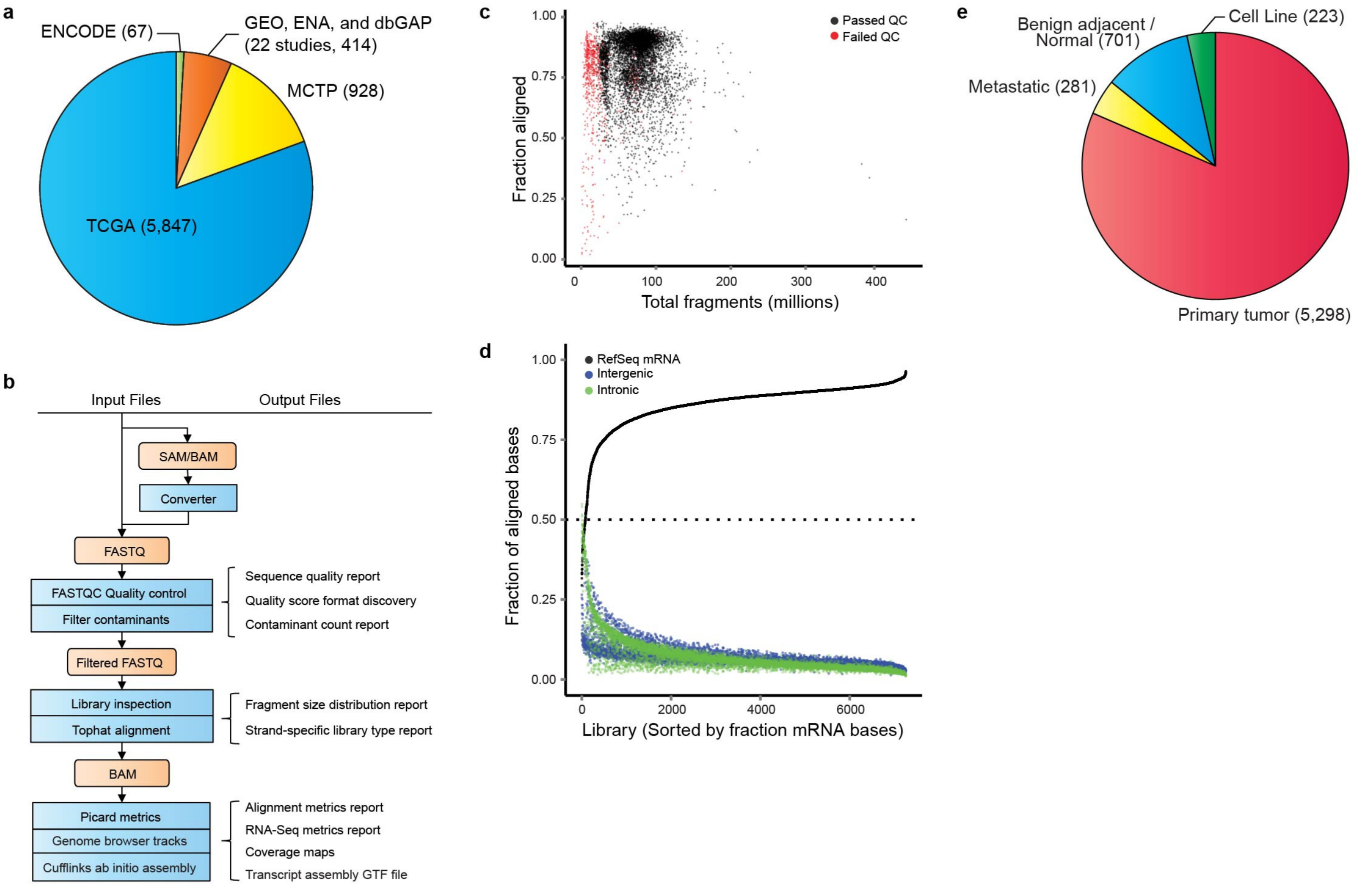
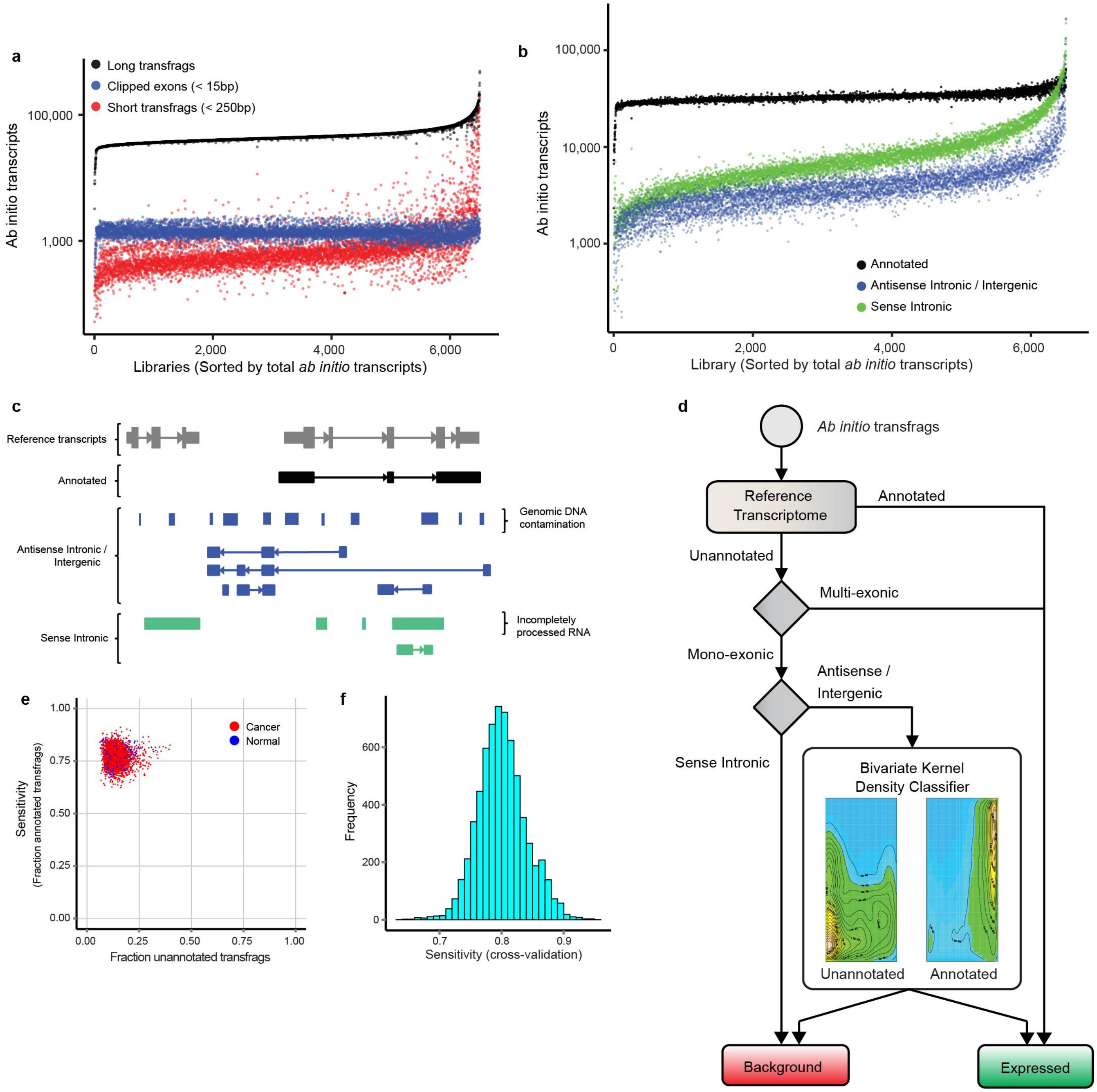


Supplementary Figure 1

Curation and processing of samples in the MiTranscriptome compendia.

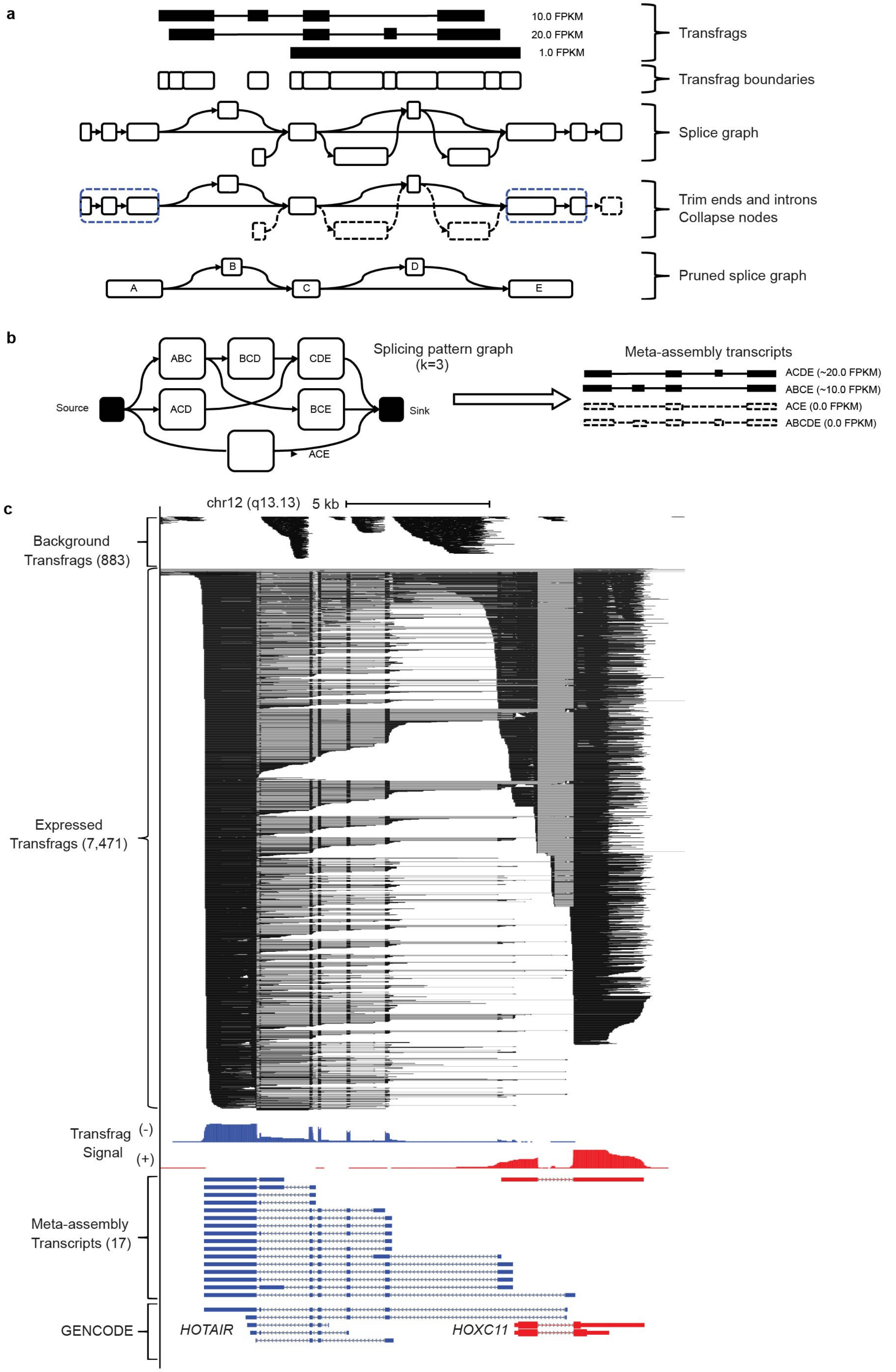
(**a**) Pie chart showing the number of studies curated from TCGA, ENCODE, MCTP and other publicly available datasets. (**b**) Workflow for bioinformatics processing of individual RNA-seq libraries. Data sets downloaded as BAM files were first converted to FASTQ format. Quality assessment of FASTQ files was performed using FASTQC. Reads mapping to mitochondria, ribosomal RNA, poly-A sequence, poly-C sequence or phiX virus (a spiked-in control) were filtered out. Fragment length distribution and orientation were determined by mapping a subset of the input reads to a set of large human exons (>500 bp). 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 the total fragments (*x* axis) and the fraction of aligned fragments (*y* axis) for each RNA-seq library. Coarse quality control filters were 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 mRNAs (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 out (dotted line). (**e**) Pie chart showing the 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.

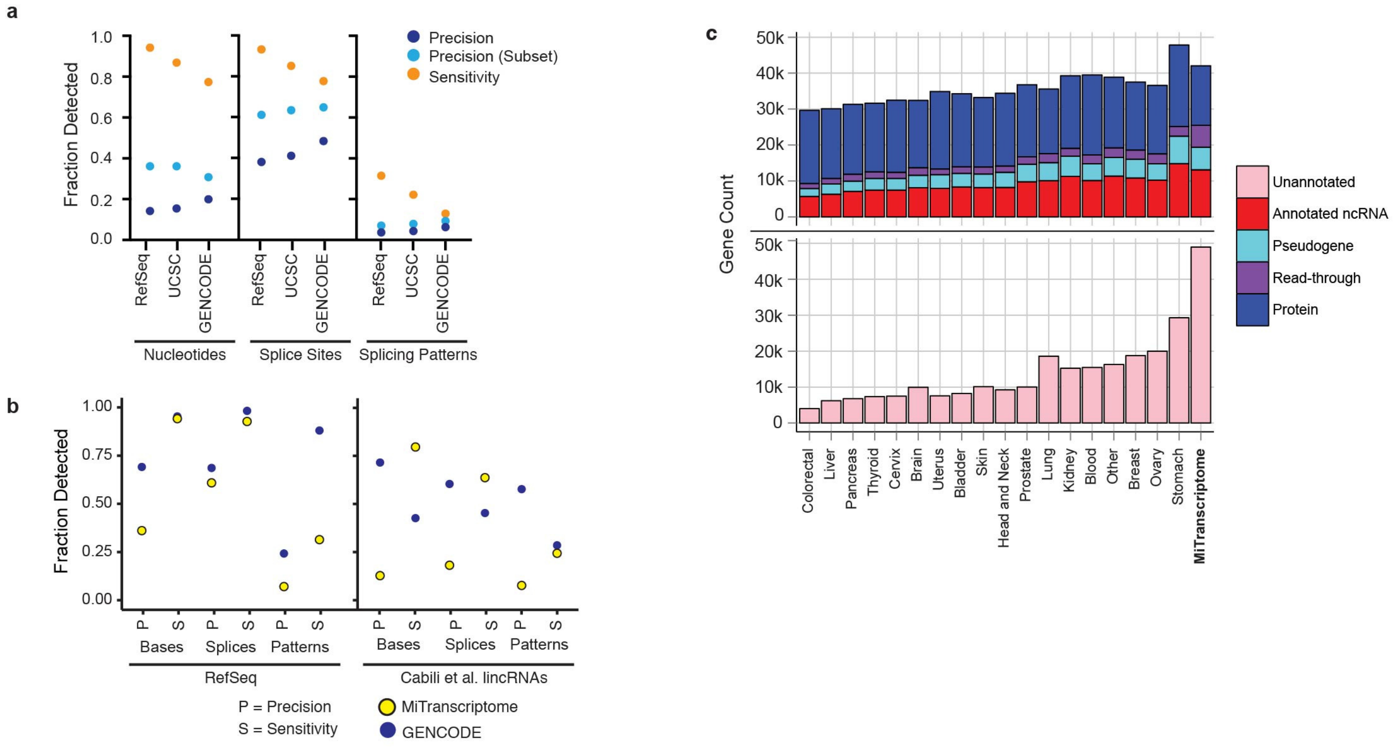
(**a**) The dot plot shows the numbers of short transfrags (red), short clipped exons (blue) and long transfrags (black) for each library. (**b**) The dot plot shows the numbers of unannotated intergenic or antisense transfrags (blue), sense intronic transfrags (green) and annotated transfrags (black) 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 contamination from both genomic DNA and incompletely processed RNA. (**d**) Decision tree depicting the transfrag filtering steps for a single library. First, transfrags were labeled ‘annotated’ or ‘unannotated’ on the basis of overlap with a reference transcriptome catalog. Annotated transfrags and unannotated multiexonic transfrags were considered expressed. Unannotated monoexonic transfrags within introns in the sense orientation of an overlapping transcript were discarded as incompletely processed RNA artifacts. Unannotated antisense or intergenic monoexonic 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 monoexonic transfrag classifier for correctly detecting annotated transcripts (*y* axis) and the fraction of unannotated transfrags predicted to be expressed (*x* axis). (**f**) Histogram demonstrating the 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
ABC
BCD
CDE
ACD
BCE
Source
Sink
ACE
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.

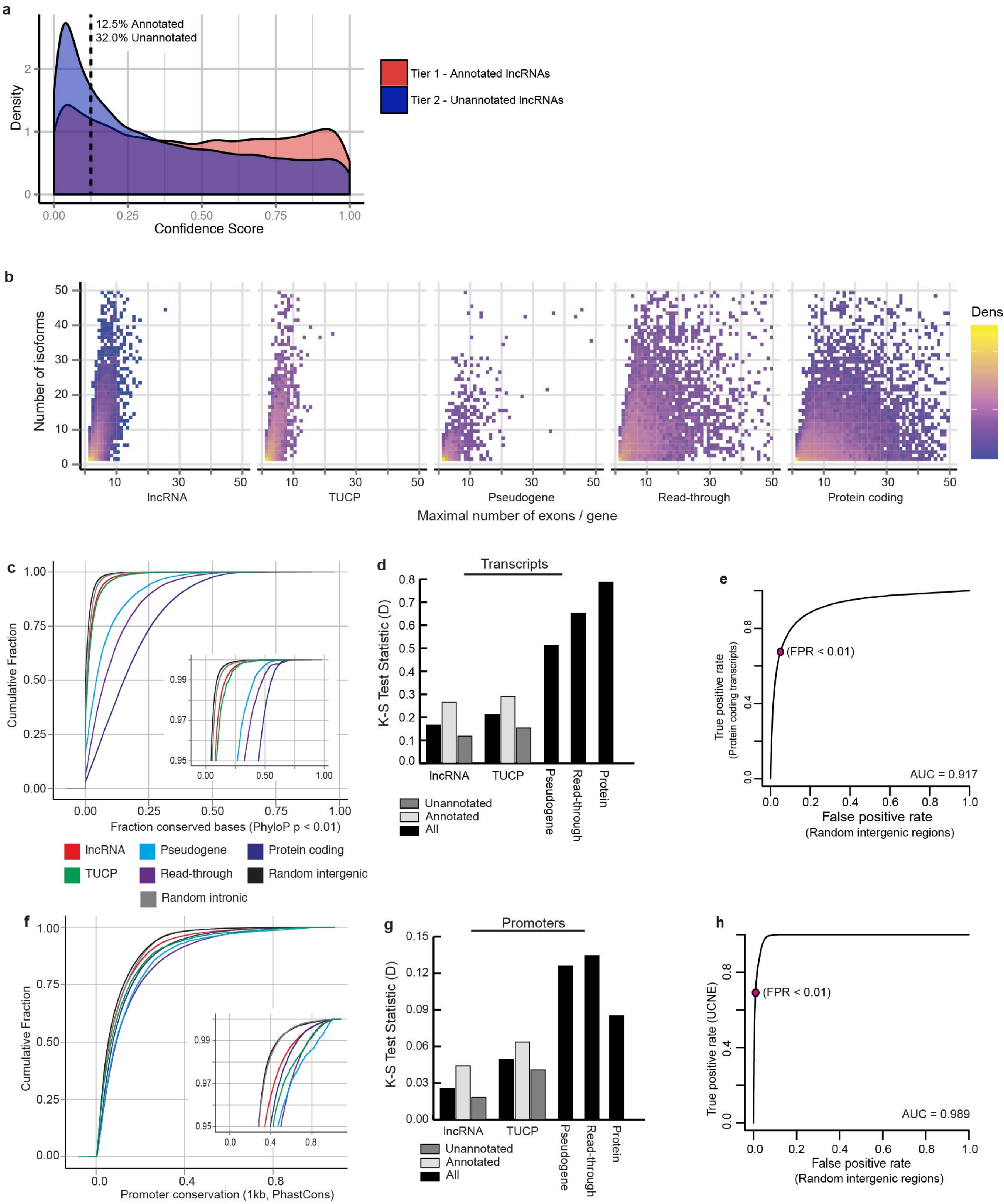
(**a**) Schematic of the transcriptome meta-assembly algorithm using a simplified example with three transfrags transcribed from left to right. The input to the meta-assembly is a list of weighted transfrags (in this case, the weights correspond to FPKM expression values). First, a splice graph is constructed 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 poorly expressed starting/ending nodes, and adjacent nodes with a degree of one are collapsed. (**b**) The pruned splice graph from **a** is subjected to meta-assembly. To encapsulate the 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 the 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 (chr. 12: 54,349,995–54,377,376, hg19). In total, 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.

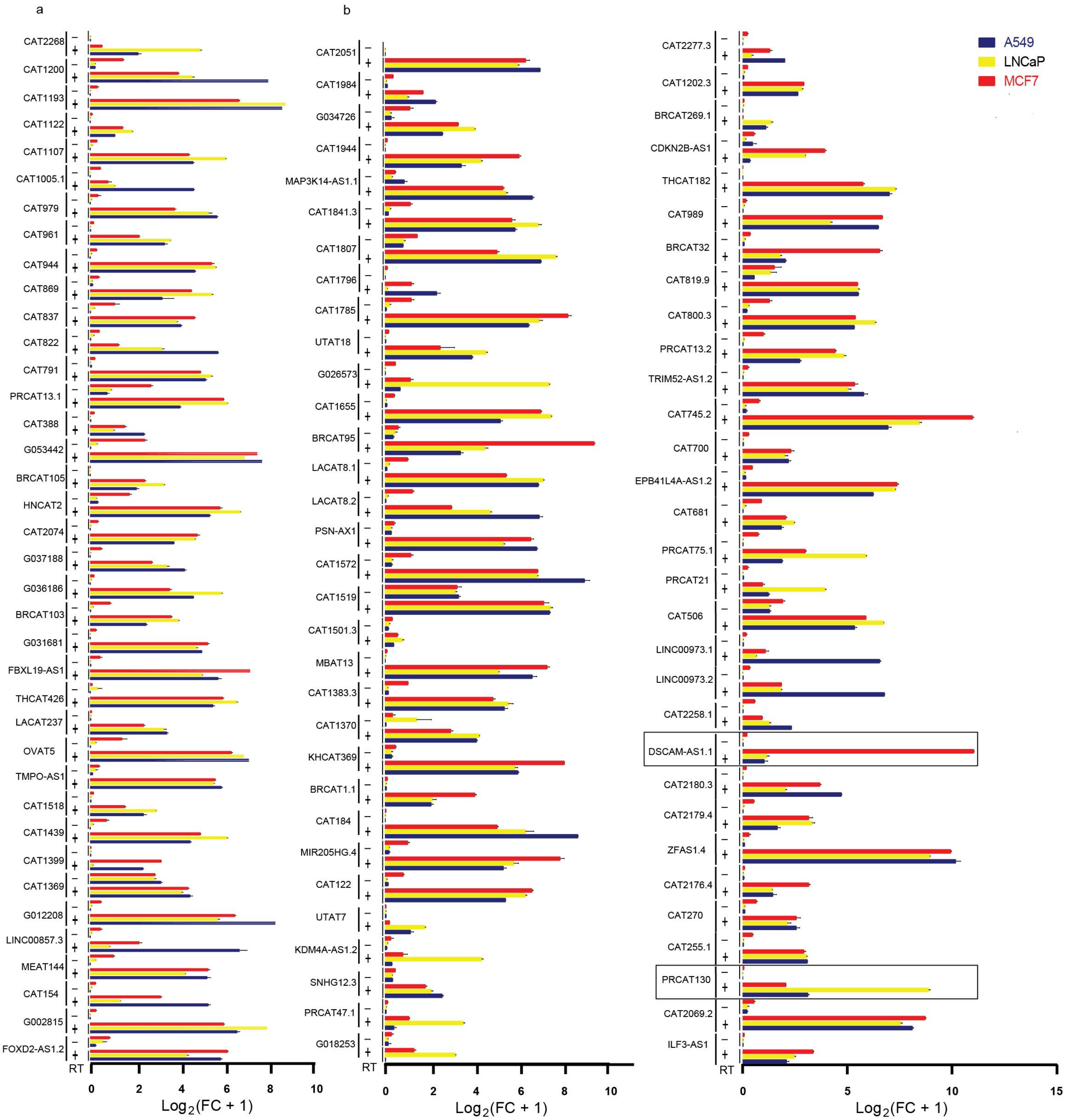
(**a**) Dot plots depicting the comparison of the MiTranscriptome with reference transcripts from RefSeq, UCSC or GENCODE. Precision (blue), precision for the subset of transcripts overlapping annotated transcripts (light blue) and sensitivity (orange) are plotted for each comparison. (**b**) Dot plots comparing the base-wise, splice-site and splicing pattern precision and sensitivity of MiTranscriptome and GENCODE using lncRNAs from RefSeq (left) or Cabili *et al*. (right). (**c**) Bar plots comparing the numbers of unannotated transcripts 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.

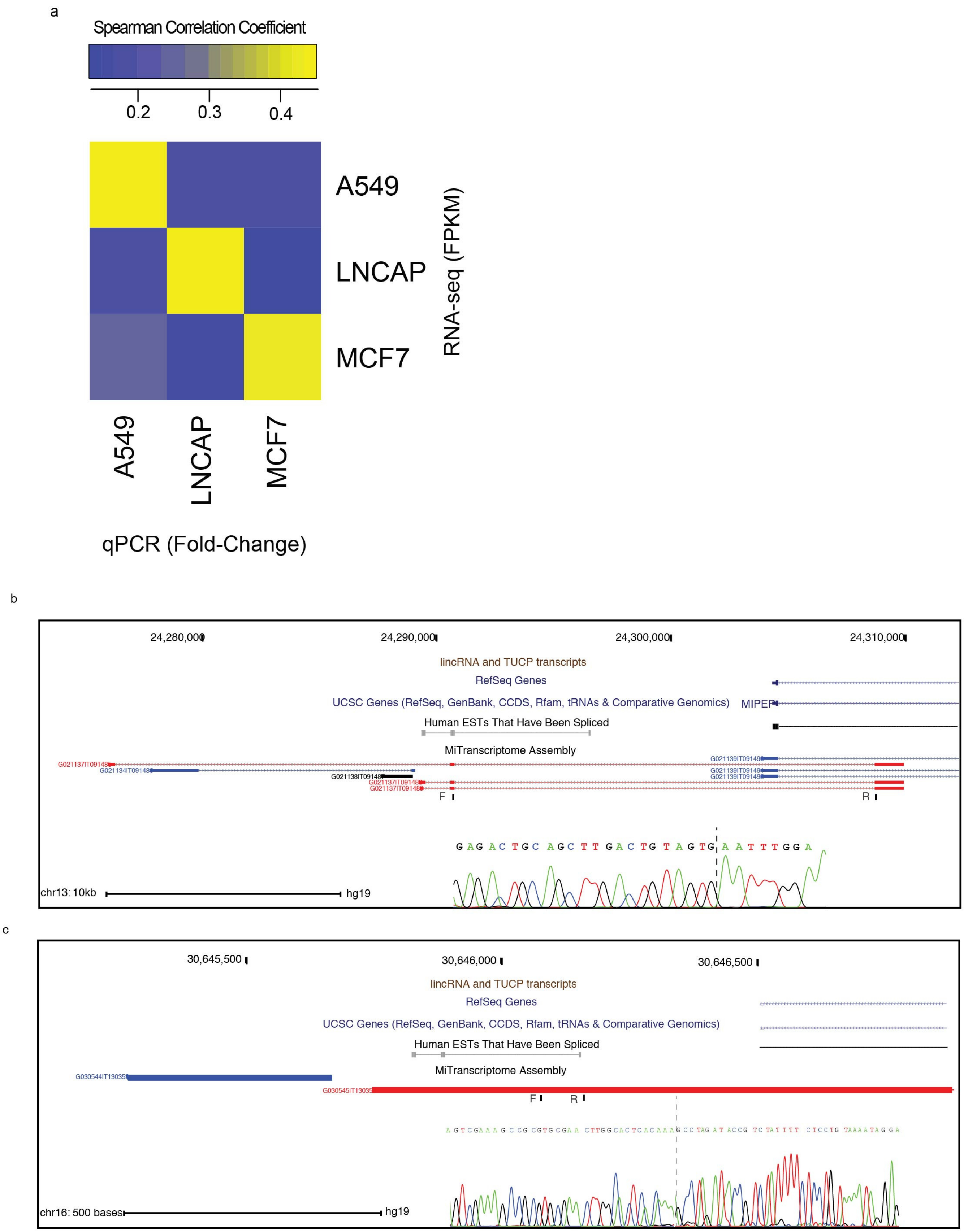
(**a**) Density histogram depicting the confidence scores for annotated and unannotated lncRNAs. (**b**) Comparison of the relationship of the maximum number of exons per gene to the number of isoforms per gene. LncRNAs tend to have fewer exons than protein-coding genes, but they have complex splicing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



Supplementary Figure 6

Validation of lncRNA transcripts.

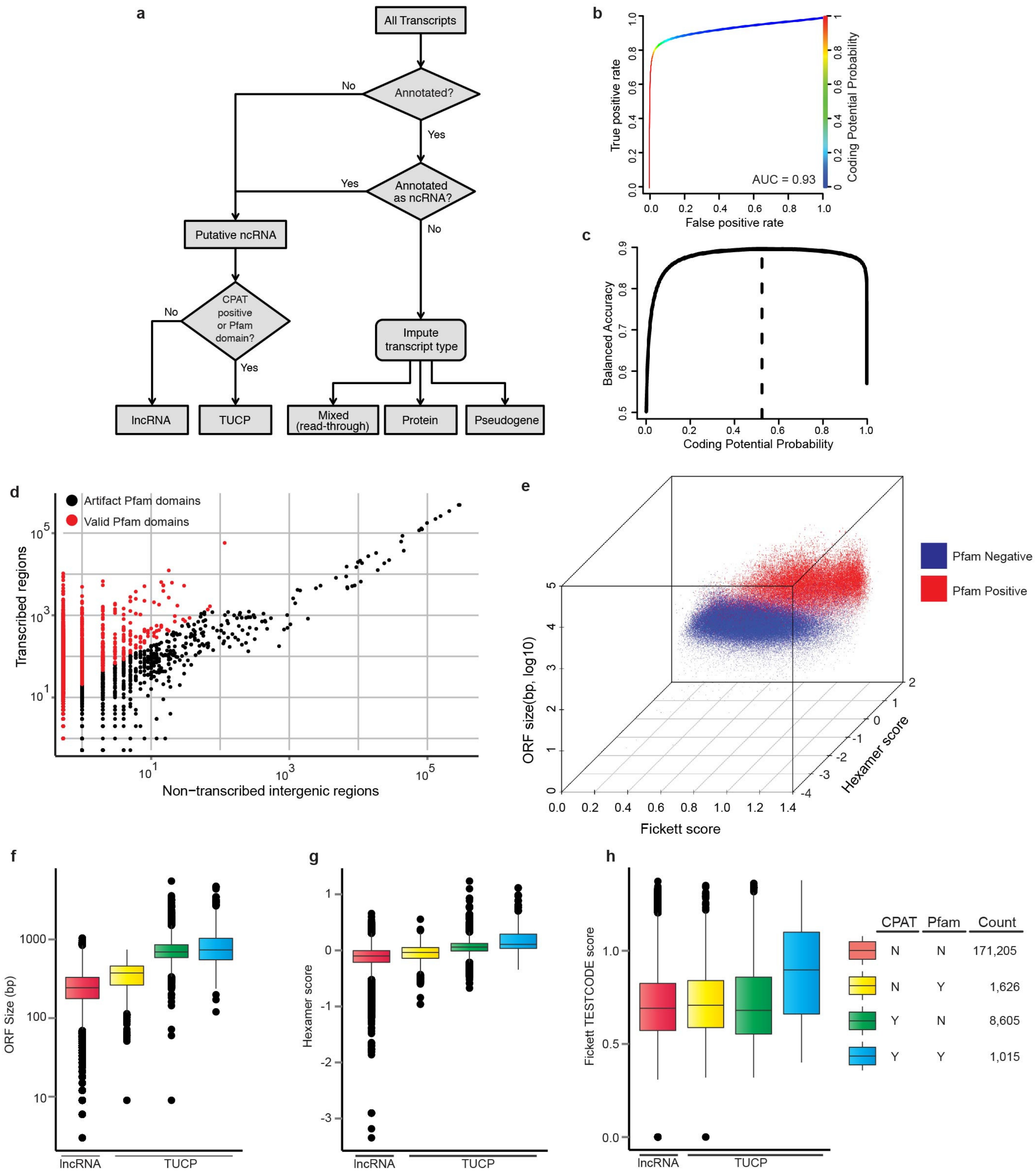
One hundred lncRNA transcripts were validated by qRT-PCR across the A549, LNCaP and MCF-7 cell lines using an approach with or without revers transcriptase. C_t values were first normalized to housekeeping genes (*CHMP2A*, *EMC7*, *GPI*, *PSMB2*, *PSMB4*, *RAB7A*, *REEP5*, *SNRPD3*) and then to the median value of all samples using the $\Delta\Delta C_t$ method. Here data are plotted as a logirithmic of fold change over the median with s.e.m. Validation was performed on (**a**) 38 monoexonic transcripts and (**b**) 62 multiexonic transcripts. The boxed transcripts are two representative examples of lncRNAs with lineage/cancer specificity 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|T091485
G021134|T091480
G021138|T09148
G021137|T091486
G021137|T091480
G021139|T091490
G021139|T09149
G021139|T091490
F
R
G A G A C T G C A G C T T G A C T G T A G T G A A T T T G G A
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
A G T C G A A A G C C G C G T G C G A A C T T G G C A C T C A C A A A G C C T A G A T A C C G T C T A T T T T C T C C T G T A A A A T A G G A
chr16: 500 bases
hg19

Supplementary Figure 7

Further validation of lncRNA transcripts.

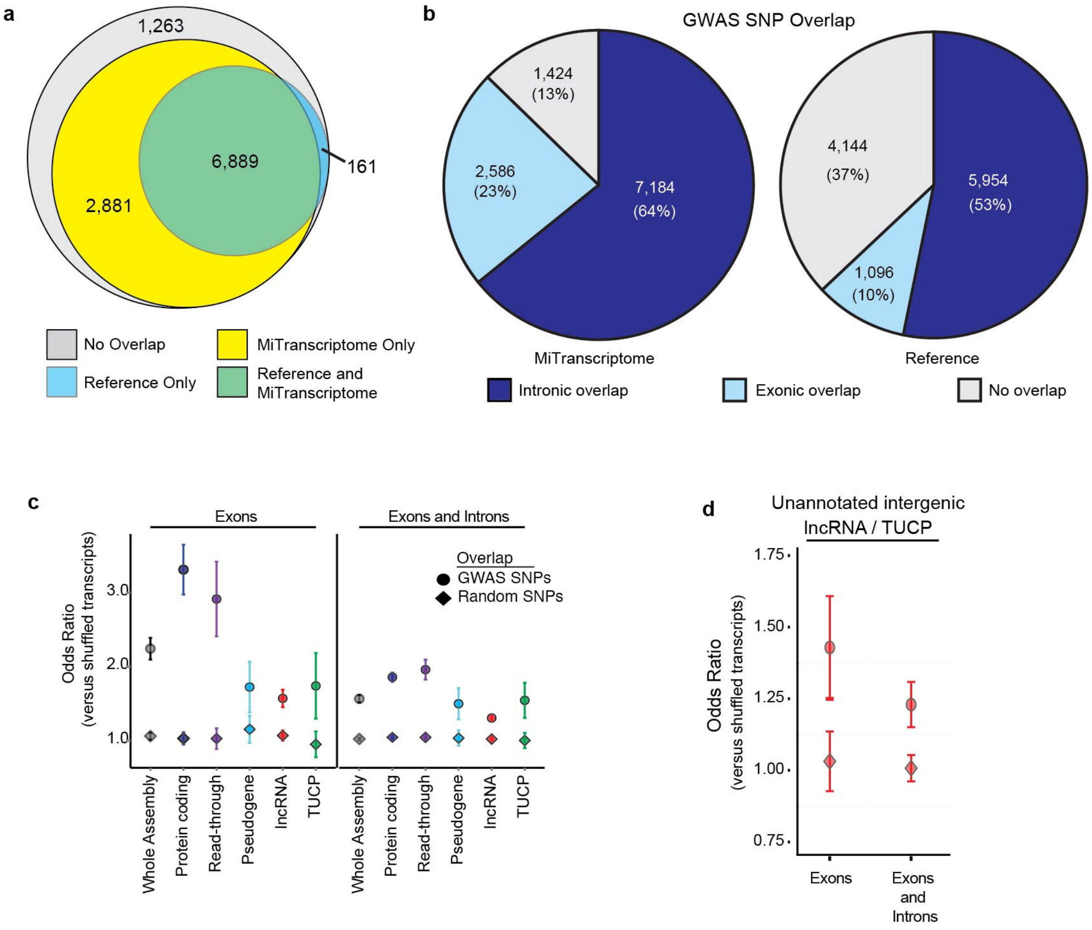
(**a**) Heat-map representation of the correlation between qPCR (fold change over the median) with RNA-seq (FPKM) of 100 selected transcripts in the A549, LNCaP and MCF-7 cell lines. (**b**,**c**) Representative example of 2 of 20 previously unannotated lncRNA transcripts that were analyzed by Sanger sequencing to ensure primer specificity with their associated chromatograms. As seen in the UCSC Genome Browser View, a (**b**) multiexonic lncRNA (Gene ID: G021137) and (**c**) monoexonic lncRNA (Gene ID: G030545).



Supplementary Figure 8

Classification of transcripts of unknown coding potential.

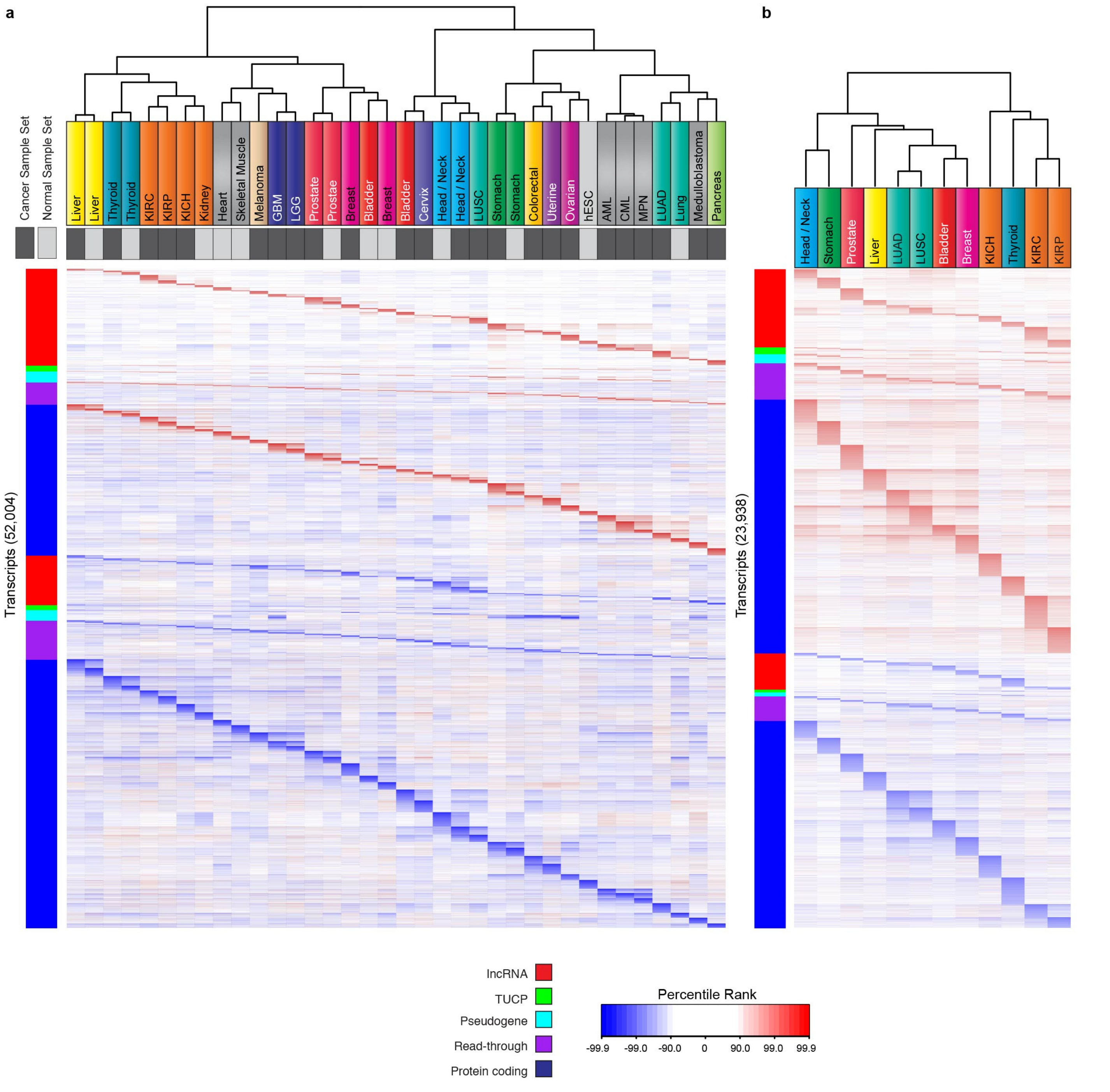
(**a**) Decision tree showing the categorization of *ab initio* transcripts. Unannotated transcripts and annotated noncoding RNAs were classified as either lncRNA or TUCP. Transcript categories for protein-coding genes, pseudogenes and read-throughs were imputed from overlapping reference annotations. (**b**) ROC curve comparing the false positive rate (*x* axis) with the true positive rate (*y* axis) for CPAT coding potential predictions of noncoding RNAs versus protein-coding genes. (**c**) Curve comparing the probability cutoff (*x* axis) with balanced accuracy (*y* axis). The dotted line shows the cutoff used to call TUCP transcripts. (**d**) Scatter plot comparing the frequency 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. Transcripts represented by red points contain valid Pfam domains, while blue do not. (**f**–**h**) Box plots 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 the MiTranscriptome assembly for disease-associated regions.

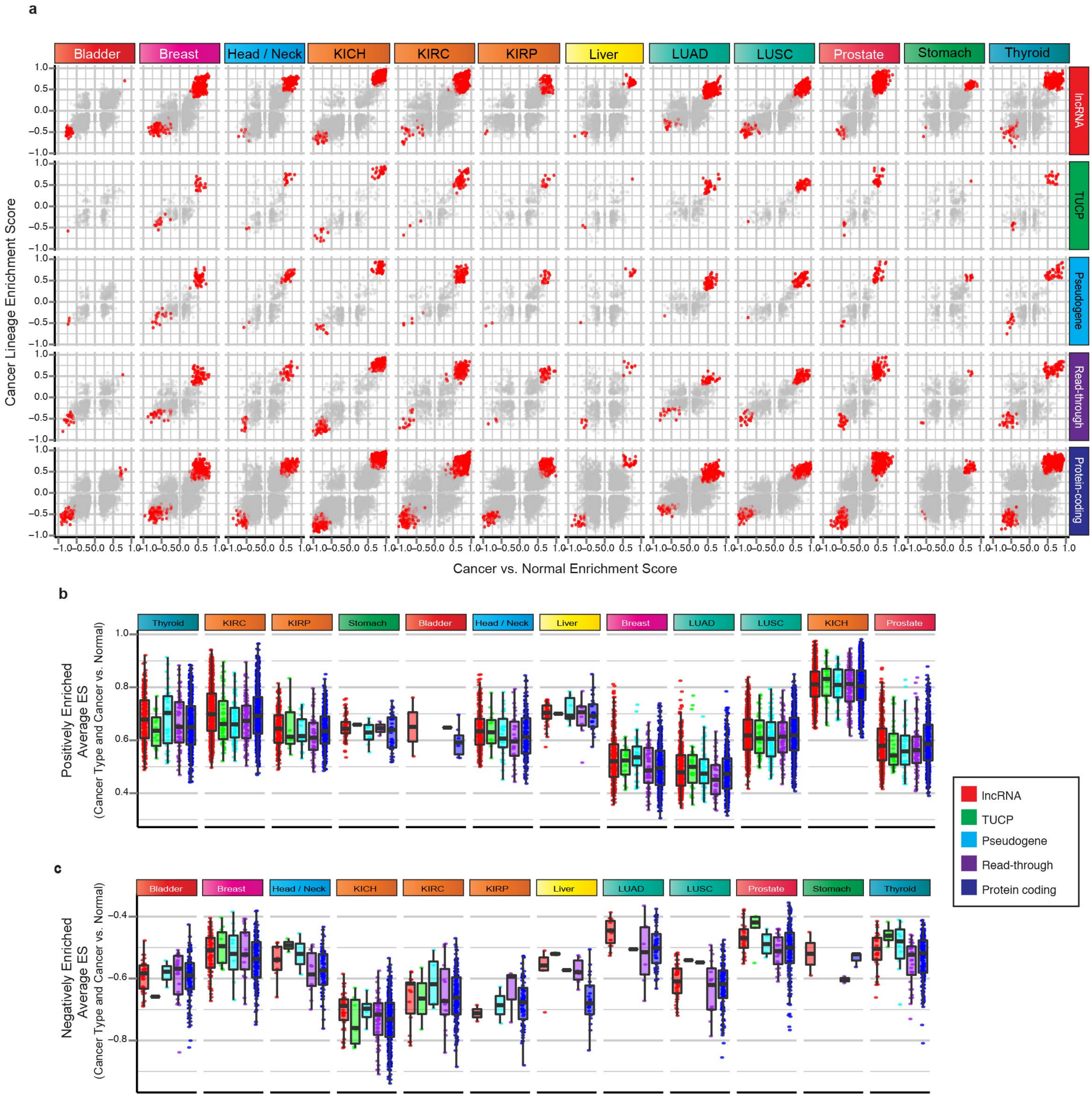
(**a**) Venn diagram comparing the coverage of disease- or trait-associated genomic regions (i.e., GWAS SNPs) for the MiTranscriptome assembly (yellow) in comparison to reference catalogs (blue). (**b**) Pie charts comparing the distributions of intronic and exonic GWAS SNP coverage of the MiTranscriptome assembly (left) and reference catalogs (right). (**c**) Dot plot displaying the enrichment of GWAS SNPs versus random SNPs for different transcript categories. Enrichment odds ratios (transcript-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 s.d. of the distribution of odds ratios. Both exonic and whole-transcript enrichment is reported. (**d**) Dot plot showing the enrichment of GWAS SNPs (circle) versus random SNPs (diamond) for novel intergenic lncRNAs and TUCPs. Enrichment odds ratios (transcript-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 s.d. 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.

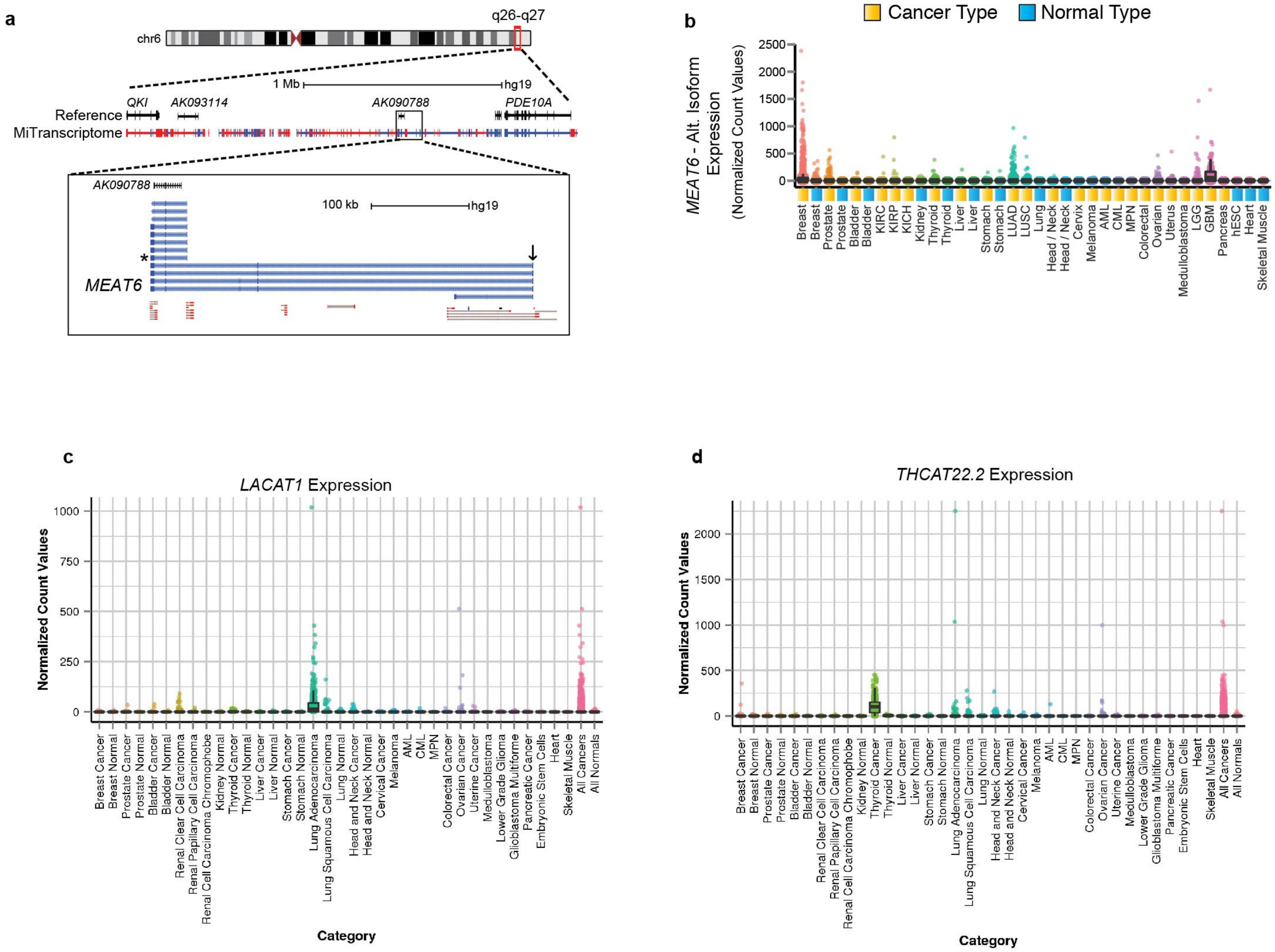
(**a**) Heat map of lineage-specific transcripts nominated by SSEA. Each column represents a sample set from 1 of 25 cancer (dark gray) and 13 normal (light gray) lineages, and each row represents an individual transcript. Colored labels above columns reflect the organ system cohorts used in assembly. 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 < 1 × 10^{-7}) and ranked in the top 1% of the most positively or negatively enriched transcripts within at least one sample set. The heat-map color spectrum corresponds to percentile ranks, with underexpressed transcripts colored blue and overexpressed transcripts colored red. The column dendrogram shows unsupervised hierarchical clustering of the sample sets. (**b**) Heat map of cancer-specific transcripts (CATs) nominated by SSEA. Columns represent 12 cancer types, and colored column labels reflect the organ system cohorts used in assembly. All transcripts were statistically significant (FDR $< 1 \times 10^{-3}$) and ranked in the top 1% of the most positively or negatively enriched transcripts within at least one sample set. The column dendrogram shows unsupervised clustering results. The row side color and heat-map color schemes are identical to those in **a**.



Supplementary Figure 11

Lineage-specific and cancer-specific transcripts.

(**a**) Scatter plot grid showing lineage-specific and cancer-specific transcripts nominated by SSEA. A row of scatter plots for each transcript category is plotted across 12 cancer types. Each plot shows the cancer versus normal enrichment score (*x* axis) and the cancer lineage enrichment score (*y* axis). Red points indicate cancer and lineage associated transcripts within the respective cancer types, and gray points indicate all other cancer and lineage associated transcripts. (**b**,**c**) Box plots comparing the performance of (**b**) positively enriched cancer and lineage associated transcripts and (**c**) negatively enriched transcripts for each 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 the chromosome 6q26-q27 locus. The 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. The 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 an asterisk in **a**). This isoform variant does not use the alternate start site used by *MEAT6* and closely resembles *AK090788*. (**c**,**d**) Expression profiles for cancer- and lineage-associated transcripts across all MiTranscriptome tissue cohorts are shown for (**c**) lung adenocarcinoma and (**d**) thyroid cancer.

Supplementary Note

Table of Contents

Methods

RNA-Seq data processing details

Software versions were managed effectively using the Modules Environment Management system (http://modules.sourceforge.net). Computational analysis was performed in a 64-bit Linux environment (Red Hat Enterprise Linux 6). Pre-compiled 64-bit Linux binaries were downloaded when available.

Initial sequence quality control metrics were calculated using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Next, filtering was performed to remove reads mapping to mitochondrial DNA, ribosomal RNA, poly-A, poly-C, Illumina sequencing adaptors, and the spiked-in phiX174 viral genome. Sequences were downloaded from the Illumina iGenomes server (2012, March 9). Mapping was performed using bowtie2 (2.0.2).

The fragment size distribution (for paired-end libraries) and fragment layout of each library was determined automatically by mapping a subset of the reads to a reference consisting of the 15,868 unique Ensembl v69 exons larger than 500bp that had no other overlapping features on either strand. These exons represent contiguous genomic regions where both paired-end reads from a single fragment could confidently be aligned. An alignment index was prepared from this reference using the bowtie-build utility.

Reads were mapped using TopHat2 (2.0.6 and 2.0.8) using default parameters[1]. Reference genome annotation files were downloaded from the Illumina iGenomes FTP server (ftp://ussd-ftp.illumina.com/Homo_sapiens). A human genome reference was constructed from UCSC version Feb 2009 (GRCh37/hg19) chromosomes 1-22, X, Y, and mitochondrial DNA. References from alternate haplotype alleles were omitted. Alignment index files for Bowtie versions 0.12.8 and 2.0.2 were built from this reference using the bowtie-build and bowtie2-build programs, respectively. The Ensembl version 69 transcriptome reference gene set was downloaded from the Ensembl FTP server (ftp://ftp.ensembl.org/pub/release-69/gtf/homo_sapiens). Chromosome names were converted from GRCh37 format to UCSC format (e.g. “1” converted to “chr1”). Genes found on alternate haplotype alleles were omitted. The cuffcompare utility (http://cufflinks.cbcb.umd.edu) was used as specified in the Cufflinks user’s manual to assign promoter and transcription start site attributes to the gene features in the Ensembl reference. Alignment index files for Bowtie versions 0.12.8 and 2.0.2 were prepared from this reference using the --transcriptome-index option in TopHat version 2.0.6 (http://tophat.cbcb.umd.edu).

Sequence alignment metrics were computed using the Picard tools CollectMultipleMetrics and CollectRnaSeqMetrics (http://picard.sourceforge.net). The Picard CollectRnaSeqMetrics diagnostic utility required gene annotation and ribosomal interval files as input. The “refFlat” table provided by the Illumina iGenomes download package (2012, March 9) was used as the gene annotation reference. Ribosomal DNA intervals were curated from the RepeatMasker table downloaded from the UCSC table browser[2]. This table of repeat elements was originally provided for hg19 by UCSC on 4/27/2009. Tracks for visualization on genome browsers were generated using the

BEDTools ‘genomecov’ utility and the UCSC bedGraphToBigWig utility[3,4].

Ab initio assembly was performed using Cufflinks (2.0.2) with multi-read correction enabled[5]. Gene features with the ribosomal RNA biotype ‘rRNA’ were added to a mask file for use with the --mask-file option in Cufflinks.

Filtration of noise contamination from aligned reads and *ab initio* assembly

To discriminate genomic DNA contamination from robust transcription we developed a classification method that utilizes both relative transcript abundance and recurrence across independent biological samples. The method requires a known transcript catalogue (Ensembl version 69) to determine the annotation status of *ab initio* transfrags. Transfrags that overlapped known transcripts in the sense orientation were denoted “annotated”, and the remaining transfrags were categorized as either “Sense Intronic” or “Antisense / Intergenic” based on their relationship to annotated transcripts. Relative abundance was determined by using the empirical distribution of FPKM values to converting transcript FPKM values into quantiles. Recurrence levels were first computed per base by counting independent biological samples with evidence of transcription (replicates of identical cell lines or tumor tissues from the same patient were not counted towards recurrence). A single recurrence value was then computed for each transfrag by averaging the recurrence values of all bases of the transfrag. After computing relative abundance and recurrence for all transfrags, we trained a classifier to discriminate annotated from unannotated transfrags as a surrogate for classifying true transcription from background noise. Specifically, we compute bivariate kernel density estimates using the abundance-recurrence axes separately for annotated and unannotated transfrags.

These densities were mapped onto a square grid (50 x 50). We then divided the annotated density by the unannotated density at each grid point after adding a nominal value to avoid floating point overflow errors. This resulted in a new grid containing likelihood ratios for annotated versus unannotated transfrags along the abundance-recurrence axes. To account for the total noise present in the library we weighted the likelihood estimates by the relative ratio of unannotated versus annotated transfrags in the library being classified. This weight equaled the ratio of the fraction of known to unannotated transcripts in a library divided by the ratio of the medians of these fractions in all libraries. Finally, for each transfrag in an *ab initio* assembly we computed the weighted log-likelihood of the transfrag being annotated by linearly interpolating the transfrag abundance and recurrence onto the grid. For each library we determined a likelihood ratio cutoff by optimizing the balanced accuracy (average of sensitivity and specificity) of the classifier performance. Transfrags with likelihood below this cutoff were labeled 'background' and the remainder 'expressed'. Results from individual libraries were then concatenated to produce separate background and expressed transfrag catalogues as output. Transcripts classified as background noise were discarded and meta-assembly was carried out on the expressed fraction. To assess the sensitivity of our classification method we ran the filtering approach after leaving out 10% of annotated transfrags as 'test' data. The ability to detect these genes was then assessed using likelihood cutoffs determined without the test data included. The classifier achieved remarkable performance (average AUC of 0.89, range 0.77-0.96) and displayed no bias for cancer versus normal samples. Moreover, the classifier recovered test transcripts left out of the training process with 80% mean sensitivity (range 0.64-0.95).

Transcriptome meta-assembly details

We developed AssemblyLine (manuscript in preparation, see source code at http://assemblyline.googlecode.com) as a software package written in Python and R to (1) characterize and filter sources of background noise in RNA-Seq assemblies and (2) perform meta-assembly to coalesce large-scale RNA-Seq datasets. AssemblyLine accepts as input a set GTF files containing transfrags assembled from individual libraries. Transfrags of length less than 250bp were omitted from meta-assembly, and the remaining transfrags were labeled as 'annotated' or 'unannotated' relative to a reference GTF file (GENCODE version 16). An *ab initio* transfrag was considered 'annotated' if its exons overlapped any reference transcript exons on the identical strand. A recurrence score for each *ab initio* transfrag was computed as the average number of samples (replicate libraries from a single cell line or tissue were considered a single sample) per nucleotide with same-stranded transcription.

We performed classification and filtering of 'background' and 'expressed' transfrags by modeling the abundance (FPKM) and recurrence of 'annotated' and 'unannotated' transcripts using bivariate kernel density estimation on a square grid (grid size 50x50, bandwidth determined by Silverman's rule of thumb). A grid of likelihood ratios was derived from the 'annotated' and 'unannotated' grids by element-wise division at each grid point. The probability of each transfrag being 'annotated' was then determined by linear interpolation onto this grid, and this probability was used as a surrogate measure for the chance that a transcript represented background noise. A likelihood ratio of less than or equal to one was used as a cutoff for filtering ‘background’ transcripts.

Filtered transcripts were subjected to the AssemblyLine meta-assembly algorithm. First, we trim low scoring ends in the graph that correspond to extraneously long exons or overhanging exons that extend into introns. Second, nodes within introns are trimmed when their scores are less than a fraction of neighboring exons. Weakly connected components of the pruned splicing graphs are then extracted and processed independently. A splicing graph encompasses the milieu of possible isoforms that could be transcribed. Enumerating all possible paths through splicing graphs is impractical; many graphs have millions of paths of which only minute fractions are observed *in vivo*. The initial input transfrags provide partial paths through the splicing graph and also indicate which parts of the graph are more abundant. Our approach incorporates this partial path information by building a splicing pattern graph that subsumes the original splice graph. The splicing pattern graph is a type of *De Bruijn* graph where each node represents a contiguous path of length *k* through the splice graph, and edges connect paths with *k-1* nodes in common. As *k* increases so does the amount of correlative path information retained in the graph at the cost of losing short transfrags with length less than *k*. Each node in the graph carries a weight equal to the summed weights from all transcripts that share the node. Thus for each splice graph the partial path length *k* is optimized to maximize the number of nodes in the path graph with the constraint that the summed node weights of transfrags with path length greater than or equal to *k* is above a user-specified fraction of the total score of all transfrags. After the path graph has been constructed, we effectively extend every partial path transfrag into a full-length transcript by transmitting the transfrag's weight along incoming and outgoing edges. This weight is allocated proportionally at nodes with multiple incoming or outgoing edges. This approach effectively extends all partial transcript fragments into full-length transcripts and assures that the sum of incoming and outgoing node weights are equivalent. Finally, a set of isoforms is predicted from the graph using a greedy algorithm. The algorithm finds and reports the highest abundance transcript by traversing the graph using dynamic programming. The weight of the transcript equals the minimum weight of all nodes in the path. The transcript weight is then subtracted from every node in the path and the dynamic programming procedure is repeated. Suboptimal transcripts are enumerated until a path weight falls below a fraction of the highest weighted transcript (*e.g.* the major isoform). The total number of isoforms produced from each gene can also be explicitly constrained. The meta-assembled isoforms are then reported in GTF and/or BED format. A genome track with summed node weights can optionally be reported in BedGraph format as well.

To limit transcript output for complex loci, isoforms with abundance less than 10% of the major transcript isoform were excluded (--*fraction-major-isoform 0.10*), a maximum of 20 isoforms were allowed for each gene (--*max-paths* 20). During splicing pattern graph creation an optimal *De Bruijn* graph parameter *k* was determined to maximize the number of graph nodes. A maximum value of *k* was limited to 20 to improve the computational tractability of the optimization approach (--*kmax 20*). The output of meta-assembly was a GTF-formatted file as well as BED and BEDGraph-formatted files (--*gtf –bed* --*bedgraph*).

Classification of transcripts of unknown coding potential (TUCP)

We ran Coding Potential Assessment Tool (CPAT) version 1.2.1 (https://code.google.com/p/cpat) with default parameters and used the human hexamer table and logit model provided by the authors[6]. We scanned for Pfam 27.0 (March 2013) A and B hits using the pfam_scan.pl utility (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools) built on HMMER 3.1b[7,8]. We performed receiver operating characteristic (ROC) analysis using the ROCR package[9].

To control for false positives we also scanned non-transcribed intergenic regions in the same manner. We observed 3,781,935 hits to 12,430 unique Pfam domains in transcribed regions compared with 1,774,937 hits to 1,277 unique domains in non-transcribed intergenic space. We compared the occurrences of each Pfam domain in transcribed versus non-transcribed regions using Fisher's Exact Test and flagged 750 domains with an odds ratio of less than 10.0 or p-value greater than 0.05 as likely artifacts. The remaining 11,726 Pfam domains were considered valid. This procedure filtered 2,972,629 artifact hits and retained 809,306 valid hits. Putative non-coding transcripts harbored only 4,674 (0.40%) of the valid Pfam domains.

Conservation analysis

Genomic conservation profiles generated by the phyloP (phylogenetic p-values) and PhastCons algorithms (http://compgen.bscb.cornell.edu/phast/) for multiple alignments of 45 vertebrate genomes to the human genome (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/vertebrate) were downloaded from the UCSC genome browser[10-12]. The ‘wigFix’ files were converted into ‘bigWig’ files using the ‘wigToBigWig’ binary utility program provided by the UCSC genome browser[3]. For each transcript a vector of exon-wise conservation scores was extracted using the ‘bigWigToBedGraph’ utility and concatenated into a single vector. Conservation metrics were then computed from these vectors.

GWAS analysis

Intersections of GWAS SNPs with transcripts or exons was performed using the BEDtools ‘intersect’ tool, with the ‘-split’ option invoked for quantification of exonic overlap[4].

The number of GWAS SNPs overlapping the entire assembly and individual transcript categories (i.e. lncRNA,TUCP, pseudogene, protein-coding, and read-through) was determined by BEDTools ‘intersect’ for both the whole transcript and for exonic regions (n_{GWAS}). Subsequently, a set of all the SNPs from two popular SNP arrays (Illumina HumanHap550 and Affymetrix SNP6) was created, which we term the “SNP background”. The amount of SNPs from the SNP background overlapping the MiTranscriptome was calculated ($n_{background}$), and the fraction of the number of overlapping GWAS SNPs to the number of overlapping SNPs from the SNP background ($frac_{GWAS} = \frac{n_{GWAS}}{n_{background}}$) was then reported for each category. This fraction was also calculated using random shuffling of the MiTranscriptome and its components into non-coding regions of the genome ($frac_{shuffle}$). One hundred shuffles were performed for each condition, and an odds ratio ($OR_{GWAS} = \frac{frac_{GWAS}}{frac_{shuffle}}$) was determined for each shuffle. The purpose of using $frac_{GWAS}$instead of simply using n_{GWAS} in this analysis is to control for the possibility that during the shuffle, transcripts could be shuffled into regions not represented on SNP arrays (i.e. regions unable to possess GWAS SNPs), falsely lowering the amount of GWAS SNP overlap by the shuffle. If transcripts are shuffled into regions that are not represented by the SNP background, both n_{GWAS} and $n_{background}$ will decrease together, with $frac_{GWAS}$ relatively unchanged.

Shuffling was performed using the BEDTools 'shuffle' tool. MiTranscriptome transcripts were grouped by transcription locus (i.e. regions of the genome that have contiguous transcription) prior to shuffling. Shuffling of transcript loci was performed to control for the fact that transcripts within a locus are spatially linked to one another. Shuffling without locus clustering would falsely elevate the amount of genome covered by transcripts, and subsequently elevate the number of SNPs overlapping the shuffled regions. A concatenation of the UCSC hg19 gaps file (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/gap.txt.gz) and the MiTranscriptome protien-coding transcripts was used as an exclusion file for these shuffles.

As a negative control, the entire above analysis was repeated using and equal number randomly selected SNPs (chosen from the Illumina HumanHap550 and Affymetrix SNP6 background) in place of the GWAS SNPs. The significance of enrichment for GWAS SNPs versus random SNPs was measured across identical shuffles of the transcript loci using paired Student's t-tests comparing the set of odds ratios for all shuffles.

Similar analysis was performed to determine enrichment for novel intergenic lncRNAs and TUCPs. The intergenic space was defined as all regions not covered by the merged reference. For this analysis, the shuffles were performed into the intergenic space, instead of all non-coding space. The exclusion file used by BEDtools ‘shuffle’ was a concatenation of the UCSC gaps file and the merged reference.

Transcript expression estimation

We estimated the transcript abundances for all transcripts in the MiTranscriptome assembly using Cufflinks version 2.1.1[5] with the following parameters: ‘--max-frag-multihits=1’, ‘--no-effective-length-correction’, ‘--max-bundle-length 5000000’, ‘--max-bundle-frags 20000000’. To convert normalized transcript abundance estimates (FPKM) to approximate fragment count values we multiplied each FPKM by the transcript length (in kilobases) and by the “Map Mass” value (divided by 1.0e6) reported in the Cufflinks log files. Through reverse engineering and some assistance from the seqanswers online forum (seqanswers.com) we determined that this factor was utilized in the normalization process. Abundance estimation for 28 libraries failed for technical reasons (corrupt BAM files) and these libraries were discarded from the expression analysis. Expression estimation for 2,246 transcripts yielded errors and/or zero-valued counts and hence discarded.

Transcript expression enrichment testing

The method adapts the weighted Kolmorgorov-Smirnoff (KS) tests proposed by Gene Set Enrichment Analysis (GSEA). In contrast to GSEA, which tests for associations with gene sets, Sample Set Enrichment Analysis (SSEA) tests for associations between individual gene expression observations (which could be transcript or gene expression) and sample sets. Thus, SSEA is analogous to performing GSEA on a 'transposed' input dataset. However, SSEA incorporates important features not provided by GSEA: (1) methodology for non-parametric analysis of discrete count data (*e.g.* RNA-Seq count datasets), (2) engineering improvements to enable analysis of big datasets (here, we analyze a matrix of 381,731 rows and 6,475 columns using less than 1 Gb of RAM), and (3) parallelization of the algorithm for use in high-performance computing environments. For a discussion of the potential advantages of SSEA relative to other RNA-Seq differential expression analysis tools please refer to **Supplementary Discussion**.

Differential expression testing was performed using the Sample Set Enrichment Analysis method developed as part of this study (http://ssea.googlecode.com). We ran SSEA with 100 iterations of count resampling and 1,000 null permutations for each transcript (--resampling-iterations=100, --perms=1000). These parameters yielded a minimum FDR resolution of approximately 1e^{-7} for all sample sets. Weights for the KS-test were log(x + 1)-transformed normalized count values (--weight-hit=log, --weight-miss=log, --weight-param=1).

Discussion

Approaches for mitigating background noise in RNA-Seq data

To circumvent the added complexity of background noise previous transcriptome studies restricted themselves to multi-exonic transcripts, intergenic regions or both[13-17]. However, given that over 5% of RefSeq transcripts longer than 200nt are mono-exonic[18] and that lncRNAs generally have fewer exons than protein-coding genes[19], a significant population of expressed mono-exonic transcripts may be missing from gene catalogs. Furthermore, the high degree of overlapping and interleaving transcription observed in humans demands analyses that include intronic regions as well. Apart from excluding areas of the genome, previous studies contended with background noise by designing filtering strategies. Ramskold *et al*. compared the expression levels of exons and intergenic regions to determine an empirical threshold for calling a gene expressed[20]. Similarly, Cabili *et al.* derived empirical detection thresholds by comparing the coverage of full length versus partial length transcripts corresponding to known genes, and further defined a high-confidence set of transcripts that were detected in multiple samples or by independent *ab initio* assembly programs[17]. A study to incorporate zebrafish RNA-Seq data into the Ensembl genebuild discarded exon regions with relatively low coverage[14]. In contrast to empirical filtering methods, Guttman *et al.* developed a statistical approach that models background noise as though read alignments were randomly permuted throughout the genome[16]. Although all of these strategies enriched for highly expressed genes, they do not account for classes of transcripts that are expressed at relatively low levels[17]. In their study of human transcription the ENCODE consortium employed a statistic called the non-parametric irreproducible detection rate (*npIDR*)[21,22]. This statistic embodies the notion that purposeful transcription should be observable by independent experiments. The study filtered unannotated transcripts that were less than 90% recurrent ($npIDR < 0.1$) between biological replicates of the same sample but still detected a large number of unannotated mono-exonic transcripts. The authors acknowledged the possibility of artifacts due to low levels of DNA contamination but did not compare *npIDR* values between unannotated and annotated transcripts to credential their chosen detection threshold. Altogether the aforementioned schemes establish the use of noise thresholds based on expressed levels and reproducibility, but no previous study suggested a rigorous method for filtering background noise in large datasets.

Accuracy for classification of expressed versus background transfrags

Our transfrag classification method conservatively assumes that all unannotated transfrags are background noise and invariably overestimates the manifestations of genomic DNA contamination in libraries. Indeed, our goal in designing the classifier was to conservatively estimate true transcription in order to maximize confidence in unannotated transcript assemblies.

Motivation for meta-assembly approach

Meta-assembly refers to merging together multiple *ab initio* assemblies to produce a consensus assembly. Establishing a consensus assembly is vital to downstream analysis because it provides a common foundation for comparing transcriptional dynamics[23,24]. Previously, we developed a merging approach that clustered isoforms into a single set of exon regions per gene[13]. This strategy facilitated the discovery of unannotated cancer-associated loci but abolished isoform-level information and relied upon additional assays such as Rapid Amplification of cDNA Ends RACE for precise delineation of transcript structures. An earlier generation of algorithms was developed for EST assembly and introduced splicing graphs as an effective representation of the isoform problem[25,26]. Building on these approaches, Trapnell *et al.* released a meta-assembly utility within the Cufflinks package called Cuffmerge[24]. Cuffmerge converts transcripts from *ab initio* assemblies into faux read alignments and reruns Cufflinks on these alignments in a modified mode. Cufflinks then emits a minimal set of merged transcripts that explains the input transcripts. In our experience, the use of Cuffmerge on large datasets induced scalability issues even when we limited the allowable minor isoform fraction levels (data not shown). Thus, we believe that the most recent Cuffmerge version we had access to (Cufflinks version 2.0.2) required further optimization before it can be used effectively on large datasets. Alternatively, aggregating the raw sequences from multiple RNA-Seq samples before running standard *ab initio* or *de novo* assembly programs can produce a consensus assembly[27]. However, naively aggregating raw sequences compounds background noise, forcing a choice of a single set of filtering parameters for all samples. Most importantly, transcripts specific to a subset of samples may pose as minor isoforms and be unintentionally pruned.

Performance of transcriptome reconstruction relative to reference catalogs

In the recent assessment of transcriptome reconstruction methods for RNA-Seq, Steijger *et al*. observed relatively poor sensitivity for observed complete splicing patterns[28]. On the human test regions assessed, Cufflinks (the algorithm used as a foundation for this study) achieved 39% detection of all transcript exons. Although the Steijger *et al*. comparisons were limited to genes expressed in HepG2 liver cells, their result is roughly consistent with our observed splicing pattern sensitivity of 31% for RefSeq genes. Given these results, we offer three explanations for the low sensitivity and precision for splicing pattern detection: (1) RNA-Seq protocols tend to capture incompletely processed RNAs that may not have undergone complete splicing and/or poly-adenylation. Bioinformatics tools, including Cufflinks and the meta-assembler used in this study, attempt to account for this problem by trimming first and last exons and clipping out retained introns. However, the ability of these tools to correct for incompletely processed RNA artifacts is limited, especially for loci where exonic and intronic RNA abundance levels are similar. (2) Illumina RNA-Seq protocols sequence libraries with fragments of approximately 200-300bp, which is often much smaller than full-length transcripts. Thus, transcriptome assembly methods are hampered by the intrinsic nature of Illumina RNA-Seq data; specifically, the lack of long reads with full-length transcript splicing patterns. (3) The reference annotations that are used as a benchmark for evaluating the performance of transcriptome assembly tools may themselves be inaccurate or incomplete. Studies by the ENCODE consortium and others have revealed splicing complexity far beyond what has been catalogued[21,28], and the results of this study assembly suggest an even greater level of splicing complexity than previously observed. Thus, perhaps the lack of detection of known splicing patterns may not be a fault of RNA-Seq or computational tools but rather at attribute of the reference catalogs.

Assessing coding potential

Although codon substitution frequency (CSF) metrics can be a powerful predictor of coding potential[29], algorithms that employ CSF require multiple sequence alignments. In our transcriptome assembly we observed multitudes of putative lncRNAs in regions with poor evolutionary conservation where CSF analysis would lack sensitivity. Furthermore, we found the process of extracting and concatenating blocks of multiple sequence alignment data to be computationally cumbersome and poorly optimized. Therefore, for this study we opted to use the alignment-free Coding Potential Assessment Tool (CPAT) version 1.2.1 (https://code.google.com/p/cpat)[6]. CPAT determines the coding probability of transcript sequences using a logistic regression model built from ORF size, Fickett TESTCODE statistic, and hexamer usage bias. Although CPAT does not utilize CSF information, it was nevertheless shown to have superior discriminatory ability for human transcriptome assembly data. We believe that the MiTranscriptome assembly offers an important benchmark for subsequent testing of coding potential prediction tools. Our approach of combining Pfam results with CPAT score produced confident calls for coding potential. The presence of Pfam domains provided a strong support for CPAT coding predictions. The presence or absence of a Pfam domain stratified transcripts by the three features modeled by CPAT as well as overall coding probability. Transcripts possessing Pfam domains were much more likely to be predicted positive by CPAT than those lacking a Pfam domain (p-value < 2.2e-16, odds ratio=90.3, Fisher's Exact Test). Given the complementary aspects of Pfam domain and CPAT prediction we designated putative non-coding transcripts with either a Pfam domain or a positive CPAT prediction as TUCP. In total 11,603 uncharacterized transcripts were flagged as TUCPs, including 5,248 transcripts previously annotated as lncRNAs. There were 2,729 uncharacterized transcripts with at least one Pfam domain, including 1,700 that did meet the CPAT criteria. By contrast, 8,874 CPAT positive transcripts lacked a valid Pfam domain. Intriguingly, transcripts predicted by CPAT that also harbored valid Pfam domains had longer ORFs, higher hexamer scores, and higher Fickett TESTCODE scores than other TUCPs, suggesting that the Pfam and CPAT calls may be complementary.

Differential expression testing via enrichment analysis

An established paradigm for the differential expression analysis of RNA-Seq data is the parametric modeling of the data using the negative binomial distribution. Current prominent algorithms that operate in this manner include CuffDiff[23], edgeR[30], and DEseq[31]. With many thousands of RNA-seq libraries becoming available for analysis, there is a pressing need for a tool to analyze differential expression on large datasets. Besides overcoming important engineering considerations (such as storing expression data for thousands of samples and hundreds of thousands of genes in RAM), existing methods implicitly assume that the two conditions being analyzed (e.g. cancer versus normal) are homogenous (i.e. the samples within each condition are being drawn from the same population). Preliminary analyses of the TCGA data have discovered that there is substantial intracancer heterogeneity[32,33], with many aberrations known to be present in only a subset of tumors from any given cancer type. The availability of large numbers of samples makes discovery of subtype-specific or ‘outlier’ events tractable. Of particular interest is the identification of transcript expression unique to aggressive cancers with poor prognosis. For example, the lncRNA *SChLAP1* is expressed in only ~20% of prostate cancers and strongly predicts cancer-related mortality[34]. With these observations in mind, we designed SSEA to leverage the statistical power afforded by large numbers of samples by employing non-parametric, semi-supervised methodology. Validation of SSEA on known cancer outlier genes such as *SChLAP1* illustrated its remarkable sensitivity to detect subtype-specific transcription. Head-to-head comparisons with parametric RNA-Seq methods are pending, but we anticipate that SSEA will offer unique discovery power for large-scale datasets.

Supplementary References

1. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
2. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-6 (2004).
3. Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204-7 (2010).
4. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
5. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-5 (2010).
6. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* **41**, e74 (2013).
7. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).
8. Finn, R.D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42**, D222-30 (2014).
9. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940-1 (2005).
10. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42**, D764-70 (2014).
11. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-21 (2010).
12. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-50 (2005).
13. Prensner, J.R. *et al.* Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol* **29**, 742-9 (2011).
14. Collins, J.E., White, S., Searle, S.M. & Stemple, D.L. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. *Genome Res* **22**, 2067-78 (2012).
15. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res* **42**, D749-55 (2014).
16. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503-10 (2010).
17. Cabili, M.N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915-27 (2011).
18. Pruitt, K.D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* **42**, D756-63 (2014).
19. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775-89 (2012).

20. Ramskold, D., Wang, E.T., Burge, C.B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* **5**, e1000598 (2009).
21. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-8 (2012).
22. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
23. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**, 46-53 (2013).
24. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-78 (2012).
25. Heber, S., Alekseyev, M., Sze, S.H., Tang, H. & Pevzner, P.A. Splicing graphs and EST assembly problem. *Bioinformatics* **18 Suppl 1**, S181-8 (2002).
26. Haas, B.J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654-66 (2003).
27. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-52 (2011).
28. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **10**, 1177-84 (2013).
29. Lin, M.F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275-82 (2011).
30. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-40 (2010).
31. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).
32. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).
33. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* **45**, 1127-1133 (2013).
34. Prensner, J.R. *et al.* The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat Genet* **45**, 1392-8 (2013).