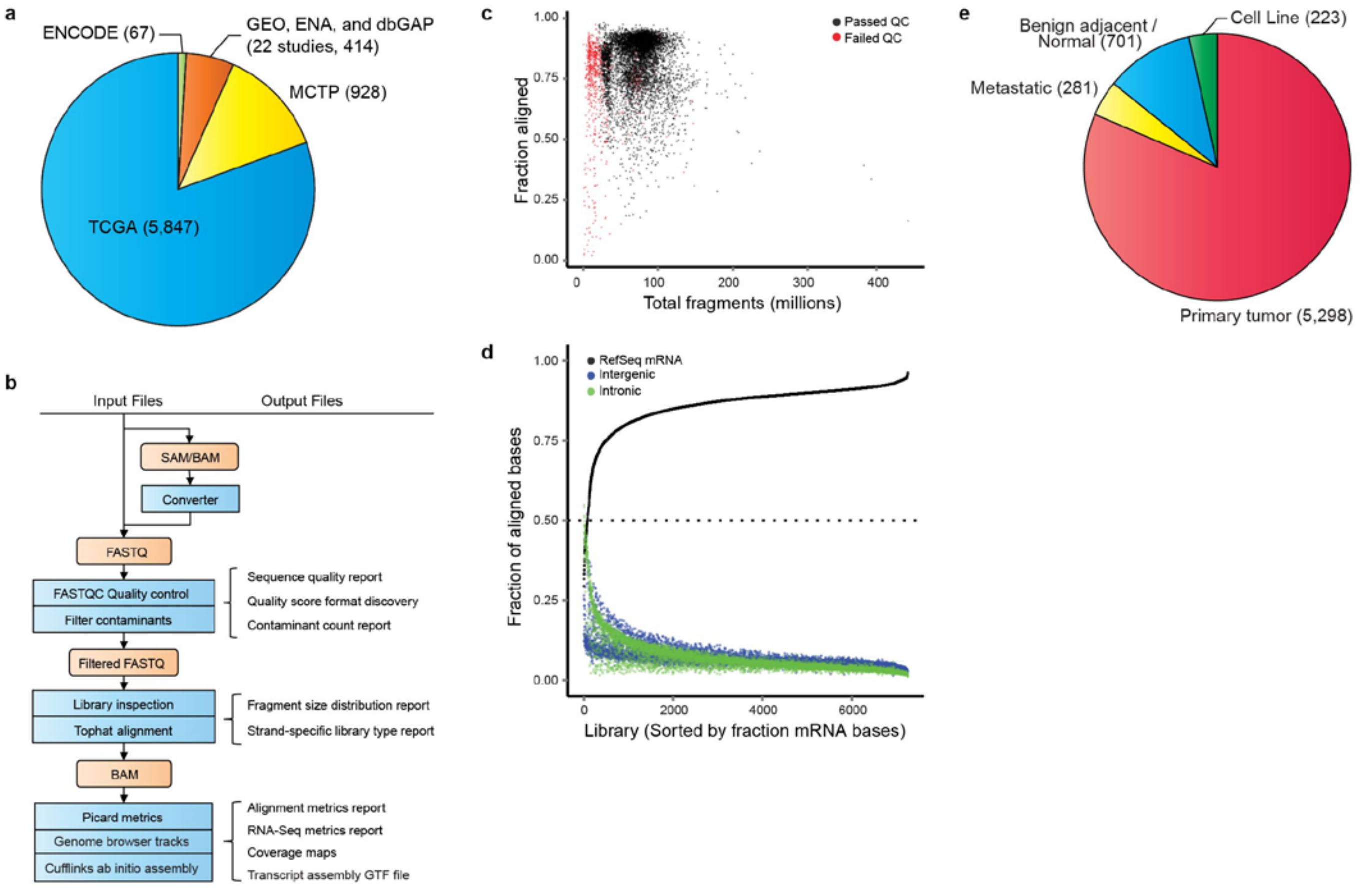
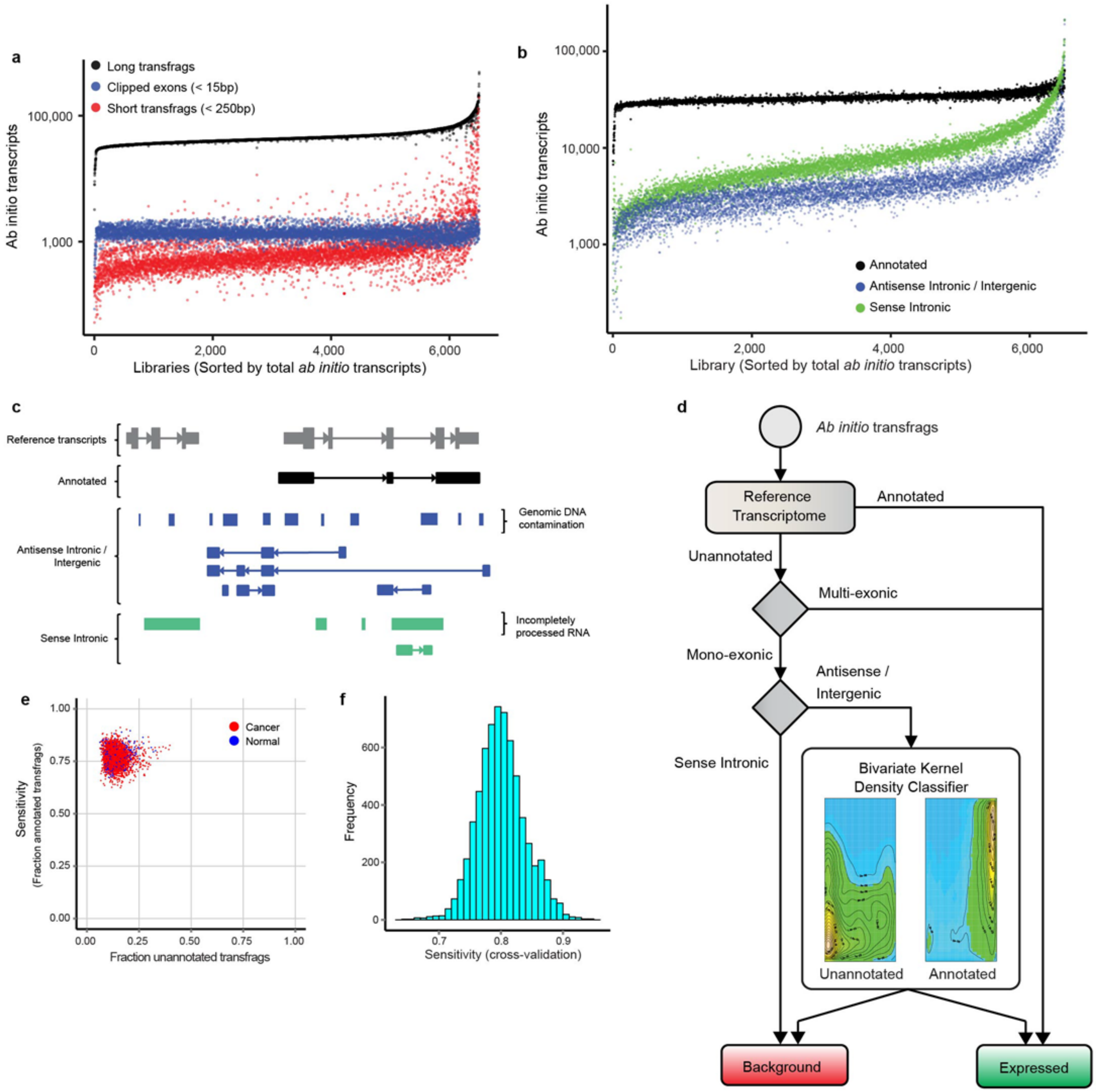


Supplementary Figure 1

Curation and processing of samples in the MiTranscriptome compendia

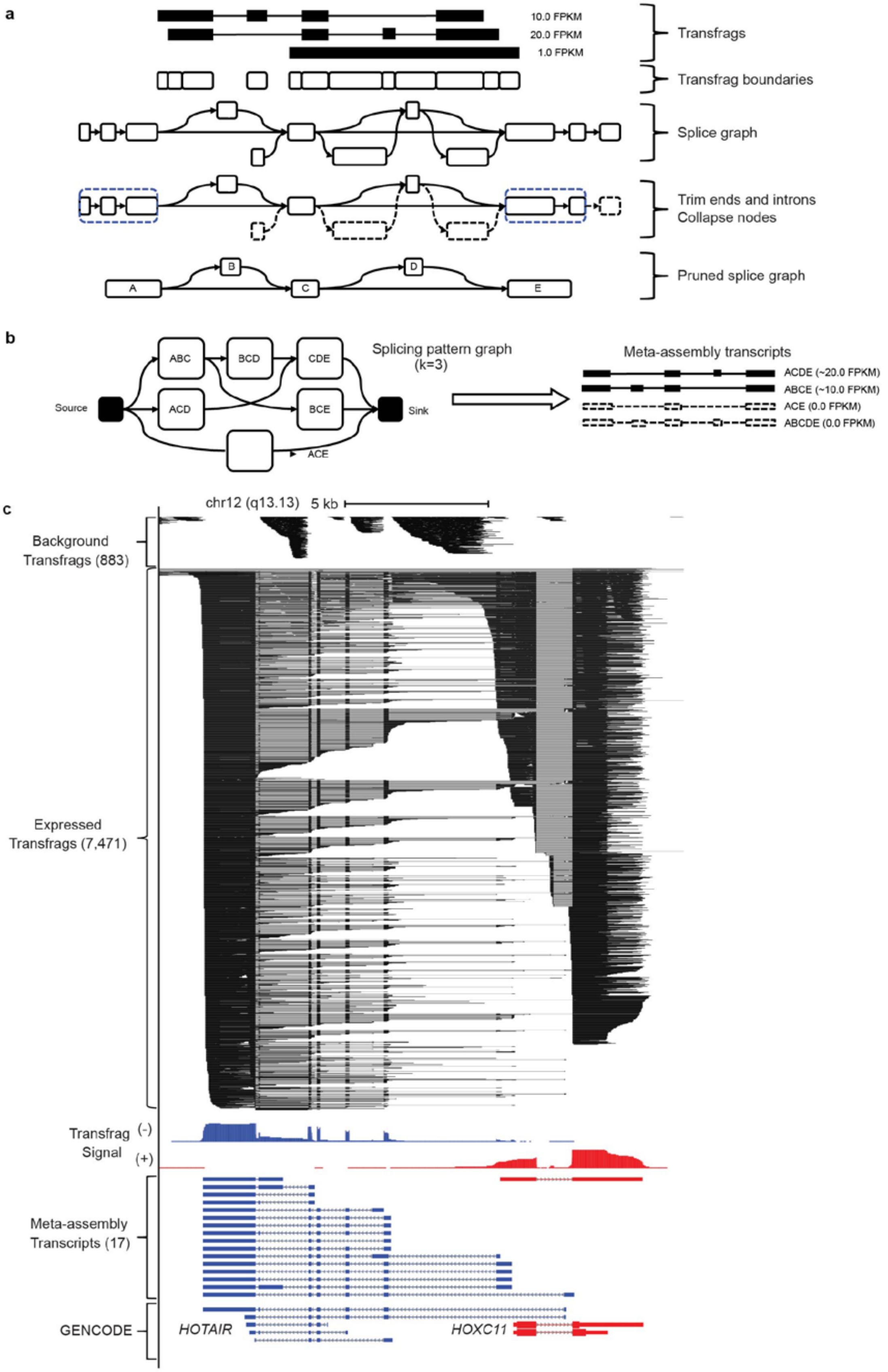
a, Pie chart showing the number of studies curated from TCGA, ENCODE, MCTP, and other publicly available datasets. **b**, Workflow for bioinformatics processing of individual RNA-Seq libraries. Datasets downloaded as BAM files were first converted to FASTQ format. Quality assessment of FASTQ files was performed using FASTQC. Reads mapping to mitochrondria, ribosomal RNA, poly-A sequence, poly-C sequence, or phiX virus (a spiked-in control) were filtered. Fragment length distribution and orientation were determined by mapping a subset of the input reads to a set of large human exons (>500bp). Reads were aligned using TopHat (v2.0.6) with Bowtie2 (v2.1.0). Gene fusion calling was performed using Tophat-Fusion (v2.0.6) with bowtie1 (v0.12.9). Read alignment metrics were computed using Picard Tools, and genome track information was generated using BEDTools and UCSC binary utilities. Finally, ab initio transcriptome assembly was performed using Cufflinks version 2.0.2. **c**, Scatter plot showing total fragments (x axis) and the fraction of aligned fragments (y axis) for each RNA-Seq library. Coarse quality control filters used to remove libraries with fewer than 20 million total fragments or 20 million alignments (red point). **d**, Dot plot showing for each library the fraction of aligned bases corresponding to Refseq mRNA (black points), intronic regions (green points), or intergenic regions (blue points) on the y axis. Libraries with fewer than 50% of aligned bases corresponding to RefSeq mRNA were filtered (dotted line). **e**, Pie chart showing numbers of primary tumors (red), metastatic tumors (yellow), benign adjacent tissues or tissues from healthy individuals (blue), or cell lines (green) for 6,503 RNA-Seq libraries that passed coarse quality control filters.



Supplementary Figure 2

Transfrag filtering

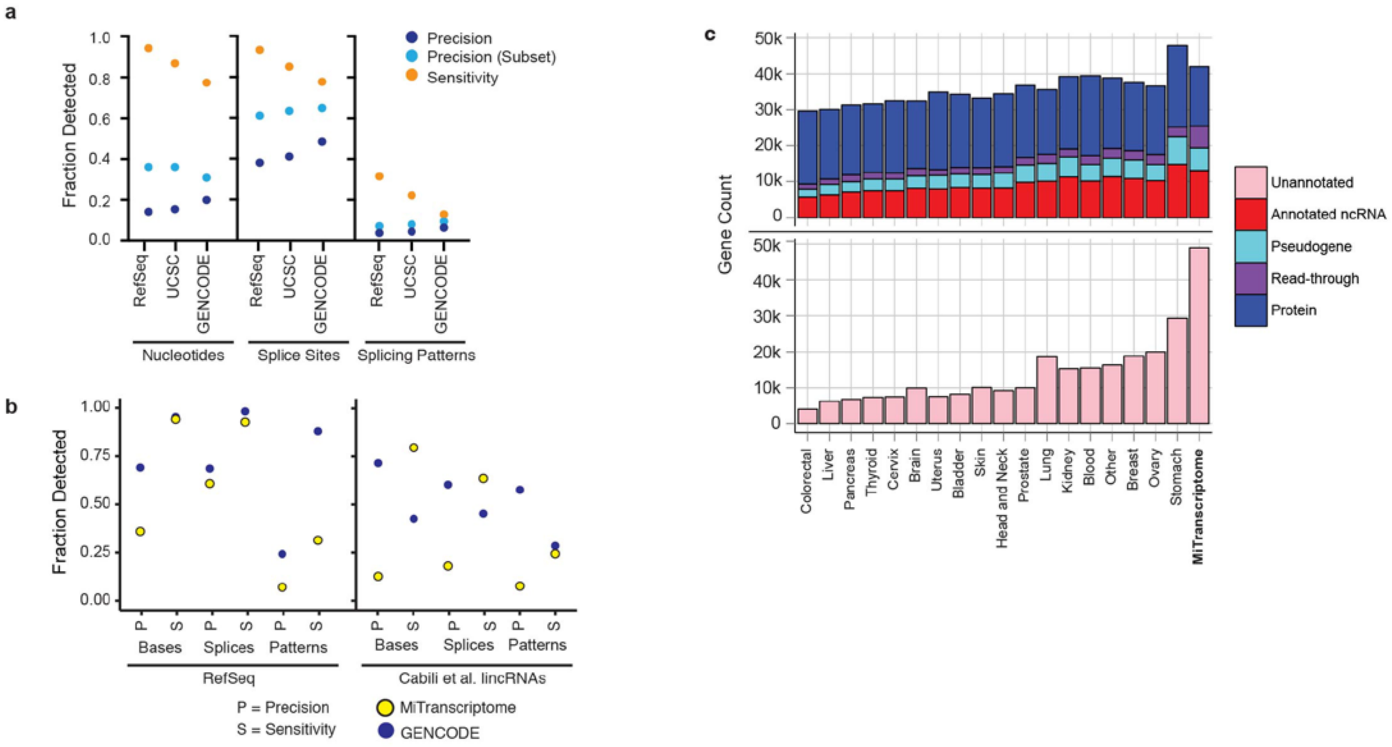
a, Dot plot shows the numbers of short transfrags (red), short clipped exons (blue), and long transfrags (black) for each library. **b**, Dot plot shows the numbers of unannotated intergenic or antisense transfrags (blue), sense intronic transfrags (green), and annotated transfrags (black) for each library. **c**, Example transcript models illustrating categories of ab initio transcripts and sources of background noise. Annotated transfrags (black) overlap reference transcripts on the same strand. Unannotated antisense intronic or intergenic transfrags (blue) may be confounded by genomic DNA contamination. Unannotated sense intronic transfrags (green) may be confounded by both genomic DNA and incompletely processed RNA contamination. **d**, Decision tree depicting transfrag filtering steps for a single library. First, transfrags were labeled ‘Annotated’ or ‘Unannotated’ based on overlap with a reference transcriptome catalog. Annotated transfrags and unannotated multi-exonic transfrags were considered expressed. Unannotated mono-exonic transfrags within introns in the sense orientation of an overlapping transcript were discarded as incompletely processed RNA artifacts. Unannotated antisense or intergenic mono-exonic transfrags were subjected to a bivariate kernel density classification method to discriminate recurrent, reliable transcription from genomic DNA contamination artifacts. Transfrags predicted as ‘expressed’ were incorporated into meta-assemblies. **e**, Scatter plot comparing the sensitivity of the mono-exonic transfrag classifier for correctly detecting annotated transcripts (y axis) and the fraction of unannotated transfrags predicted to be expressed (x axis). **f**, Histogram demonstrating sensitivity for correctly detecting annotated test transcripts held out of the classifier training process.


a
10.0 FPKM
20.0 FPKM
1.0 FPKM
Transfrags
Transfrag boundaries
Splice graph
Trim ends and introns
Collapse nodes
A
B
C
D
E
Pruned splice graph
b
Source
ABC
BCD
CDE
ACD
BCE
ACE
Sink
Splicing pattern graph
(k=3)
Meta-assembly transcripts
ACDE (~20.0 FPKM)
ABCE (~10.0 FPKM)
ACE (0.0 FPKM)
ABCDE (0.0 FPKM)
c
chr12 (q13.13)
5 kb
Background
Transfrags (883)
Expressed
Transfrags (7,471)
Transfrag
Signal
(-)
(+)
Meta-assembly
Transcripts (17)
GENCODE
HOTAIR
HOXC11

Supplementary Figure 3

Meta-assembly

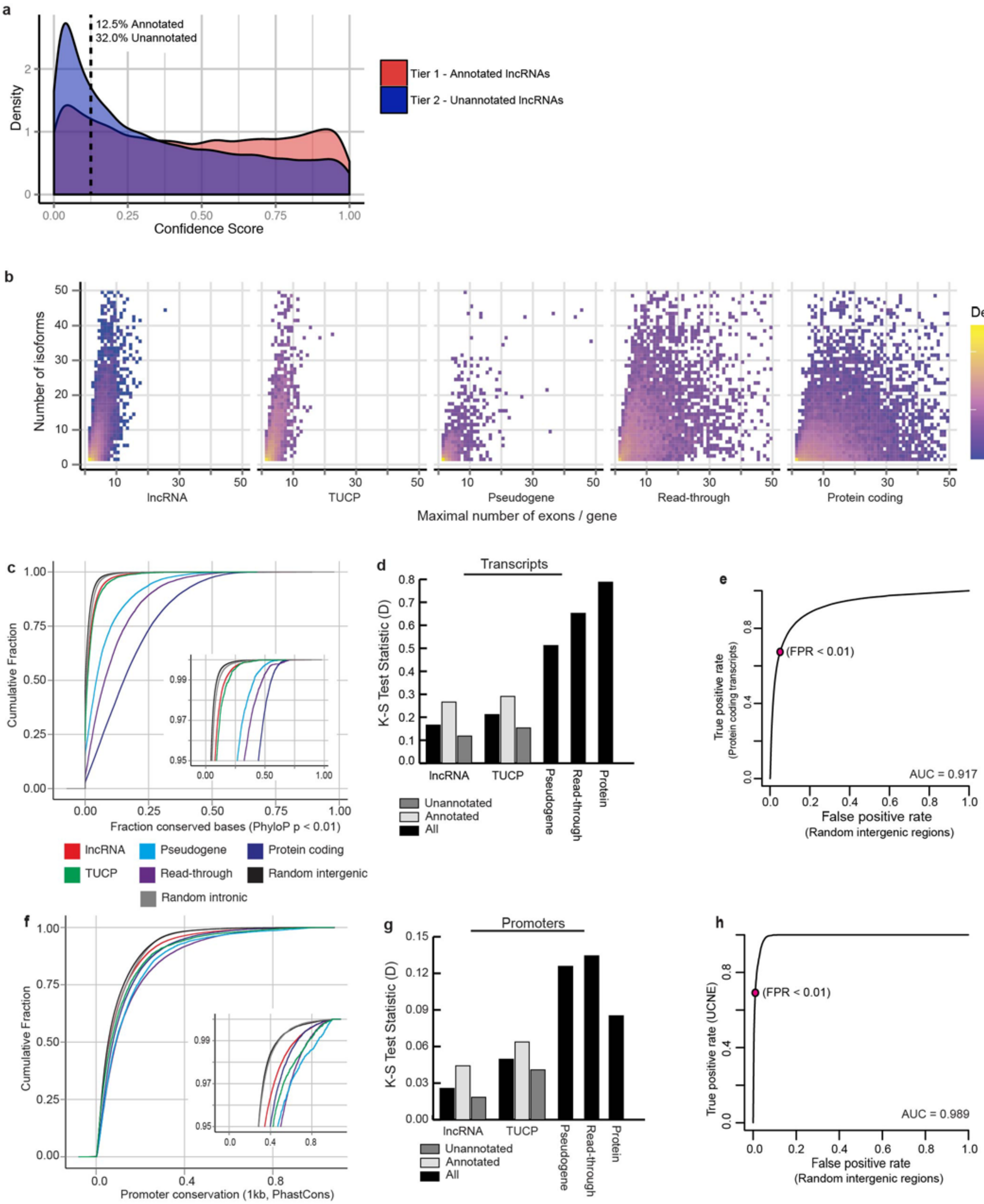
a, Schematic of transcriptome meta-assembly algorithm using a simplified example with three transfrags transcribed from left-to-right. The input to the meta-assembly is a list of weighted transfrags (in this case, the weights correspond to FPKM expression values). First, a splice graph is constructed using the transfrag exon boundaries. The splice graph is a directed acyclic graph (DAG) with nodes (rounded rectangular boxes) representing contiguously transcribed genomic bases and edges (arrows) corresponding to possible alternative splicing and promoter usage. The splice graph is then trimmed to remove lowly expressed starting/ending nodes, and adjacent nodes with a degree of one are collapsed. **b**, The pruned splice graph from panel a is subjected to meta-assembly. To encapsulate splicing pattern information present in the original transfrags the pruned splice graph is converted into a splicing pattern graph. A splicing pattern graph is a *De Bruijn* graph where each node represents a group of *k* consecutive connected nodes from the splice graph (in this example *k*=3), and edges connect adjacent node groups. In real cases *k* is automatically chosen to optimize the number of nodes in the splicing pattern graph. Finally, the splicing pattern graph is repeatedly traversed using a greedy dynamic programming algorithm to determine the set of most highly abundant isoforms from the graph. In this example, isoforms ACDE and ABCE recapitulate input transfrags with nearly identical FPKM values, and invalid isoform combinations ACE and ABCDE are discarded. **c**, Genome view showing an example of the meta-assembly procedure for breast cohort transfrags in a chromosome 12q13.3 locus containing the lncRNA *HOTAIR* and the protein-coding gene *HOXC11* on opposite strands (chr12:54,349,995-54,377,376, hg19). 883 transfrags were considered background noise and not used for meta-assembly. A dense cluster of 7,471 expressed transfrags from 1,076 breast RNA-Seq libraries was used as input. The aggregated transfrag signal on the positive (+) and negative (-) strands is shown below. Meta-assembly produced 17 transcripts from the transfrags, including transcripts that matched GENCODE *HOTAIR* and *HOXC11* splicing patterns as well as *HOTAIR* transcripts with unannotated splice sites.



Supplementary Figure 4

Characterization of unannotated transcripts

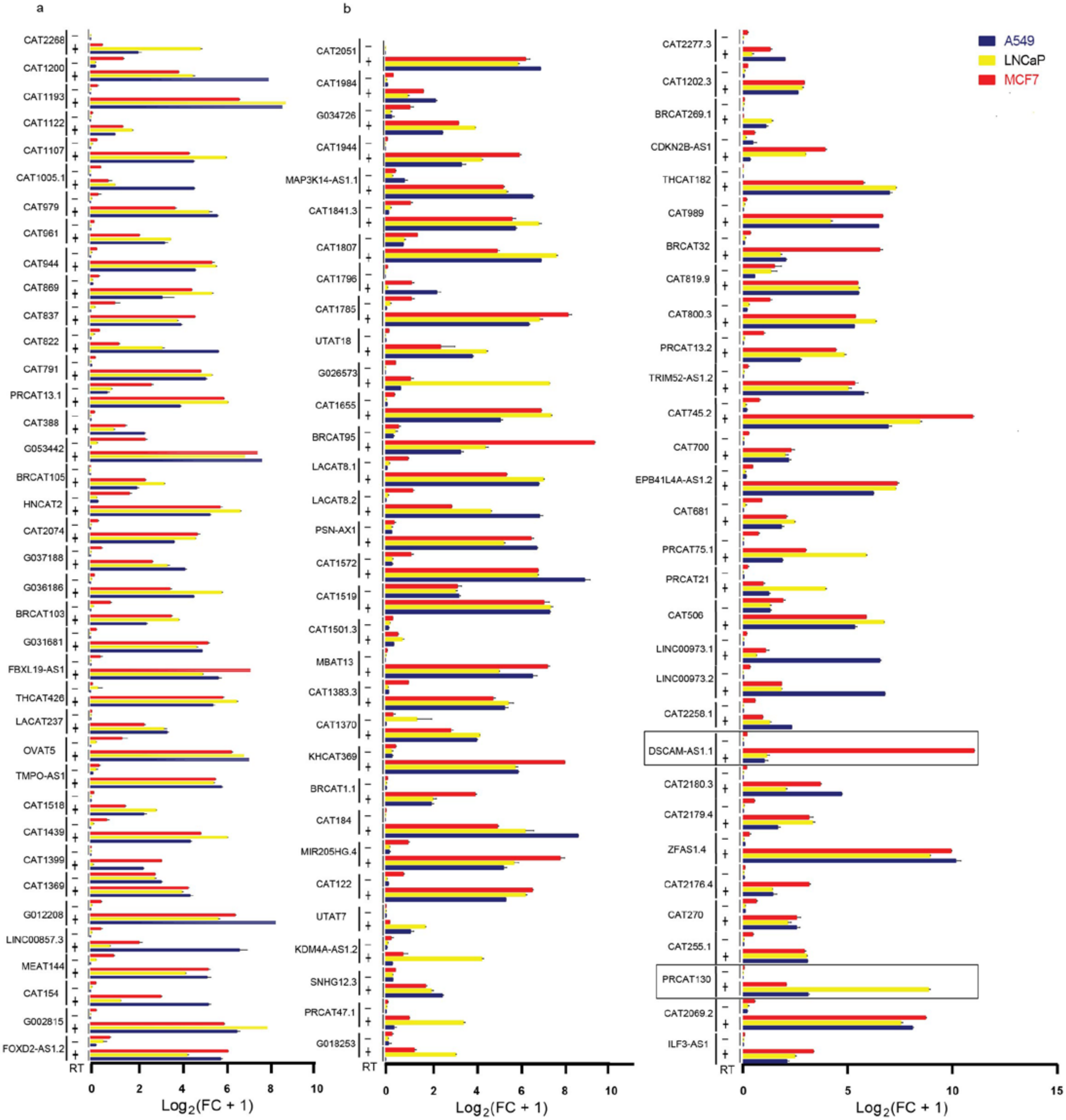
a, Dot plots depicting comparison of MiTranscriptome with reference transcripts from RefSeq, UCSC, or GENCODE. Precision (blue), precision for the subset of transcripts overlapping annotated transcripts (light blue), and sensitivity (orange) are plotted for each comparison. **b**, Dot plots comparing the basewise, splice site, and splicing pattern precision and sensitivity of MiTranscriptome and GENCODE using RefSeq (left) or Cabili et al. lncRNAs (right). **c**, Bar plots comparing numbers of unannotated versus different classes of annotated transcripts for each of the 18 cohorts. (top) Stacked bar plot showing annotated ncRNAs (red), pseudogenes (cyan), read-throughs (purple), and protein-coding genes (blue). (bottom) Bar plot showing unannotated transcripts (pink).



Supplementary Figure 5

MiTranscriptome characterization

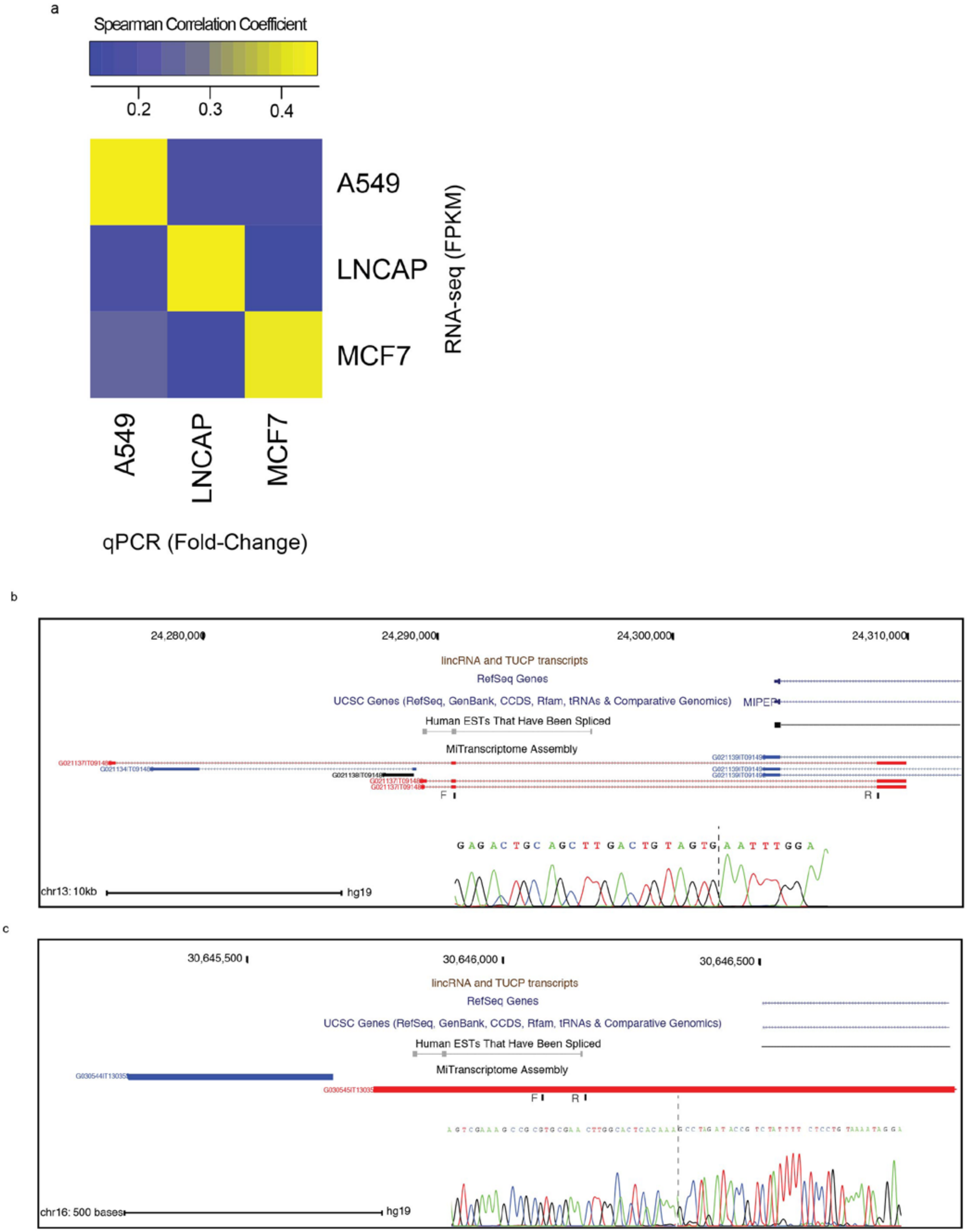
a, Density histogram depicting the confidence scores for annotated and unannotated lncRNAs. **b**, Comparison of the relationship of maximum number of exons per gene to the number of isoforms per gene. LncRNAs tend to have fewer exons than protein-coding genes, but they have complex splicing patterns that yield multiple transcript isoforms. **c**, Cumulative distribution plot for basewise conservation fraction of (blue) proteins, (purple) read-throughs, (cyan) pseudogenes, (green) TUCPs, (red) lncRNAs. Random intergenic (black) and intronic (grey) regions are plotted as controls. Inset plot highlights upper 5th percentile of distribution. **d**, Bar plot showing KS test statistics for classes of transcripts versus random intergenic controls. **e**, ROC curve for predicting conservation of protein-coding genes versus random intergenic controls. The cutoff (pink point) chosen for calling highly conserved transcripts is plotted. **f**, Cumulative distribution plot for promoter conservation (legend shared with **c**). Inset plot highlights upper 5th percentile of distribution. **g**, Bar plot showing KS tests for promoter conservation versus random intergenic regions. **h**, ROC curve for predicting ultra-conserved non-coding elements versus random intergenic regions. Cutoff (pink point) chosen for nominating ultraconserved lncRNAs is plotted.



Supplementary Figure 6

Validation of lncRNA transcripts

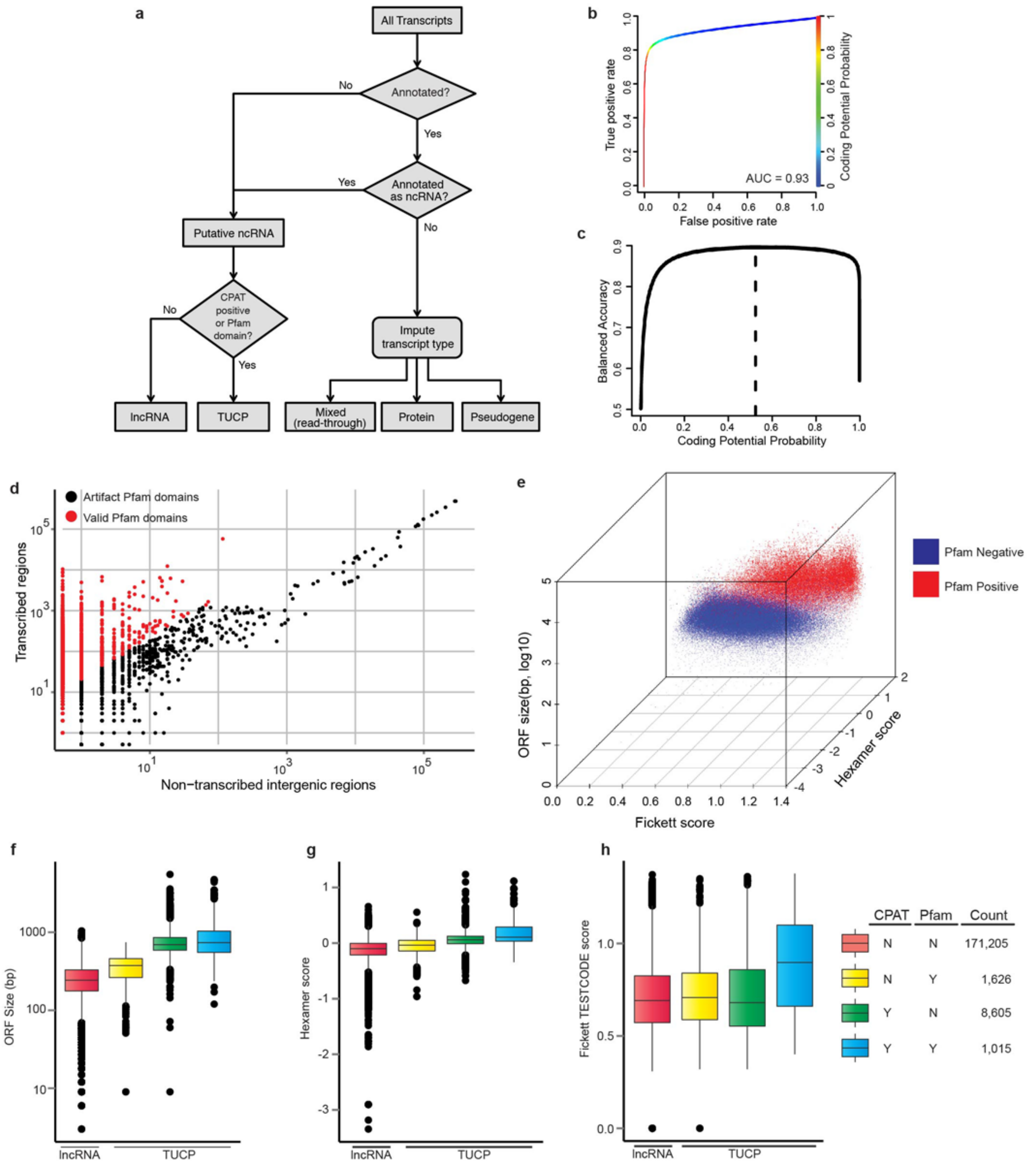
One hundred lncRNA transcripts were validated by qRT-PCR across A549, LNCaP, and MCF7 cell lines using a +/- RT approach. Ct values were first normalized to housekeeping genes (*CHMP2A, EMC7, GPI, PSMB2, PSMB4, RAB7A, REEP5, SNRPD3*) and then to the median value of all samples using the delta-delta Ct method. Here, data is plotted as a logirithmic of fold change over median with standard error of the mean. Validation was performed on **a**, thirty-eight mono-exonic transcripts and **b**, sixty-two multi-exonic transcripts with associated MiTranscriptome name or Gene ID. Boxed transcripts are two representative examples of lncRNAs with lineage/cancer specficity in breast, or prostate according to SSEA analysis (Supplementary Table 10) whose cell line expression profile (by qRT-PCR) reflects what is expected from tissue analysis.


a
Spearman Correlation Coefficient
0.2
0.3
0.4
A549
LNCAP
MCF7
RNA-seq (FPKM)
A549
LNCAP
MCF7
qPCR (Fold-Change)
b
24,280,000
24,290,000
24,300,000
24,310,000
lincRNA and TUCP transcripts
RefSeq Genes
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
MIPEP
Human ESTs That Have Been Spliced
MiTranscriptome Assembly
G021137|T09148
G021134|T09148
G021138|T09148
G021137|T09148
G021137|T09148
G021139|T09149
G021139|T09149
G021139|T09149
F
R
GAGACTGCAGCTTGACTGTAGTGAATTTGGA
chr13: 10kb
hg19
c
30,645,500
30,646,000
30,646,500
lincRNA and TUCP transcripts
RefSeq Genes
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
Human ESTs That Have Been Spliced
MiTranscriptome Assembly
G030544|T13035
G030545|T13035
F
R
AGTCGAAAGCCGCGTGCGAACTTGGCACTCACAAAGCCTAGATACCGTCTATTTTCTCCTGTAAAATAGGA
chr16: 500 bases
hg19

Supplementary Figure 7

Validation of lncRNA transcripts, continued

a, Heatmap representation of the correlation between qPCR (fold change over median) with RNA-seq (FPKM) of 100 selected transcripts in cell lines A549, LNCaP, and MCF7. **b** and **c**, Representative example of two of twenty previously unannotated lncRNA transcripts that were analyzed by Sanger sequencing to ensure primer specificity with their associated chromatograms. As seen in the UCSC Genome Browser View a (**b**) multi-exonic lncRNA (Gene ID: G021137) and (**c**) mono-exonic lncRNA (Gene ID: G030545).

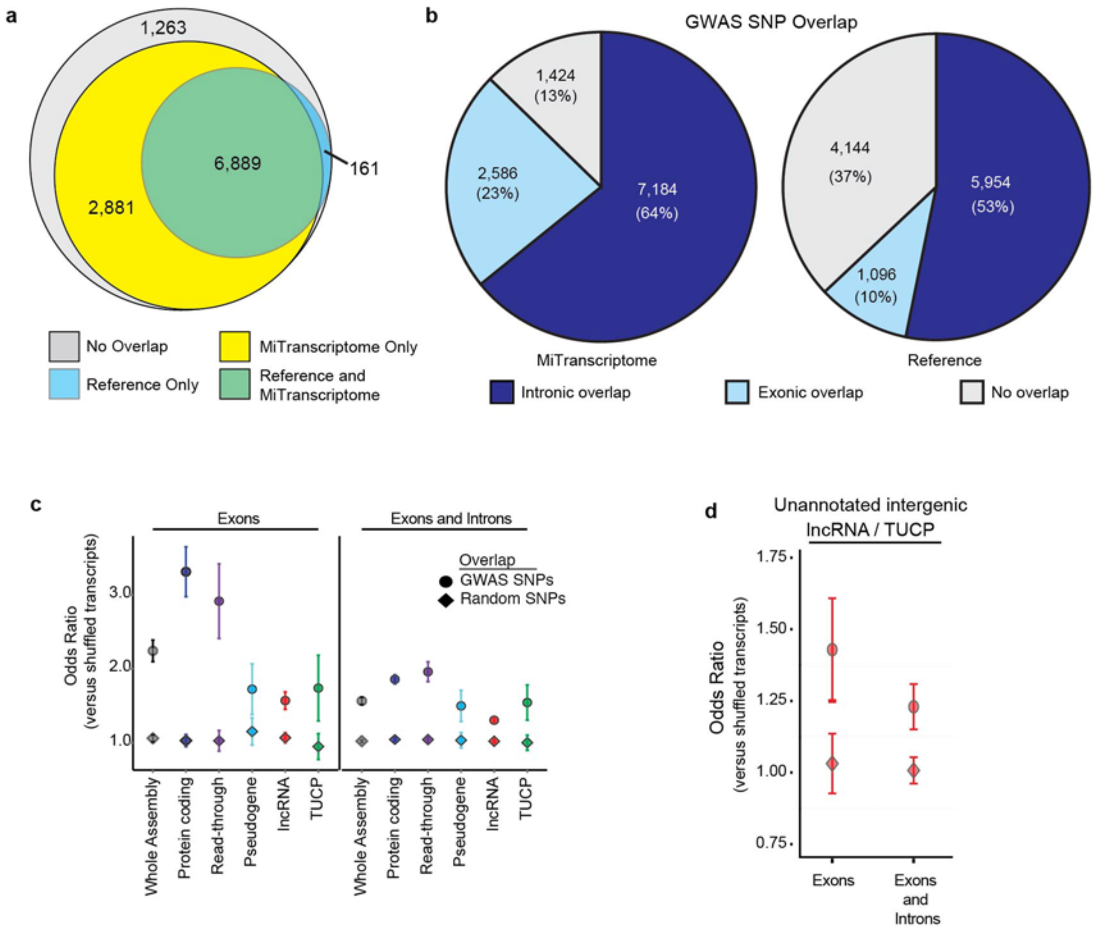


	CPAT	Pfam	Count
🟥	N	N	171,205
🟨	N	Y	1,626
🟩	Y	N	8,605
🟦	Y	Y	1,015

Supplementary Figure 8

Classification of transcripts of unknown coding potential

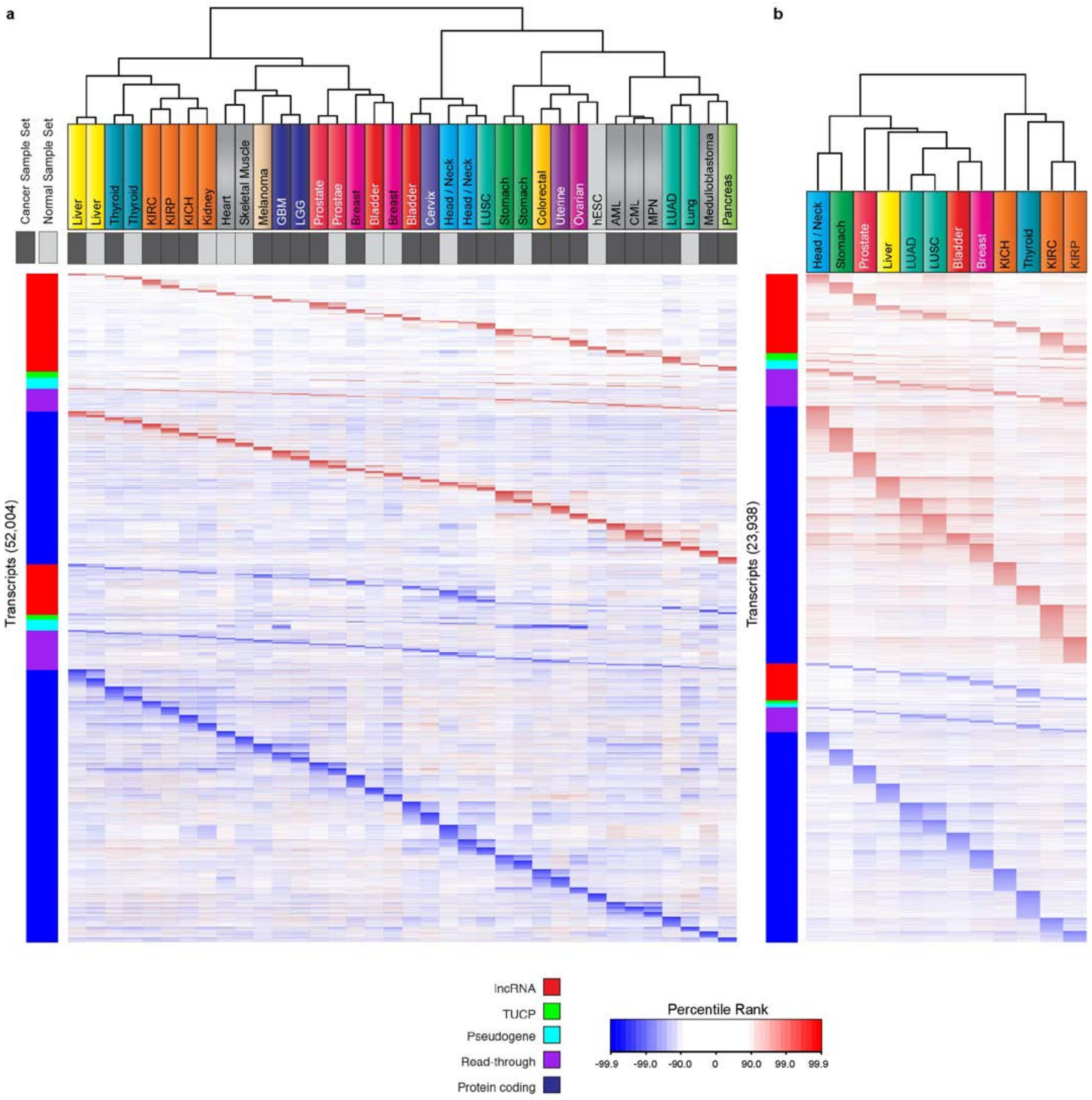
a, Decision tree showing categorization of *ab initio* transcripts. Unannotated transcripts and annotated ncRNAs were classified as either lncRNA or TUCP. Transcript categories for protein-coding genes, pseudogenes, and read-throughs were imputed from overlapping reference annotations. **b**, ROC curve comparing false positive rate (*x* axis) with true positive rate (*y* axis) for CPAT coding potential predictions of ncRNAs versus protein-coding genes. **c**, Curve comparing probability cutoff (*x* axis) with balanced accuracy (*y* axis). Dotted line show cutoff used to call TUCP transcripts. **d**, Scatter plot comparing frequencies of Pfam domain occurrences in non-transcribed intergenic space versus transcribed regions. Points in red were considered valid Pfam domain hits and points in black were considered artifacts. **e**, Three-dimensional scatter plot comparing Fickett score (*x* axis), ORF size (*y* axis), and Hexamer score (*z* axis) for all transcripts. Red points contain valid Pfam domains and blue points do not. **f**,**g**,**h** Boxplots comparing ORF size (**f**), Hexamer score (**g**), and Fickett score (**h**) for lncRNAs (red), TUCPs predicted by Pfam only (yellow), TUCPs predicted by CPAT (green), and TUCPs predicted by both Pfam and CPAT (blue).



Supplementary Figure 9

Enrichment of MiTranscriptome assembly for disease-associated regions

a, Venn diagram comparing coverage of disease- or trait-associated genomic regions (i.e. GWAS SNPs) for the MiTranscriptome assembly (yellow) in comparison to reference catalogs (blue), with the area of intersection is shaded green. **b**, Pie charts comparing distributions of intronic and exonic GWAS SNP coverage of the MiTranscriptome assembly (left) and reference catalogs (right). **c**, Dot plot displaying enrichment of GWAS SNPs versus random SNPs for different transcript categories. Enrichment odds ratios (transcripts-SNP overlaps versus shuffled transcript-SNP overlaps) are plotted on the *y* axis. Points indicate the mean of 100 permutations for tests of enrichment with GWAS SNPs (circle) or random SNPs (diamond), and error bars depict +/- 2 standard deviations of the distribution of odds ratios. Both exonic and whole-transcript enrichment is reported. **d**, Dot plot showing enrichment of GWAS SNPs (circle) versus random SNPs (diamond) for novel intergenic lncRNAs and TUCPs. Enrichment odds ratios (transcripts-SNP overlaps versus shuffled transcript-SNP overlaps) are plotted on the *y* axis. Points indicate the mean of 100 shuffles for comparisons with GWAS SNPs (circle) or random SNPs (diamond), and error bars depict +/- 2 standard deviations of the distribution of odds ratios. Both exonic and whole-transcript enrichment is reported.

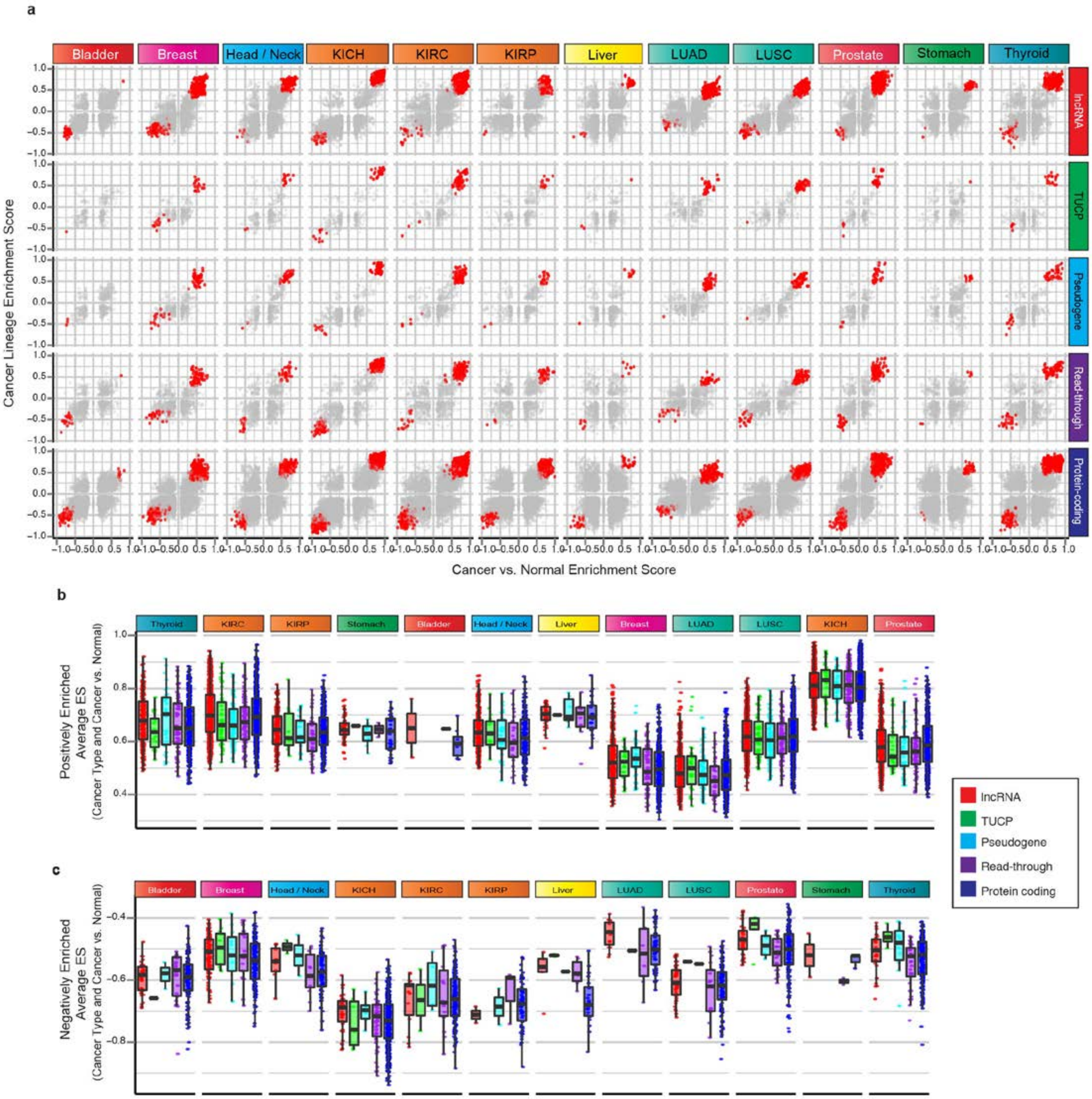


Supplementary Figure 10

Discovery of lineage-associated and cancer-associated transcripts

a, Heatmap of lineage-specific transcripts (LATs) nominated by SSEA. Each column represents a sample set from one of 25 cancer (dark grey) and 13 normal (light grey) lineages and each row represents an individual transcript. Colored labels above columns reflect organ system cohorts used in assembly. Row side colors correspond to lncRNAs (red), TUCPs (green), pseudogenes (cyan), read-throughs (purple), and protein coding transcripts (blue). All transcripts were statistically significant (FDR < $1e^{-7}$) and ranked in the top

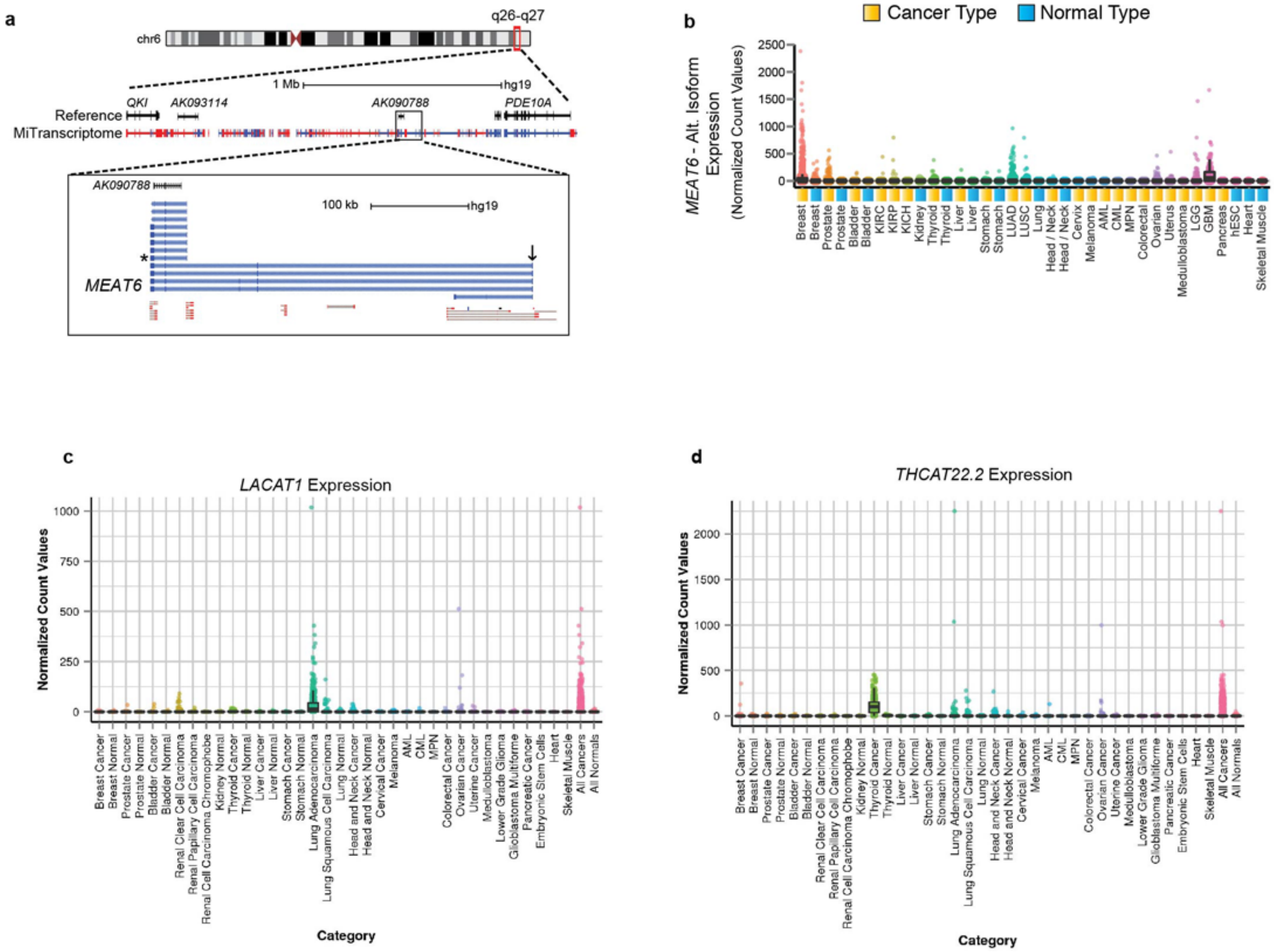
1% most positively or negatively enriched transcripts within at least one sample set. The heatmap color spectrum corresponds to percentile ranks, with under-expressed transcripts colored blue and over-expressed transcripts colored red. The column dendrogram shows unsupervised hierarchical clustering of sample sets. **b**, Heatmap of cancer-specific transcripts (CATs) nominated by SSEA. Columns represent 12 cancer types, and colored column labels reflect organ system cohorts used in assembly. All transcripts were statistically significant (FDR < $1e^{-3}$) and ranked in the top 1% most positively or negatively enriched transcripts within at least one sample set. Column dendrogram shows unsupervised clustering results. Row side color and heatmap color schemes are identical to (**a**).



Supplementary Figure 11

Lineage-specific and cancer-specific transcripts

a, Scatter plot grid showing lineage-specific and cancer-specific transcripts (CLATs) nominated by SSEA. A row of scatter plots for transcript category is plotted across 12 cancer types. Each plot shows Cancer vs. Normal enrichment score (*x* axis) and the Cancer Lineage enrichment score (*y* axis). Red points indicate CLATs within the respective cancer types, and grey points indicate CLATs for other cancer types. **b** and **c**, Boxplots comparing the performance of (**b**) positively enriched CLATs and (**c**) negatively enriched CLATs for each transcript category across 12 cancer types. The average of the lineage and cancer versus normal ES is plotted on the y axis.



Supplementary Figure 12

Examples of cancer and/or lineage associated transcripts

a, Genomic view of chromosome 6q26-q27 locus. Protein coding genes *QKI* and *PDE10A* flank an intergenic region with two annotated lncRNAs, *AK093114* and *AK090788*. MiTranscriptome transcripts are shown in a dense view populating this intergenic space. Most zoomed view (bottom) depicts *MEAT6*, a melanoma associated lncRNA. *AK090788* overlaps a portion of *MEAT6*, but the full *MEAT6* transcript uses an alternate start site (black arrow). **b**, Expression data for *MEAT6* (demarcated by asterisk in **a**). This isoform variant does not use the alternate start site used by *MEAT6*, and closely resembles *AK090788*. Expression profile for cancer and lineage associated transcripts across all MiTranscriptome tissue cohorts are shown for (**c**) lung adenocarcinoma and (**d**) thyroid cancer.