given step-wise *k*-nearest stranded reads of a query unstranded read mapped to a genome locus using maximum likelihood estimation (MLE) as in the following equation.

$$L_r^* = \underset{L\in\{+,-\}}{\operatorname{argmax}} \left(\prod_{1}^{k} m_{i,j=k} \right)$$

where *i* is the direction of the *k*th-nearest read, *L* is a set of possible directions that are, and L_r^* is the maximum likelihood direction of the unstranded read *r*. Using the MLE, all maximum likelihood directions were predicted for all unstranded reads. If an unstranded read was paired-end, then its direction was determined differently, as follows. If a fragment of a paired-end read spanned an exon-junction, the direction of the read was directly determined by the splice signal without MLE. If the directions of two fragments of a paired-read were inconsistent, the direction with greater likelihood was chosen for the read.

2.2.3. Prediction of read directions using MLE

To facilitate stranded assemblies with additional stranded reads, we predict the directions of unstranded RNA-seq reads using *k*MC models whose transition probabilities were estimated with the directions of a current read x and its *k*-nearest stranded reads, x_k. In the prediction step, the direction of a read with an unknown direction, y, was determined using MLE (Fig. 4A). A read with a predicted direction (RPD) was treated as a pseudo-stranded read and was used in the downstream assembly. Performing systematic analyses while increasing *k*, we found the optimum to be *k*=3, a value at which the accuracy is maximized and the computational cost is minimized (Figs. 5A and B). Compared to a simple majority voting method with *k*-nearest stranded reads, *k*MC performed better as *k* increased (Figs. 5C and D). Thus, we predicted the directions of all unstranded RNA-seq data using the Markov chain model with the optimum *k*-order and assembled all stranded read-like RPDs. Compared to the original assembly (unstranded assembly), those that were re-assembled from RPDs (RPD assembly) were significantly improved by 9.3–10.7% in their precision without compromising their recall (Fig. 4B for HeLa cells and Fig. 4C for mES cells). For instance, unstranded reads from a genomic locus where *LOC148413* and *MRPL20* are convergently transcribed were assembled into an erroneous annotation but their RPDs led to correction of the erroneous gene structure (Fig. 4D).

To test the general usage of the *k*MC model, we predicted the directions of unstranded reads from HeLa cells using the *k*MC model trained in mES cells, and vice versa. The species-mismatched models were comparable to the species-matched models (Fig. 6), suggesting that the *k*MC model can be generalized to other cell types and species.

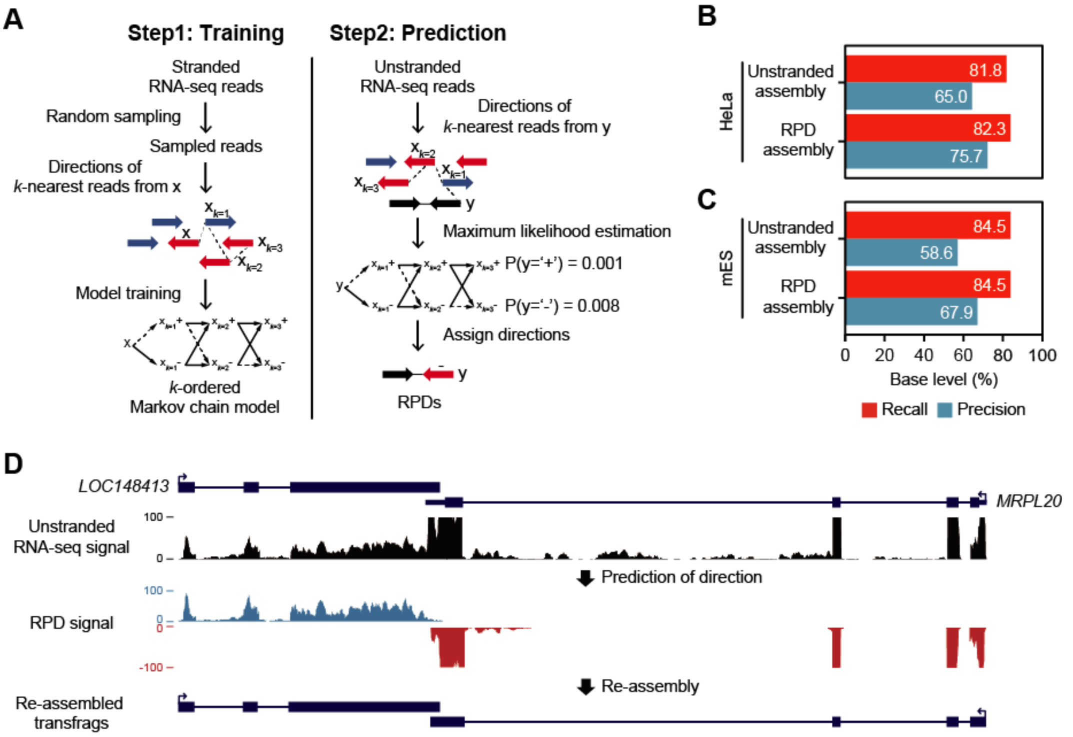


Figure 4. Prediction of read directions using MLE. (A) Overview of *k*MC training and MLE of read direction. (Left) *S* base reads randomly sampled from stranded RNA-seq reads and their matched step-wise *k*-nearest reads ($x_{k=1}$, $x_{k=2}$, $x_{k=3,\dots}$) were used for training *k*MC. Blue arrows are reads in the forward (+) direction and red arrows are reads in the reverse (-) direction. (Right) Prediction of read direction using MLE. Step-wise *k*-nearest stranded reads ($x_{k=1}$, $x_{k=2}$, $x_{k=3,\dots}$) from a query unstranded read (black arrow) were extracted and used to calculate two likelihoods at (+) and (-). A direction with the maximum likelihood is finally assigned to the query read. (B-C) Accuracies of transcriptomes assembled with RPDs (*k*=3) and unstranded reads in HeLa (B) and mES cells (C). (D) An example of resulting transfrags re-assembled with RPDs. *LOC148413* and *MRPL20* are convergently overlapped at a locus where unstranded RNA-seq signals (black) are not separated but blue and red RPD signals are clearly separated in the forward and reverse directions, respectively.

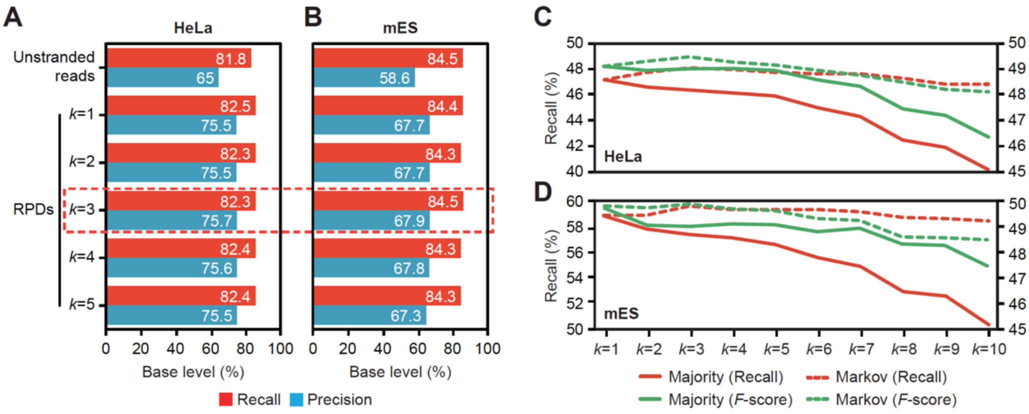


Figure 5. Systematic analysis of MC *k*-order. (A-B) Accuracies of RPDs across different *k*-orders. The optimal *k*-order (dotted box) was set as the *k* at which the precision was maximized for HeLa (A) and mES cells (B). (C-D) Accuracies (*F*-score and recall) of RPDs by *k*MCs and *k*-nearest majority voting across different *k* values (from 1 to 10) for the antisense-overlapping loci in HeLa (C) and mES cells (D).

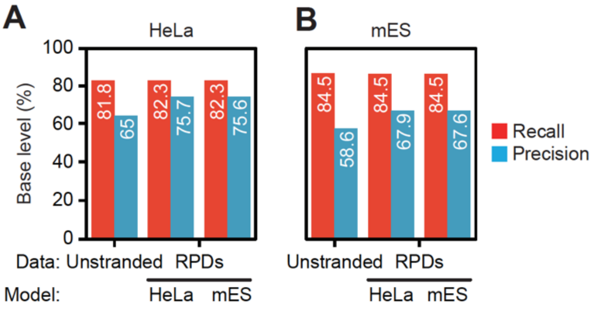


Figure 6. Generality of the *k*MC model. (A-B) Comparisons of unstranded, species-mismatched, and species-matched models. Accuracies of unstranded, species-mismatched (by the *k*MC models trained in mES cells), and species-matched models (by the *k*MC models trained in HeLa cells) in HeLa cells (A) and vice versa (B).

2.2.4. The benefits of expression quantification from RPDs

The use of stranded RNA-seq data leads not only to better transcriptome assembly, but also in principle to better gene expression quantification. To test whether the expression quantification benefits from the prediction of strand information, the gene expression values were calculated with unstranded and corresponding RPDs, and then were compared to those calculated with stranded reads (Fig. 7). Overall, the unstranded reads over-estimated the expression level of genes in the loci with antisense-overlapping transcripts but RPDs corrected the over-estimation, leading to better correlation with those of stranded reads.

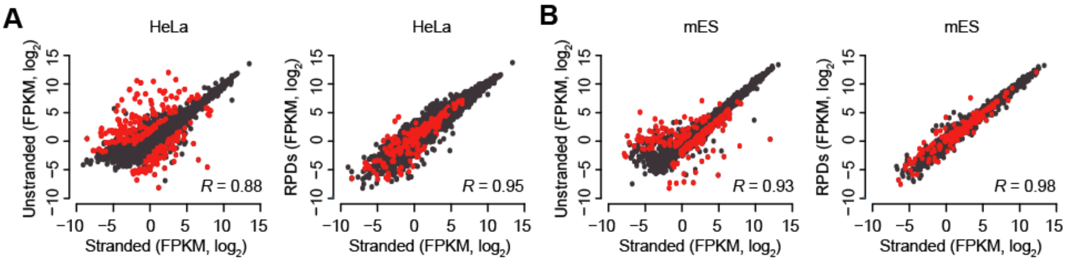


Figure 7. The expression quantification benefits from RPDs*.* (A-B) Comparisons of gene expression values (FPKM, $\log_2$) estimated by stranded (X-axis) and unstranded reads (Y-axis, left) or RPDs (Y-axis, right) in HeLa (A) and mES cells (B). The correlation coefficients were calculated with Pearson's correlation between the X- and Y-axis values. The red dots indicate genes with antisense-overlapped genes.

2.3. Refining boundaries and finding new exon-junctions

2.3.1. Updating exon-junctions

Shallow sequencing depth and short read length often cause transcript fragmentation in transcriptome assembly, mainly due to missing exon-junction reads and discontinuity of read overlaps. To update exon-junction signals missed in the original assembly, all pairs of neighboring transfrags on the same strand within a distance ranging from 50 bp to the 99th percentile of the lengths of all known introns (50-265,006 bp for human and 50-240,764 bp for mouse) were re-examined. The neighboring transfrags within a distance of 50 bp were combined. If more than two exon-junction reads in at least two samples were detected, the neighbored transfrags were connected by the junction. Otherwise, the gaps between two neighboring transfrags were further scrutinized to detect *cis*-splicing signals. The gaps including splice donor ‘GU’ and acceptor ‘AG’ signals, but not CAGE-seq or 3P-seq tags, between two neighboring transfrags were scanned by MaxEntScan (version 20040420) (*64*), which calculates entropy scores for splice donor and acceptor sites. If the maximum entropy scores of both the splice donor and acceptor sites were above 0.217, a cutoff used in previous studies (*65*), then the interspace between the ‘GU’ and ‘AG’ was regarded as an intron and the two transfrags were connected by the intron.

2.3.2. Refining TSSs and CPSs

RNA-seq-based transcriptome assembly often results in imprecise transcript boundaries (Fig. 8). To annotate the accurate transcription start sites (TSSs) and cleavage and polyadenylation sites (CPSs), transfrags were updated by Cap analysis gene expression sequencing (CAGE-seq) (*35*) and poly(A) position profiling by sequencing (3P-seq) tags (*66*). The method for TSS identification from CAGE-seq tags was modified from the method for CPS identification from 3P-seq tags (*66*). Of the identified sites, those located in either the first exon or in the 3kb upstream region of a gene, without overlapping the upstream gene, were regarded as TSSs of the gene. Similarly, of the CPSs identified from 3P-seq tags, those assigned to either the 3’ UTR or the 5kb downstream region of a gene, without overlapping the downstream gene, were regarded as CPSs of the gene. After updating TSSs and CPSs, we removed all redundant transcripts or inclusive transfrags.

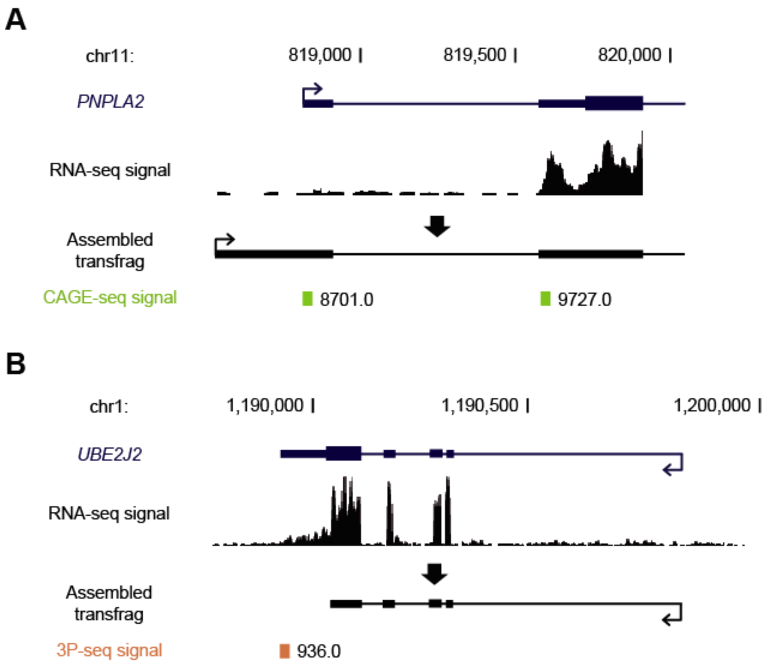


Figure 8. Erroneous transfrag boundaries. (A) Shown is an example of mis-assembly of *PNPLA2* at its 5' end, evident by CAGE-seq signals. (B) Shown is an example of mis-assembly of *UBE2J2* at its 3' end, evident by 3P-seq signals.

2.3.3. Full-length transcripts with updated exon-junctions and boundaries

To improve the integrity of the assembled transcriptome, the missed exon-junctions were examined by either experimental or computational approaches (Fig. 9; see Section 2.3.1). Of 51,270 potential exon-junctions, 1,506 (3%) were additionally supported by the experimental approach in HeLa cells (Fig. 10A) and a similar fraction of potential junctions were supported in mES cells (Fig. 10A). Of the newly connected exon-junctions, 91.0-94.4% were present in GENCODE annotations and the remainder were novel (Fig. 10A). The unconnected potential exon-junctions were examined further with the program MaxEntScan to determine whether the most likely putative splicing signal, 'GU-AG,' existed in the region between two neighboring transfrags (Fig. 9). Using that approach, 11,153 potential junctions for HeLa cells and 7,634 for mES cells were newly connected (Fig. 10A); 84.7–85.2% were present in GENCODE gene annotations and the remainder were novel (Fig. 10A).

To improve transfrag boundary annotation, TSSs, determined from CAGE-seq, and CPSs, determined from 3P-seq, were incorporated into relevant transfrags (Fig. 9; see Section 2.3.2). For TSSs and CPSs, respectively, 93-94% and 96-98% of transfrags were either confirmed or revised (Fig. 10B). Transfrags updated for both TSS and CPS (91-92%) were regarded as full-length transcripts (Fig. 10B). Updating TSSs improved the definition of the upstream promoter regions in which transcription factor binding sites (TFBSs) are significantly enriched (Fig. 10C). Similarly, transfrags with CPSs displayed an enriched poly(A) signal, AAUAAA within 15–30nt upstream of the cleavage site, compared to those without CPS updates (Fig. 10D).

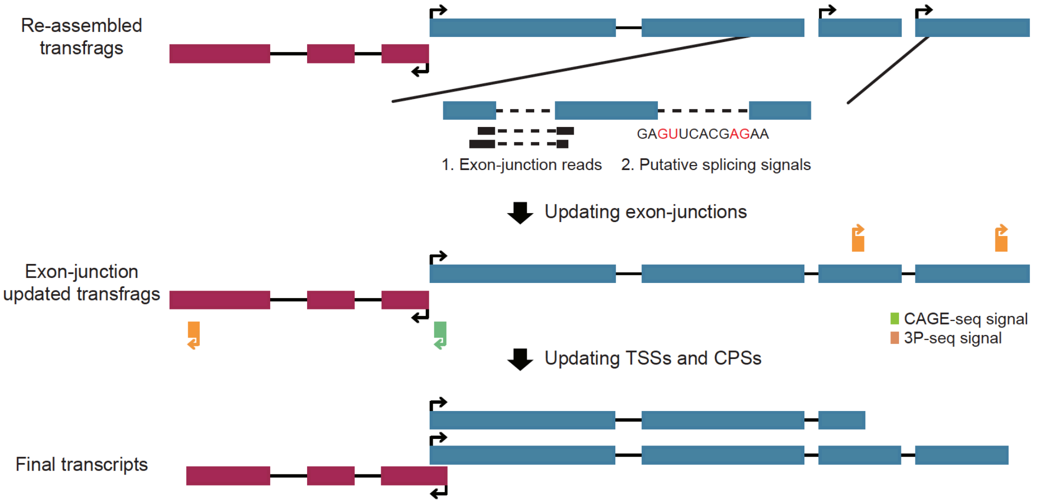
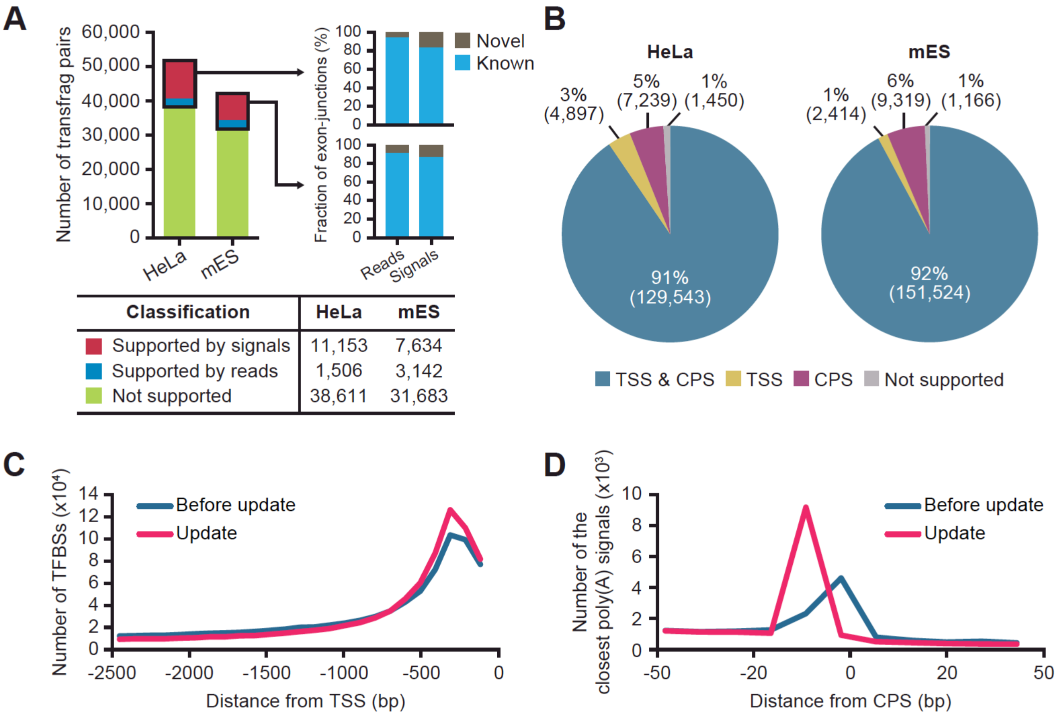


Figure 9. Updating exon-junctions, TSSs, and CPSs in transfrag models. Shown is a workflow for updating transfrag models, which comprises two steps: i) updating exon-junctions and ii) refining TSSs and CPSs.



Classification	HeLa	mES
🟥 Supported by signals	11,153	7,634
🟦 Supported by reads	1,506	3,142
🟩 Not supported	38,611	31,683

Figure 10. Transcript models with new exon-junctions and accurate ends. (A) The number of neighboring transfrag pairs supported by putative splicing signals (red), by exon-junction reads (navy), and by neither (green) in HeLa and mESC cells. The numbers in parentheses in the key indicate the number of pairs in each group. Among exon-junctions supported by either exon-junction reads or putative splicing signals, the fractions of known (cyan) and novel (brown) exon-junctions in GENCODE annotations are shown in the inset. (B) The fraction of transfrags updated with both TSS and CPS (blue), with only TSS (yellow), with only CPS (magenta), and with neither TSS or CPS (grey) in HeLa and mESC cells. (C) The number of TFBSs upstream of the original 5' end (blue) and of the 5' end updated with a TSS (pink) in HeLa cells. (D) The number of transfrags with a close poly(A) signal, AAUAAA, over the relative distances from the original 3' end (blue) and the 3' end updated with a CPS (pink) of transfrags in HeLa cells.

2.4. Integrating multimodal RNA-seq data

2.4.1. Co-assembly Followed by End-correction (CAFE) pipeline

We developed a transcriptome assembly pipeline, called CAFE (Co-assembly Followed by End-correction), which utilizes both stranded and unstranded RNA-seq data to reconstruction full-length transcripts by integrating CAGE-seq and 3P-seq data (Fig. 11). The CAFE pipeline consists of three main modules: (1) MAXIM - Prediction of directions of unstranded reads using MLE (see Section 2.2), (2) COCOA - Either re-assembly of transfrags with RPDs or co-assembly with both RPDs and stranded reads, (3) BEX - Construction of full-length transcripts by updating exon-junctions, TSSs and CPSs from transfrag models (see Section 2.3).

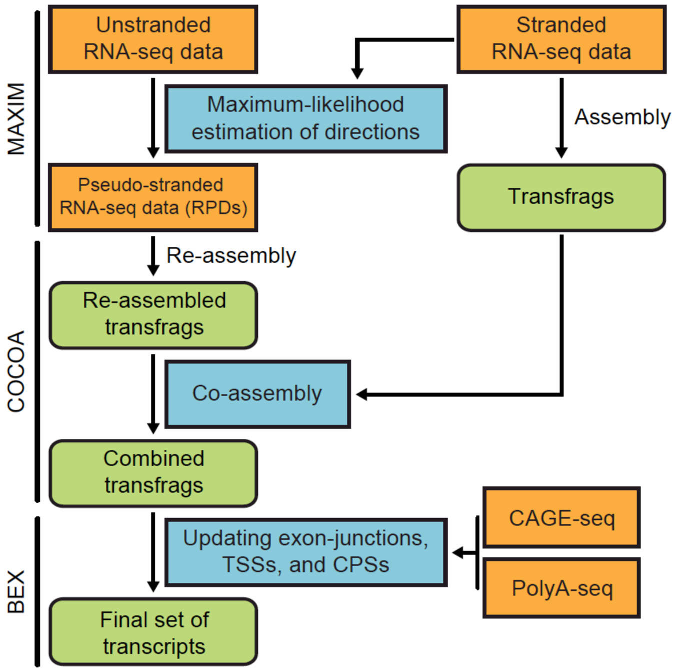


Figure 11. A schematic flow of the CAFE pipeline. Shown is the schematic flow of the CAFE pipeline according to the combined, pseudo-stranded (RPD) and stranded assembly. If there are both stranded and unstranded reads in the same cell type, the MAXIM, COCOA, and BEX steps are all executed. If there are only unstranded reads, the MAXIM step is carried out with the pooled stranded RNA-seq data.

2.4.2. CAFE improves transcriptome annotations

To evaluate the pipeline, we first sought to re-assemble only RPDs (named pseudo-stranded assembly) from HeLa and mES cells, and measured the accuracy of intermediate assembly at each step by comparing our results to GENCODE protein-coding genes in the base level (Fig. 12A). After updating TSSs and CPSs, the evaluation was proceeded with only transfrags with a major TSS and CPS while the count of transfrags took account of all isoforms. In total, 143,129 transfrags from 25,118 loci were assembled from HeLa cells; the quality of the resulting assembly for protein-coding genes was improved by about 14% for precision and about 1.6% for recall, compared to the original unstranded assembly (Fig. 12A). Similarly, CAFE assembled 164,423 transfrags from 24,605 loci in mES cells and improved the quality of protein-coding gene assembly by 18.4% for precision and 1.3% for recall (Fig. 12A). Although the resulting transfrags that overlapped with GENCODE lncRNAs were relatively less accurate than those of protein-coding genes partly because of their low and tissue-specific expression patterns, CAFE also improved the quality of such transfrags by 22.1% and 8.3% for precision in HeLa and mES cells, respectively, without compromising recall. A major factor behind the increased precision for both protein-coding and lncRNA genes was the prediction of read direction and re-assembly (Fig. 12A).

We next performed combined assembly (co-assembly) of both stranded reads and RPDs using CAFE. The resulting assemblies included 166,227 transfrags from 25,591 loci in HeLa cells and 244,085 transfrags from 26,332 loci in mES cells (Fig. 12B). Both the recall and precision of the final resulting transcriptome were greatly improved in the base level, compared to that in the original assembly (Fig. 12).

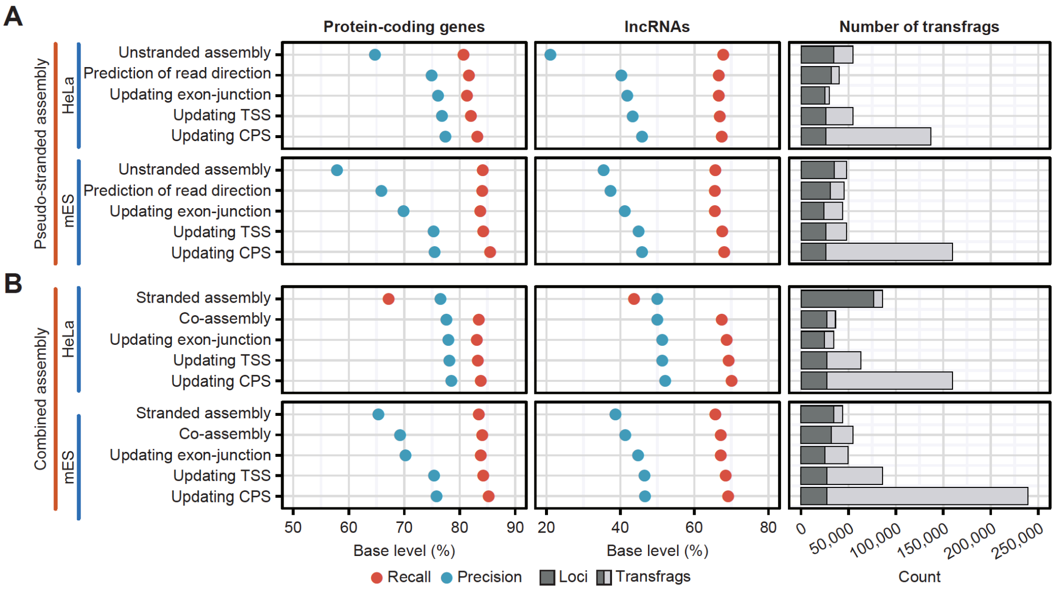


Figure 12. Step-wise evaluation of transcriptomes re-assembled by CAFE. (A) Shown are the accuracies and sizes of pseudo-stranded transcriptomes (RPD assembly) at each step of CAFE in HeLa (top) and mES cells (bottom). The recall (red solid circle) and precision (blue) of the assemblies are measured by comparing to GENCODE protein-coding genes (left panel) and lncRNAs (middle panel). The number of assembled transfrags and their loci are indicated at each step (right panel). (B) Shown are the accuracies and sizes of combined transcriptome assemblies of both stranded reads and RPDs. The low recall of the stranded assembly from HeLa cells is presumably because the stranded reads are of the single-end type and are 36 or 72 nt long. Otherwise, as in (A).

2.5. Benchmarking other transcriptome assemblers

To check whether the improvement in transcriptome assembly depends on a specific base assembler (originally, Cufflinks+CAFE), other reference-based assemblers, Scripture (*11*) and StringTie (*15*), were benchmarked using the same dataset (Scripture+CAFE and StringTie+CAFE). The resulting assemblies were more accurate for both HeLa (Fig. 13A; 8.6~9.9% greater recall and 11.4~12.9% greater precision) and mES cells (Fig. 13B; 3.2~4.9% greater recall and 10.2~10.6% greater precision) than the original assemblies in the base level. Additionally, two available *de novo* assemblers, Trinity (*19*) and Velvet (*17*), were also benchmarked by predicting the strand information of unstranded reads using CAFE and the resulting *de novo* assemblies of RPDs and stranded reads were more accurate than the original *de novo* assemblies (Figs. 13A and B). Taken together, CAFE was able to improve initial assemblies robustly regardless of the base assembler used.

The number of full-length transcripts is another important aspect in the quality of transcriptome assembly. We thus compared the number of full-length transcripts assembled by CAFE which include both a TSS and a CPS in the first and the last exons to the number in the original and de novo assemblies. Trinity+CAFE and Velvet+CAFE assembled 8.8~10.4% more full-length transcripts than in the original de *novo* assemblies (Fig. 13C). Cufflinks+CAFE, StringTie+CAFE, and Scripture+CAFE assembled 14.6%, 10.1%, and 13.9% more full-length transcripts than in the original assembly, respectively (Fig. 13D). Similarly, CAFE constructed more full-length transcripts than in the original and *de novo* assemblies from mES cells (Fig. 13D).

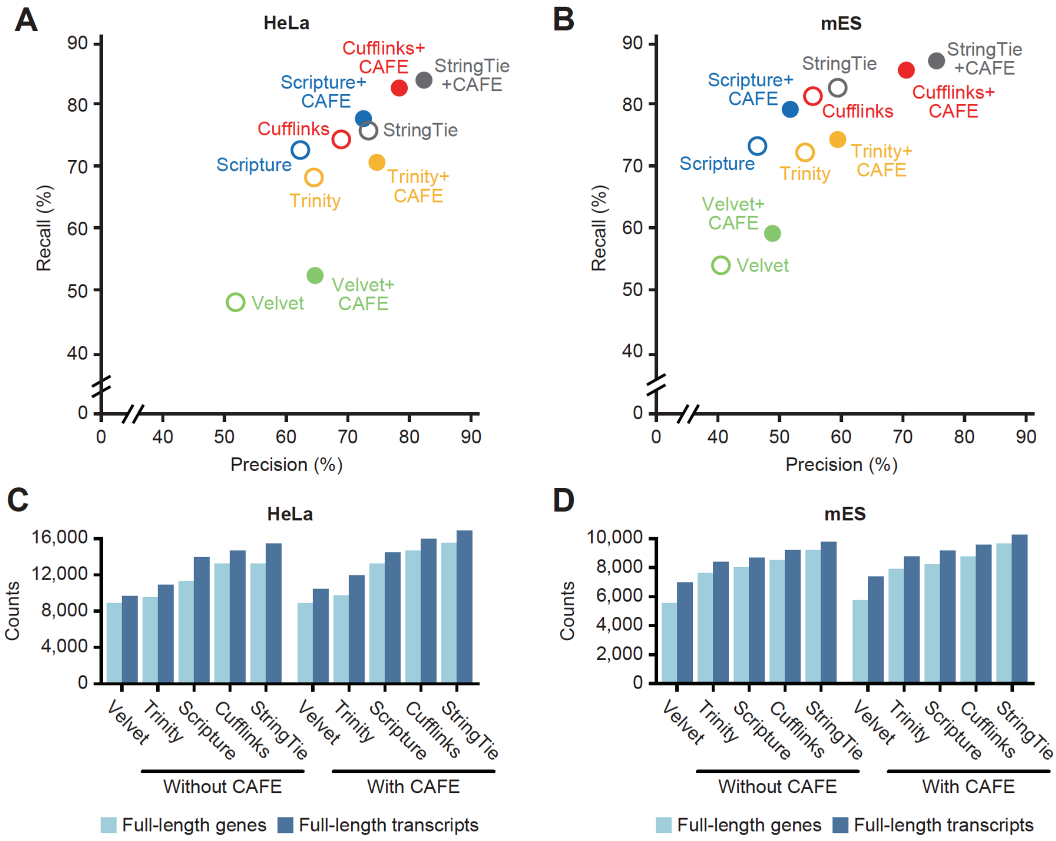


Figure 13. Benchmarking other base assemblers. (A-B) The accuracies of combined transcriptome assemblies (solid circles) reconstructed by CAFE with base assemblers and of the original transcriptome assemblies (open circles) reconstructed by respective base assemblers, such as Cufflinks (red), Scripture (blue), StringTie (grey), Velvet (green), and Trinity (yellow), in HeLa (A) and mES cells (B). The accuracies of the original assemblies were calculated by averaging the accuracies of stranded and unstranded assemblies reconstructed by each base assembler. Velvet and Trinity were used as *de novo* assemblers, and Scripture, StringTie and Cufflinks were used as reference-based assemblers. (C-D) The numbers of full-length genes (light blue) and transcripts (blue) in the co-assemblies were compared to those in the original assemblies from HeLa (C) and mES cells (D). For the original assemblies, the higher number of full-length genes in the stranded and unstranded original assemblies was chosen.

2.6. High-confidence human transcriptome map

2.6.1. Construction of high-confidence transcriptome map

To construct a comprehensive human transcriptome map, large-scale transcriptome data were collected from the ENCODE Project, the Human BodyMap 2.0 Project and NCBI GEO human cell lines; these data included 65 unstranded and 104 stranded RNA-seq data, TSS profiles across 17 human tissues, and CPS profiles from four human cell lines. We first predicted the directions of approximately six billion reads from 62 unstranded RNA-seq datasets using 60 cell-type-matched stranded RNA-seq datasets from 35 different cell types (Fig. 14A). The transcriptome assembly of the RPDs was more accurate than the unstranded transcriptome assembly in the base level (Fig. 14B), suggesting that the prediction of read directions significantly reduced erroneous transfrag assemblies. The co-assembly of RPDs and stranded reads with TSS and CPS profiles (Fig. 14A) reconstructed 338,359 transcripts from 46,634 loci, named BIGTranscriptome.

To examine their quality, BIGTranscriptome were compared to those of RefSeq, GENCODE (manual), GENCODE (automatic), Pacific Biosciences (PacBio) long read assembly (PacBio), and MiTranscriptome in terms of the number of full-length independent transcripts. Although BIGTranscriptome reconstructed fewer transcripts than did MiTranscriptome (Table 1A), it contained more (16,376, 35%) independent genes that had at least one transcript with boundaries defined by TSSs and CPSs than did MiTranscriptome (5,741, 6%) and GENCODE (manual: 6,522, 14%; automatic: 1,301, 7%) (Table 1B). Moreover, BIGTranscriptome included six thousand full-length independent single-exonic transcripts with a direction (~32.24% of single-exonic transcripts), whereas other annotations included tens of thousands of single-exonic transcripts, only 1~21% of which were full-length independent single-exonic transcripts (Table 2A). Thousands of those that remained appeared to be partial fragments that were included in BIGTranscriptome annotations (Table 2B).

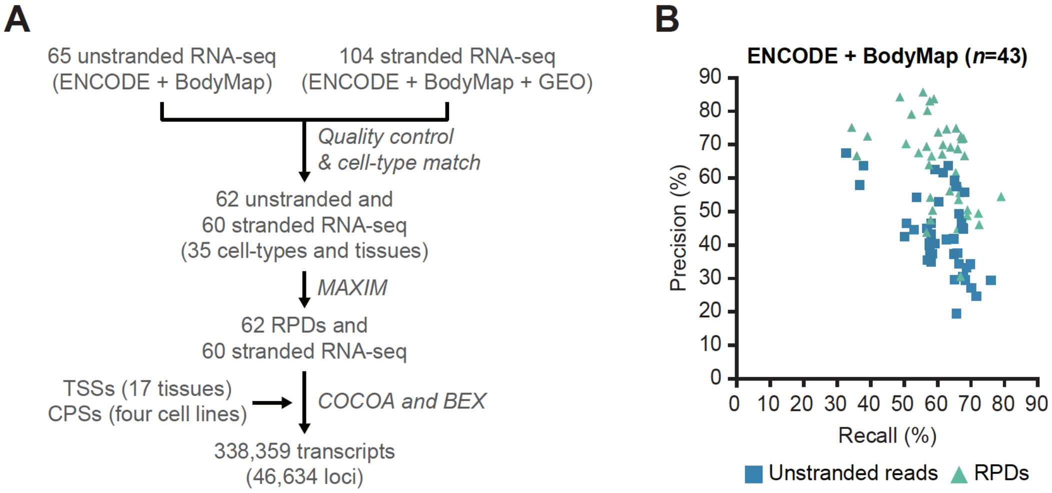


Figure 14. Comprehensive human transcriptome map. (A) A schematic flow for the reconstruction of the BIGTranscriptome map using large-scale RNA-seq samples from human cell lines, ENCODE, and Human BodyMap 2.0 Projects. (B) Accuracies of unstranded (blue) and RPD assemblies (mint) from the ENCODE and Human BodyMap Projects 2.0.

Annotations	Genes	Transcripts	Exons	Introns
RefSeq	23,982	42,709	230,903	212,045
GENCODE (manual)	47,971	175,025	534,198	336,853
GENCODE (automatic)	17,720	21,391	106,407	88,146
MiTranscriptome	91,013	384,016	730,761	524,525
PacBio (MCF7)	15,688	47,416	139,273	119,321
BIGTranscriptome	46,634	338,359	580,429	378,707

Table 1A. Statistics of transcriptome annotations.

Annotations	Genes	TSS and CPS	TSS or CPS	Not supported
RefSeq	23,982	7,702 (32%)	9,992 (42%)	6,288 (26%)
GENCODE (manual)	47,971	6,522 (14%)	17,562 (37%)	23,887 (50%)
GENCODE(automatic)	17,720	1,301 (7%)	4,831 (27%)	11,588 (65%)
MiTranscriptome	91,013	5,741 (6%)	21,851 (24%)	63,421 (70%)
PacBio (MCF7)	15,688	3,735 (24%)	8,749 (56%)	3,204 (20%)
BIGTranscriptome	46,634	16,376 (35%)	11,219 (24%)	19,039 (41%)

Table 1B. Independent genes supported by TSS and CPS tags.

Annotations	Single-exonic transcripts	Undetermined transcripts	Independent transcripts
RefSeq	3,937	0	458 (12%)
GENCODE (manual)	14,740	0	989 (7%)
GENCODE (automatic)	10,087	0	105 (1%)
MiTranscriptome	39,901	14,102 (35%)	1,216 (3%)
PacBio (MCF7)	5,107	0	1,068 (21%)
BIGTranscriptome	18,642	0	6,011 (32%)

Table 2A. Single-exonic transcripts from annotations.

Annotations	Putative partial fragments	Partial fragments confirmed by BIGTranscriptome
RefSeq	3,479	50 (1.4%)
GENCODE (manual)	13,751	55 (0.4%)
GENCODE (automatic)	9,982	364 (3.7%)
MiTranscriptome	38,685	1,434 (3.7%)
PacBio (MCF7)	4,039	67 (1.7%)

Table 2B. Partial fragments from annotations.

2.6.2. The annotation accuracy of BIGTranscriptome map

The accuracy of BIGTranscriptome annotations was evaluated at the base level in terms of recall and precision based on RefSeq, GENCODE (manual), GENCODE (automatic), or PacBio (MCF7) annotations. BIGTranscriptome annotations were found to be 14.7 ~ 36.7% more precise for the RefSeq and GENCODE (manual) transcripts than were MiTranscriptome annotations, without compromising recall (Fig. 15A). We also checked if the intron structures of BIGTranscriptome agreed with those of the RefSeq, GENCODE, expression sequence tags (ESTs), PacBio, and combined annotations (RefSeq + GENCODE + EST + PacBio), and compared the results to those of MiTranscriptome. Overall, our BIGTranscriptome annotations were superior to those of MiTranscriptome for both recall (22.6% greater) and precision (40.5% greater) in the combined annotations (Fig. 15A), indicating that BIGTranscriptome transcripts are less likely to be fragmented.

87.0% of the 29,274 putative BIGTranscriptome introns, not detected in the combined annotations, included a canonical splicing signal, ‘GU’-‘AG’, two nucleotides away from both ends; the remainder lacked the canonical splice signal. Although the putative introns of MiTranscriptome also included the canonical splice signals at a similar level as BIGTranscriptome, the putative splice sites of MiTranscriptome showed significantly lower maximum entropy scores than those of BIGTranscriptome at both splice donor and acceptor sites (Figs. 15B and C).

To evaluate the accuracy of BIGTranscriptome transcript boundaries, we counted TFBSs in the regions upstream of the TSSs and canonical poly(A) signals in the regions around the CPSs. A higher fraction of TFBSs within 500nt upstream of a TSS (Fig. 15D) and poly(A) signals within 15-30nt upstream of a CPS (Fig. 15E) were observed for BIGTranscriptome transcripts than for MiTranscriptome and GENCODE (automatic), indicating that BIGTranscriptome includes transcripts with more precise ends. However, because the CPS information was profiled from only four human cell types, we additionally updated the cell-type specific 3’ ends of transcripts using GETUTR, which predicts the 3’ end of a transcript from RNA-seq data (*67*).

A

Annotations (%)	BIGTranscriptome				MiTranscriptome			
	Base level		Intron level		Base level		Intron level	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
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

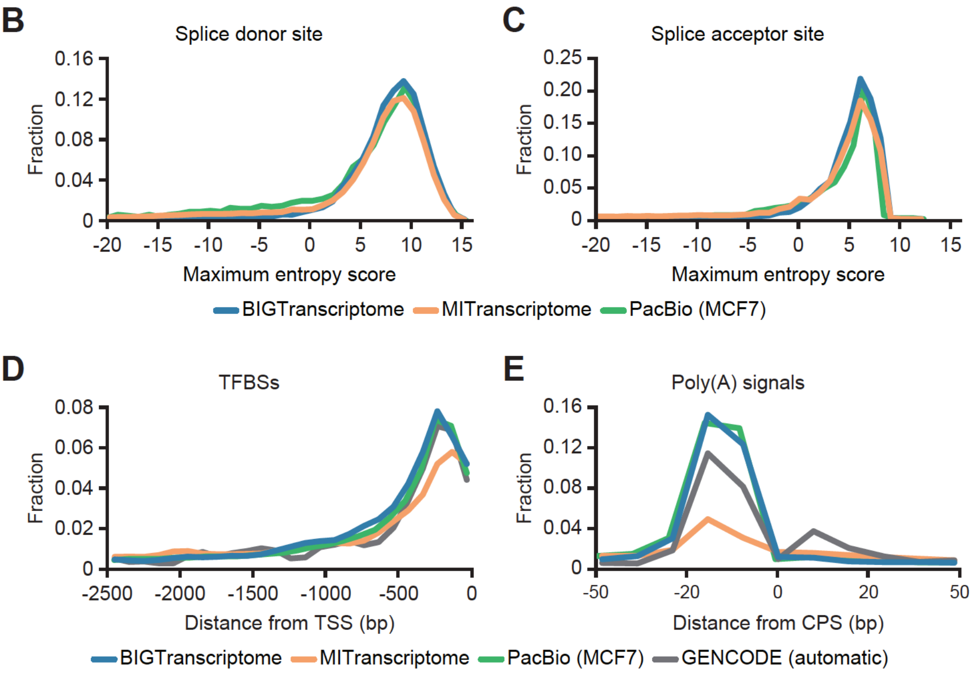


Figure 15. The annotation accuracy of BIGTranscriptome. (A) Shown are the accuracies of BIGTranscriptome and MiTranscriptome at the base and intron levels based on four different sets of annotations (RefSeq, manual and automatic GENCODE, PacBio, and EST), and a combined set of annotations. (B-E) Maximum entropy scores of the putative splice donor sites (B) and of putative splice acceptor sites (C). Blue lines are from BIGTranscriptome, green lines are from PacBio assembly, and orange lines are from MiTranscriptome. (D) The fraction of TFBSs upstream of the 5' end of BIGTranscriptome transcripts (blue) was compared to those of MiTranscriptome (orange), GENCODE (automatic) (black), and PacBio assembly (green). (E) The fraction of the closest poly(A) signals, AAUAAA, in the region just upstream of the 3' end of BIGTranscriptome annotations (blue) compared to those of MiTranscriptome (orange), GENCODE (automatic) (black), and PacBio assembly (green).

2.6.3. Impact of downstream analysis based on accurate annotation

We sought to examine whether our BIGTranscriptome annotations could benefit the expression profiling of genes and their downstream analysis. *T222734* was annotated as a single form in MiTranscriptome but this sequence turned out to be an independent protein-coding gene, *PRPF6*, and a lncRNA, *LINC00176*, evident with CAGE-seq and 3P-seq, in BIGTranscriptome (Fig. 16A). Using the single and the two independent forms of the genes, we performed Kaplan-Meier survival analyses for 164 liver cancer samples from the TCGA Project. We found that the *PRPF6* gene is a more significant marker (log rank test *P* = 0.0006; Fig. 16C) for the prognosis of the liver cancer patients than *T222734* (log rank test *P* = 0.003; Fig. 16B), whereas *LINC00176* is expressed at a low level and is not significant marker (log rank test *P* = 0.3; Fig. 16D). Similarly, *AC15645* (lncRNA) and *MLXIP* (protein-coding gene) were annotated in BIGTranscriptome but they were annotated as a single form (*T087998*) in MiTranscriptome (Fig. 17A). The *MLXIP* annotated in our BIGTranscriptome appeared to be a more significant prognosis marker (Fig. 17E) than *T087998* and *T088004* annotated in MiTranscriptome (Fig. 17B and C) but the lncRNA, *AC15645* turned out to be expressed at a low level (Fig. 17D).

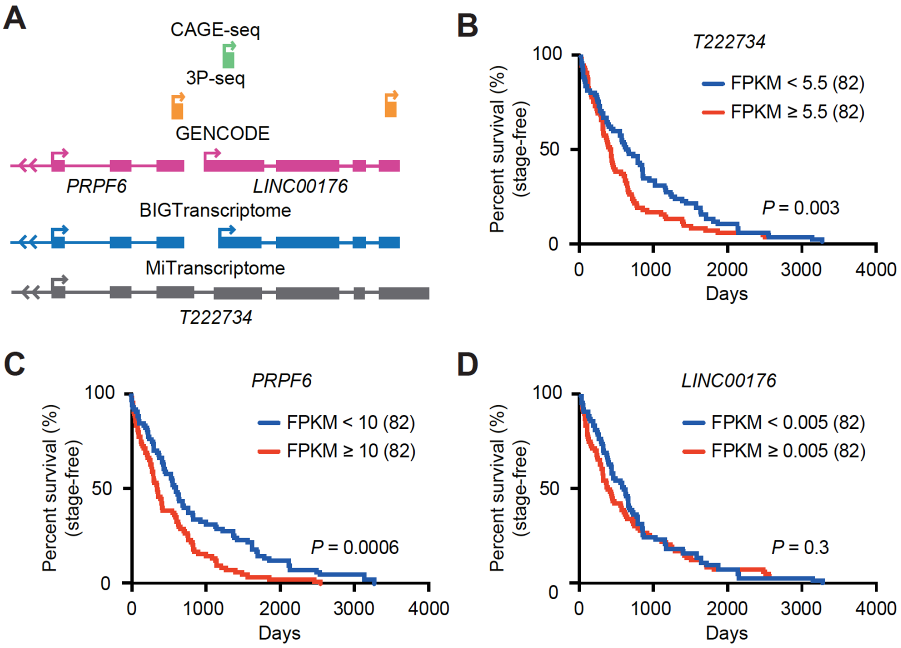


Figure 16. Mis-annotated gene model in MiTranscriptome. (A) Examples of mis-annotated gene model in MiTranscriptome. (A) Gene models of BIGTranscriptome and MiTranscriptome, and CAGE-seq and 3P-seq data, at a locus. A fused single form, *T222734*, was annotated in MiTranscriptome whereas two independent genes, *PRPF6* and *LINC00176*, were annotated in BIGTranscriptome. (B-D) Survival analyses for TCGA liver cancer samples based on the resulting gene models. 164 patient samples including termination events were divided into two groups, the top 50% (red) and bottom 50% (blue), by the median FPKM values of *T222834* (B), *PRPF6* (C), and *LINC00176* (D).

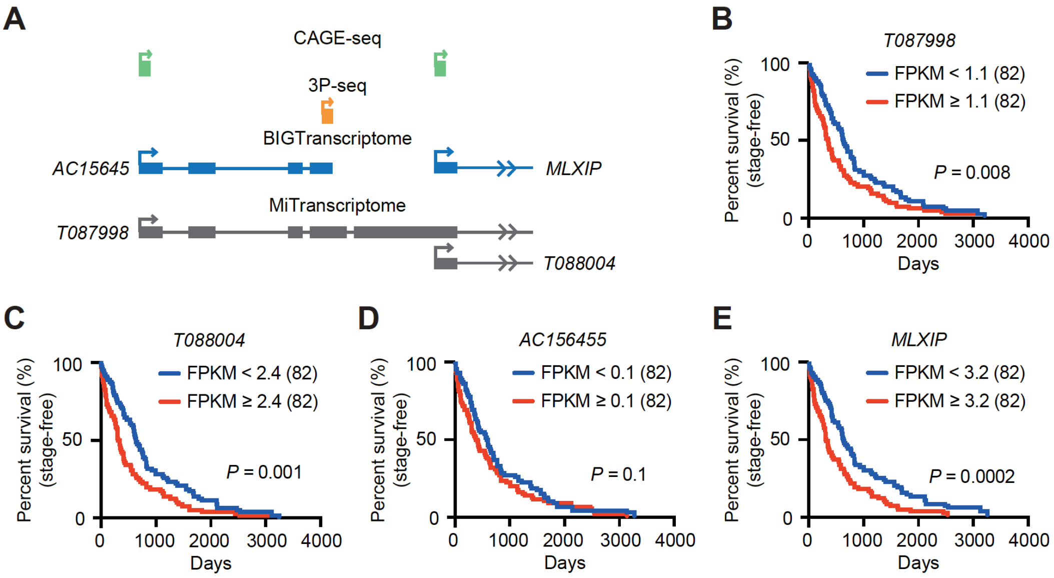


Figure 17. Survival analysis for liver cancer samples based on mis-annotated transcripts. (A) Gene models of a protein-coding gene, *MLXIP*, and a lncRNA, *AC156455*, in BIGTranscriptome and MiTranscriptome. (B-E) Survival analyses for TCGA liver cancer patient samples based on the gene models. 164 samples including termination events were divided into two groups, the top 50% (red) and bottom 50% (blue), according to the median FPKM values of *T087998* (B), *T088004* (C), *AC156455* (D), and *MLXIP* (E).

Chapter 3. Identifying ESCC-driving lncRNAs based on reconstructed transcriptome map

3.1. A comprehensive set of dysregulated lncRNAs in ESCC

Contribution Statement: In this session, Bo-Hyun You performed the all sequencing data and bioinformatics analyses, supervised by Jin-Wu Nam.

3.1.1. Reconstruction of transcriptome map and annotation of lncRNAs

To construct a comprehensive set of lncRNAs in esophageal squamous cell carcinoma (ESCC), RNA-seq performed from paired cancerous and non-cancerous tissues of 13 ESCC patients in the Yonsei Severance Hospital (YSH) cohort and was subjected to transcriptome assembly, and lncRNA annotation (Fig. 18). Transcriptome assembly was performed from RNA-seq data in the YSH cohort using the CAFE pipeline (version 1.0.1) (*29*). A total of 50,474 transcripts from 24,228 loci were constructed and classified as known or novel transcripts based on their overlap with transcripts from GENCODE v19 annotation. To annotate lncRNA genes, we performed the following filtration steps: [1] transcripts shorter than 200nt in length were discarded and [2] transcripts sense-overlapping with exons of known genes were excluded. We assessed the coding potential of the remaining transcripts using two independent coding potential calculators: [1] CPC, which is a BLASTX-based method that performs similarity searches against all non-redundant protein sequences from multiple species (*68*) and [2] CPAT, which is an alignment-free method for calculating coding potential using sequence-based features (*69*). Of the remaining transcripts, those with a CPC score < 0 and a CPAT score < 0.364 were defined as novel lncRNAs. The final lncRNA catalogue was made by combining the novel and known lncRNAs from GENCODE v19 annotation. In total, 6411 lncRNAs from 4842 known loci and 1924 from 1657 novel loci were annotated.

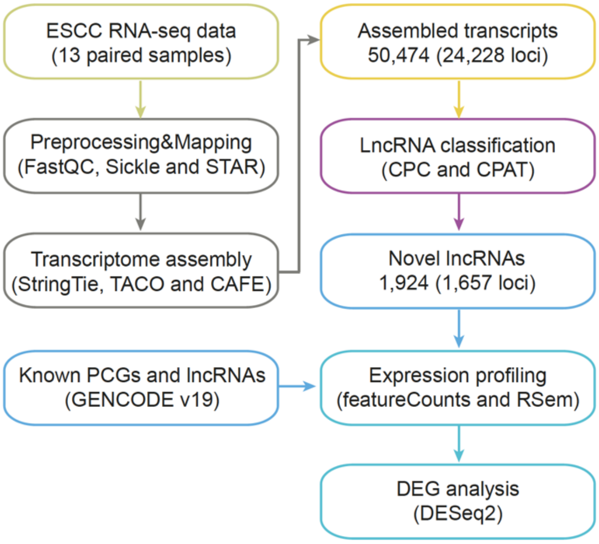


Figure 18. Annotation and expression analysis of lncRNAs in ESCC*.* A schematic flow for lncRNA annotation and expression profiling of novel and known lncRNAs from the YSH ESCC cohort.

3.1.2. Expression profiling of lncRNAs from multiple cohorts

Using the resulting annotations of known and novel lncRNAs, lncRNA expression levels were measured over the YSH cohort, publicly available ESCC cohorts (95 tumor samples from TCGA and 15 paired samples from a Chinese ESCC cohort) (*61, 70*), and GTEx Esophagus mucosa datasets (328 samples) (*24*). Because the RNA-seq data from the TCGA ESCC and GTEx Esophagus mucosa datasets were unstranded type, the strand information was predicted and the unstranded reads were converted to RPDs using CAFE (*29*).

3.1.3. An ethnically independent set of DE lncRNAs

Of the total of 8335 lncRNAs, 465 (305 upregulated and 160 downregulated) were significantly dysregulated in ESCC from YSH cohort, exhibiting greater than two-fold differences in expression, with a false discovery rate (FDR) $\leq$ 0.01 (Fig. 19A). Then, to identify an ethnically independent set of DE lncRNAs in ESCC, 113 DE lncRNAs commonly dysregulated in all three ESCC cohorts (YSH, GTEx+TCGA, and Chinese) were selected (Fig. 19A). Of the 113 confidence DE lncRNAs, 20 were newly annotated, 32 were upregulated, and 81 were downregulated in ESCC (Fig. 20). A majority of the confidence DE lncRNA genes were either located in intergenic regions or were antisense to other genes (Fig. 19B); their genomic and clinical features, such as subcellular localization of the lncRNA (Fig. 19C), associations with enhancers (Fig. 19D*)*, and DNA methylation (Fig. 19E), were systematically examined. As previously reported, many overlapped with enhancers (Fig. 19D) and seemed to be associated with epigenetic markers of other genes (Fig. 19E).

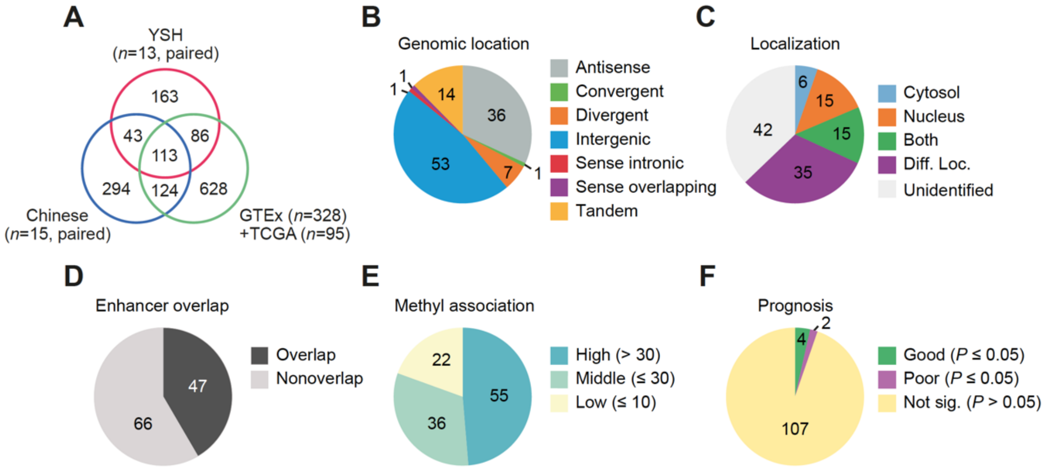


Figure 19. Characterization of a common set of DE lncRNAs in ESCC*.* (A) Venn diagram of the DE lncRNAs detected in three ESCC cohorts (YSH, Chinese, and GTEx+TCGA). The pie charts (B-F) show the number of DE lncRNAs categorized according to their genomic location (B), subcellular localization (C), enhancer overlap (D), association with DNA methylation (E), and prognostic power (F).

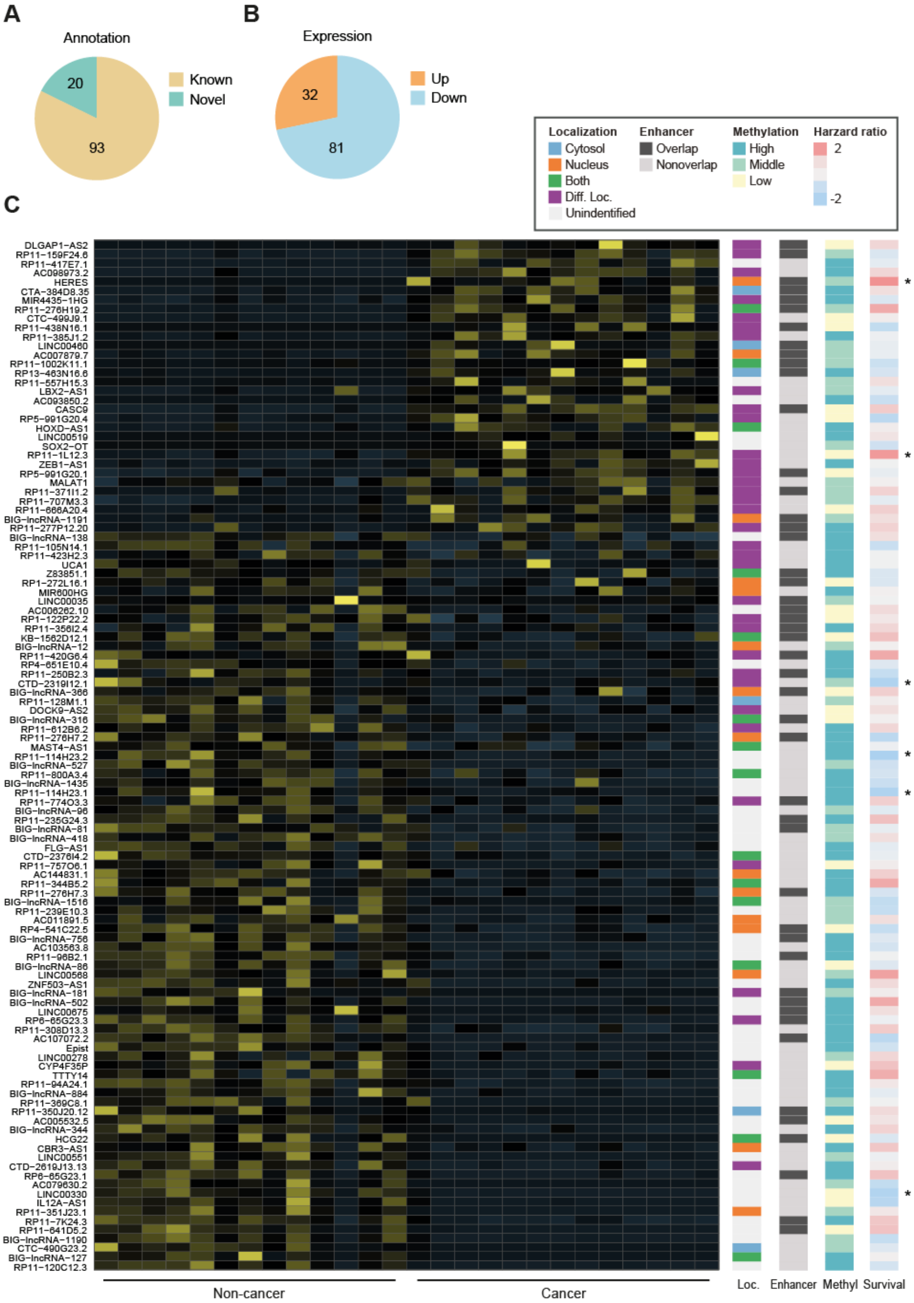


Figure 20. Characteristics of common DE lncRNAs in distinct ESCC cohorts. (a) The numbers of known and novel DE lncRNAs are shown in a pie chart. (b) The expression change patterns of these lncRNAs are shown in a pie chart. (c) The heatmap represents the expression levels of 113 DE lncRNAs in non-cancer and cancer samples from the YSH cohort (left panel). The right panels show the subcellular localization (Loc.), enhancer overlap (Enhancer), DNA methylation association (Methyl), and associated hazard ratio (Survival) of 113 DE lncRNAs. *$P \leq 0.05$.

3.1.4. Association of DE lncRNAs with clinical outcomes

To find clinically relevant lncRNAs that are associated with survival outcomes of patients, Kaplan-Meier survival analyses for all 113 DE lncRNAs were performed with TCGA ESCC datasets comprising 95 patient samples (Fig. 19F). Six DE lncRNAs were significantly associated with survival rates, two (HERES and RP11-1L12.3) of which were associated with a high hazard ratio (HHR; $P \leq 0.05$) and four (RP11-114H23.1, RP11-114H23.2, CTD-2319I12.1, and LINC00330) of which were associated with a low hazard ratio (LHR; $P \leq 0.05$). To delineate how the expression of the six lncRNAs stratifies ESCC patients, we clustered samples from the TCGA and YSH cohorts including additional 10 RNA-seq samples where the clinical values were available based on the binary expression patterns (high and low) of the six lncRNAs, revealing four distinct classes of patients: class L1~L4 (Fig. 21). Noticeably, class L1, in which only the HHR markers are highly expressed, showed a worse survival rate than class L3 ($P < 0.05$) and the other classes ($P = 0.01$; Fisher's exact test). Class L3 tended to display a greater overall survival rate than other classes ($P < 0.05$; Fisher's exact test). Importantly, class L1 appeared to be significantly associated with smoking ($P < 0.05$; Fisher's exact test), compared to other classes. Taken together, these results indicate that the six lncRNAs represent prognostic signature genes that can stratify ESCC patients based on clinical outcomes.

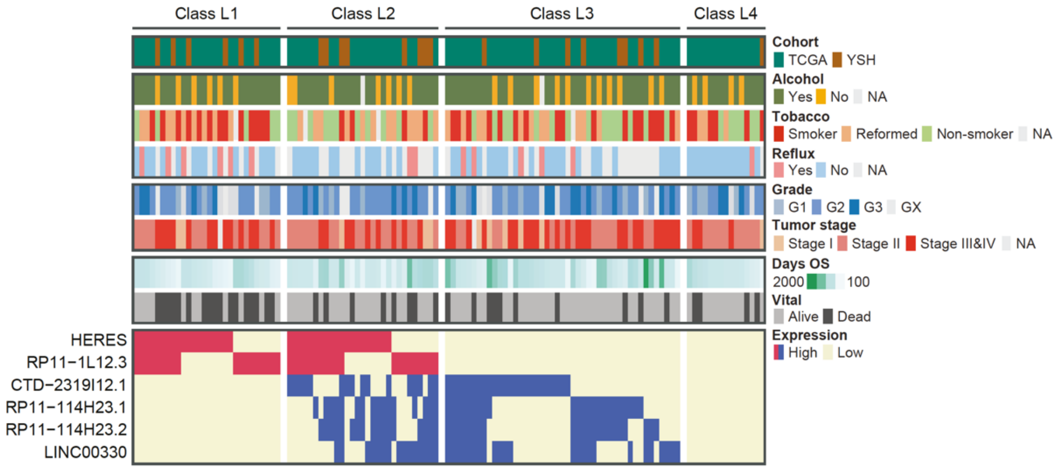


Figure 21. The four ESCC subclasses with different clinical outcomes. The four ESCC subclasses based on the six prognostic marker lncRNAs in Fig. 19F. The top section presents cohort information, clinical history, pathological features, and survival information from the YSH and TCGA ESCC patients. The various categories are represented as different colors, as shown in the legend on the right. The expression patterns of the six prognosis-related lncRNAs in the RNA-seq datasets are shown with a colored heatmap in the bottom section (red indicates the top 33% highly expressed lncRNAs associated with a HHR; blue indicates the top 33% highly expressed lncRNAs associated with a LHR).

3.2. A novel lncRNA, HERES, is upregulated in ESCCs

Contribution Statement: In this session, Bo-Hyun You performed the all sequencing data, survival, and conservation analyses, supervised by Jin-Wu Nam. Jung-Ho Yoon (Yonsei University) performed all experiments, supervised by Sang Kil Lee.

3.2.1. *HERES* encodes alternative splicing isoforms

Since HERES, one of the HHR markers (see Section 3.1.4), was greatly upregulated in ESCCs compared to paired adjacent non-cancerous samples (Fig. 22) and most strongly associated with poor vital status (Fig. 21), we investigated whether HERES might be an ESCC-driving lncRNA. The lncRNA gene encodes two isoforms with the same transcription start sites (TSSs) (*35*) and cleavage and polyadenylation sites (CPSs) (*66*) (Fig. 23A). Isoform #1, HERES.1, is 2160nt and contains two exons, whereas isoform #2, HERES.2, is an intron-retained, single-exonic transcript that is 6675nt. Both isoforms were confirmed to lack coding potential (Fig. 23A). Analysis of ESCC RNA-seq datasets (Fig. 24A) and quantitative RT-PCR (qRT-PCR) in an ESCC cell line (KYSE-30) (Fig. 23B) showed that HERES.1 is the major isoform. Only a short region in *HERES* exon 1 displays sequence conservation with a region in the mouse genome; the intergenic *HERES* locus between the *GLS* and *NAB1* genes on chromosome 2 is syntenically conserved in mouse (Fig. 24B).

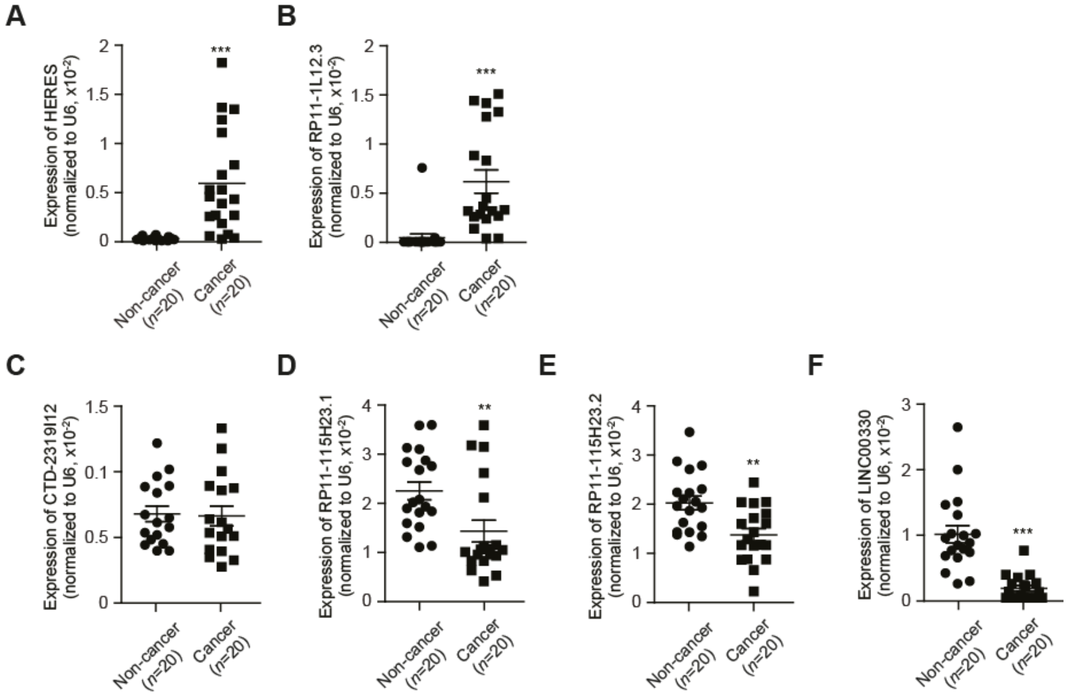


Figure 22. Expression levels of six prognostic lncRNAs in ESCC tissues. (A-F) Expression levels of six prognostic lncRNAs were evaluated using qRT-PCR in paired cancer and non-cancer frozen tissue samples. T-test was used to estimate significance. ***P* ≤ 0.01, ****P* ≤ 0.001.

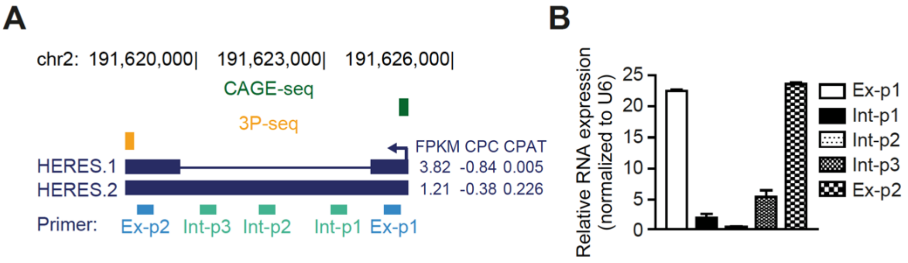


Figure 23. HERES encodes two isoforms. (A) The *HERES* genomic locus with CAGE-seq and 3P-seq signals. qRT-PCR primer sets were designed to recognize exonic (two) and intronic regions (three). The coding potentials calculated by CPC and CPAT are indicated on the right. (B) qRT-PCR results using the five primer sets in KYSE-30 cells. (B) Error bars represent the mean ± SD from three independent experiments.

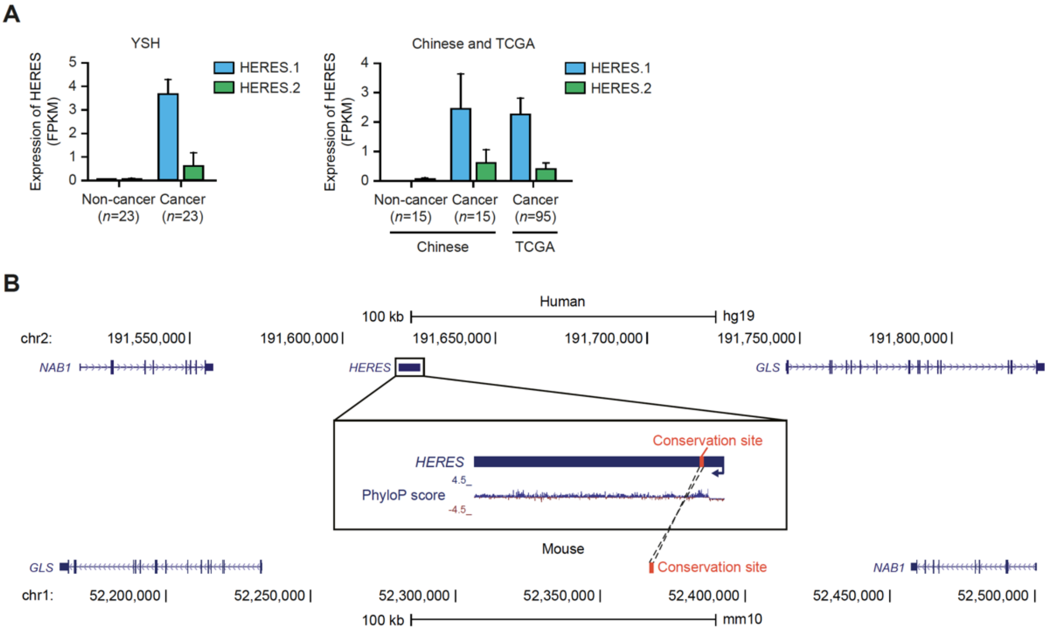


Figure 24. Gene locus and isoform-level expression of HERES. (A) Expression levels of two HERES isoforms (HERES.1 and HERES.2) in non-cancer and cancer tissues from YSH (*n*=23, paired), Chinese (*n*=15, paired), and TCGA ESCC cohorts (*n*=95). (B) The HERES genomic locus with sequence and positional conservation between human and mouse shown. The region of sequence conservation is indicated with a red box.

3.2.2. HERES upregulated in squamous-cell-type cancers

Elevated HERES expression was then validated in other ESCC cell lines. Compared to that in a normal esophageal epithelial cell line (Het-1A), HERES expression appeared to be upregulated greater than 10-fold in all tested ESCC cell lines (Fig. 25A), as observed in ESCC samples (Fig. 25B). HERES was significantly upregulated not only in ESCC, but also in esophageal adenocarcinoma (ESAD) and other squamous carcinomas (head and neck, and lung squamous cell carcinoma; HNSC and LUSC), but not in lung adenocarcinoma (LUAD) (Fig 26A). These expression changes were further confirmed in 66 ESCC samples from the YSH cohort using qRT-PCR, which revealed an elevated level of HERES in cancers compared to adjacent non-cancerous samples (Fig. 25C, left panel). HERES expression in the adjacent non-cancerous samples was higher than that in normal mucosa from the normal population (Fig. 25C, right panel). As observed in the YSH cohort (Fig. 25D, left panel), HERES levels were significantly correlated with stage-free survival rates in the TCGA ESCC cohort (Fig. 25D; right panel), and multivariate analysis with clinical information revealed that the HERES level was strongly associated with tumor grade (*P* = 0.004; Fisher’s exact test) but not with other clinicopathological factors (Fig 26B).

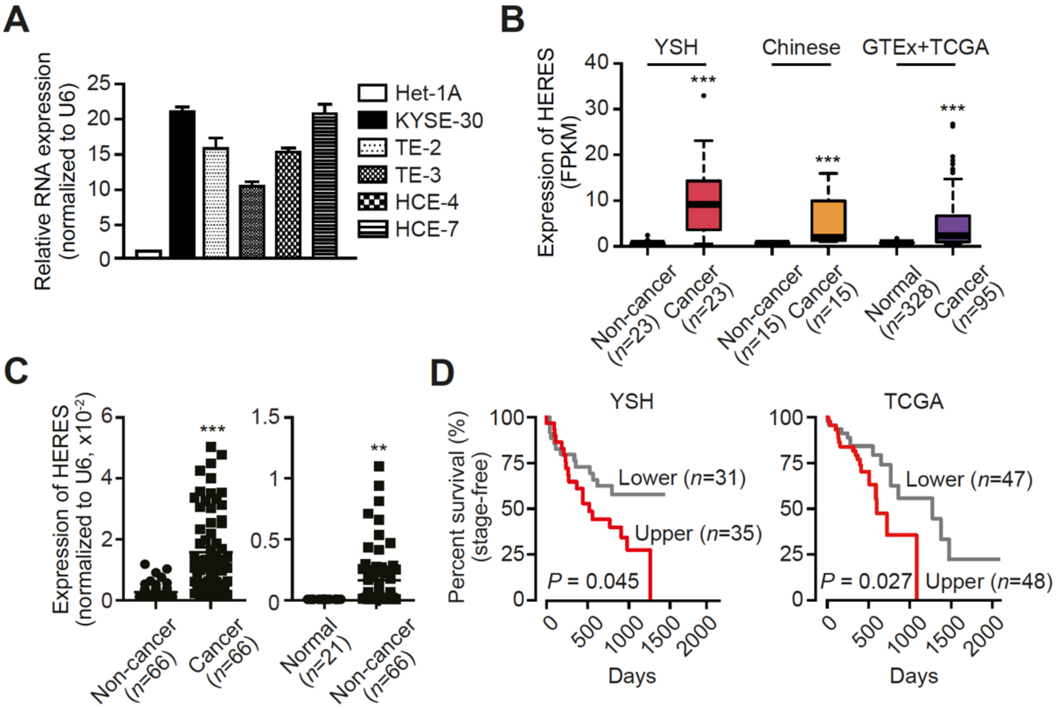
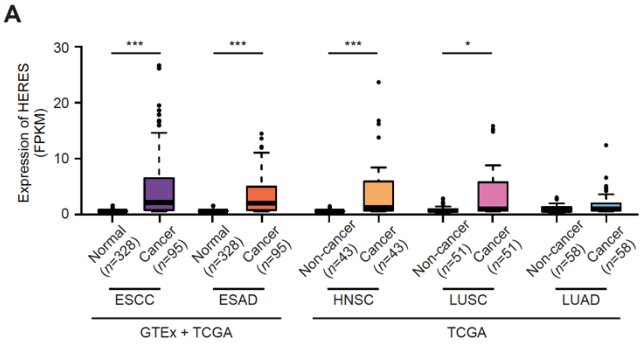


Figure 25. HERES is a highly expressed lncRNA in ESCC. (A) The HERES expression level was measured in five ESCC cell lines and a normal esophageal cell line (Het-1A). (A) Error bars represent the mean ± SD from three independent experiments. (B) The box plots show the HERES expression levels in normal, non-cancerous, and cancerous tissues from the YSH cohort (paired), the Chinese cohort (paired), and the TCGA cohort. (C) HERES expression levels measured by qRT-PCR in additional frozen tissue samples including YSH ESCC (*n*=66) and adjacent samples (non-cancer; *n*=66) (left panel) and in normal mucosa tissues (*n*=21) from reflux symptom patients (right panel). (D) Survival analyses of YSH and TCGA patients from whom the ESCC samples were obtained based on the HERES expression level. ***P* ≤ 0.01, ****P* ≤ 0.001.



B

Factors	Values	# of patients	*P*
Gender	Male / Female	95	0.386
Race	White / Asian	86	0.829
Reflux history	Yes / No	78	0.774
Smoking history	1,2 / 3,4	91	0.659
Alcohol history	Yes / No	93	0.229
Pathologic history	I,II / III,IV	92	0.826
Tumour grade	G1,G2 / G3	95	0.004

Figure 26. HERES upregulated in squamous-cell-type cancers. (A) The box plots show the HERES expression levels in normal, non-cancer, and cancer tissues from the GTEx and TCGA cohorts. (B) Association of clinicophathological features and HERES expression in TCGA ESCC patients. Fisher's exact test was used to estimate significance. **P* ≤ 0.05, ****P* ≤ 0.001.

3.3. HERES promotes cancer development and progression

Contribution Statement: In this session, Jung-Ho Yoon (Yonsei University) performed all experiments, supervised by Sang Kil Lee.

To investigate whether HERES is involved in cancer development and progression, the effects of HERES knockdown on cell proliferation, migration, invasion, and colony formation were explored with siControl- and siHERES-treated cells. The proliferation indices (Optical Density [O.D.] values) were significantly reduced in both siHERES_1 and siHERES_2-treated cells compared to siControl-treated cells (Fig. 27A), and the introduction of a HERES pcDNA expression construct partly rescued the proliferation activity (Fig. 27A), indicating that HERES can regulate cell proliferation. Migration and invasion assays showed that both cell migration and invasion were greatly reduced in siHERES-treated cells compared to siControl-treated cells (Figs. 27B and C). In addition, HERES knockdown also reduced colony formation measured at 14 days after siRNA transfection (Fig. 27D). A role for HERES in malignant ESCC progression was confirmed by the reduction of N-cadherin and vimentin levels in siHERES-treated cells and by the rescue of these levels by the introduction of the HERES pcDNA construct to the cells (Fig. 27E). These results suggest that HERES can promote cancer progression and metastasis.

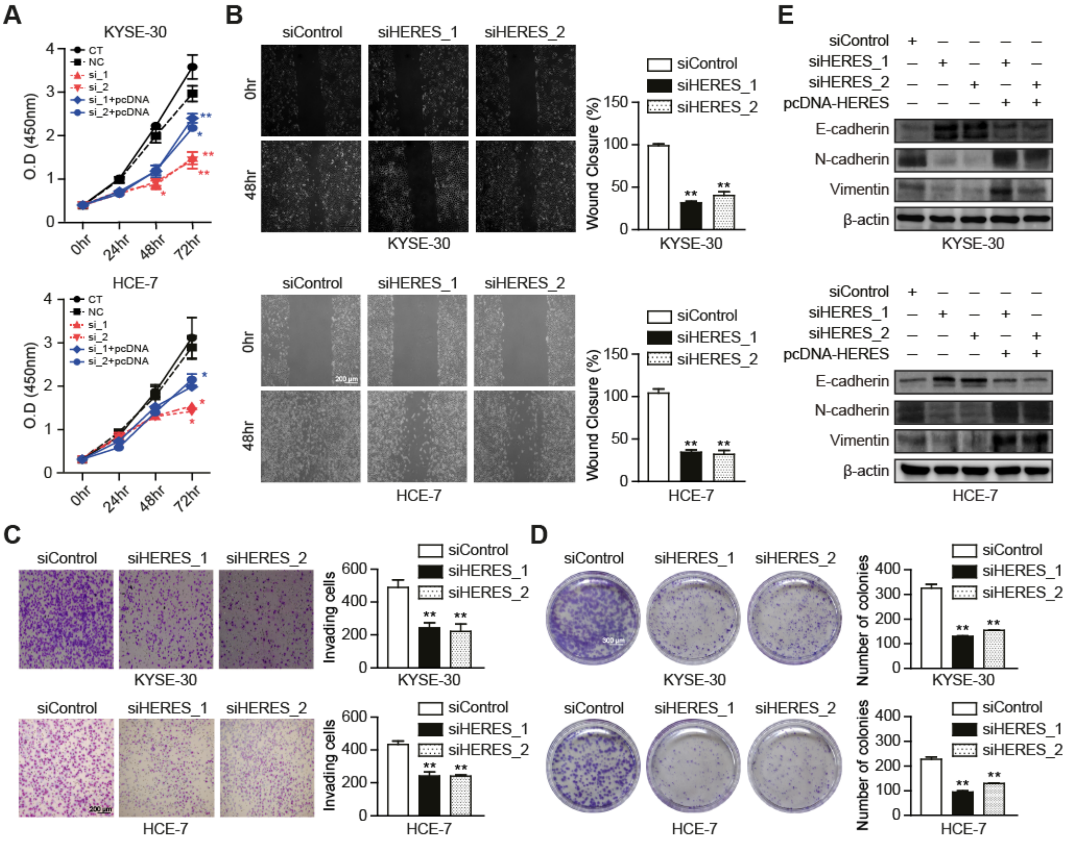


Figure 27. HERES modulates cell proliferation, migration, invasion, and colony formation. (A) Cell viability was measured using an MTS assay in KYSE-30 and HCE-7 cells transfected with siControl (NC), siHERES (si_1 and si_2), or siHERES followed by pcDNA-HERES (si_1+pcDNA, si_2+pcDNA). Growth curves were compared between siHERES- and siControl-transfected cells, and between pcDNA-HERES+siHERES- and siHERES-transfected cells. Wound healing assays (B), invasion assays (C), and colony formation assays (D) were performed in KYSE-30 and HCE-7 cells after HERES knockdown. The bar graphs represent the frequency of wound closure (B) and the number of invading cells (C) and colonies formed (D). Data represent the mean ± SD from three independent experiments (A-D). *$P \leq 0.05$, **$P \leq 0.01$. (E) Expression of EMT markers in KYSE-30 and HCE-7 cells transfected with siControl or the indicated combinations of siHERES and pcDNA-HERES as determined by immunoblot.

3.4. HERES regulates Wnt signaling pathways

Contribution Statement: In this session, Bo-Hyun You performed NanoString and sequencing data analyses, supervised by Jin-Wu Nam. Jung-Ho Yoon (Yonsei University) performed all experiments, supervised by Sang Kil Lee.

3.4.1. HERES affects Wnt signaling pathway-related genes

To study the means by which HERES promotes cancer development and progression, the changes in expression of ~730 cancer-related pathway genes were analyzed in siHERES-treated and siControl-treated KYSE-30 cells using the NanoString nCounter PanCancer Pathways Panel (Fig. 28A). 77 cancer-related genes (34 for up-regulation and 43 for down-regulation) were dysregulated greater than two-fold in siHERES-treated cells compared to siControl cells (Fig. 28A); the expression changes of the two most upregulated genes (*CACNA2D3* and *SFRP2*) and the two most downregulated genes (*BMP7* and *GRIN1*) in this group were confirmed by qRT-PCR (Fig. 29). Noticeably, among the genes dysregulated by HERES reduction, 14 belong to the Wnt signaling pathway, and half of the 10 most upregulated genes (*CACNA2D3*, *SFRP2*, *CACNA1E*, *CXXC4*, and *SFRP4*) are involved in the Wnt signaling pathway. *CACNA2D3*, which encodes a subunit of the calcium channel protein complex, was previously shown to be induced in ESCC (*71*) and other cancers (*72, 73*) via epigenetic mechanisms, and its downregulation led to inactivation of Wnt/Ca^{2+} signaling pathway (*73*). *SFRP2* encodes a member of the SFRP family that modulates the Wnt signaling pathway; *SFRP2* hypermethylation is known to enhance cell invasiveness in both cancers and non-cancerous diseases (*74, 75*). The enrichment of canonical and noncanonical Wnt signaling pathway-related genes among the genes that respond to HERES depletion, together with results from previous studies, suggest that HERES may regulate cancer development via control of Wnt signaling pathways.

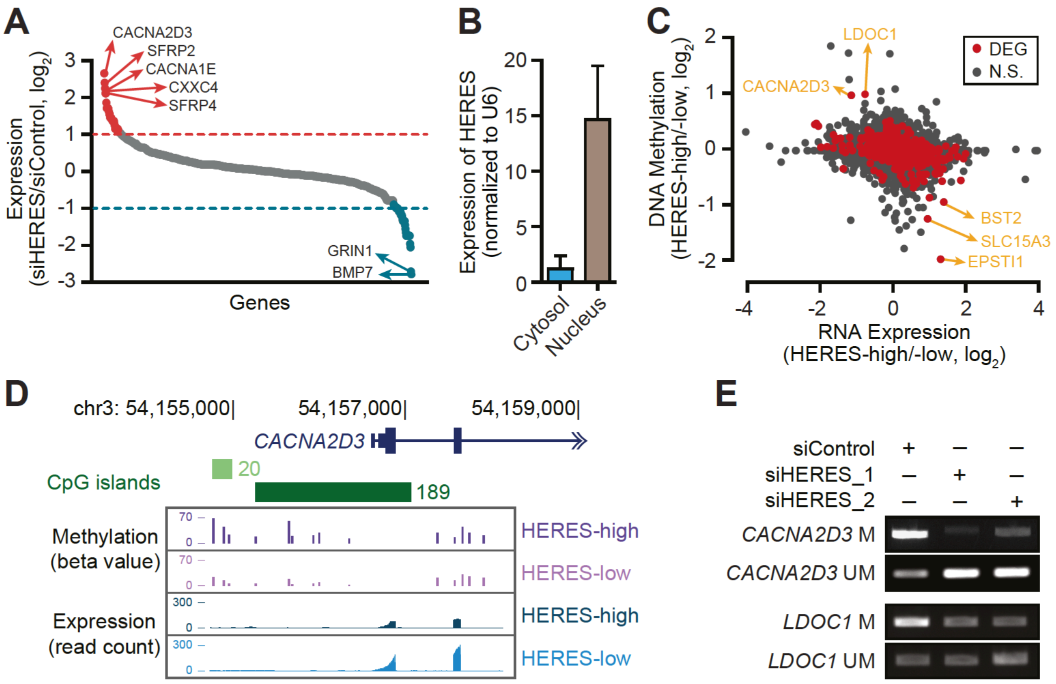


Figure 28. HERES affects Wnt signaling pathway-related genes. (A) Changes in the expression of cancer-related genes in response to siHERES treatment compared to siControl are shown. The colored circles indicate genes that are upregulated (red) or downregulated (blue) under HERES-depleted conditions. Changes in the expression of the highlighted genes were experimentally confirmed by qRT-PCR. (B) HERES expression in the nuclear and cytoplasmic fractions of KYSE-30 cells as determined by qRT-PCR. (C) Log-scaled fold-changes of expression (X-axis) and DNA methylation (Y-axis) of each gene in the HERES-high versus HERES-low sample groups from the TCGA dataset. The red dots indicate DE genes in ESCC. The highlighted genes are those for which there is anti-correlation between expression and DNA methylation. (D) The *CACNA2D3* genomic locus with CpG island tracks, DNA methylation (beta value), and RNA expression (read count) in the HERES-high and HERES-low groups. (E) *CACNA2D3* and *LDOC1* DNA methylation patterns (methylation (M) and unmethylation (UM)) were measured by MS-PCR in KYSE-30 cells transfected with siControl or siHERES.

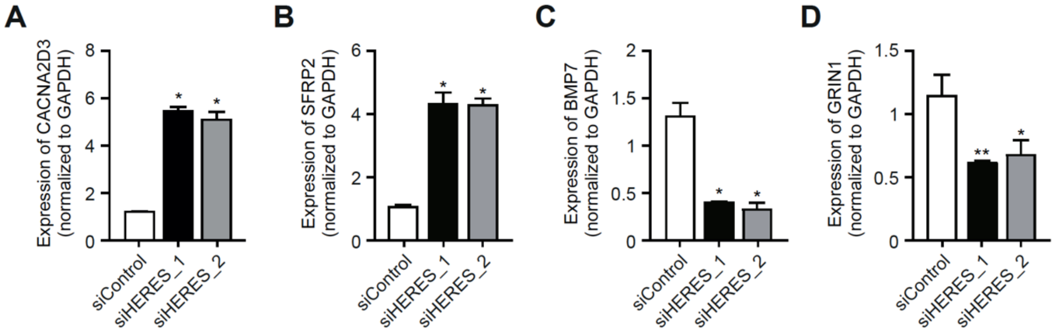


Figure 29. Expression patterns of putative target genes of HERES. (A-D) Expression levels of the four most dysregulated genes, as determined from nanoString results, were validated by qRT-PCR in KYSE-30 cells. Data represent the mean ± SD from three independent experiments (A-D). **P* ≤ 0.05, ***P* ≤ 0.01.

3.4.2. HERES epigenetically regulates target genes

As HERES appeared to be enriched in the nucleus rather than the cytoplasm (Fig. 28B) and nucleus-localized lncRNAs are often reported to be epigenetic regulators, a potential epigenetic role for HERES was investigated by analyzing publicly available array-based DNA methylation and RNA-seq data from TCGA ESCC samples. Based on their HERES expression level, the ESCC samples were first divided into two subgroups, HERES-high and HERES-low, and the changes in expression and DNA methylation of the protein-coding genes were then compared between the subgroups (Fig. 28C). Of the genes, *CACNA2D3* and *LDOC1* (Figs. 28D and 30A) were downregulated and hypermethylated in the HERES-high group, whereas *EPSTI1*, *SLC15A3*, and *BST2* were upregulated and hypomethylated in the HERES-high subgroup. The expression and DNA methylation changes were confirmed in HERES-depleted KYSE-30 cells by qRT-PCR and methylation-specific (MS) PCR. Only two downregulated genes (*CACNA2D3* and *LDOC1*) were confirmed to have both expression and DNA methylation changes (Figs. 28E and 30B-H). On the other hand, although *SFRP2* and *CXXC4* did not display DNA methylation changes in the analysis of the TCGA ESCC samples, the expression and DNA methylation signals of the Wnt signaling-related genes were changed similarly to those of *CACNA2D3* and *LDOC1* in the siHERES-treated cells compared to the siControl-treated cells (Figs 29B and 31A-C).

Because DNA methylation is often associated with histone modifications {Cedar, 2009 #25;Vire, 2006 #24}, global changes in histone modification markers in response to HERES knockdown were examined, revealing a marked decrease in H3K27me3 levels (Fig. 32A). We then investigated where the H3K27me3 signal was depleted in the genomic regions of three Wnt signaling pathway genes in siHERES-treated cells using chromatin immunoprecipitation (ChIP)-qPCR analysis. Significantly reduced H3K27me3 signals were observed at specific sites in the genes (recognized by primer 5 for *CACNA2D3*, primers 3, 4, and 5 for *SFRP2*, and primer 9 and 10 for *CXXC4*) in siHERES-depleted cells (Figs. 32B-D).

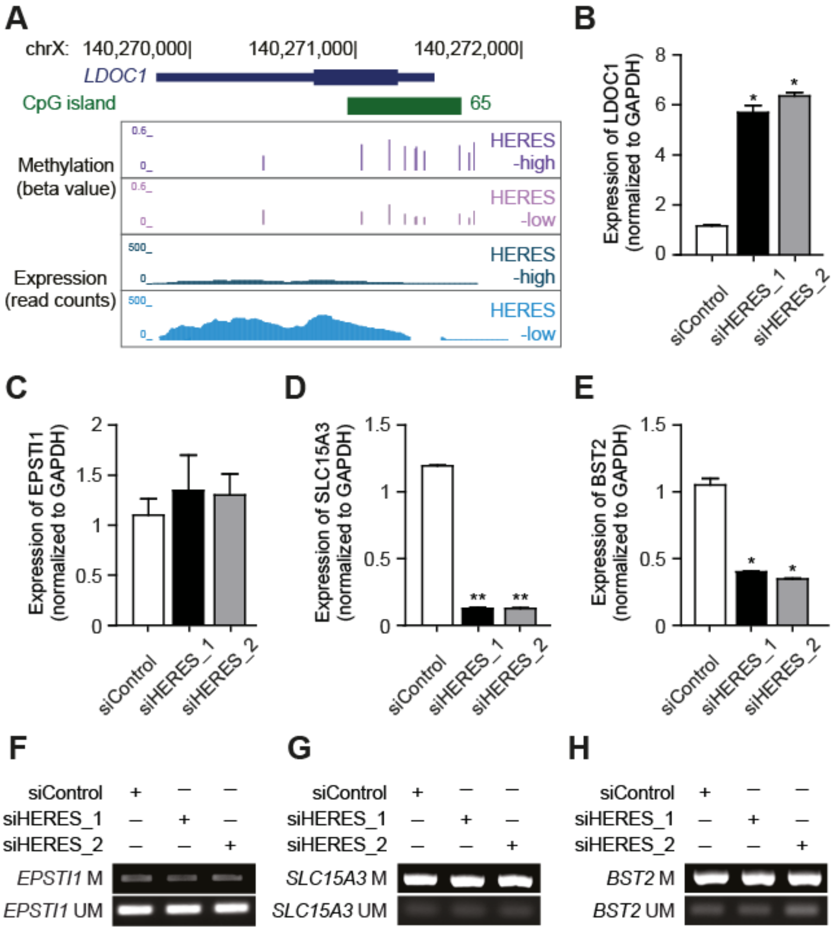


Figure 30. Expression and methylation patterns of putative target genes of HERES. (A) The *LDOC1* genomic locus with CpG island tracks, DNA methylation (beta value), and RNA expression (read count) in the HERES-high and HERES-low groups. (B-E) Expression levels of LDOC1, EPSTI1, SLC15A3, and BST2 were measured by qRT-PCR in KYSE-30 cells. (F-H) DNA methylation patterns (methylation (M) and unmethylation (UM)) of *EPSTI1*, *SLC15A3*, and *BST2* were measured by MS-PCR in KYSE-30 cells. Data represent the mean ± SD from three independent experiments. **P* ≤ 0.05, ***P* ≤ 0.01.

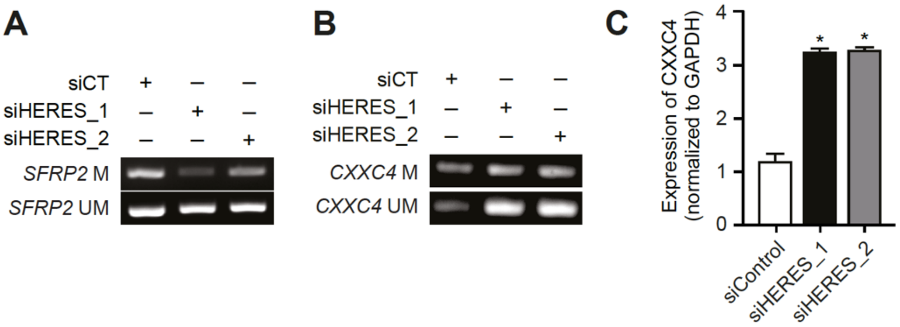


Figure 31. HERES epigenetically regulates genes involved in the Wnt signaling pathway. (A-B) DNA methylation patterns (methylation (M) and unmethylation (UM)) of *SFRP2* and *CXXC4* were measured by MS-PCR in KYSE-30 cells. (C) The CXXC4 expression level was confirmed by qRT-PCR in KYSE-30 cells.

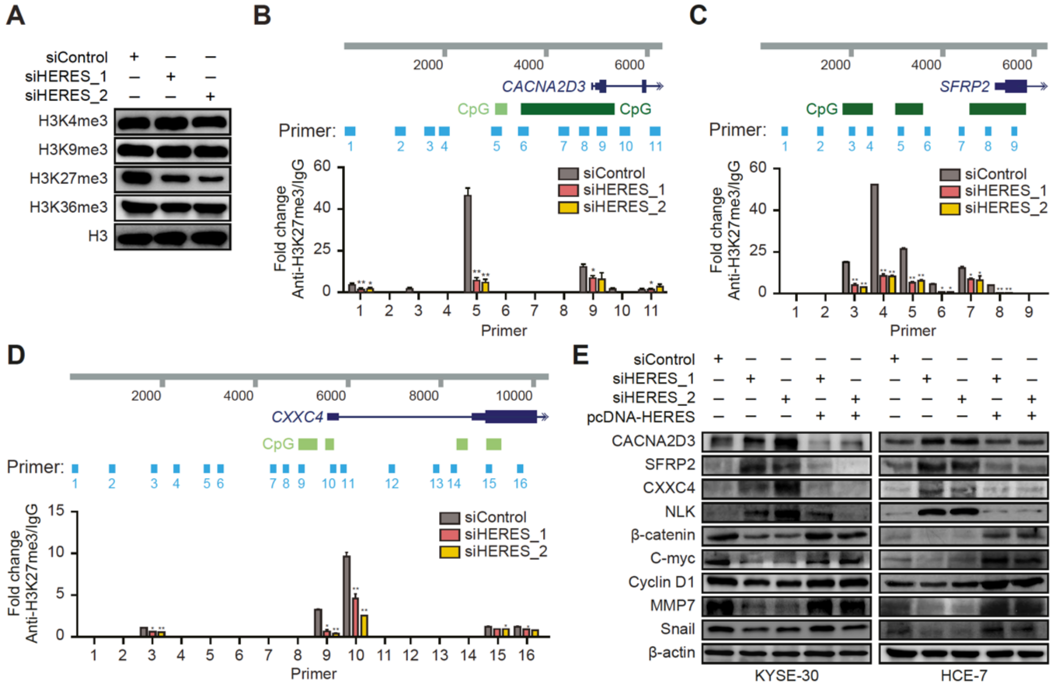


Figure 32. HERES regulates canonical and noncanonical Wnt signaling pathways. (A) Immunoblots of histone modification markers in siControl- or siHERES-transfected KYSE-30 cells. (B-D) ChIP-qPCR analysis of the H3K27me3 levels of *CACNA2D3* (B) *SFRP2* (C), and *CXXC4* (D) in siControl- or siHERES-transfected KYSE-30 cells. Data represent the mean ± SD from three independent experiments. *$P \leq 0.05$, **$P \leq 0.01$. (E) Expression of EMT markers in KYSE-30 and HCE-7 cells transfected with siControl or the indicated combinations of siHERES and pcDNA-HERES as determined by immunoblot.

3.4.3. HERES regulates canonical and noncanonical Wnt signaling pathways

Previous studies reported that CACNA2D3 downregulation inhibited the non-canonical Wnt/Ca^{2+} signaling pathway by decreasing the intracellular calcium level and NLK expression (*73*) and that SFRP2 and CXXC4 play roles as negative regulators of the canonical Wnt signaling pathway {Chung, 2009 #23;Kojima, 2009 #28}. We thus examined changes in the expression of two Wnt signaling-related factors, NLK and β-catenin, in siHERES-treated cells (Fig. 32E). As expected, HERES reduction increased the NLK level and decreased β-catenin in KYSE-30 and HCE-7 cells. In addition, changes in the expression of Wnt downstream targets were also confirmed (Fig. 32E). In contrast, introducing the pcDNA-HERES construct to cells reverted the expression levels of NLK, β-catenin, and Wnt downstream targets (Fig. 32E). Taken together, these results suggest that HERES downregulation in cancers perturbs and promotes canonical and non-canonical Wnt signaling pathways via epigenetic regulation, resulting in the inhibition of cancer progression.

3.5. The effect of HERES on the cell cycle and apoptosis

Contribution Statement: In this session, Jung-Ho Yoon (Yonsei University) performed all experiments, supervised by Sang Kil Lee.

Because two of the downstream targets of HERES, *CCND1* and *CACNA2D3*, are known to regulate the cell cycle and apoptosis (*71, 73*), the effect of the loss of HERES on the cell cycle and apoptotic processes was examined. Cell counting showed that siHERES-treated cells were arrested at G0/G1 (Fig. 33A). Flow cytometry showed that siHERES-treated cell populations exhibited significantly increased levels of apoptosis compared to siControl-treated cells (Fig. 33B). An induction of apoptotic factors, such as cleavage of poly (ADP-ribose) polymerase (PARP), cleaved caspase-9, and Bax, and a reduction of the anti-apoptotic factor, Bcl-2, were confirmed in siHERES-treated cells. However, the rescue of HERES expression reverted the expression of these factors to levels in control cells (Fig. 33C).

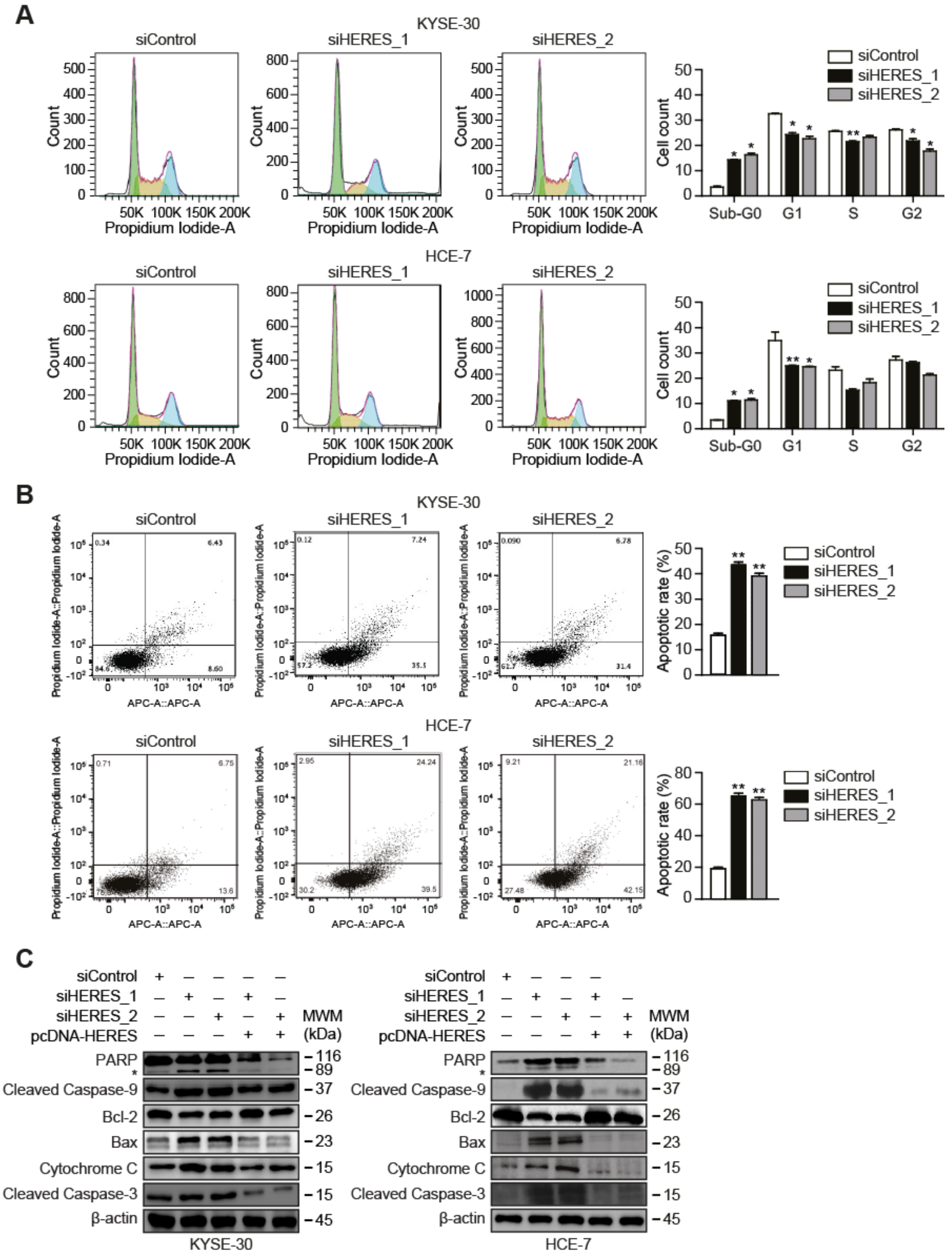

A
KYSE-30
siControl
siHERES_1
siHERES_2
Count
Propidium Iodide-A
50K 100K 150K 200K
Cell count
Sub-G0
G1
S
G2
HCE-7
B
APC-A::APC-A
Propidium Iodide-A::Propidium Iodide-A
Apoptotic rate (%)
C
siControl
siHERES_1
siHERES_2
pcDNA-HERES
MWM (kDa)
PARP
Cleaved Caspase-9
Bcl-2
Bax
Cytochrome C
Cleaved Caspase-3
β-actin
116
89
37
26
23
15
15
45
KYSE-30
HCE-7

Figure 33. The effect of HERES on the cell cycle and apoptosis. Cell cycle (A) and apoptosis (B) assays were performed on siRNA-transfected KYSE-30 and HCE-7 cells. (A) Cell cycle analysis of siRNA-transfected KYSE-30 and HCE-7 cells by flow cytometry. The bar graph shows the percentage of cells in sub-G0, G1, S, and G2 phases in siRNA-transfected KYSE-30 and HCE-7 cell populations. (B) Apoptosis was measured by flow cytometry using PI/Annexin V staining. The bar graph represents the percentage of apoptotic cells in each population. Data represent the mean ± SD from three independent experiments. (C) Apoptosis markers were assessed by immunoblot in KYSE-30 and HCE-7 cells transfected with siControl or siHERES and/or pcDNA-HERES. **P* ≤ 0.05, ***P* ≤ 0.01.

3.6. HERES interacts with EZH2 to regulate *CACNA2D3*

Contribution Statement: In this session, Bo-Hyun You performed the sequence analysis, supervised by Jin-Wu Nam. Jung-Ho Yoon (Yonsei University) performed all experiments, supervised by Sang Kill Lee.

3.6.1. Interaction of HERES with EZH2

We then asked how HERES regulates the expression of target genes at the epigenetic level. To address this question, binding sites for possible epigenetic modulators that can drive the histone methylation of target genes were first examined using publicly available ChIP-seq datasets from the ENCODE Project. We found that all three HERES target genes contained enhancer of EZH2 binding sites in their promoter regions (Fig. 34). Because EZH2, a subunit of the PRC2, has a well-known role in histone methylation to generate H3K27me3 and is known to interact with nuclear lncRNA, we suspected that EZH2 would be a binding partner of HERES. To examine the molecular relationship between HERES and EZH2, EZH2 RNA and protein levels were quantified in siControl- and siHERES-treated KYSE-30 cells, showing that HERES reduction decreased the EZH2 protein level but not the RNA level (Figs. 35A and 36A). Subsequently, RNA immunoprecipitation (RIP) (Fig. 35B) and EZH2 IP (Fig. 35C) assays showed the interaction of HERES and *CACNA2D3* with EZH2.

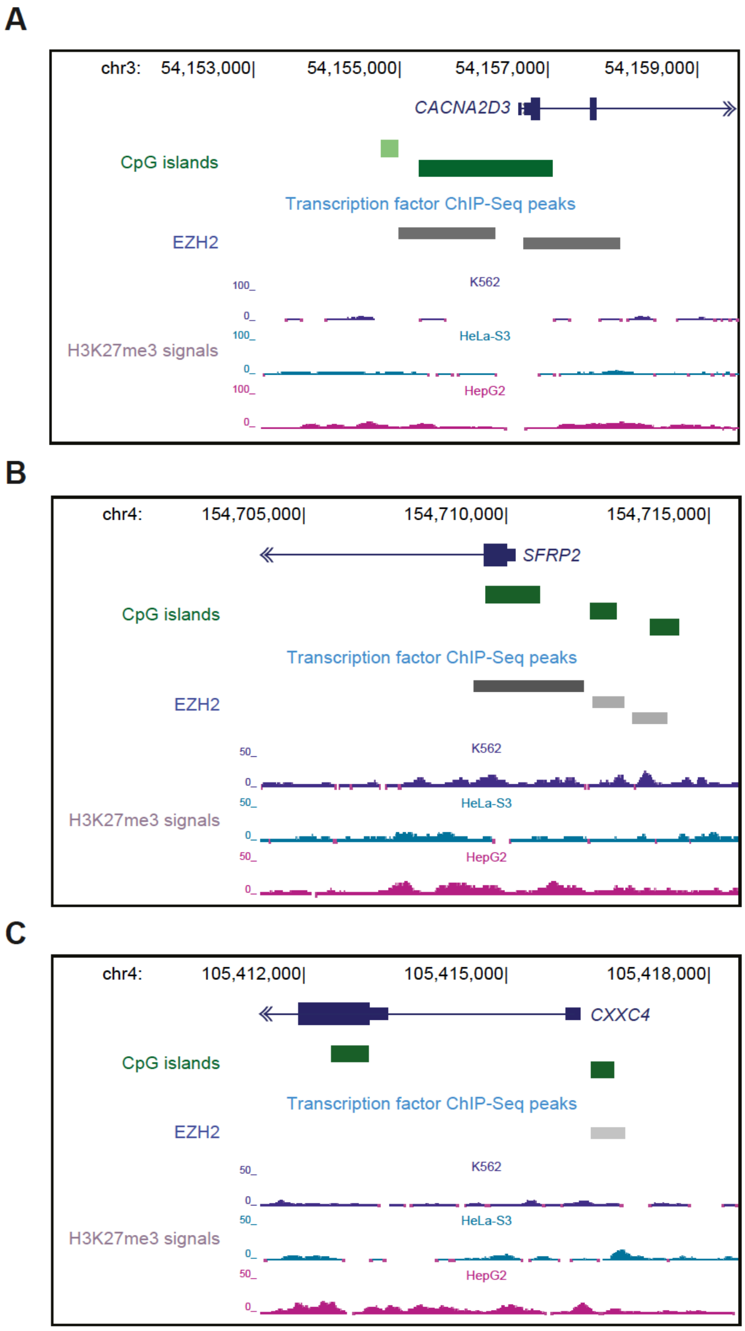


Figure 34. EZH2 ChIP-seq signals identified in *CACNA2D3*, *SFRP2*, and *CXXC4*. *CACNA2D3* (A), *SFRP2* (B), and *CXXC4* (C) genomic loci with CpG island tracks, EZH2 ChIP-seq peaks, and H3K27me3 signals.

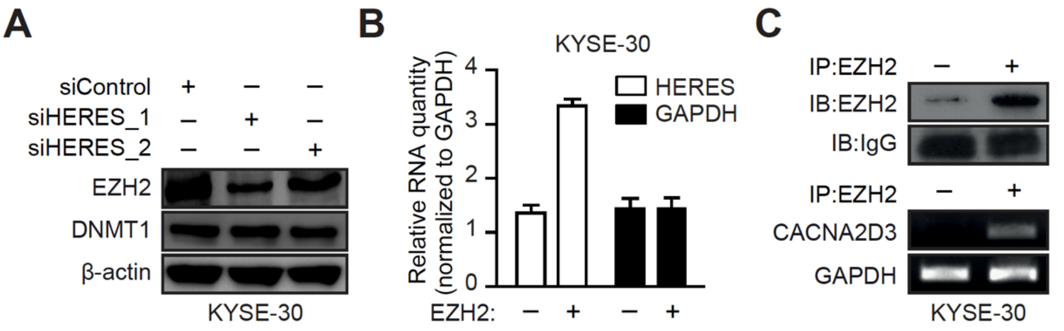
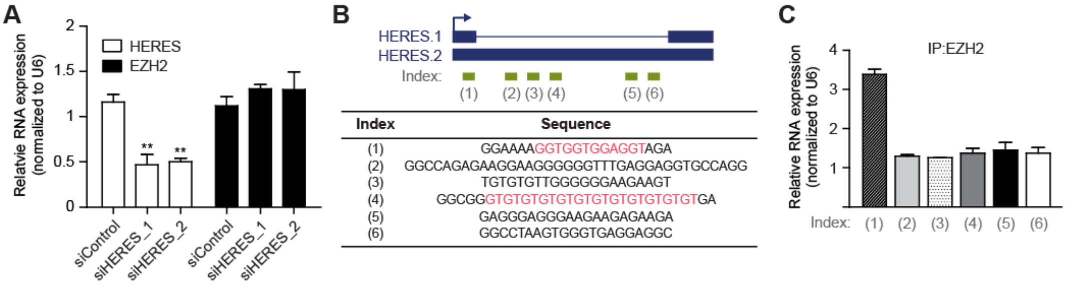


Figure 35. Interaction of HERES with EZH2*.* (A) Immunoblots of EZH2 and DNMT1 in KYSE-30 cells transfected with either siControl or siHERES. (B) RIP assays were performed with anti-EZH2 in KYSE-30 cell lysates. The quantity of HERES in the cell lysates (input) and the immunoprecipitates was measured by qRT-PCR. (C) IP assays were performed with anti-EZH2 in KYSE-30 cell lysates. The quantity of CACNA2D3 in the cell lysates (input) and the immunoprecipitates was measured by immunoblot. Data represent the mean ± SD from three independent experiments (B).



Index	Sequence
(1)	GGAAAAGGTGGTGGAGGTAGA
(2)	GGCCAGAGAAGGAAGGGGGGTTTGAGGAGGTGCCAGG
(3)	TGTGTGTTGGGGGGAAGAAGT
(4)	GGCGGGTGTGTGTGTGTGTGTGTGTGTGTGA
(5)	GAGGGAGGGAAGAAGAGAAGA
(6)	GGCCTAAGTGGGTGAGGAGGC

Figure 36. Putative binding sites of EZH2 on HERES. (A) Relative expression of HERES and EZH2 in KYSE-30 cells transfected with a control siRNA (siControl) or HERES siRNAs (siHERES_1 or siHERES_2). (B) Six putative G-rich regions in HERES transcripts. (C) Relative HERES expression determined using the six primer sets referred to (B) in EZH2-IP. Data represent the mean ± SD from three independent experiments. $^{**}P \leq 0.01$.

3.6.2. HERES regulates CACNA2D3 via direct interaction with EZH2

To validate a direct interaction between HERES and PRC2-EZH2, we then searched for PRC2-EZH2 binding motifs in the HERES sequence. Because the PRC2 complex including EZH2 is known to be recruited by G-rich motif, we scanned for G-rich regions in HERES transcripts, leading to the identification of six regions including two potential g-quadruple structure motifs (Fig. 36B). EZH2-IP and qRT-PCR showed that a single region with four GGW repeats (index 1) was significantly enriched in EZH2-IP (Fig. 36C). To further investigate if the HERES GGW repeat sequence (index 1) is necessary for an interaction with EZH2, we constructed a plasmid vector that harbors a HERES sequence that lacks the GGW repeat region (HERES-Mut) (Fig. 37A). A RIP assay confirmed that EZH2 failed to interact with HERES-Mut in KYSE-30 cells (Fig. 37B). RNA fluorescence in situ hybridization (FISH) of HERES and fluorescein isothiocyanate (FITC) staining of CACNA2D3 validated that HERES was principally localized to the nucleus and that the GGW sequence (index 1) is necessary for the interaction with EZH2 to downregulate CACNA2D3 (Fig. 37C). Cells transfected with HERES-Mut exhibited significantly increased CACNA2D3 at both the RNA and protein level, whereas HERES overexpression (pcDNA-HERES) reduced the CACNA2D3 level (Figs. 37D and E).

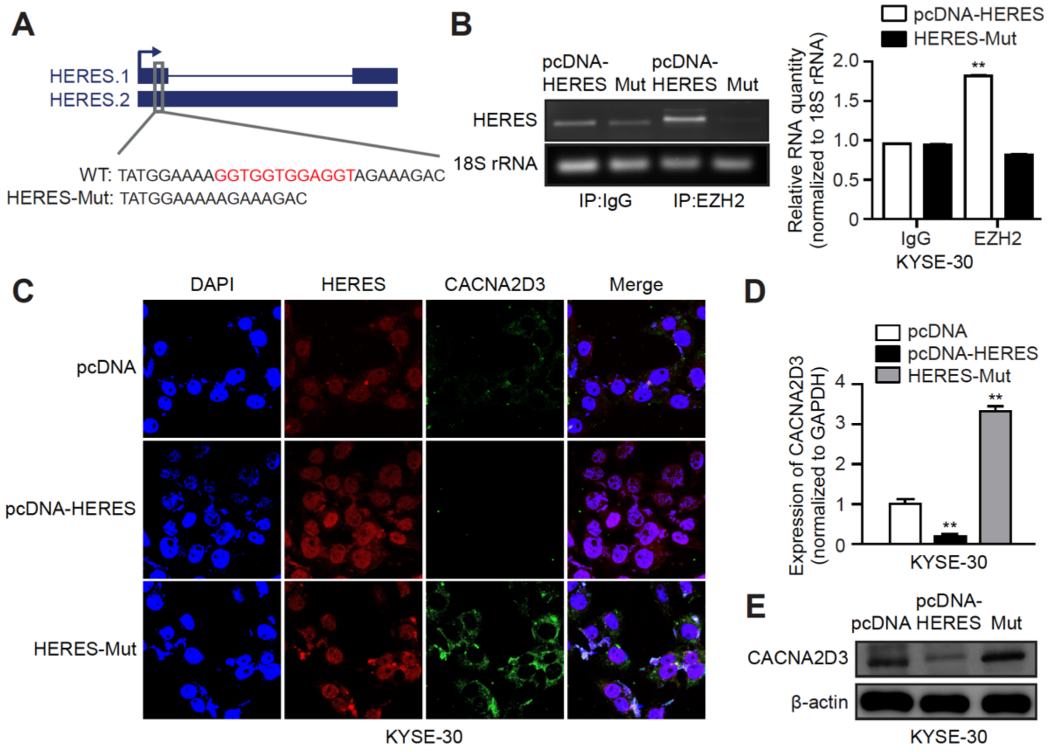


Figure 37. HERES directly interacts with EZH2 via G-rich motif*.* (A) Representation of the WT and mutated (HERES-Mut) HERES sequences used for IP with anti-EZH2. HERES-Mut contains a deletion of the G-rich sequence (index 1) presented in Fig. 36B. (B) RIP assays were performed with anti-EZH2 in lysates of KYSE-30 cells transfected with either pcDNA-HERES or HERES-Mut (left panel). The bar graph shows the relative amount of HERES after anti-EZH2 IP using lysates of cells transfected with either pcDNA-HERES or HERES-Mut (right panel). (C) RNA FISH to visualize HERES (red) and FITC staining of CACNA2D3 (green) in KYSE-30 cells transfected with pcDNA (upper panel), pcDNA-HERES (middle panel), or HERES-Mut (lower panel). Nuclei were stained with 4′,6-diamidino-2-phenylindole (DAPI) (blue). CACNA2D3 RNA (D) and protein (E) levels were measured in KYSE-30 cells transfected with pcDNA, pcDNA-HERES, or HERES-Mut by qRT-PCR and immunoblot, respectively. Data represent the mean ± SD from three independent experiments (B and E). ***P* ≤ 0.01.

3.7. HERES as a candidate therapeutic target

Contribution Statement: In this session, Hoin Kang (Catholic University) performed the xenograft assay, supervised by Eun Kyung Lee.

To investigate whether HERES controls tumor growth *in vivo*, we carried out xenograft assays with siControl- and siHERES-treated cancer cell lines (Fig. 38). Both the volume and weight of tumors derived from HERES-depleted samples were significantly reduced compared to tumors derived from control cells four weeks after the injection (Figs. 38A-C). We further examined changes in the expression of HERES and its target genes, finding that the reduction of HERES was maintained for four weeks after siHERES injection (Fig. 39), whereas the levels of HERES targets were significantly increased in tumor samples derived from HERES-depleted cells compared to control cells (Fig. 38D). We also confirmed that global H3K27me3 and EZH2 levels were decreased in HERES-depleted tumor samples (Fig. 38D), suggesting that HERES is a promising candidate therapeutic target that controls tumor growth through the regulation of canonical and non-canonical Wnt signaling pathways *in vivo* (Fig. 38D).

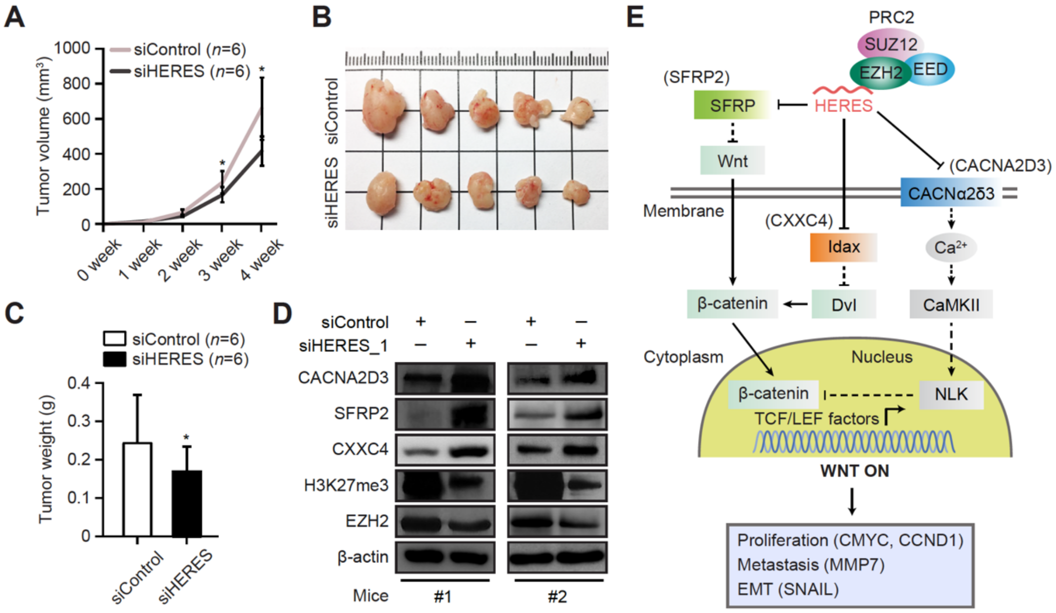


Figure 38. Expression of HERES regulates tumorigenicity in xenograft models. KYSE-30 cells transfected with either siControl or siHERES were injected into nude mice (6 mice for each group). The resulting xenograft tumor volumes (A, B) and weights (C) are shown. (A) Tumor growth curves showing that tumors in the siHERES group grew markedly slower than those in the siControl group. (B) Images of tumor volumes from the xenograft models. (C) Tumor weights in the siHERES and siControl groups 4 weeks after cell injection. Data represent the mean ± SD. (D) Immunoblot analysis (#1 and #2) of levels of key components (CACNA2D3, SFRP2, and CXXC4) from the canonical and non-canonical Wnt signaling pathways and of H3K27me3 and EZH2 in the siHERES and siControl xenograft models. (E) A graphic illustration of HERES-regulated canonical and non-canonical Wnt signaling pathways in ESCC.

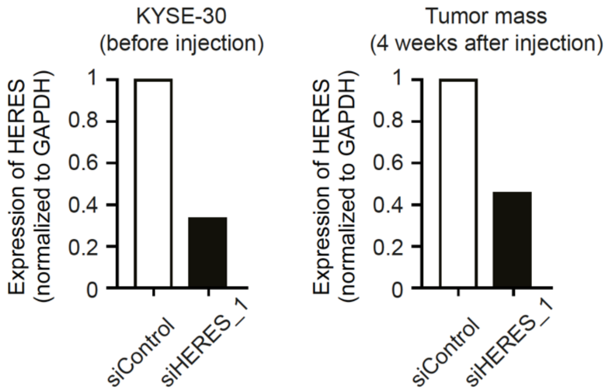
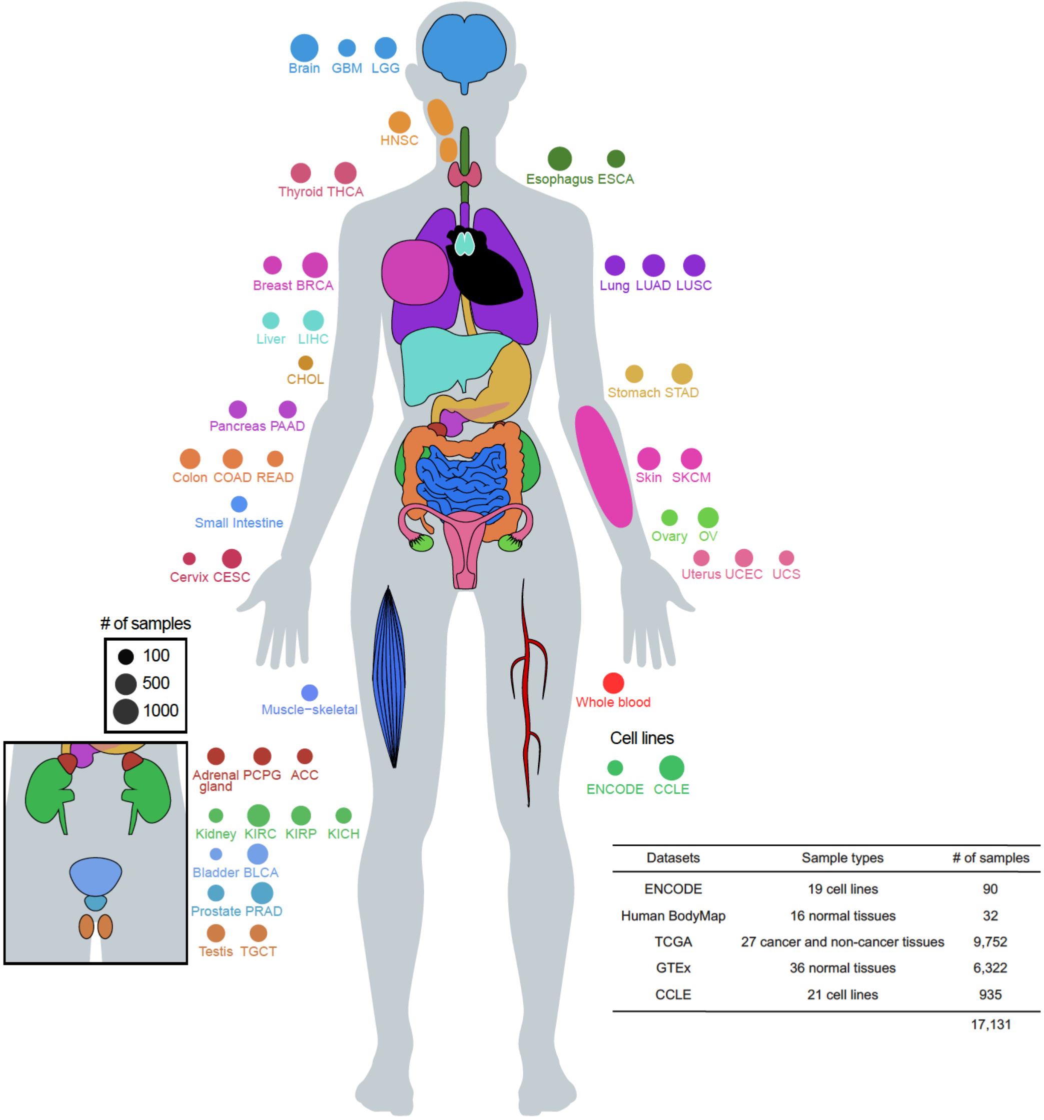


Figure 39. Expression of HERES before and 4 weeks after injection into nude mice. Relative HERES expression before (left panel) and 4 weeks after (right panel) injection into nude mice. HERES expression was measured in KYSE-30 cells transfected with a control siRNA (siControl) or HERES siRNA (HERES_1) or in the tumor mass formed by injection of the cells.

Chapter 4. Pan-cancer annotations of lncRNAs

4.1. Construction of integrative transcriptome maps

To construct integrative transcriptome maps, large-scale RNA-seq datasets were obtained from the ENCODE Project (*n*=90), the Human BodyMap 2.0 Project (*n*=32), the TCGA Project (*n*=9,752), the GTEx Project (*n*=6,322), and the CCLE Project (*n*=935). A total of 17,131 RNA-seq samples were composed of 36 normal tissue types, 27 cancer paired with non-cancer tissue types, and 21 cell lines were collected (Fig. 40). To control RNA-seq data quality, samples with less than 40 million uniquely mapped reads were excluded from downstream analysis (Fig. 41). Because most of RNA-seq samples were unstranded, the reads were converted to RPDs (pseudo-stranded) using CAFE (*29*). Remaining RNA-seq samples were each used to assemble individual transfrags with StringTie (*15*). Among the resulting transfrags, ones with an abnormal number of transfrags (smaller than 50,000 transfrags or larger than 200,000 transfrags) were considered to be unreliable and were thus discarded. Meta-assembly was performed from 16,784 RNA-seq assemblies according to types of cancer and tissue with TACO. Finally, cancer-/tissue-specific transcriptome maps were updated using CAFE pipeline (*29*).



Datasets	Sample types	# of samples
ENCODE	19 cell lines	90
Human BodyMap	16 normal tissues	32
TCGA	27 cancer and non-cancer tissues	9,752
GTEx	36 normal tissues	6,322
CCLE	21 cell lines	935
		17,131

Figure 40. RNA-seq datasets for construction of transcriptome maps. A total of 17,131 RNA-seq samples comprising of 36 normal tissue types, 27 cancer paired with non-cancer tissue types, and 21 cell lines were used for constriction of transcriptome maps.

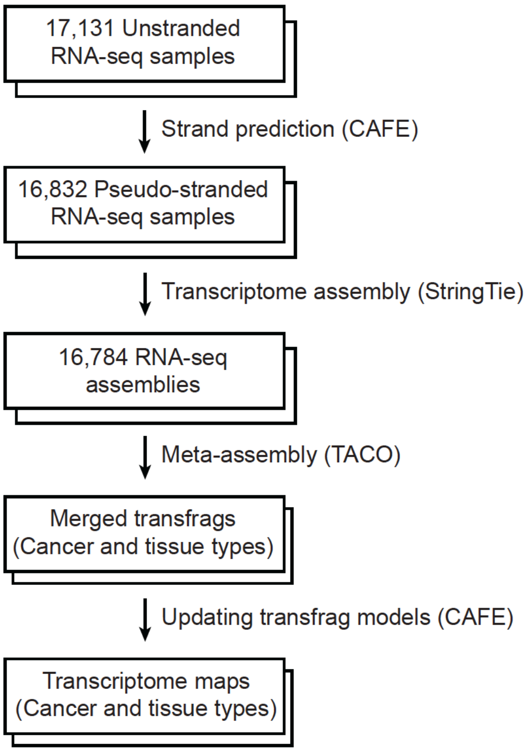

17,131 Unstranded RNA-seq samples
Strand prediction (CAFE)
16,832 Pseudo-stranded RNA-seq samples
Transcriptome assembly (StringTie)
16,784 RNA-seq assemblies
Meta-assembly (TACO)
Merged transfrags (Cancer and tissue types)
Updating transfrag models (CAFE)
Transcriptome maps (Cancer and tissue types)

Figure 41. Construction of integrative transcriptome maps. A schematic flow for construction of cancer-/tissue-specific transcriptome maps from large-scale RNA-seq datasets.

4.2. Accuracies of transcriptome maps

Normal tissue transcriptome maps originated from the GTEx Project contained 110,000 transcripts with 2.5 isoforms per loci on average. Data from bladder generated the smallest transcriptome with 78,449 transcripts, while the one from testis had the largest number of transcripts (276,339) (Fig. 42). Testis transcriptome also shows abnormally high number of assembled loci and isoform per loci, while transcriptome build from other tissues mostly have more reasonable number. On the contrary, cancer transcriptomes created from TCGA datasets distinctively varied in terms of loci and transcripts. CHOL had the smallest transcriptome comprised of 90,319 transcripts (35,790 loci), while STAD had the largest transcriptome with 223,765 transcripts (95,663 loci). It is also worthy of note that unlike GTEx transcriptome, those built from TCGA datasets showed considerably different number of isoforms per loci, with the number being 2.2 for COAD and 3.2 for TGCT.

To assess the quality of GTEx and TCGA transcriptome maps, recall and precision were calculated at base, intron, and splicing level against GENCODE v32 annotation (Fig. 43). Apart from those generated from testis, GTEx transcriptome maps showed recall and precision rates of 91.4-97.9% and 72.7-86% on base level, respectively, which are higher than that of MiTranscriptome and BIGTranscriptome annotations. Most of TCGA transcriptome maps appeared to have recall rates similar to that of GTEx transcriptome maps (89.4-96.7%). However, transcriptome maps from BRCA, ESCA, KIRC, LUAD, OV, and STAD had exceptionally low precision on base level, ranging from 55.5-67.7% while other had higher precision rate of 71.2-87.7%. The six transcriptome maps and those of lung and testis from

GTEx also displayed lower precision rates on intron and splicing levels. Based in the assembled transcript models, it is suspected that the quality of RNA-seq is compromised to some extent due to DNA contamination. Other samples precision rates of 80.9-89.2% on intron level and 32.8-46.5% on splicing level, which are by far higher than all other previously reported transcriptome annotations.

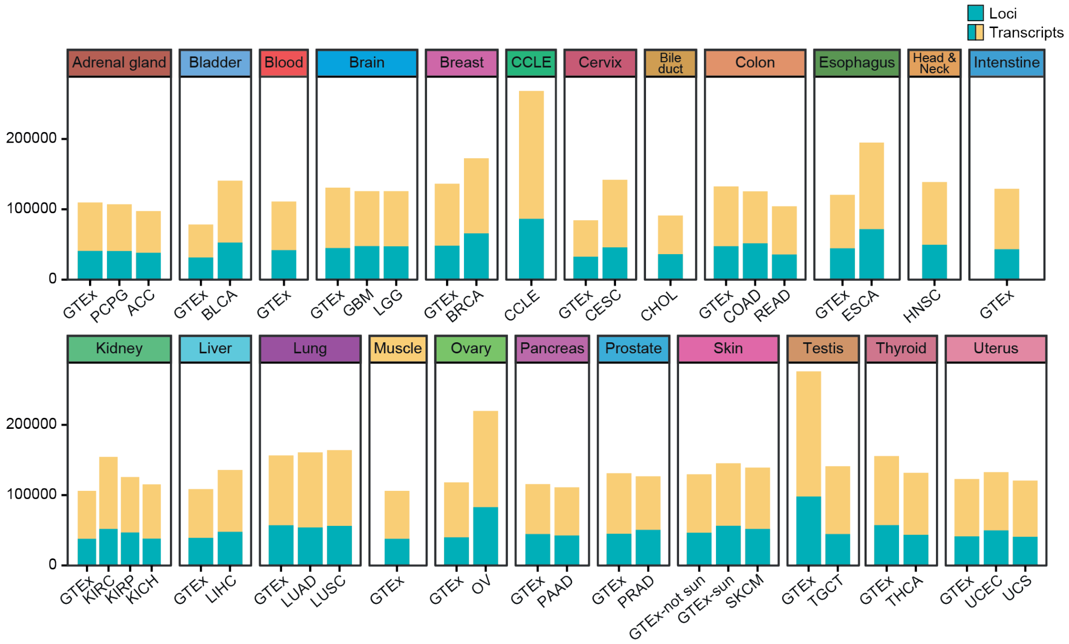


Figure 42. The GTEx and TCGA transcriptome maps. The number of reconstructed transcripts and their loci in each tissue and cancer types.

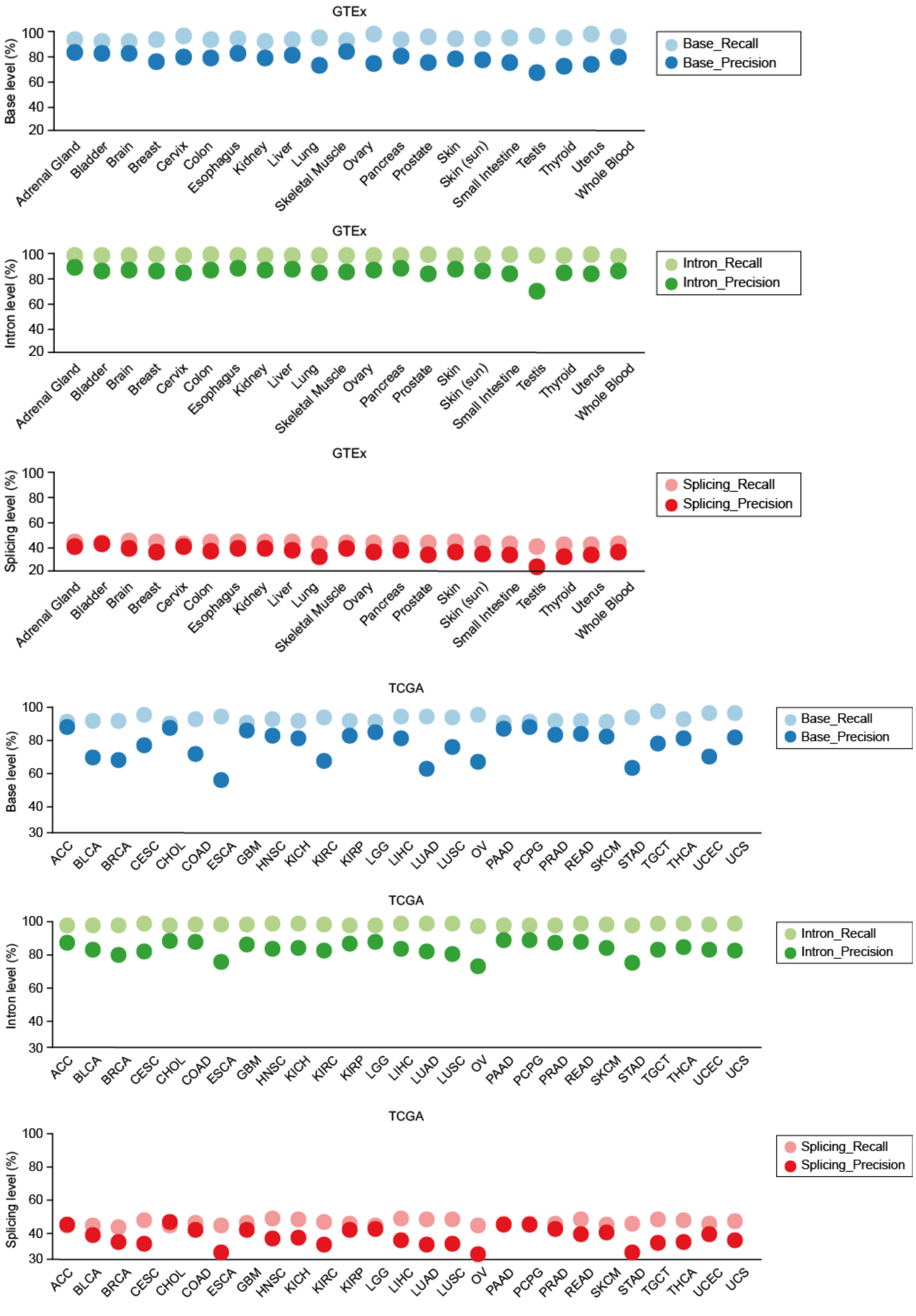


Figure 43. The annotation accuracy of transcriptome maps. Shown are the accuracies of transcriptome maps at base, intron, and splicing levels based on GENCODE v32 annotation.

4.3. Annotation of novel lncRNAs in pan-cancer

To annotated novel lncRNAs from reconstructed transcriptome maps, we performed lncRNA annotation as mentioned in the Section 3.1.1. The number of novel lncRNAs annotated was varied between the GTEx and TCGA transcriptome maps (Fig. 44). The least number of novel lncRNAs (588 and 1279 lncRNAs) were annotated from bladder and CHOL, on the other hand, the largest number of novel lncRNAs (59,142 and 49,941 lncRNAs) were annotated from testis and STAD, respectively. The latter two samples were found to have large numbers of novel lncRNAs due to abnormally enormous number of transcripts assembled in the previous step. Skin and SKCM also showed relatively huge numbers of novel lncRNAs, however, their assembled number of transcripts were comparable to the others. The average isoform per loci ratio in novel lncRNAs were 1.2 and 1.3 in GTEx and TCGA, respectively. LUAD from TCGA and brain from GTEx showed the highest isoform per loci ratio compare to the other samples.

The completeness, or the proportion of full-length genes was checked for both GTEx- and TCGA- originated novel lncRNAs (Fig. 45). LncRNA genes supported by both CAGE-seq (TSS) and 3P-seq (CPS) tags were considered as full-length genes. Same analysis was also done for lncRNA annotations for GENCODE, MiTranscriptome, BIGTranscriptome, FANTOM CAT, and CHESS, 13%, 4%, 21%, 18%, 14% of which were full-length lncRNA genes, respectively. Meanwhile, except for testis novel lncRNAs, 18-24% of GTEx novel lncRNA maps appeared to be full-length lncRNA genes, indicating their completeness. TCGA novel lncRNAs also had considerably higher proportion of full-length lncRNA genes, from 16~25%, with the exception for six cancer types (BRCA, ESCA, KIRC, LUAD, OV, and

STAD). Overall, it can be concluded from the results that both TCGA and GTEx novel lncRNA maps are so far the most accurate annotation.

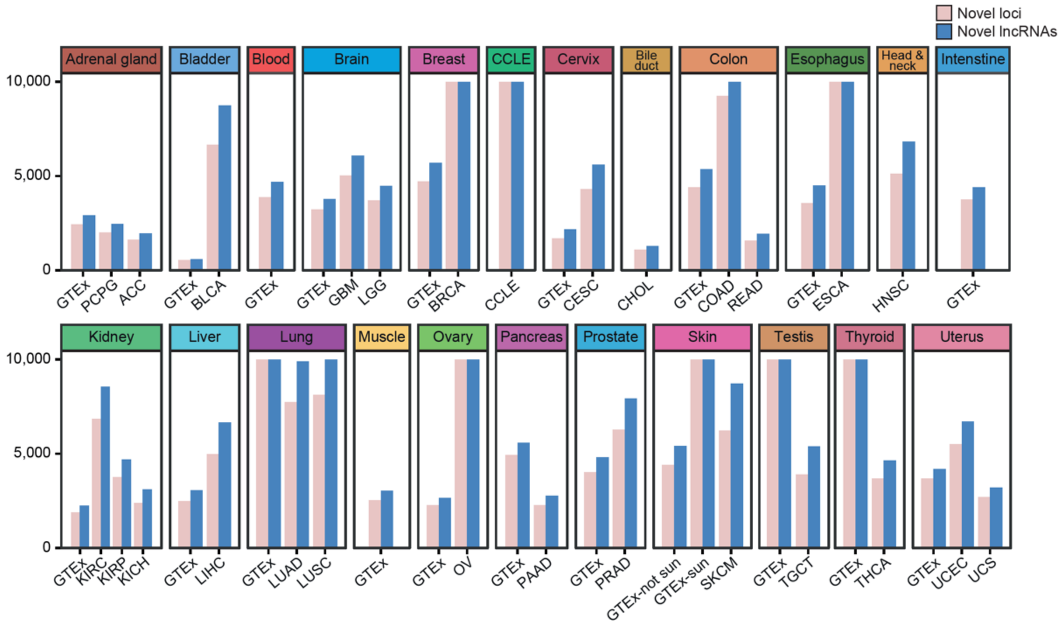


Figure 44. Pan-cancer atlas of novel lncRNAs. The number of annotated novel lncRNAs and their lncRNA genes from reconstructed transcriptome maps.

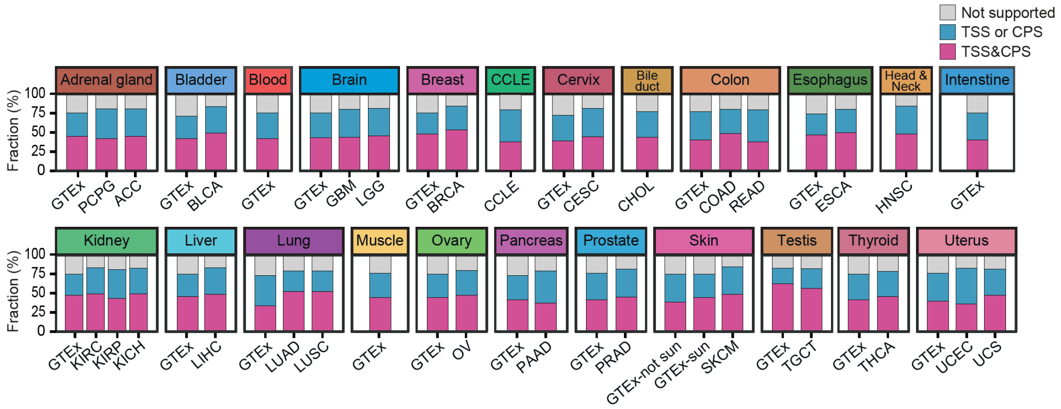


Figure 45. The completeness of novel lncRNAs. The fraction of novel lncRNAs supported with both TSS and CPS (purple), with either TSS or CPS (blue), and with neither TSS or CPS (grey).

Chapter 5. Discussion

A high-performing transcriptome assembly pipeline, CAFE, enabled us to significantly improve the quality of the resulting assemblies by resurrecting large-scale unstranded RNA-seq data, which was formerly used for less informative or less specific transcriptome assembly. The re-use of the large-scale unstranded RNA-seq data is valuable in following three reasons. For example, public transcriptome databases, such as the TCGA Project (*25, 26*), the GTEx Project (*24*), the CCLE Project (*23*), and NCBI GEO, include large-scale unstranded RNA-seq data. Hence, determining the directions of unstranded reads enables the construction of highly accurate transcriptome maps, which is necessary for highly qualitative downstream analyses. Although determining the directions of unstranded reads requires stranded data in the corresponding cell type or tissue, the use of pooled stranded data can still be of benefit to the prediction of transcript direction and the following assembly. In fact, the RPDs of unstranded TCGA, GTEx and CCLE data were predicted using pooled stranded RNA-seq data and showed high accuracy (Fig. 43). Secondly, in the case of genes with low expression such as those encoding lncRNAs, additional RPDs benefit transcriptome assembly by increasing the read-depth of those genes. Although the targeted capture of low-abundant transcripts like lncRNAs using antisense oligonucleotides enabled an increase in the copy number of the target transcripts (*76*), this approach is only applicable to known transcripts. Thirdly, additional RPDs could increase the detection of missed exon-junctions, resulting in the connection of fragmented transfrags.

We utilized CAGE-seq and 3P-seq data to profile transcript TSSs and CPSs, which detect unambiguous ends at single base resolution as well as transcript alternative forms. However, the assignment of multiple TSSs and CPSs raises a question: which pairs of ends, in all possible combinations, are relevant? Moreover, if a gene has alternative splicing isoforms, the number of possible isoforms is exponentially increased by multiple TSSs and CPSs. CAFE now generates all possible but unique isoforms, some of which would not actually exist in cells. Therefore, a precise way to determine a TSS-CPS pair simultaneously would provide biologically relevant isoforms directly. One approach is to integrate paired-end ditag (PET) data that contains both 5' and 3' end sequence tags of transcripts (*6*) and an alternative is to sequence full-length RNAs using third-generation sequencing methods such as Iso-seq (*77*).

Applying CAFE pipeline to RNA-seq datasets of ESCC patients successfully identified a novel ESCC-driving lncRNA, HERES. A series of computational and experimental analyses showed that HERES transcriptionally controls multiple target genes in the Wnt signaling pathways at the epigenetic level by interacting with the EZH2-PRC2 complex. Because the targets and *HERES* are generally located on different chromosomes, HERES appears to act via EZH2-PRC2 *in trans* rather than *in cis*. Intriguingly, HERES RNA contains some repeat elements including GGW and Alu repeats, which extensively match sequences in the upstream regions of the target genes including *CACNA2D3*. Particularly, the Alu repeats would provide complementary base-pairing between HERES its target DNA sequences as well as it might be related to the nuclear localization of HERES, as previously reported (*78*).

Albeit we reported in this study that a G-rich motif in HERES is important for binding to EZH2, it remains unclear which part of EZH2 interacts with HERES. A previous study showed that the N-terminal region of EZH2 is important for RNA binding through a G-rich motif (*79*). A series of deletion mutants of human PRC2 revealed that the basic N-terminal helix of EZH2, particularly residues 32-42 in the helix, are the most critical for RNA binding through a G-rich motif. Given these results, the G-rich motif embedded in HERES probably also interacts with the basic N-terminal helix of EZH2, although such a direct interaction needs to be verified.

Transcription factor ChIP-seq data from the ENCODE Project revealed that the first HERES exon contains some enhancer-related transcription factor binding sites (TFBSs) for CEBPB, EP300, and AP-1 subunits (JUN, FOS), suggesting that HERES could be regulated epigenetically by modulation of the chromatin state at its locus. On the other hand, the expression of HERES near enhancer-related TFBSs raises the possibility that HERES is an enhancer RNA (eRNA) that regulates neighboring genes *in cis*. However, two observations argue against this idea: first, HERES is highly abundant and includes both a 5’ cap and 3’ polyadenylation, unlike eRNAs, and second, the genes neighboring *HERES* were only marginally affected by siHERES transfection.

Although there have been reports that some lncRNAs participate in regulating the Wnt signaling pathway, their targets appear limited (*80*). Our results suggest that HERES could be a master regulator of the Wnt signaling pathway, because it controls key components of both canonical and Ca^{2+}-related non-canonical pathways (Fig. 38E). Our results highlight the potential significance of HERES in terms of targeted therapy.

References

1. Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).
2. B. J. Haas, M. C. Zody, Advancing RNA-Seq analysis. *Nat Biotechnol* **28**, 421-423 (2010).
3. J. A. Martin, Z. Wang, Next-generation transcriptome assembly. *Nat Rev Genet* **12**, 671-682 (2011).
4. J. Harrow *et al.*, GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774 (2012).
5. The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
6. S. Djebali *et al.*, Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).
7. M. N. Cabili *et al.*, Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915-1927 (2011).
8. M. K. Iyer *et al.*, The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47**, 199-208 (2015).
9. C. C. Hon *et al.*, An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**, 199-204 (2017).
10. M. Pertea *et al.*, CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol* **19**, 208 (2018).
11. M. Guttman *et al.*, Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503-510 (2010).
12. C. Trapnell *et al.*, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515 (2010).
13. N. Boley *et al.*, Genome-guided transcript assembly by integrative analysis of RNA sequence data. *Nat Biotechnol* **32**, 341-346 (2014).

14. L. Maretty, J. A. Sibbesen, A. Krogh, Bayesian transcriptome assembly. *Genome Biol* **15**, 501 (2014).

15. M. Pertea *et al.*, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-295 (2015).

16. M. Shao, C. Kingsford, Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat Biotechnol* **35**, 1167-1169 (2017).

17. D. R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829 (2008).

18. J. Martin *et al.*, Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics* **11**, 663 (2010).

19. M. G. Grabherr *et al.*, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652 (2011).

20. M. H. Schulz, D. R. Zerbino, M. Vingron, E. Birney, Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086-1092 (2012).

21. Z. Safikhani, M. Sadeghi, H. Pezeshk, C. Eslahchi, SSP: an interval integer linear programming for de novo transcriptome assembly and isoform discovery of RNA-seq reads. *Genomics* **102**, 507-514 (2013).

22. Y. Xie *et al.*, SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660-1666 (2014).

23. J. Barretina *et al.*, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).

24. The GTEx Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585 (2013).

25. G. Ciriello *et al.*, Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* **45**, 1127-1133 (2013).

26. C. Kandoth *et al.*, Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339 (2013).

27. J. W. Nam, D. P. Bartel, Long noncoding RNAs in C. elegans. *Genome Res* **22**, 2529-2540 (2012).

28. S. Zhao *et al.*, Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics*

16, 675 (2015).

29. B. H. You, S. H. Yoon, J. W. Nam, High-confidence coding and noncoding transcriptome maps. *Genome Res* **27**, 1050-1062 (2017).
30. B. Uszczynska-Ratajczak, J. Lagarde, A. Frankish, R. Guigo, R. Johnson, Towards a complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet* **19**, 535-548 (2018).
31. T. Steijger *et al.*, Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **10**, 1177-1184 (2013).
32. I. Ulitsky, A. Shkumatava, C. H. Jan, H. Sive, D. P. Bartel, Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537-1550 (2011).
33. R. Yamashita *et al.*, Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res* **21**, 775-789 (2011).
34. J. B. Brown *et al.*, Diversity and dynamics of the Drosophila transcriptome. *Nature* **512**, 393-399 (2014).
35. The FANTOM Consortium and the RIKEN PMI and CLST (DGT), A promoter-level mammalian expression atlas. *Nature* **507**, 462-470 (2014).
36. R. Nusse, H. E. Varmus, Wnt genes. *Cell* **69**, 1073-1087 (1992).
37. P. Polakis, Wnt signaling and cancer. *Genes Dev* **14**, 1837-1851 (2000).
38. F. Deng, K. Zhou, W. Cui, D. Liu, Y. Ma, Clinicopathological significance of wnt/beta-catenin signaling pathway in esophageal squamous cell carcinoma. *Int J Clin Exp Pathol* **8**, 3045-3053 (2015).
39. T. Zhan, N. Rindtorff, M. Boutros, Wnt signaling in cancer. *Oncogene* **36**, 1461-1473 (2017).
40. T. Grigoryan, P. Wend, A. Klaus, W. Birchmeier, Deciphering the function of canonical Wnt signals in development and disease: conditional loss- and gain-of-function mutations of beta-catenin in mice. *Genes Dev* **22**, 2308-2341 (2008).
41. H. Clevers, Wnt/beta-catenin signaling in development and disease. *Cell* **127**, 469-480 (2006).
42. M. V. Semenov, R. Habas, B. T. Macdonald, X. He, SnapShot:

Noncanonical Wnt Signaling Pathways. *Cell* **131**, 1378 (2007).

43. R. T. Moon, A. D. Kohn, G. V. De Ferrari, A. Kaykas, WNT and beta-catenin signalling: diseases and therapies. *Nat Rev Genet* **5**, 691-701 (2004).

44. J. N. Anastas, R. T. Moon, WNT signalling pathways as therapeutic targets in cancer. *Nature Reviews Cancer* **13**, 11 (2012).

45. V. Conteduca *et al.*, Barrett's esophagus and esophageal cancer: an overview. *Int J Oncol* **41**, 414-424 (2012).

46. K. J. Napier, M. Scheerer, S. Misra, Esophageal cancer: A Review of epidemiology, pathogenesis, staging workup and treatment modalities. *World J Gastrointest Oncol* **6**, 112-120 (2014).

47. S. Ohashi *et al.*, Recent Advances From Basic and Clinical Studies of Esophageal Squamous Cell Carcinoma. *Gastroenterology* **149**, 1700-1715 (2015).

48. K. Higuchi *et al.*, Current management of esophageal squamous-cell carcinoma in Japan and other countries. *Gastrointest Cancer Res* **3**, 153-161 (2009).

49. M. W. Wiedmann, J. Mossner, New and emerging combination therapies for esophageal cancer. *Cancer Manag Res* **5**, 133-146 (2013).

50. X. Kang *et al.*, Personalized targeted therapy for esophageal squamous cell carcinoma. *World J Gastroenterol* **21**, 7648-7658 (2015).

51. C. P. Ponting, P. L. Oliver, W. Reik, Evolution and functions of long noncoding RNAs. *Cell* **136**, 629-641 (2009).

52. I. Ulitsky, D. P. Bartel, lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**, 26-46 (2013).

53. J. J. Quinn, H. Y. Chang, Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet* **17**, 47-62 (2016).

54. B. Kleaveland, C. Y. Shi, J. Stefano, D. P. Bartel, A Network of Noncoding Regulatory RNAs Acts in the Mammalian Brain. *Cell* **174**, 350-362 e317 (2018).

55. J. H. Yuan *et al.*, A long noncoding RNA activated by TGF-beta promotes the invasion-metastasis cascade in hepatocellular carcinoma. *Cancer Cell* **25**, 666-681 (2014).

56. M. M. Ali *et al.*, PAN-cancer analysis of S-phase enriched lncRNAs identifies oncogenic drivers and biomarkers. *Nat Commun* **9**, 883 (2018).

57. R. A. Gupta *et al.*, Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071-1076 (2010).

58. F. J. Chen *et al.*, Upregulation of the long non-coding RNA HOTAIR promotes esophageal squamous cell carcinoma metastasis and poor prognosis. *Mol Carcinog* **52**, 908-915 (2013).

59. Y. Wu *et al.*, Up-regulation of lncRNA CASC9 promotes esophageal squamous cell carcinoma growth by negatively regulating PDCD4 expression through EZH2. *Mol Cancer* **16**, 150 (2017).

60. J. H. Yoon *et al.*, The long noncoding RNA LUCAT1 promotes tumorigenesis by controlling ubiquitination and stability of DNA methyltransferase 1 in esophageal squamous cell carcinoma. *Cancer Lett* **417**, 47-57 (2018).

61. C. Q. Li *et al.*, Integrative analyses of transcriptome sequencing identify novel functional lncRNAs in esophageal squamous cell carcinoma. *Oncogenesis* **6**, e297 (2017).

62. C. Trapnell *et al.*, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578 (2012).

63. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).

64. G. Yeo, C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377-394 (2004).

65. X. Jian, E. Boerwinkle, X. Liu, In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res* **42**, 13534-13544 (2014).

66. J. W. Nam *et al.*, Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol Cell* **53**, 1031-1043 (2014).

67. M. Kim, B. H. You, J. W. Nam, Global estimation of the 3' untranslated region landscape using RNA sequencing. *Methods* **83**, 111-117 (2015).

68. L. Kong *et al.*, CPC: assess the protein-coding potential of transcripts using

sequence features and support vector machine. *Nucleic Acids Res* **35**, W345-349 (2007).

69. L. Wang *et al.*, CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* **41**, e74 (2013).

70. The Cancer Genome Atlas Research Network, Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169 (2017).

71. Y. Li *et al.*, Investigation of tumor suppressing function of CACNA2D3 in esophageal squamous cell carcinoma. *PLoS One* **8**, e60027 (2013).

72. A. Wanajo *et al.*, Methylation of the calcium channel-related gene, CACNA2D3, is frequent and a poor prognostic factor in gastric cancer. *Gastroenterology* **135**, 580-590 (2008).

73. A. M. Wong *et al.*, Characterization of CACNA2D3 as a putative tumor suppressor gene in the development and progression of nasopharyngeal carcinoma. *Int J Cancer* **133**, 2284-2295 (2013).

74. M. T. Chung *et al.*, SFRP1 and SFRP2 suppress the transformation and invasion abilities of cervical cancer cells through Wnt signal pathway. *Gynecol Oncol* **112**, 646-653 (2009).

75. H. Zou *et al.*, Aberrant methylation of secreted frizzled-related protein genes in esophageal adenocarcinoma and Barrett's esophagus. *Int J Cancer* **116**, 584-591 (2005).

76. M. B. Clark *et al.*, Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nat Methods* **12**, 339-342 (2015).

77. D. Sharon, H. Tilgner, F. Grubert, M. Snyder, A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**, 1009-1014 (2013).

78. Y. Lubelsky, I. Ulitsky, Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* **555**, 107-111 (2018).

79. Y. Long *et al.*, Conserved RNA-binding specificity of polycomb repressive complex 2 is achieved by dispersed amino acid patches in EZH2. *Elife* **6**, (2017).

80. V. Zarkou, A. Galaras, A. Giakountis, P. Hatzis, Crosstalk mechanisms between the WNT signaling pathway and long non-coding RNAs. *Noncoding RNA Res* **3**, 42-53 (2018).

Supplementary Materials

Datasets. RNA-seq data used throughout this study were either from newly sequenced material or downloaded from public databases. A pair of stranded and unstranded RNA-seq datasets from mES cells were newly sequenced (GSE84946) and 169 publicly available samples of stranded and unstranded RNA-seq data from diverse cell types, including HeLa cells, were downloaded from the ENCODE Project (www.encodeproject.org) and Human BodyMap 2.0 Project (www.broadinstitute.org). To collect data produced by a similar library construction method, samples were selected with the following criteria: 1) samples with poly-A selection applied; 2) samples with stranded RNA-seq data passing a quality control filter (such that the precision of the stranded assembly is equal to or greater than 45%); 3) samples with unstranded RNA-seq and stranded RNA-seq data from the same cell type. After filtration, 122 RNA-seq samples across 35 cell types were analyzed for transcriptome assembly. In addition, 17,009 unstranded RNA-seq samples from 36 normal tissue types, 27 cancer paired with non-cancer tissue types, and 21 cell lines were downloaded from the GTEx Project (http://www.gtexportal.org) and the GDC data portal (the TCGA and CCLE Projects; https://portal.gdc.cancer.gov). The data were filtered with the same criteria above. RNA-seq data from paired non-cancer and cancer samples from 23 Korean esophageal squamous cell carcinoma (ESCC) patients were obtained during this study (GSE130078) and RNA-seq data from 30 Chinese ESCC samples were downloaded from the NCBI Sequence Read Archive (SRA) (www.ncbi.nlm.nih.gov/sra/) under accession number SRP064894. CAGE-seq processed data across 17 tissues for human and 23 tissues for mouse were downloaded from the FANTOM Project (www.fantom.gsc.riken.jp). CPS data from different cell types (HeLa, HEK293, Huh7, and IMR90 for human; mES, 3T3, liver, muscle, heart, white adipose tissue, and kidney for mouse) were downloaded from NCBI Gene Expression Omnibus (GEO) (GSE52531).

Processing of RNA-seq data. To check RNA-seq data quality, RNA-seq reads were mapped to the corresponding reference genomes (hg19 for human and mm9 for mouse) using Bowtie (version 1.0.0) with default parameters. We calculated mismatch rates across the mapped read positions. If the raw read end(s) had a mismatch rate higher than 10%, they were trimmed off using Seqtk (version 1.0-r31). In addition, the trimmed reads with Phred base quality ≤ 20 were filtered using Sickle (version 1.200). The remaining reads were mapped to the corresponding reference genomes using STAR (version 2.5.2b) with mapping parameters "--alignIntronMin 61 --alignIntronMax 265006 --outFilterMultimapNmax 20" for human and "--alignIntronMin 52 --alignIntronMax 240764 –outFilterMultimapNmax 20" for mouse.

Base transcriptome assembly. Base transcriptome assemblies were performed using Cufflinks (version 2.1.1) with assembly parameters "--min-isoform-fraction 0.15 --pre-mrna-fraction 0.2 --junc-alpha 0.001 --small-anchor-fraction 0.06 --min-frags-per-transfrag 12 --max-multiread-fraction 0.65" for unstranded reads and "--min-isoform-fraction 0.05 --pre-mrna-fraction 0.2 --junc-alpha 0.01 --small-anchor-fraction 0.09 --min-frags-per-transfrag 8 --max-multiread-fraction 0.65" for stranded reads.

Benchmarking base assembly. To evaluate the performance of CAFE with other base assemblers, we performed reference-based assembly using Scripture (a beta version) with the default parameter and StringTie (version 1.3.0) with the parameter "-m 0 -j 2 -g 50 -M 0.75" and *de novo* transcriptome assembly using Trinity (version 20140717) with the parameter "--min_contig_length 200" and Velvet (version 1.2.10) with the parameter "-hash_length 25 -min_contig_lgth 50". For benchmarking of reference-based assemblers (Scripture, Cufflinks, and StringTie) and *de novo* assemblers (Trinity and Velvet), each base assembler assembled transcriptomes using stranded and unstranded reads, respectively, and their averaged performance (recall and precision) was measured and then compared to the performance of CAFE in the co-assembly of RPDs and stranded reads.

PacBio transcriptome assembly. To assemble the transcriptome from PacBio Iso-seq data, we downloaded data from human MCF7 cell lines sequenced with a total of 119 SMRT cells from the PacBio website (*81*). For subsequent analysis, we used the 'Iso-seq' protocol from the SMRT Portal provided by PacBio. Using the filtering module in the 'Iso-seq' protocol, we acquired 1,857,590 reads of insert from Iso-seq data. In this step, we set parameters as "Minimum_Full_Passes 0 Minimum Predicted Accuracy 75". Next, the classify module filtered about two-thirds of the reads of insert from the above with parameters set as "Minimum_Sequence_Length 300 Full-Length_Reads_Do_Not_Require_PolyA_ Tails False", leaving 524,084 full-length reads. The last module was the cluster module, which left 80,010 polished isoforms with "Predict_Consensus_Isoforms_ Using_The_ICE_Algorithm True Call_Quiver_To_Polish_Consensus_Isoforms True Minimum_Quiver_Accuracy_To_Classify_An_Isofrom_AS_HQ 0.99" parameters. Finally, GMAP (version 2015-07-23) was used with parameters "-f samse -n 0” to map to the human genome, hg19. The final assembled set contained 47,416 non-redundant transcripts (15,688 genes).

Reference gene annotations. Among the protein-coding and lncRNA genes from GENCODE annotations (human: version 19 for Sections 2 and 3, or version 32 for Section 4, mouse: version M1), those with over 1 fragments per kilobase of exons per million mapped reads (FPKM) in the corresponding types of cell, tissue, or cancer were selected. To build a bona fide lncRNA gene set, we performed the following filtration steps: (1) transcripts shorter than 200nt in length were discarded, and (2) lncRNAs sense-overlapping with exons of known protein-coding genes and noncoding genes (small ncRNAs including miRNAs, snRNAs, and snoRNAs and structural ncRNA genes including rRNAs and tRNAs) were excluded. The references were used to evaluate the quality of the transcriptome assemblies.

Evaluation of transcriptome assembly. To evaluate the quality of transcriptome assembly, we compared the resulting assembly with the reference gene annotations (protein-coding and lncRNA genes, respectively) using in-house script. The recall and precision were estimated at the base, intron and splicing levels of the assembled transfrags.

Evaluation of full-length genes and isoforms. To evaluate how many full-length genes and isoforms were assembled, we collected the transcripts that simultaneously included a TSS in the first exon and a CPS in the last exon of the resulting transfrags. In addition, the transcripts aligned to the reference transcripts with at least a 95% match were regarded as full-length transcripts. At the gene level, gene models that unified all isoform exons were compared.

Expression profiling of lncRNAs. Gene-/isoform-level expression values of the known and novel lncRNAs were calculated with BAM files from the YSH (23 paired samples) and Chinese (15 paired samples) cohorts and with RPD- converted BAM files from the TCGA (95 tumor samples) and GTEx (328 normal samples) cohorts using featureCounts (version 1.5.1) with parameters "--minOverlap 60 -s 2 -p -B -C" and RSEM (version 1.3.0) with parameters "--strandness reverse --mapper star --estimate-rspd".

Differentially expressed (DE) lncRNAs in multiple ESCC cohorts. To identify lncRNAs that were commonly dysregulated in ESCCs from ethnically independent cohorts, we first identified lncRNAs that were significantly DE [≥ two-fold difference with a false discovery rate (FDR) ≤ 0.01] between cancerous and non-cancerous samples in each cohort using DESeq2. Tumor sample data from TCGA and normal esophagus data from GTEx were compared. DE lncRNAs common to all three ESCC cohorts were selected and were considered to be ethnically independent DE lncRNAs.

Cellular and genomic characterization of lncRNAs. To investigate lncRNA subcellular localization patterns, RNA-seq data from nucleus and cytoplasm-fractionated samples from eight different cell lines (A549, GM12878, H1-hESC, HeLa-S3, HepG2, K562, MCF-7, and SK-N-SH) were downloaded from the ENCODE Project (www.encodeproject.org) and processed to produce BAM files, which were then used to calculate lncRNA expression levels. The subcellular localization ratio (SLR; the ratio of expression in the nucleus versus cytosol) was calculated for lncRNAs that are expressed at greater than 1 FPKM in either the nucleus or cytosol in each cell type. LncRNAs were divided into five classes according to the SLR in the eight cell types. LncRNAs that were enriched in the nucleus (SLR ≥ 1.5; “nucleus”) or cytosol (SLR ≤ 0.5; “cytosol”) were classified accordingly. LncRNAs with a comparable ratio in the nucleus and cytoplasm (0.5 < SLR < 1.5) over all cell types were classified as “both”. If a lncRNA showed a different SLR in at least one cell type versus the others, it was labeled as “differentially localized”. LncRNAs that were not expressed in all cell types were classified as “unidentified”.

To annotate enhancer-associated lncRNAs, enhancer annotations from normal esophageal and cancer cell lines (A549, HeLa-S3, HepG2, K562 and MCF-7) were downloaded from EnhancerAtlas (http://www.enhanceratlas.org). Among the lncRNAs that were expressed at greater than 1 FPKM in each cell line, those that overlapped with an enhancer region in at least one cell type were classified as enhancer-associated lncRNAs.

DNA methylation analysis. DNA methylation data from 95 ESCC samples were downloaded from the GDC data portal (the TCGA Project; https://portal.gdc.cancer.gov) in processed form (Data level 3). Beta values of DNA methylation in each sample were reassigned to the transcriptome newly constructed from RNA-seq data from the YSH cohort. Then, for each gene, the median beta values observed in promoter regions were measured for each ESCC sample.

Survival analysis. Clinical information from the YSH cohort was obtained from YSH and information from the TCGA cohort was downloaded from the GDC data portal (https://portal.gdc.cancer.gov). For association with the clinical information for the YSH cohort, RNA levels were measured from 66 cancerous tissue samples, including 23 for RNA-seq, by qRT-PCR. Stage-free survival analyses were performed for YSH and TCGA patients with termination events using the Kaplan-Meier estimate. *P* values were estimated using the log-rank (Mantel-cox) test.

List of Publications

- **Bo-Hyun You*,** Jung-Ho Yoon*, Hoin Kang, Eun Kyung Lee, Sang Kil Lee, and Jin-Wu Nam. HERES, a lncRNA that Regulates Canonical and Noncanonical Wnt Signaling Pathways via Interaction with EZH2. ***Proceedings of the National Academy of Sciences (PNAS),*** *116(49):24620-24629, 2019.*

- Jung-Ho Yoon, **Bo-Hyun You,** Chan Hyuk Park, Yeong Jin Kim, Jin-Wu Nam, and Sang Kil Lee. The long noncoding RNA LUCAT1 promotes tumorigenesis by controlling ubiquitination and stability of DNA methyltransferase1 in esophageal squamous cell carcinoma. ***Cancer Letters,*** *417:47-57, 2018.*

- **Bo-Hyun You,** Sang-Ho Yoon and Jin-Wu Nam. High-confidence coding and noncoding transcriptome maps. ***Genome Research,*** *27(6):1050-1062, 2017.*

- Yong-Hee Rhee, Tae-Ho Kim, A-Young Jo, Mi-Yoon Chang, Chang-Hwan Park, Sang-Mi Kim, Jae-Jin Song, Sang-Min Oh, Sang-Hoon Yi, Hyeon Ho Kim, **Bo-Hyun You,** Jin-Wu Nam, and Sang-Hun Lee. LIN28A enhances the therapeutic potential of cultured neural stem cells in a Parkinson's disease model. ***Brain,*** *139(Pt 10):2722-2739, 2016.*

- Jin-Wu Nam, Seo-Won Choi, and **Bo-Hyun You.** Incredible RNA: Dual Functions of Coding and Noncoding. ***Mol. Cells,*** *39(5):367-374, 2016.*

- MinHyeok Kim*, **Bo-Hyun You*,** and Jin-Wu Nam. Global estimation of the 3' untranslated region landscape using RNA sequencing. ***Methods,*** *83:111-117, 2015.*

* denotes equal contribution

국문요지

고정밀 전사체 지도 작성 및 다양한 암종에서 긴 비번역 RNA 분석

유보현

자연과학대학 생명과학과

한양대학교

RNA 염기서열분석 (high-throughput RNA sequencing; RNA-seq) 기술의 출현과 발전으로, 이를 기반으로 하여 전사체 지도를 재구성하고 새로운 유전자를 동정하는 연구가 지속적으로 이루어지고 있다. 전사체의 상당 부분은 비번역 RNA 로 구성되어있고, 이들 중 긴 비번역 RNA (long noncoding RNA; lncRNA)는 다양한 생물학적 현상에서 중요한 기능을 하는 것으로 알려졌다. 기존의 전사체 지도들은 새로운 lncRNA 동정을 통해 전사체의 다양성과 복잡성을 이해하는데 많은 기여를 했지만, 방향성이 없는 RNA-seq (unstranded RNA-seq) 데이터를 사용하여 여전히 불완전하고 많은 오류를 포함하고 있다. 따라서 우리는 전사체 지도의 정확도를 높이기 위해 고성능 전사체 재구성 파이프라인 CAFE 을 개발했다. CAFE 파이프라인은 최대우도추정을 통해 unstranded RNA-seq 데이터의 방향성을 예측하고, 재구성된 전사체의 5'과 3' 끝부분을 교정하여 전사체 지도를 재구성한다. ENCODE 프로젝트의 대규모 RNA-seq 데이터에 CAFE 파이프라인을 적용하여 이전의 전사체 지도들과 비교하여 더 정확하고 다양한 고정밀 전사체 지도 BIGTranscriptome 을 재구성했다.

CAFE 파이프라인을 다양한 집단 (한국인, 중국인 그리고 서양인)의 식도편평상피세포암 (ESCC) 환자들에서 생산된 RNA-seq 데이터에 활용하여 전사체 지도를 재구성했다. 그 결과, 1,924 개의 새로운 lncRNA 들을 동정하고, 이들 중 다양한 집단에서 공통적으로 비정상적 발현을 하는 113 개의

lncRNA 들을 찾았다. 6 개의 lncRNA 들은 ESCC 환자들의 임상 정보와 유의미한 관련성을 나타냈고, 이들을 통해 ESCC 환자들을 예후에 따른 4 개의 그룹으로 분류할 수 있었다. 본 연구에서 새롭게 동정한 lncRNA 을 HERES 로 명명하고, *in vitro* 와 *in vivo* 실험을 통해 HERES 가 ESCC 의 발생과 분화를 촉진하는 것을 확인했다. 또한, HERES 가 EZH2 와 상호작용을 통해 Wnt 신호전달 체계에 중요한 요소인 CACNA2D3, SFRP2 와 CXXC4 의 발현을 동시에 조절하는 것을 밝혔다.

마지막으로 16,000 개 이상의 대규모 RNA-seq 데이터를 확보하여 암과 조직에 특화된 전사체 지도를 재구성했다. 전사체 지도에는 새롭게 동정한 lncRNA 들과 암과 조직에서 특이적으로 발현되는 lncRNA 들이 다수 포함되어 있었다.

결과적으로 CAFE 파이프라인을 통해 정확한 전사체 지도의 재구성이 가능하고, 이를 기반으로 번역과 비번역 전사체의 다양성과 복잡성을 이해하며, 이는 HERES 와 같이 암의 진단과 치료에 활용할 수 있는 바이오마커와 치료 타겟을 발굴하는데 큰 도움을 줄 것 이다.

감사의 글

지난 6 년의 학위 과정 동안 정말 많은 분들의 도움을 받았습니다. 많은 분들의 가르침과 격려가 있었기 때문에 부족한 제가 무사히 학위 과정을 마칠 수 있었습니다. 그 동안 제게 도움을 주셨던 모든 분들께 이 자리를 빌어 감사의 말씀을 드리고 싶습니다.

우선 지도 교수님이신 남진우 교수님께 진심으로 감사드립니다. 교수님의 연구에 대한 열정과 긍정적인 에너지를 보며 연구자의 마음가짐과 자세를 배울 수 있었습니다. 항상 옆에서 많은 가르침을 주시고, 너그럽게 이해하고 기다려 주셔서 감사합니다. 앞으로 교수님의 가르침에 보답할 수 있도록 항상 노력하겠습니다.

박사 학위 심사를 맡아주신 심사위원장 김태민 교수님, 백대현 교수님, 심지원 교수님, 최준호 교수님께 진심으로 감사드립니다. 심사위원 교수님들의 아낌없는 지도와 조언으로 학위 논문을 완성할 수 있었습니다.

석사 학위 심사위원장을 맡아주신 송영수 교수님께도 감사드립니다. 힘든 시기에 교수님의 따뜻한 말이 제게 큰 힘이 되었습니다. 바쁘신 일정에도 석사, 박사 학위 심사를 모두 맡아주신 심지원 교수님께 다시 한번 감사드립니다.

HERES 연구에 많은 도움을 주신 이상길 교수님, 이은경 교수님, 윤정호 박사님, 강호인 박사님께 깊이 감사드립니다. 교수님들과 박사님들의 지도와 도움으로 연구의 결실을 맺을 수 있었습니다.

연구를 함께 시작하고, 지금은 각자의 길을 걷고 있는 현주, 현윤, 좌원, 서원, 경우, 효선에게 감사의 말을 전합니다. 연구실에서 함께 보낸 시간들은 제게 큰 힘이 됐고, 평생 잊지 못할 순간입니다. 학위 과정 동안 연구실에서 많은 도움을 주신 손장일 박사님, Vipin, 경태, 상호, 석주, 동은 선생님, 민학, 도헌, 성진, 은경, 우희에게 감사드립니다. 연구실에서 함께해준 동료들이 있었기 때문에 학위 과정을 즐겁게 보낼 수 있었습니다.

마지막으로 어떤 상황에서도 절 믿고 응원해주시는 아버지, 어머니, 누나, 부족한 저를 따뜻하게 가족으로 받아주신 장인어른, 장모님, 희애, 그리고 오랜 시간 동안 변함없이 곁에 있어준 희수에게 감사합니다.

Declaration of Ethical Conduct in Research

I, as a graduate student of Hanyang University, hereby declare that I have abided by the following Code of Research Ethics while writing this dissertation thesis, during my degree program.

"First, I have strived to be honest in my conduct, to produce valid and reliable research conforming with the guidance of my thesis supervisor, and I affirm that my thesis contains honest, fair and reasonable conclusions based on my own careful research under the guidance of my thesis supervisor.

Second, I have not committed any acts that may discredit or damage the credibility of my research. These include, but are not limited to : falsification, distortion of research findings or plagiarism.

Third, I need to go through with Copykiller Program(Internet-based Plagiarism-prevention service) before submitting a thesis."

NOVEMBER 13, 2019

Degree :	Doctor	
Department :	DEPARTMENT OF LIFE SCIENCE	
Thesis Supervisor :	Nam, Jin-Wu	
Name :	You Bo-Hyun	(Signature)

연구 윤리 서약서

본인은 한양대학교 대학원생으로서 이 학위논문 작성 과정에서 다음과 같이 연구 윤리의 기본 원칙을 준수하였음을 서약합니다.

첫째, 지도교수의 지도를 받아 정직하고 엄정한 연구를 수행하여 학위논문을 작성한다.

둘째, 논문 작성시 위조, 변조, 표절 등 학문적 진실성을 훼손하는 어떤 연구 부정행위도 하지 않는다.

셋째, 논문 작성시 논문유사도 검증시스템 "카피킬러"등을 거쳐야 한다.

2019년11월13일

학위명 : 박사

학과 : 생명과학과

지도교수 : 남진우

성명 : 유보현 (서명)

한 양 대 학 교 대 학 원 장 귀 하