Supplementary Data

Supplementary Methods:

Description of pipeline parameters that are adjusted in order to change the stringency during the pipeline run:

- Percent identity cutoff: Minimum percent identity to call an alignment hit as high-scoring (default value is 95%). This cutoff is also the minimum identity required by each alignment pair in the chimeric transcript detection step.
- Percent coverage cutoff: Minimum alignment query coverage to call an alignment as high-scoring (default value is 90%).
- Deletion cutoff: Minimum number of skipped exonic bases from the reference genome due to the gapped alignment of query read before the read is characterized as exon-deletion (default value is 10 bp).
- Exon extension cutoff: Minimum number of extended bases outside the exon boundaries to characterize it as internal-exon-extension (if extension is in intron) or alternativeTSS or alternativePolyadenylation (if the extension is out of the transcriptional boundaries) (default value is two bp).
- Gene radius: Maximum extension of the known transcript in the upstream or downstream region to include intergenic region mapped reads in the known gene models (default value is 5000 bp). This value can be set to zero if all the intergenic mapped reads need to be characterized as gene-desert.
- Gap tolerance: Maximum number of bases required on the query read before it is designated as chimeric transcript and searched for the alignment pair (default value is 20 bp). This cutoff is also the maximum allowed query bases between the fragmented alignments of the query sequence.
- Annotation mode: Two possible values; “unique”, “multi”. “Unique” is the default setting that will cause the pipeline to characterize each high-scoring read to only one best fitting known transcript. If set as “multi”, reads will be characterized with all the known transcripts the read mapped to.

Obtaining data from ChimerDB 2.0:
ChimerDB 2.0 (http://ercsb.ewha.ac.kr:8080/FusionGen) (1) is a database of chimeric transcripts (or fusion gene transcripts) that are detected from the publicly available nucleotide sequences available in databases such as GenBank and SRA (short read archive). It also reports gene fusion pairs that are previously reported in literature. We downloaded GenBank accession IDs and corresponding fusion gene pairs for the chimeric transcripts present in ChimerDB 2.0. Downloaded fusion gene pairs were cross-referenced against the fusion gene pairs that were reported in literature and also available at Chimer DB 2.0. We retained only those GenBank accession IDs that had fusion gene pairs that were also reported previously in the literature. Nucleotide sequences for the retained GenBank accession IDs were obtained using nucleotide database (http://www.ncbi.nlm.nih.gov/nuccore) at NCBI. The resulting 206 sequences were defined as high-confidence dataset of chimeric transcripts that were used as the test data for testing the chimer detection module in R-SAP.
Raw RNA-Seq data cleaning:
454 sequencing reads for MAQC Reference Human dataset were masked for low-complexity repeats (including simple repeats) using DustMakser (2) program before aligning them to the reference genome. Masked regions were trimmed using in-house perl scripts. Trimmed reads shorter than 20bp were excluded because of the BLAT’s (3) limitation to align such sort reads with high accuracy.

Alignment screening and top-scoring hit selection:
Command line BLAT (downloaded from http://users.soe.ucsc.edu/~kent/src/) generated psl and pslx output files don’t have alignment percent identity, percent query sequence coverage and alignment score. These values were necessary for each alignment hit in order to sort and prioritize all possible reference genome hits for each sequencing read. For alignment score and alignment percent identity calculations, we incorporated the code available at http://www.genome.ucsc.edu/FAQ/FAQblat.html#blat4 in our alignment-screening module.

Percent query coverage was calculated as:

\[ ((\text{alignment end position in query- alignment start position in query})/\text{(query length)}) \times 100 \]

In order to obtain the best possible alignment (top-hit), all the alignment hits were sorted hierarchically first on alignment score, then on percent query coverage (if scores of the two alignment hits were equal) and finally on percent alignment identity (if coverage were equal).

Gene expression microarray data analysis:
Cel files from Affymetrix Human U133Plus2.0 and Affymetrix Human Exon 1.0 ST V2 obtained from GEO (Gene Expression Omnibus at: http://www.ncbi.nlm.nih.gov/geo/) were analyzed using Affymetrix Expression Consol 1.1 provided on Affymetrix website www.affymetrix.com. Samples were RMA normalized, and transcript/gene level expression analysis and probe-set annotation was done using annotation files (hg18) downloaded from http://www.affymetrix.com/support/index.affx.

TaqMan qRT-PCR data analysis:
Four replicates of TaqMan qRT-PCR measurements for MAQC Human Reference sample were obtained from GEO. Original dataset consisted of normalized expression values and their presence/absence calls for 1044 probes. Expression value for each probe was taken as the mean expression values across the four replicates. A probe was considered expressed if it had at least 75% presence call (present call in at least three replicates). 973 expressed probes were retained after the filtering. Expressed probes were then assigned to RefSeq transcripts (hg18) using TaqMan probe annotations provided under GEO’s platform record for TaqMan (platform: GPL4097). 962 probes were successfully assigned to RefSeq transcripts and 727 of them were also detected (R-SAP RPKM > 0) in MAQC Human Reference RNA-Seq data.
ENCODr E Gm12878 cell line RNA-Seq data analyses:

Running TopHat:
TopHat v1.3.1 (4) (available at http://tophat.cbcb.umd.edu/) was used for the human reference genome (hg18) alignment of ENCODE Gm12878 RNA-Seq reads. TopHat was run as:

tophat -m 1 -F 0 -g 1 --coverage-search <bowtie_index> <input_fastq_file>

Using “-g 1” parameter, we allowed only uniquely mapped reads to be reported by TopHat. Overall 38,524,540 (out of 87,929,372) reads were mapped to the reference genome. TopHat outputs alignments in bam format that is used as input for Cufflinks (5).

Running Cufflinks:

a. Assembly mode:
Aligned reads from TopHat were assembled using Cufflinks v1.1.0 (available at http://cufflinks.cbcb.umd.edu/). Cufflinks was run as:

cufflinks -u -F 0.0 <tophat-alignments>

Transcriptome assembly resulted in 76,101 assembled transcripts. Cufflinks assembled transcripts are reported in GTF format that contains genomic coordinates of transcript and putative exons.

b. Abundance quantification mode:
In order to estimate the expression values of RefSeq (hg18) transcripts, we ran Cufflinks v1.1.0 in its quantification mode. Aligned reads from TopHat and RefSeq transcripts in GTF format were used as input for Cufflinks. Cufflinks was run as:

cufflinks -G Hg18RefSeq.gtf -u -F 0.0 --overhang-tolerance 3 <tophat-alignments>

Parameter “--overhang-tolerance 3” was used to match R-SAP’s default cutoff for “exon-extension” (3 bp). Cufflinks reports abundance estimates as FPKM (fragments per kilobase of transcript per million fragments mapped) that are comparable to RPKM (reads per kilobase of transcript per million reads mapped) for single end read data.

Running Cuffcompare:
Cufflinks assembled transcripts (from previous step) were compared with RefSeq (hg18) transcripts using Cuffcompare. Cuffcompare is a module in Cufflinks program that compares multiple transcript sets (including reference transcripts) in order to generate transcript structural variant classifications. Cuffcompare was run as:

cuffcompare -r Hg18RefSeq.gtf -R -C <cufflinks-transcript-assembly>
Running RSEM:
RSEM v1.1.13 (6) (available at http://deweylab.biostat.wisc.edu/rsem/) estimates transcript expression values by aligning RNA-Seq reads to reference transcript sequences. RSEM uses BowTie as an aligner that is run inherently from RSEM. In order to run RSEM, we supplied original fastq files for the Gm12878 RNA-Seq data and RefSeq (hg18) transcript sequences that were obtained using UCSC Table browser (http://genome.ucsc.edu/cgi-bin/hgTables).

RSEM was run in two steps:

1. Preparing reference transcript sequence index:
   rsem-prepare-reference <bowtie-path> <Hg18RefSeq.fa>

2. Estimating expression:
   rsem-calculate-expression --out-bam --seed-length 28 <bowtie-path> --bowtie-n 3 --bowtie-e 200 --bowtie-m 1 --phred33-quals --fragment-length-mean 200 --fragment-length-sd 80 <input-fastq-file> <reference-transcript-sequence-index>

Cufflinks default values for fragment length distribution mean and standard deviation were used for RSEM run. RSEM estimated RSEM reports two measures of abundance estimates: Expected read count from each transcript and estimated fraction of transcripts made up by a given isoform or gene (τ value). τ value is generally converted to TPM (transcripts per million) by multiplying by $10^6$ in order to get the expression value of transcripts. TPM value is not directly comparable to RPKM or FPKM value. We converted TPM values to comparable RPKM values using the conversion formula provided in (7):

$$\text{RPKM}_i = \frac{10^6 \times \tau_i}{\sum_j \tau_j l_j}$$

Here $i$ corresponds to the $i^{th}$ RefSeq transcript and $j$ varies from 1 to total number of RefSeq transcripts. $l_j$ is the length (in bp) of $j^{th}$ transcript.

Supplementary References:

1. Kim, P., Yoon, S., Kim, N., Lee, S., Ko, M., Lee, H., Kang, H. and Kim, J. (2010) ChimerDB 2.0—a knowledgebase for fusion genes updated. Nucleic acids research, 38, D81-85.
2. Morgulis, A., Gertz, E.M., Schaffer, A.A. and Agarwala, R. (2006) A fast and symmetric DUST implementation to mask low-complexity DNA sequences. Journal of computational biology : a journal of computational molecular cell biology, 13, 1028-1040.
3. Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. Genome research, 12, 656-664.
4. Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics (Oxford, England), 25, 1105-1111.
5. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology, 28, 511-515.
6. Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12, 323.
7. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. and Dewey, C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics (Oxford, England)*, 26, 493-500.

**Supplementary Tables:**

**Supplementary Table 1:** Data sources and types of datasets that were used for the demonstration of R-SAP’s application as well as its performance assessment and testing.

| Data                        | Data type                  | Platform                  | Sample                                | Database               | Database ID/Accession                      |
|-----------------------------|----------------------------|---------------------------|---------------------------------------|------------------------|-------------------------------------------|
| MAQC RNA-seq                | Single end sequencing reads| Roche 454                 | MAQC Universal Human Reference RNA (Sample A) | SRA (Short read archive at NCBI) | SRX002934 (Runs: SRR013995, SRR013996, SRR013997, SRR013998, SRR013999) |
| MAQC expression microarray  | Cel files (intensities)    | Affymetrix Human U133Plus2.0 | MAQC Universal Human Reference RNA (Sample A) | GEO (Gene expression omnibus at NCBI)  | GSM589512                                  |
| MAQC TaqMan qRT-PCR         | Normalized expression values| TaqMan Human MAQC (TAQ, platform id: GPL4097) | MAQC Universal Human Reference RNA (Sample A) | GEO (Gene expression omnibus at NCBI)  | GSM129638, GSM129639, GSM129640, GSM129641, |
| ENCODE RNA-seq              | Paired end 75 bp long reads (insert length 200 bp) | Illumina Genome Analyzer | Gm12878 cell line complete Poly-A selected | UCSC DCC (UCSC ENCODE Data coordination center) | wgEncodeCaltechRnaSeqGm12878R2x75Nall200Fast.qRd1Rep1 |
| ENCODE expression microarray| cel files (microarray intensities) | Affymetrix Human Exon 1.0 ST arrays | Gm12878 cell line | GEO (Gene expression omnibus at NCBI)  | GSM472901                                  |
**Supplementary Table 2**: GenBank accession IDs for the 206 EST and mRNA sequences that were used as the high confidence test dataset (see supplementary method section) for testing the chimera-detection module of R-SAP. 164 (~80% of 206) Accession IDs (in red color) were reported as chimera-transcripts by R-SAP.

| Accession IDs | GenBank Accession IDs | GenBank Accession IDs | GenBank Accession IDs | GenBank Accession IDs |
|---------------|-----------------------|-----------------------|-----------------------|-----------------------|
| AB000268.1    | AF492832.1             | AY624559.1             | EU446645.1             | AB300355.1            |
| AB010342.1    | AJ131466.1             | AY624560.1             | FM165197.1             | AB300356.1            |
| AB038155.1    | AJ131467.1             | AY633656.1             | FM165198.1             | AB300357.1            |
| AB127472.1    | AJ251843.1             | AY803272.1             | L03357.1               | AF047022.1            |
| AB275889.1    | AJ251844.1             | BC008826.1             | L22179.1               | AF125093.1            |
| AF024541.1    | AJ251845.1             | D90075.1               | M13096.1               | AF143407.1            |
| AF031404.1    | AJ251846.1             | DQ084494.1             | M19730.1               | AF186109.1            |
| AF041811.2    | AJ297349.1             | DQ204770.1             | M25946.1               | AF230662.1            |
| AF060927.1    | AY624571.1             | M30829.1               | AF231996.1             |
| AF060928.1    | AJ299261.1             | DQ204772.1             | M30832.1               | AF254086.1            |
| AF060929.1    | AJ299262.1             | DQ204773.2             | M31213.1               | AF272376.1            |
| AF060930.1    | AJ301611.1             | DQ437654.1             | M73779.1               | AF295356.1            |
| AF060931.1    | AJ301612.1             | DQ437655.1             | S50916.1               | AF297746.1            |
| AF102845.1    | AJ303089.1             | DQ451148.1             | S72478.1               | AF297747.1            |
| AF113911.1    | AJ417079.1             | DQ831522.1             | S72604.1               | AF310722.1            |
| AF123094.1    | AJ438986.1             | DQ841178.1             | S75763.1               | AF390893.1            |
| AF125808.1    | AJ549094.1             | DQ845345.1             | S77574.1               | AF524261.1            |
| AF125809.1    | AJ549095.1             | DQ845346.1             | U02308.1               | AF533988.1            |
| AF177236.1    | AJ549096.1             | DQ886024.1             | U02368.1               | AFY186998.1           |
| AF177237.1    | AJ972402.1             | DQ898313.1             | U41814.1               | AFY380223.1           |
| AF177238.1    | AM491359.1             | DQ898314.1             | X03541.1               | AFY380226.1           |
| AF177239.1    | AM491360.1             | DQ912588.1             | X06418.1               | AFY54915.1            |
| AF186140.1    | AM491361.1             | DQ912589.1             | X62947.1               | AFU327511.1           |
| AF231995.1    | AM491362.1             | DQ912590.1             | X98708.1               | AFM165196.1           |
| AF254087.1    | AM491363.1             | EF051633.1             | X98709.1               | AFM82827.1            |
| AF254088.1    | AY040324.1             | EF158045.1             | X98710.1               | AS71225.1             |
| AF272374.1    | AY040555.1             | EF374064.1             | Y08643.1               | AS72479.1             |
| AF272375.1    | AY043457.1             | EF406122.1             | Y15913.1               | AS72621.1             |
| AF272383.1    | AY138857.1             | EF423615.1             | Y15914.1               | AS72865.1             |
| AF272384.1    | AY138858.1             | EF525170.1             | Y15915.1               | AS74529.1             |
| AF272385.1    | AY138859.1             | EF632110.1             | Y15916.1               | AS79325.1             |
| AF297748.1    | AY138860.1             | EU090248.1             | Y15917.1               | AS79332.1             |
| AF297749.1    | AY186997.1             | EU090249.1             | Y15918.1               | AS81242.1             |
| AF364037.1    | AY187920.1             | EU216066.1             | Y15919.1               | ASU35622.2            |
| AF373587.1    | AY187921.1             | EU216066.1             | Y15920.1               | ASU41743.1            |
| AF395885.1    | AY187922.1             | EU216070.1             | Y15921.1               | ASX07537.1            |
| AF422798.1    | AY380222.1             | EU216071.1             | Y16346.1               | AX79200.1             |
| AF477006.1    | AY380224.1             | EU236680.1             | Z35761.1               | AX85960.1             |
| AF487522.1    | AY380225.1             | EU236948.1             | AB000267.1             |
| AF487905.1    | AY624556.1             | EU314929.1             | AB001343.1             |
| AF487906.1    | AY624557.1             | EU364772.1             | AB012575.1             |
| AF492831.1    | AY624558.1             | EU432099.1             | AB300354.1             |
**Supplementary Table 3:** Intron-retention events detected in MAQC Reference Human dataset using R-SAP from high-scoring reads that were also characterized as internal-exon-extension.

| Characterization sub-category       | Reads (Represented RefSeq transcripts) |
|------------------------------------|----------------------------------------|
| Internal-exon-extension            | 18,419 (7,648)                         |
| Complete intron-retention          | 361 (275)                              |
| Total number of retained introns   | 305                                    |

**Supplementary Table 4:** Distribution of “multiple-annotations” reads that were detected in MAQC Reference Human dataset using R-SAP. Since more than one type of novel transcriptional event was detected in each of the “multiple-annotations” reads, reads representing characterization sub-categories here may be overlapping. “Exon-skipping” and “intron-retention” events are already included in “exon-deletion” and “internal-exon-extension” events.

| Sub-categories (characterization) | Reads |
|-----------------------------------|-------|
| Exon-deletion                     | 2889  |
| (Exon-skipping)                   | (579) |
| Internal-exon-extension           | 1867  |
| (Intron-retention)                | (13)  |
| Alternative TSS                   | 563   |
| Alternative Polyadenylation       | 782   |
Supplementary Table 5: Distribution of Trans-ABySS characterized reads that were also classified as “high-scoring” by R-SAP previously using MAQC Human Reference RNA-Seq data. RefSeq transcripts (hg18) were used as annotated set of transcripts. “novel-transcript” are those that could not be mapped to any of the known RefSeq exon by Trans-ABySS. Novel transcriptional event sub-categories (AS3, AS, AS53, novel_exon, novel_intron, novel_utr, retained_intron and skipped_exon) have 121 overlapping reads between them. Filtered-out reads are those that were not reported in any category by Trans-ABySS.

| Sub-categories (characterization) | Reads (% total “high-scoring”) | Associated RefSeq transcripts |
|----------------------------------|--------------------------------|-------------------------------|
| AS3                              | 1447 (0.29%)                   | 960                           |
| AS5                              | 1503 (0.30%)                   | 1031                          |
| AS53                             | 24 (0.004%)                    | 22                            |
| novel_exon                       | 608 (0.12%)                    | 462                           |
| novel_intron                     | 259 (0.05%)                    | 212                           |
| novel_utr                        | 357 (0.072%)                   | 190                           |
| retained_intron                  | 2 (0.0004%)                    | 1                             |
| skipped_exon                     | 768 (0.15%)                    | 568                           |
| Total novel-transcriptional      | 4847 (0.98%)                   | 2548                          |
| events (121 overlapping          |                                |                               |
| subtracted from the total sum    |                                |                               |
| Mapping within known exons       | 123066 (25.05%)                | 16211                         |
| Total exon-associated            | 127913 (26.04%)                | 18759                         |
| Novel-transcripts                | 144173 (29.35%)                | NA                            |
| Total reported                   | 272086 (55.4%)                 | 18759                         |
| Filtered-out                     | 219031 (44.6%)                 |                                |
| Total high-scoring               | 491117                         |                               |
**Supplementary Table 6:** Distribution of transcripts that were assembled from ENCODE Gm12878 RNA-Seq data using Cufflinks and then characterized by R-SAP. RefSeq transcripts (hg18) were used as reference set of transcripts.

| Sub-categories (characterization) | Reads (% Cufflinks assembled transcripts) | Represented RefSeq transcripts |
|----------------------------------|------------------------------------------|--------------------------------|
| Exon-skipping                     | 1,389 (1.82%)                            | 1,186                          |
| Exon-deletion                     | 597 (0.78%)                              | 548                            |
| AlternativeTSS                    | 1,957 (2.57%)                            | 4,382                          |
| Alternative Polyadenylation       | 2,848 (3.74%)                            | 5,361                          |
| Internal-exon-extension           | 15,416 (20.25%)                          | 6,744                          |
| Intron-retention                  | 3,563 (4.68%)                            | 2,059                          |
| Multiple-annotations (additional exon-skipping, exon-deletion, alternativeTSS, alternativePolyA, internal-exon-extension and intron-retention) | 14,255 (18.73%) | 7,667 |
| Total novel-transcripts           | 40,025 (52.59%)                          | 13,638                         |
| Exon-only                         | 8,940 (11.74%)                           | 8,275                          |
| Intron-only                       | 10,172 (13.36%)                          | 3,772                          |
| Neighboring-exons                 | 3,166 (4.16%)                            | 2,282                          |
| Gene-desert                       | 8,582 (11.27%)                           |                                |
| Uncharacterized                   | 5,216 (6.85%)                            |                                |
| Total Cufflinks assembled transcripts | 76,101                 |                                |
**Supplementary Table 7:** Distribution of transcripts that were assembled from ENCODE Gm12878 RNA-Seq data using Cufflinks and then classified by Cuffcompare into structurally variant classes of RefSeq transcripts (hg18).

| Cuffcompare classification | Classification code | Number of reads (% of assembled transcripts) | Associated RefSeq transcripts |
|----------------------------|----------------------|---------------------------------------------|------------------------------|
| Potential novel-isoform    | j                    | 24,752 (32.52%)                            | 9240                         |
| Full match + contained     | =, c                 | 14,015 (18.41%)                            | 11046                        |
| Falling entirely within intron | i                  | 10,149 (13.33%)                            | 3765                         |
| Possible polymerase run (2kb away from reference transcript) | p                  | 1,772 (2.32%)                             | 1361                         |
| Unknown, intergenic        | u                    | 10,131 (13.31%)                            | 0                            |
| Generic reference exon overlap | o                  | 1,389 (1.82%)                             | 1257                         |
| Exonic overlap (on opposite strand) | x                  | 765 (1%)                                  | 586                          |
| Intron overlap (on opposite strand) | s                  | 306 (0.4%)                                | 297                          |
| Single exon with partial intron overlap (pre-mRNA fragment) | e            | 12,822 (16.84%)                           | 4016                         |
| **Total assembled transcripts** | | **76,101** | |
**Supplementary Table 8:** Comparison between R-SAP’s characterizations and Cuffcompare’s classification of transcripts that were previously assembled from ENCODE Gm12878 RNA-Seq dataset using Cufflinks (also see Table 5 and Supplementary Table S6 and S7 for detailed distribution and comparison of other sub-categories).

| R-SAP characterization | Cuffcompare classification (code) | Number of associated assembled transcripts | Overlap |
|-----------------------|-----------------------------------|---------------------------------------------|---------|
|                       |                                   | R-SAP | Cuffcompare | #Reads | %Cuffcompare | %R-SAP |
| Exon-only             | Complete match + contained (=,c) | 8940  | 14015       | 8699   | 62%         | 97.3%  |
| Intron-only           | Intron-only (i)                   | 10172 | 10149       | 10137  | 99.9%       | 99.6%  |
| Neighboring-exon      | Polymerase run (p)                | 3166  | 1772        | 1772   | 100%        | 55.6%  |
| Gene-desert           | Unknown-intergenic (u)            | 8582  | 10131       | 8582   | 84.71%      | 100%   |

**Supplementary Table 9:** Sequencing reads, reference genome alignment and R-SAP characterization statistics for the ENCODE RNA-seq data for Gm12878 cell line. RPKM values were estimated using exon-only reads that were also classified as high-scoring reads by R-SAP.

| Category                              | Reads     |
|---------------------------------------|-----------|
| Raw sequencing reads                   | 87,929,372|
| Total reference genome mapped          | 54,095,800|
| Uniquely mapped                        | 49,533,283|
| Multi-hits                             | 4,562,517 |
| High-scoring                           | 32,815,685|
| Exon-only                              | 22,390,589|
| Total RefSeq transcripts detected using Exon-only reads | 24,080 |
Supplementary Figures:

Supplementary Figure 1: Frequency of exon skipping in high-scoring reads from MAQC Reference Human dataset. A total of 893 sequencing reads resulted from the skipping of 1191 exons corresponding to 645 Hg18 RefSeq known transcripts.
A. Exon-only

B. Exon-deletion
C. **Exon-deletion**

![Exon-deletion diagram](image)

D. **Exon-skipping**

![Exon-skipping diagram](image)
E. **Alternative TSS**

F. **Alternative Polyadenylation**
G. **Internal-exon-extension**

![Diagram showing internal-exon-extension]

H. **Intron-retention**

![Diagram showing intron-retention]
I. **Intron-only**

![Intron-only Diagram]

J. **Potential new intronic-exon**

![Potential new intronic-exon Diagram]
Supplementary Figure 2: Examples of various sub-categories characterized by R-SAP from the test MAQC Reference Human dataset as they are displayed in the UCSC genome browser (hg18) snap-shots. Reference genome alignment of the sequence is shown under the track “Your Sequence from Blat search” (in black). RefSeq gene tracks are displayed under “RefSeq genes” track (in blue). (A) Exon-only. (B-C) Exon-deletion events in RefSeq transcripts resulting from the skipping of exonic bases and detected from the discontinuous blocked alignment of sequencing read. (D) Exon-skipping event. (E) Gene boundary expansion event at 5’UTR (alternativeTSS) (F) Gene boundary expansion event at 3’UTR (alternativePolyadenylation), respectively. (G) Internal-exon-extension. (H) Intron-retention. (I) Intron-only. (J) Intron-only read sequences were further used for the detection of potential new “intronic-exon” events that had comparable (same order of magnitude) expression value (estimated with as RPKM) with the expression value (RPKM) value of the transcript itself. 3^{rd} intron of gene U2AF2 is shown in the figure with a potential new exon (pointed by blue arrow). RPKM values are displayed under the track “MAQC Human Reference” (in orange). (K-L) Sequencing reads that do not fall within any of the RefSeq transcript boundaries but map within the pre-specified gene radius (‘d’, 5kb default) are characterized as neighboring-exon (K) while reads that map outside gene-radius are characterized as gene-desert reads (L). (M) Shows an “uncharacterized” read (under “Use Supplied Track”) where alignment blocks of the read are separated by more than the cutoff value (100kb default). Such reads could not be associated or characterized with any RefSeq transcript and very likely be resulting from cDNA library or sequencing artifact, or alignment artifact.
Supplementary Figure 3: Correlation plots of RefSeq transcripts (hg18) quantification estimates from ENCODE Gm12878 RNA-Seq data using three different methods: R-SAP, Cufflinks and RSEM. Log2 transformation expression estimates shown here. A Spearman correlation of A. 0.84 (p<0.0001) and B. 0.65 (p<0.0001) was observed when R-SAP’s expression estimates (Y-axis) were compared with those from Cufflinks (A.) and RSEM (B.) (X-axis). C. Shows a correlation of .040 (p<0.0001) expression estimates from Cufflinks and RSEM.