Identification of the Cellular Targets of the Transcription Factor TCERG1 Reveals a Prevalent Role in mRNA Processing.

James L. Pearson1,5, Timothy J. Robinson 2,3, Manuel J. Muñoz6, Alberto R. Kornblihtt6, and Mariano A. Garcia-Blanco1,4,5*

1Department of Molecular Genetics and Microbiology, 2Department of Molecular Cancer Biology, 3Medical Scientist Training Program, 4Department of Medicine, 5Center for RNA Biology, Duke University Medical Center, Durham, North Carolina, USA, 27710, and 6Laboratorio de Fisiología y Biología Molecular, Departamento de Fisiología, Biología Molecular y Celular, IFIBYNE-CONICET, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. Ciudad Universitaria, Pabellón 2, (C1428EHA) Buenos Aires, Argentina.

* To whom correspondence should be addressed. Tel: 919-613-8636; Fax: 919-613-8646; Email: garci001@mc.duke.edu

SUMMARY:
The transcription factor TCERG1 (TCERG1) associates with RNA polymerase II holoenzyme and alters the elongation efficiency of reporter transcripts. TCERG1 is also found as a component of highly purified spliceosomes and has been implicated in splicing. In order to elucidate the function of TCERG1 we used siRNA mediated knock-down followed by en masse gene expression analysis to identify its cellular targets. Analysis of data from HEK293 and HeLa cells identified high confidence targets of TCERG1. We found that targets of TCERG1 were enriched in miRNA binding sites, suggesting the possibility of post-transcriptional regulation. Consistently, RT-PCR analysis revealed that many of the changes observed upon TCERG1 knock-down were due to differences in alternative mRNA processing of the 3' UTRs. Furthermore, a novel computational approach, which can identify alternatively processed events from conventional microarray data, showed that TCERG1 led to widespread alterations in mRNA processing. These findings provide the strongest support to date for a role of TCERG1 in mRNA processing and are consistent with proposals that TCERG1 couples transcription and processing.

INTRODUCTION:

TCERG1, which was previously known as co-activator of 150kD (CA150), was originally identified as a component of an active cellular fraction that supported Tat activated transcription from the HIV-LTR(1,2). Subsequent cloning and characterization determined that TCERG1 is composed of multiple protein domains, most notable of which are three WW domains in the N-terminal half and six FF repeats in the C-terminus(1). Immunodepletion of TCERG1 from HeLa nuclear extract results in the loss of Tat-transactivation of the HIV-LTR, with little effect on basal transcription(1). Over-expression of TCERG1 in cell culture represses expression from HIV-LTR and alpha-4 integrin reporter constructs by inhibiting of transcription elongation(3). Inhibition of these minimal reporter constructs is promoter specific and TATA box dependent(3). Consistent with a role in elongation, TCERG1 is found associated with elongation factors, Tat-SF1 and P-TEFb(4). TCERG1 is also present in a complex with RNA Polymerase II (RNAPII) holoenzyme and via the FF domains TCERG1 preferentially associates with the hyper-phosphorylated form (II0) (1,5). This experimental evidence demonstrates a tight and functional association of TCERG1 with elongation competent RNAPII.

Accumulating evidence also implicates TCERG1 in the process of RNA splicing. The WW domain 2 (WW2) of TCERG1 interacts with the splicing factors, SF1, U2AF, and components
of the SF3 complex (6,7). TCERG1 has been identified in highly purified spliceosomes in multiple studies (8-10) and was recently identified as a substrate of CARM1, an arginine methyltransferase whose activity is known to affect alternative splicing (11). Over-expression studies demonstrate that TCERG1 can affect splicing of β-globin and β-tropomyosin minimal splicing reporters (7).

The processes of transcription and splicing are known to be coordinated by the CTD of RNAPII. In addition to binding TCERG1, the CTD is known to interact with factors involved in capping, splicing, and polyadenylation (12-16). The CTD is widely accepted as the critical site for the assembly of the machinery responsible for transcription coupled mRNA processing, and is required for the efficient splicing, polyadenylation, and termination of transcription in vivo (13,17). The modular structure of TCERG1, with splicing factor associating-WW domains present in the N-terminus and CTD-associating FF repeats in the C-terminus, offers the ideal structure for a protein involved in coupling transcription and splicing. Consistent with this model, both halves of TCERG1 have been shown to be critical for the assembly of higher order transcription/splicing complexes (4). Fittingly, the C. tentans TCERG1 homolog (hrp130) accumulates at the intron-rich Balbiani ring 3 gene (18).

Attempts to elucidate the function of TCERG1 have been limited to biochemical analysis and transient over-expression studies utilizing artificial transcription and splicing reporters (1,3,6,7,19,20). An important gap in our knowledge is the identity of TCERG1-responsive cellular genes. This study combines RNAi mediated knock-down using siRNAs, LUC (target-CGUACGCGGAAUACUUCGA) and TCERG1-A, using a two-hit protocol as previously described (22). Hep3B cells were plated at 10^5 cells per well in a 6-well dish. 24 hrs. after plating, siRNA duplexes siTCERG1 (CUCCAGAUGGGAAGGUUU) and siLUC (CUUACGCUGAUACUUUGA) at 40nM final concentration using Lipofectamine (Invitrogen).

RNA isolation and microarray hybridization.
For knock-down experiments, total RNA was isolated from HEK293T-EGFP and HeLa-LUC cells using the RNAeasy kit (Qiagen), and assessed for quality with an Agilent Lab-on-a-Chip 2100 Bioanalyzer. All probes for hybridization were then prepared according to standard Affymetrix protocols on the Human U133A or Human U133A_2 GeneChip arrays and scanned at a target intensity of 500 (Expression Analysis).

Microarray analysis. Genespring v7.2 (Silicon Genetics) was used to generate the list of TCERG1 responsive targets defined in Table 1, and Sup. Tables 1 and 2. The data files were GC-RMA normalized using Genespring v7.2 and all probe sets utilized in the analysis as described in Results and Eq 1, and Eq 2. All fold change values reported in Table 2, and Supplementary Tables 1, 2, 3, and 4 represent the different between the TCERG1(+)_293 (n=6) vs. TCERG1(-)_293 (n=6)
conditions. Average relative (percent) standard deviation among experimental replicates was calculated using RMA normalized (RMA Express) data including all 22115 experimental probe sets: Mock (n=3), AveStDev=10%, Median=8.7%; EGFP (n=3), AveStDev=8.7%, Median=7.4%; TCERG1-B (n=3), AveStDev=8.2%, Median=7.2%; and TCERG1-C (n=3), AveStDev=7.1%, Median=6.3%. All Affymetrix data files can be found at [http://data.cgt.duke.edu/pearson.php](http://data.cgt.duke.edu/pearson.php).

**GSEA Analysis.** Microarray data were normalized using RMA (RMA Express) before import for use by Gene set enrichment analysis (GSEA). GSEA is implemented by the software package GSEA-P(23) available for download by the Broad Institute ([http://www.broad.mit.edu/gsea/index.html](http://www.broad.mit.edu/gsea/index.html)). GSEA derived statistics were generated using 1000 permutations of gene tags.

**RT-PCR analysis.** 2.4ug of total RNA was digested with RQ1-DNase (Invitrogen) to remove any residual DNA contamination. 2.0ug of DNase treated total RNA was primed with oligodT (Invitrogen) and reverse-transcribed using MMLV-RT (Invitrogen) at 37 deg. C. for 1.5 hours. The cDNA produced from polyadenylated mRNA was then amplified by PCR using gene specific primers: RBM3, forward exon5-5’GCTATGGGAGTGGCAGGTATTA and reverse exon7-5’AGATGGGAGTCTCGCTTTC; CTTN/EMS1, forward exon2-5’CTGGGAATTCTTCTATTGGA, reverse exon4-5’ACCCCATCTTTGCTCCTTCT, and reverse inton4-5’CTGTCATGGGTATCAGGTC AAA; BUB3, forward exon7-5’CGCATCACCAGCCTTCAGTA and reverse exon8-5’AGGGGACAGAAGGGGAAATA; Beta-actin forward 5’GCTGATGTCGCTGTAACGT and reverse 5’CCTCGTCGCCACCATAGGAATC. PCR products were sequence verified. RT-PCR analysis of fibronectin EDI exon inclusion was performed as described (24).

**Bioinformatic analysis of gene expression data.** A program, Splicer AV, was written in Perl to analyze standard RMA normalized Affymetrix microarray data for evidence of alternative splicing. The inputs used to calculate the evidence of alternative processing, or Odds Score, used the log_2 fold change and signal-to-noise ratios (SNR) from each individual probe set derived from the expression data sets. The SNR was calculated as the difference of the means of two data sets divided by the sum of their standard deviations. A Gaussian mixture model was implemented to calculate the maximum likelihood that these probe set log fold changes (weighted by square root of the SNR) for a given gene were generated by a single Gaussian distribution, or by two Gaussian distributions. In this way the maximum likelihood of a single regulation event is compared to the maximum likelihood of two separate regulation events, in this case interpreted as changes in alternative processing. To avoid over-fitting, Gaussians were not allowed to have a standard deviation of less than a 0.4 log_2-fold change, which is ~28% change in expression levels. The maximum likelihood ratio of the data being described by one vs. two Gaussians is referred to as the Odds Score. This Odds Score can then be used to rank the genes in order of descending Odds Scores, creating a list of the most likely targets of alternative processing. All single probe set genes were excluded from analyses using this program. Other caveats include that a dead or inactive probe set within a gene with other functional probe sets would generate a high Odds Score, since it could appear that part of the gene is being up-regulated while the other is not. In addition, data sets with genome-wide stronger signals (i.e. higher probe set log fold change) will tend to generate higher Odds Scores. Others (25,26) have previously used single probe set level data instead of multiple probe sets as a means of detecting alternative splicing, however, such algorithms may not have detected any of the alternative processing events presented in this paper, all of which spanned multiple probe sets. For a detailed discussion of probe set discrepancies in Affymetrix microarrays, see Stalteri and Harrison (2007). A list of top targets as predicted by the program is included as Supplementary data.

**Normalized comparison of Mock vs. EGFP and Mock vs. TCERG1 knock-downs B and C.** To compare two lists of different probe set log fold change distributions, sub-distributions (subL) were first generated from each original distribution (L) that were matched for the maximum absolute...
value of each gene’s log fold change. Starting with the highest maximum absolute value of the control master list \( L \), genes were alternately drawn from each original distribution, \( L \), (i.e. Mock vs. EGFP and Mock vs. TCERG1-B) and added to that subdistribution, \( subL \), (i.e. subMock vs. EGFP or subMock vs. TCERG1-B), each time drawing the gene with the next lower absolute log fold change. In this way two \( subLs \), one from each original distribution, were drawn that could be directly compared without confounding by differences in overall log fold change magnitudes. 

**Statistical analysis of Odds Scores.** A Kolmogorov-Smirnov test was performed on the top 100 genes to examine the probability that these genes came from the same distribution (two sided KS test) or if one distribution was greater than another (one sided KS test). This analysis was performed for the maximum absolute value corrected sub-distributions.

**Statistical and experimental validation of SplicerAV.** The original RMA normalized microarray intensity values from the TCERG1(-)\(_{293} \) (n=6) experimental condition were each compared to the average of the TCERG1(+)\(_{293} \) (n=6) control condition to determine six fold change values for each probe set. The probe sets within a gene were then grouped using the groupings predicted by SPICERAV (A or B in the output shown in supplemental table 5). All normalized fold change values for each probe set within A or B were assembled into two new groups. A Welch’s t-test was performed on these two new groups to calculate the probability that the observed fold changes were the same. This probability was then corrected using the Bonferroni correction, given that \( N \) probe sets within a gene can be grouped a total of \( 2N-1 \) possible ways. Low p-values indicate that the two groups of probe sets as predicted by SPICERAV do not behave the same. This could happen because of alternative processing, poor probe set annotation, or bad probe sets. RT-PCR validation was performed under semi-quantitative conditions using radionucleotide incorporation. Products were resolved by 6% PAGE. Quantification was performed by exposure to phosphorimaging screen and analyzed by ImageQuant (Molecular Dynamics). PCR primer sequences will be made available upon request.

**RESULTS:**

**Identification of TCERG1 targets in HEK-293T cells.** We set out to identify cellular targets of TCERG1 using a combination of siRNA mediated knock-down and en masse gene expression analysis. To this end, HEK-293T cells stably expressing EGFP were used for siRNA mediated knock-down of TCERG1, allowing the heterologously expressed EGFP to be targeted by siRNA as a negative control. Mock transfected cells were an additional negative control and were considered as baseline expression. Three independent siRNA duplexes specific for TCERG1 were transfected at a final concentration of 10nM and all significantly lowered TCERG1 levels, with TCERG1-B and TCERG1-C giving the best knock-down. In mock-treated cells and those transfected with EGFP siRNA, TCERG1 levels did not change (Fig. 1A, left panel). The EGFP siRNA was fully functional as demonstrated by FACS analysis, which confirmed reduced EGFP levels after 72 hours (Fig 1A, right panel). This experiment was repeated three times with similar results. Total RNA from the controls, Mock and siEGFP, and the two siRNAs with the best knock-down, TCERG1-B and -C were used for subsequent global mRNA quantification. We chose the 72 hour time point, because TCERG1 levels had been significantly depleted for at least 24 hours.

Total RNA was prepared from Mock, siEGFP, siTCERG1-B, or siTCERG1-C treated HEK293T cells from three independent experiments. These twelve RNA samples were interrogated on Affymetrix Hu-133A_2 GeneChip arrays. Genespring v7.2 (Silicon Genetics) software was used for analysis and data were normalized using the GC-RMA method (27). The data for identical conditions, Mock (n=3), EGFP (n=3), TCERG1-B (n=3), and TCERG1-C (n=3), were averaged among the replicates (experimental variation among replicates, reported as relative standard deviation, is presented in the Experimental Procedures).

The analysis was carried out separately to derive the Down gene set (genes whose level decreased upon TCERG1 knock-down) and the Up gene set (genes whose level increased upon TCERG1 knock-down). To derive the Down gene set, we compared the Mock and EGFP conditions and excluded from the analysis any genes that
decreased 1.2 fold or greater in the EGFP condition (see Eq.1, Table 1). From the remaining genes, potential targets were identified as those genes which decreased 1.2 fold or greater when condition Mock was compared to both condition TCERG1-B and condition TCERG1-C. To derive the Up gene set we utilized the same process, varying only in the direction of the change. (see Eq. 2, Table 1). These criteria were set to cast a wide net based more on reproducibility and less on fold change. It should be noted that the 1.7 fold reduction in TCERG1 transcript, as reflected in the microarrays, resulted in an average 2.75±0.75 fold reduction in protein levels as determined by semi-quantitative western blot of the three experiments (data not shown). A more stringent criteria was used to identify probe sets that increased or decreased ≥1.5 fold and all of the examples described below (see Fig. 3) fell into this more stringent list of targets.

Utilizing two independent TCERG1 specific siRNA duplexes, and defining targets as those genes which change commonly between them allowed us to minimize false positives due to siRNA- specific off target effects. The EGFP knock down served as an additional filter to remove genes which change merely as a result of an activated the siRNA response.

The analysis described above and summarized in Table 1 resulted in the identification of 554 probe sets, representing 487 unique genes, that decreased, and 485 probe sets, representing 432 unique genes, that increased, upon TCERG1 depletion (Supplementary Tables 1 and 2).

**Utilizing TCERG1 Knock-down in HeLa cells as validation of cellular targets of TCERG1.**

In our quest to identify genuine targets of TCERG1, we performed TCERG1 knock-down experiment utilizing HeLa cells stably expressing firefly luciferase, Hela-Luc, which have a different origin from HEK293T cells. In addition to changing cell lines, the experiments in HeLa cells utilized the TCERG1 siRNA duplex, TCERG1-A, which was not used in the HEK293T analysis (Fig 1B). We reasoned that targets identified in both HEK293T and HeLa cells using different siRNAs could be considered *bona fide* TCERG1 targets.

In order to identify TCERG1 targets shared by HEK293T and HeLa cells, we used Gene Set Enrichment Analysis (GSEA)(23,28). GSEA is useful when comparing a defined gene set to the rank order of another microarray experiment. The utility of GSEA hinges on the ability to quantify and visualize the distribution of the defined gene set within the data of another microarray comparison. By relying on the distribution, GSEA dispenses with the issues of varying fold change between cell types. Specifically, the objective of the software is to determine if genes in a set S occur more frequently at the top or bottom of a list L. The program provides an enrichment score based on a weighted Kolmogorov-Smirnov statistic (23) and also defines the leading edge subset of S which is interpreted as the core subset of S responsible for the enrichment score. In our case, set S was either the Up-gene set (SU) or the Down-gene set (SD), in HEK293T cells following TCERG1 knock-down (Table 1) and the rank order list L would be a continuous ranking of all probe sets correlated to the level of TCERG1 in HeLa cells. Before performing this comparison between cell lines, we decided to carry out a test of internal consistency by analyzing the HEK293T data using GSEA parameters. As required by the method we created two conditions: TCERG1(+)293 (n=6) was derived from the control conditions, Mock (n=3) and EGFP (n=3), and TCERG1(-)293 (n=6) was derived from the knock-down conditions TCERG1-B (n=3) and TCERG1-C (n=3). These two conditions were used to construct the rank order list, L293 = TCERG1(+)293 vs. TCERG1(-)293. As expected the SU was enriched in condition TCERG1(-)293 (Fig 2A, left panel) and the SD was enriched in condition TCERG1(+)293 (right panel). This exercise gave us confidence that the GSEA could be applied to compare the results from HeLa and HEK293T cells.

We then applied GSEA to the HEK293T – HeLa comparison, keeping S = SU or SD (from HEK293T cells). In order to create a rank list $L_{HeLa}$ we carried out the following experiment. HeLa cells were transfected with TCERG1-A siRNA specific for TCERG1, or Luc siRNA, which targets the luciferase transcript, using a two-hit protocol (see Materials and Methods). At 48 hrs and 72 hrs following the second hit, total RNA and protein were harvested. This experiment was done twice and both times TCERG1 protein levels were significantly reduced at both 48hrs and 72hrs - 5 -
The RNA samples, derived from the two independent experiments, were subjected to quantification using Affymetrix HU-133A GeneChip arrays, and the data were used to create the new rank order list $L_{HeLa} = \text{TCERG1}(+)_{HeLa} \text{ VS. TCERG1}(+)_{HeLa}$. Condition TCERG1(+)_{HeLa} (n=4) combined the 48 hr and 72 hr LUC knock-downs from the two experiments, while condition TCERG1(-)_{HeLa} (n=4) combined the 48 hr and 72 hr TCERG1 knock-downs. The top of the list represents those probe sets that were positively correlated with the first condition TCERG1(+)_{HeLa}; these were the probe sets that go down upon HeLa TCERG1 knock-down (Fig. 2B). The bottom of the list represents probe sets that were negatively correlated with TCERG1(+)_{HeLa}; these were the probe sets that go up upon HeLa TCERG1 knock-down (Fig. 2B). When we applied GSEA to $L_{HeLa}$ using $S_{up}$, the 485 Up-gene set demonstrated enrichment in condition TCERG1(-)_{HeLa} with a leading edge subset of 131 probe sets (Fig 2B, left panel). When GSEA was applied to $S_{down}$, the 554 Down-gene set demonstrated significant enrichment in condition TCERG1(+)_{HeLa} (p=0.05; FDR=0.1) with a leading edge subset of 264 probe sets contributing to the core enrichment (Fig 2B, right panel). Heat maps displaying the correlation of the 50 most enriched of each output are shown to the right of each panel of Fig 2. These 131 probes sets, representing 123 gene targets, up-regulated upon TCERG1 depletion (i.e., require TCERG1 for decreased expression), and 264 probe sets, representing 226 down-regulated gene targets (i.e., require TCERG1 for increased expression) are defined here as the “highest confidence” targets of TCERG1 and we refer to these as belonging to our target list (Table 2, and Supplementary Tables 3 and 4).

TCERG1 depletion results in changes in mRNA processing.

Whereas in some cases (e.g., RBM3, which was downregulated by 2.1 fold) we noted changes in overall level of transcripts, we also noticed several instances where multiple probe sets assaying the same gene did not behave consistently. In the case of $EMS1/CTTN$, which was the most up-regulated TCERG1 responsive target in HEK293T cells and was present among the 131 member highest confidence list defined by GSEA, there are four probe sets. While three probe sets, which queried exonic sequences did not respond appreciably to TCERG1 knock-down, the probe set which identified $EMS1(CTTN)$ as the most affected (4.8 fold upregulated) by TCERG1 knock-down was found to query sequences within intron 4. As shown in Fig 3, RT-PCR amplification of $CTTN$ mRNA using oligodT priming for the RT step and PCR primers designed to sequences in exon 2 and intronic sequences downstream of exon 4 resulted in production of a product that increased upon TCERG1 knock-down. This product was sequenced and identified as a $CTTN$ transcript with retained intron 4 sequences. The product of amplification from exon 2 to exon 4 of $CTTN$ did not change upon TCERG1 knock-down, demonstrating the specificity of the effect of TCERG1 on one isoform of $CTTN$ mRNA (Fig 3).

$BUB3$ is interrogated by four Affymetrix probe sets, however, of these, only two changed upon TCERG1 knock-down in HEK293T cells, with one of these (downregulated by 1.7 fold) passing through the HeLa GSEA filter. Careful examination of the $BUB3$ sequences revealed that the two probe sets most affected by TCERG1 knock-down interrogated sequences present only when a particular 3' splice site is utilized. Alternate 3' splice site utilization would result in a decrease in the signal from these probe sets upon TCERG1 knock-down. Indeed, amplification of $BUB3$ transcripts with primers designed to visualize this event revealed a change in 3' splice site usage upon TCERG1 knock-down in HEK293T cells (Fig 3).

These data suggested that many changes in mRNA levels of TCERG1 targets, as reported by Affymetrix microarray analysis, could represent changes in RNA processing.

TCERG1 knock-down affects the inclusion of the Fibronectin EDI exon.

In order to obtain independent confirmation of these observations we directly evaluated the effect of TCERG1 depletion on the splicing of the Fibronectin EDI exon. Although the EDI exon is not interrogated directly by the microarray experiments described above, splicing for this exon has been previously shown to be sensitive to alterations in transcription elongation (29,30). Skipping of this exon is stimulated by high elongation rates. Depletion of TCERG1 by siRNA treatment of Hep3B cells transfected with reporter minigenes provoked an increase in EDI
inclusion independently of the promoter used (CMV or mFN) (Fig. 4). These data with a well-characterized alternative splicing reporter provided additional confirmation of the effects of TCERG1 depletion on alternative processing. **TCERG1 knock-down results in prevalent changes in mRNA processing.**

The Affymetrix H133A series of GeneChip arrays have 4,642 genes with two or more probe sets. The presence of multiple probe sets provides the possibility to observe isoform specific changes. To this end, we developed a program, SplicerAV, to predict genes with a high likelihood of alternative processing by analyzing the behavior of their probe sets using a phenotype correlated expression dataset.

SplicerAV determined if the log fold changes for the group of probe sets for a given gene varied in their distribution (see methods for determination of log fold change and signal-to-noise ratio). In other words, whether or not the probe sets distribute into one or two groups. If the log fold changes for all probes sets for a given gene distributed in one group, then we concluded that there was no change in processing detected by these probe sets. If, however, the distribution of the log fold changes for all probe sets for a given gene was best described by two groups, we suspected an alternative processing event. To identify and rank the genes suspected of alternative processing, we generated an Odds Score. This was done using the log fold change in expression for each probe set weighted by a function of its signal-to-noise ratio. The Odds Score was defined as the ratio of the likelihood that the probes sets were described by two events versus the likelihood that the probe sets were described by one event. The lowest possible Odds Score for a gene was 1, which indicated that all probe sets for a given gene behaved identically and provided no evidence of alternative processing. An Odds Score >1 indicated some discrepancy in the behavior of the probe sets, which could be caused by an alternative processing event. The greater the value of the Odds Score the higher that gene ranked in the list of possible alternative processing candidates.

Comparison of HEK-293T knock-down TCERG1(+)293 vs. TCERG1(-)293 was used to generate and rank Odds Scores for the 4,642 genes on the array with two or more probes. **CTTN** and **BUB3**, which we had shown are alternatively processed in response to CA150 depletion, were ranked first and second on the list (Supplementary Table 5), providing validation that SplicerAV could identify genes that were alternatively processed from Affymetrix gene based microarray data.

We examined our top 12 predictions using two approaches, statistical (generation of p-values) and experimental (semi-quantitative RT-PCR) and the results are summarized in Table 3. The statistical approach derived a p-value for the predicted probe set distributions using the microarray expression values (see Methods). Ten of the top twelve predictions had p-values < 0.01 demonstrating the robust nature of the program (Table 3). Of these ten top significant predictions, eight generated readily testable hypotheses. In addition to CTTN and BUB3, three additional genes among these eight were experimentally shown to undergo the alternative processing predicted. ACACA (2.3 fold upregulated) demonstrated alternative exon inclusion, PPP3CB (1.6 fold downregulated) and SYNCRIP (1.5 fold upregulated) changes could be explained by alternate polyadenylation sites (Fig. 5). Of the three remaining genes, MTC1P1 was unamenable to RT-PCR, while ASAH1 and APPBP2 did not appear to be alternatively processed. The predicted alternative processing of RABGGTB, which had a probe set that was downregulated by 1.6 fold and was ranked number 43 by SplicerAV, was also validated. The change in RABGGTB expression upon TCERG1 knock-down could be best explained by alternative polyadenylation site usage (Fig 5). Two of the top ten significant predictions did not generate a testable hypothesis; MAP2K5 probe set behavior was unintelligible and one of two RBM3 probe sets was poorly annotated and not specific for any curated RBM3 transcript.

We also used SplicerAV to ask whether or not the effects of TCERG1 knock-down were widespread. If this were true, knock-down would result in a significant change in the number of genes predicted to have a high Odds Scores. We compared the distribution of Odds Scores for: Mock (n=3) vs. EGFP (n=3), Mock (n=3) vs. TCERG1-B (n=3), and Mock (n=3) vs. TCERG1-C (n=3). We visualized these distributions using a Kaplan-Meier plot (survival plot), and both
TCERG1 knock-down conditions resulted in a greater number of genes displaying high Odds Score when compared to the control EGFP knock-down (Figure 6A). In order to control for the correlation between log fold change and Odds Score, we generated maximum log fold change matched sub-distributions, referred to as subLs (Fig 6B and described in Materials and Methods). To do this, all 4,642 genes from both the original Mock vs. EGFP and the original Mock vs TCERG1-B or -C were ranked by the absolute maximum log fold change of each gene. Each of these master lists L were methodically scanned for genes with similar maximum log fold changes. These similar genes were drawn from each master list L to generate a subL. These subLs were therefore closely matched by maximum absolute log fold change (Probe Score) for the pair of master distributions being examined (Fig 6B). Survival plots of the Odds Scores were generated from these matched pairs of subLs: Mock vs EGFP and Mock vs TCERG1-B (Fig 6C); and Mock vs. EGFP and Mock vs TCERG1-C (Fig 6D). In each of these two comparisons the top 100 odds scoring genes from each condition were compared using a Kolmogorov-Smirnoff (KS) test. Mock vs TCERG1-B generated a significantly higher Odds Scores compared to that of Mock vs. EGFP (one sided KS test, p = 1.39 x 10^{-11}). In the second comparison Mock vs TCERG1-C also demonstrated significantly higher Odds Scores compared to that of Mock vs. EGFP (one sided KS test, p = 3.15 x 10^{-5}). When Mock vs. TCERG1-B and Mock vs. TCERG1-C were plotted against each other we observed no significant difference in Odds Scores (two sided KS test, p=0.37) (data not shown).

This analysis demonstrated that TCERG1 knock-down resulted in a higher Odds Score when compared to EGFP knock-down, and we interpret these data as evidence for a prevalent involvement of TCERG1 in alternative processing of cellular mRNAs.

GSEA analysis identifies miRNA binding site enrichment in target genes. Using GSEA, we sought to determine if genes affected by TCERG1 levels shared any commonality that could shed additionally light on TCERG1 function. Although this manuscript has used GSEA to query one gene set at a time, GSEA was designed to query a file of many gene sets at once.

The Broad Institute has made available a motifs gene set file (c3.v2.symbols.gmt) that includes 780 gene sets that contain between 15 and 500 members, each sharing a common sequence motif. Each phenotype of the correlated data set, L_{293} = TCERG1(+)_{293} vs. TCERG1(-)_{293} was assessed for enrichment of any of these 780 motifs gene sets. The TCERG1(+){293} phenotype did not display significant enrichment for any motifs gene set, however the TCERG1(-){293} phenotype displayed enrichment of 33 gene sets with an FDR <25% and p-values of <0.01 (Table 4). Of these 33 gene sets, 21 (64%) were those defined as containing genes with a predicted mirRNA binding site.

An independent computational approach, List to List Comparison (L2L)(31), also demonstrated significant miRNA target enrichment in the 485 up set while showing none in the larger 554 dn set (Supplementary Table 6). This data demonstrate that among genes downregulated by TCERG1, there is a significant enrichment of genes predicted to bind, and presumably be regulated by, microRNAs. These data suggest that TCERG1 may regulate mRNA levels via a mechanism involving miRNAs. This may be true of the set of genes where TCERG1 alters processing of alternative 3' UTRs.

**DISCUSSION:**

TCERG1 was discovered in 1997, and despite extensive biochemical and functional characterization, its role in vivo has remained elusive. As a means to ascertain the function of TCERG1, we sought to identify the cellular genes which are responsive to alterations in TCERG1 protein levels. Our strategy, which combined RNAi mediated knock-down in both HEK293T and HeLa cells followed by microarray analysis, resulted in a list of “high confidence” cellular targets of TCERG1 and demonstrated a functional link between TCERG1 and splicing in vivo.

This study demonstrates that decreases in TCERG1 protein levels can both up-regulate and down-regulate expression of cellular gene products. While this functional analysis unambiguously identifies gene products that depend on TCERG1, it does not discriminate between several potential mechanisms. The low overall fold changes observed for the targets (ave. 1.4fold) suggests that TCERG1 may act through a mechanism not easily reported by microarrays.
designed for transcriptome based studies. It is possible that TCERG1 interacts with the nascent transcript (or RNP) and directly alters splicing decisions. This could be consistent with independent effects on transcription elongation and alternative processing. Alternatively, TCERG1 could work at the interface of RNAP II and the splicing machinery – exerting an effect on processing that is functionally coupled to effects on transcription. It is also possible that TCERG1 only affects transcription directly and that all of the processing effects are the consequence of altered transcription. Finally, TCERG1 could control other regulators that could then alter several of the targets.

TCERG1 depletion results in an increase in the levels of predicted targets of microRNAs (Table 4 and Supplementary Table 6). It is possible that TCERG1 is directly involved in the expression of miRNAs, and that upon depletion of TCERG1 there is decreased expression of miRNAs resulting in an increase in target mRNA. Alternatively, TCERG1 could regulate miRNA targets by altering the availability of the target sites. This would be accomplished by alternative mRNA processing leading to different 3’ UTRs. In fact given the bias of the A133 microarrays, which interrogate the 3’ ends of transcripts preferentially, we suggest that the CA150 targets identified here will be enriched in those with alternative 3’ UTRs. It is also possible that a target of TCERG1 could be responsible for the enrichment via an indirect mechanism. In fact, RBM3, most down-regulated gene upon TCERG1 knock-down in HEK293T cells, has been shown to affect cellular miRNA levels(32). Although the mechanism remains to be elucidated, our observations suggest that TCERG1 levels can markedly affect miRNA targets.

RBM3 is also involved in regulation of translation in neuronal cells (33)and is down-regulated by poly-g glutamine expression (34). RBM3 over-expression significantly protected cells from polyglutamine induced toxicity, suggesting a role in HD pathology (34). Interestingly TCERG1 has been suggested as a genetic modifier of Huntington’s Disease(HD)(35-37), and has been shown to be protective in models of HD neurotoxicity(38). The ability of TCERG1 to affect alternative processing of cellular mRNA, and specifically the expression of RBM3, suggests a mechanism whereby TCERG1 could influence HD progression.

Accumulating evidence suggests a role of TCERG1 in the coupling of transcription to splicing. TCERG1 fulfills a number of criteria required of such a factor. TCERG1 interacts with the CTD of RNAPII and preferentially binds a phosphorylated CTD(5). TCERG1 overexpression affects elongation in a promoter specific fashion(3). Changes in promoter context and elongation rate of transcription are known to affect splicing decisions (39). Reciprocally, addition of splice sites to a transcribed sequence has also been shown to affect transcription (40,41). TCERG1 has been defined as a spliceosome component in multiple studies(7-9,42). Immunolocalization on polytene chromosomes demonstrates a marked accumulation of the C. tentans TCERG1 homolog (hrp130) at the intron rich Balbiani ring 3, an area of active transcription and remarkably high intron density (18). The authors postulated that hrp130 was recruited to modulate elongation in order to facilitate splicing(18). The work reported here provides the strongest evidence yet that TCERG1 is involved in splicing of cellular miRNAs.

Although the gene specific Affymetrix H133 series of microarrays are not touted as having the potential to report isoform specific changes in mRNA, we have demonstrated the utility of careful analysis of these data. SplicerAV allowed the demonstration that TCERG1 levels can have prevalent effects on the levels of specific mRNA isoforms. Although limited by the number of probe sets which can report these differences, conventional Affymetrix GeneChip arrays are the predominant microarray platform used by the scientific community for comparative expression studies, and archived data derived from these studies are voluminous. SplicerAV has broad application for the reanalysis of this wealth of available microarray data for potential alternative processing.

ACKNOWLEDGEMENTS:

We are grateful to Dr. Holly K. Dressman and the Duke Microarray Facility (IGSP) for facilitating the microarray experiments and critical reading of the manuscript. We would like to thank Neal Mukherjee and Drs. Christopher Lee (UCLA), Uwe Ohler (Duke), and Caroline LeSommer for useful discussions. We would also
like to especially thank Drs. Sayan Mukherjee and Alexander Hartemink, without whom SplicerAV would not have been developed. TJR was supported by the MSTP at Duke University. This work was supported by NIH grant 1RO1GM071037 (MAGB).

REFERENCES:
1. Sune, C., Hayashi, T., Liu, Y., Lane, W. S., Young, R. A., and Garcia-Blanco, M. A. (1997) *Mol Cell Biol* **17**(10), 6029-6039.
2. Sune, C., and Garcia-Blanco, M. A. (1995) *J Virol* **69**(5), 3098-3107.
3. Sune, C., and Garcia-Blanco, M. A. (1999) *Mol Cell Biol* **19**(7), 4719-4728.
4. Sanchez-Alvarez, M., Goldstrohm, A. C., Garcia-Blanco, M. A., and Sune, C. (2006) *Molecular & Cellular Biology* **26**(13).
5. Carty, S. M., Goldstrohm, A. C., Sune, C., Garcia-Blanco, M. A., and Greenleaf, A. L. (2000) *Proc Natl Acad Sci U S A* **97**(16), 9015-9020.
6. Goldstrohm, A. C., Albrecht, T. R., Sune, C., Bedford, M. T., and Garcia-Blanco, M. A. (2001) *Mol Cell Biol* **21**(22), 7617-7628.
7. Lin, K. T., Lu, R. M., and Tarn, W. Y. (2004) *Molecular & Cellular Biology* **24**(20), 9176-9185.
8. Makarov, E. M., Makarova, O. V., Urlaub, H., Gentzel, M., Will, C. L., Wilm, M., and Luhrmann, R. (2002) *Science* **298**(5601), 2205-2208.
9. Rappsilber, J., Ryder, U., Lamond, A. I., and Mann, M. (2002) *Genome Res* **12**(8), 1231-1245.
10. Deckert, J., Hartmuth, K., Boehringer, D., Behzadnia, N., Will, C. L., Kastner, B., Stark, H., Urlaub, H., and Luhrmann, R. (2006) *Mol. Cell. Biol.* **26**(14), 5528-5543.
11. Cheng, D., Cote, J., Shaaban, S., and Bedford, M. T. (2007) *Molecular Cell* **25**(1), 71-83.
12. Kornblihtt, A. R., de la Mata, M., Fededa, J. P., Munoz, M. J., and Nogues, G. (2004) *Rna* **10**(10), 1489-1498.
13. Bird, G., Zorio, D. A., and Bentley, D. L. (2004) *Mol Cell Biol* **24**(20), 8963-8969.
14. Reed, R. (2003) *Curr Opin Cell Biol* **15**(3), 326-331.
15. Goldstrohm, A. C., Greenleaf, A. L., and Garcia-Blanco, M. A. (2001) *Gene* **277**(1-2), 31-47.
16. de la Mata, M., and Kornblihtt, A. R. (2006) *Nat Struct Mol Biol* **13**(11), 973-980.
17. McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S. D., Wickens, M., and Bentley, D. L. (1997) *Nature* **385**(6614), 357-361.
18. Sun, X., Zhao, J., Kylberg, K., Soop, T., Palka, K., Sonnhammer, E., Visa, N., Alzhanova-Ericsson, A. T., and Daneholt, B. (2004) *Chromosoma* **113**(5), 244-257.
19. Smith, M. J., Kulkarni, S., and Pawson, T. (2004) *Molecular & Cellular Biology* **24**(21), 9274-9285.
20. Goldstrohm, A. C., Duke University. Dept. of Microbiology., and Duke University. Program in Cell and Molecular Biology. (2001) Biochemical and functional analysis of the human transcription factor CA150. In., Duke University, 2001.
21. Florez, P. M., Sessions, O. M., Wagner, E. J., Gromeier, M., and Garcia-Blanco, M. A. (2005) *J Virol* **79**(10), 6172-6179.
22. Wagner, E. J., and Garcia-Blanco, M. A. (2002) *Molecular Cell* **10**(4), 943-949.
23. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005) *Proc Natl Acad Sci U S A* **102**(43), 15545-15550.
Cramer, P., Caceres, J. F., Cazalla, D., Kadener, S., Muro, A. F., Baralle, F. E., and Kornblihtt, A. R. (1999) *Molecular Cell* **4**(2), 251-258

Fan, W., Khalid, N., Hallahan, A. R., Olson, J. M., and Zhao, L. P. (2006) *Theor Biol Med Model* **3**, 19

Hu, G. K., Madore, S. J., Moldover, B., Jatkoe, T., Balaban, D., Thomas, J., and Wang, Y. (2001) *Genome Res* **11**(7), 1237-1245

Cope LM, R. I., Jaffee HA, Wu Z, Speed TP. (2004) *Bioinformatics* **20**(3), 323-331

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003) *Nature Genetics* **34**(3), 267-273

de la Mata, M., Alonso, C. R., Kadener, S., Fededa, J. P., Werbajh, S. E., Srebrow, A., and Kornblihtt, A. R. (2003) *Embo J* **20**(20), 5759-5768

Newman, J. C., and Weiner, A. M. (2005) *Genome Biol* **6**(9), R81

Dresios, J., Aschrafi, A., Owens, G. C., Vanderklish, P. W., Edelman, G. M., and Mauro, V. P. (2005) *Proc Natl Acad Sci USA* **102**(6), 1865-1870

Smart, F., Aschrafi, A., Atkins, A., Owens, G. C., Pilotte, J., Cunningham, B. A., and Vanderklish, P. W. (2007) *J Neurochem* **101**(5), 1367-1379

Kita, H., Carmichael, J., Swartz, J., Muro, S., Wyttenbach, A., Matsubara, K., Rubinsztein, D. C., and Kato, K. (2002) *Hum Mol Genet* **11**(19), 2279-2287

Holbert, S., Denghien, I., Kiechle, T., Rosenblatt, A., Wellington, C., Hayden, M. R., Margolis, R. L., Ross, C. A., Dausset, J., Ferrante, R. J., and Neri, C. (2001) *Proc Natl Acad Sci USA* **98**(4), 1811-1816.

Chattopadhyay, B., Ghosh, S., Gangopadhyay, P. K., Das, S. K., Roy, T., Sinha, K. K., Jha, D. K., Mukherjee, S. C., Chakraborty, A., Singhal, B. S., Bhattacharya, A. K., and Bhattacharyya, N. P. (2003) *Neuroscience Letters* **345**(2), 93-96

Andresen, J. M., Gayan, J., Cherny, S. S., Brocklebank, D., Alkorta-Aranburu, G., Addis, E. A., Cardon, L. R., Housman, D. E., and Wexler, N. S. (2007) *J Med Genet* **44**(1), 44-50

Arago, M., Holbert, S., Zala, D., Brouillet, E., Pearson, J., Regulier, E., Thakur, A. K., Aebrischer, P., Wetzel, R., Deglon, N., and Neri, C. (2006) *J Neurosci* **26**(17), 4649-4659

Kornblihtt, A. R. (2005) *Curr Opin Cell Biol* **17**(3), 262-268

Furger, A., O'Sullivan, J. M., Binnie, A., Lee, B. A., and Proudfoot, N. J. (2002) *Genes Dev* **16**(21), 2792-2799

Fong, Y. W., and Zhou, Q. (2001) *Nature* **414**(6866), 929-933

Neubauer, G., King, A., Rappsilber, J., Calvio, C., Watson, M., Ajuh, P., Sleeman, J., Lamond, A., and Mann, M. (1998) *Nat Genet* **20**(1), 46-50

**FIGURE LEGENDS:**

**Figure 1.** RNAi mediated TCERG1 knock-down in HEK293T and HeLa cells. A. TCERG1 knock-down in HEK293T-EGFP cells. Left Panel. HEK293T-EGFP cells were Mock transfected, or transfected with siRNA duplexes; EGFP, TCERG1-A, TCERG1-B, or TCERG1-C. 24, 48, or 72 hrs. post-transfection cell lysates were resolved by SDS-PAGE, transferred to PVDF, and immunoblotted with TCERG1 specific antiserum (top panel). Immunoblotting was also performed with PTB antiserum as a
loading control (bottom panel). Right Panel. The function of the control EGFP siRNA duplex was confirmed by FACS analysis of EGFP levels in Mock, EGFP, and TCERG1-A transfected cells at 72hrs post-transfection. B. TCERG1 knock-down in HeLa-Luc cells. HeLa-Luc cells were transfected in two independent experiments (Exp #1, left panel; and Exp #2, right panel) with siRNA duplexes; LUC or TCERG1-A using a two-hit protocol. 48 and 72 hrs. after the second siRNA transfection cell lysates were immunoblotted as in A.

**Figure 2. Identification of TCERG1 Targets by Gene Set Enrichment Analysis (GSEA).** A. Control GSEA output showing the distribution of gene sets (S), 485-up (left panel) and 554-down (right panel), within the rank order list of genes (L) derived from conditions TCERG1(+)293T and TCERG1(-)293T. The 485up list demonstrates enrichment in condition TCERG1(-)293T (right panel), while the 554down list is enriched in condition TCERG1(+). The heat maps display the 50 most enriched genes in 485up (left panel) and 554down (right panel), that correlate with each condition. B. HeLa-Luc knock-down GSEA output showing the distribution of gene sets (S), 485-up (left panel) and 554-down (right panel), within the rank order list of genes (L) derived from conditions TCERG1(+)HeLa and TCERG1(-)HeLa. The heat maps display the correlation of the 50 most enriched genes in 485up (left panel) and 554down (right panel).

**Figure 3. TCERG1 affects alternative mRNA processing.** cDNA generated by oligo-dT primed reverse transcription of total RNA from Mock, EGFP, TCERG1-B, or TCERG1-C samples from HEK293T-EGFP Exp #1, Exp #2, and Exp #3, was PCR amplified using gene specific primers. RBM3 was amplified from exon5 to exon7. CTTN message; “CTTN retained intron” was amplified using a forward primer in exon2 and a reverse primer in intron4. “CTTN” was amplified using the same exon2 primer and a reverse primer in exon4. BUB3 was amplified using a forward primer in exon7 and a reverse primer in exon8, which resulted in two products which differ in exon8 3’ splice site choice. β-actin was amplified as a control.

**Figure 4. TCERG1 influences fibronectin EDI exon inclusion.** To assess the effects of TCERG1 depletion on alternative splicing decisions, knock-down experiments were performed by transfecting siTCERG1 or siLUC as a control. A. Western blot of extracts from Hep3B cells cotransfected with T7-TCERG1 expression vector and with siTCERG1 or siLUC indicated that inhibition of TCERG1 expression by the siRNA is almost complete and specific. B. RT-PCR analysis total RNA isolated from Hep3B cells transfected with EDI exon reporter mini-gene driven by either CMV or mFN promoter and siLUC or siTCERG1 as labeled. Radionucleotides allowed visualization of EDI exon inclusion ,EDI+, and EDI exon skipping, EDI- (upper panel) as quantified in (lower panel).

**Figure 5. Experimentally validated SplicerAV targets.** The top 12 alternative splicing targets predicted by SplicerAV, as well as RABGGTB were considered for experimental RT-PCR validation. Of these 13 interrogated genes, 10 generated readily testable hypotheses. Of these, six were experimentally validated (CTTN, BUB3, ACACA, PPP3CB, SYNCRIP, RABGGTB). CTTN and BUB3 are shown in Fig 4. The remaining validated gene targets are shown above, with A-C being from the top 12 and D being 43rd. Each gene target is shown as a schematic with the predicted alternative processing hypotheses, which was generated by combining SplicerAV output with the genomic alignment of the interrogated probe sets. Arrows indicate a greater than 20% change in probe set expression. Below the predicted behavior is a schematic of the primers used for experimental RT-PCR validation, along with the predicted products. To the right are quantifications of both the predicted hypotheses and the experimental RT-PCR. The quantifiable predictions were made by averaging the expression of the probe sets which interrogated regions corresponding to the predicted product. Both the microarray data and RT-PCR data were obtained using the TCERG1(+)293 (n=6) vs. TCERG1(-)293 (n=6) experimental conditions.
Figure 6. TCERG1 knock-down results in global changes in mRNA processing. A. Kaplan-Meier survival plot of Odds scores for Mock vs. EGFP, Mock vs. TCERG1-B, and Mock vs. TCERG1-C siRNA knock-down. The cumulative number of genes that have the given Odds score or greater is plotted for each distribution. B. Schematic showing the method used to generate subdistributions, (subLs) with similar distributions of maximum log fold change scores. Both original distributions, in this case Mock vs EGFP and Mock vs TCERG1-C, have any genes with only one probe set removed. Within each remaining, multiple probe set gene, the probe set with the highest absolute change in expression is identified as that gene’s maximum log fold change (MLFC), shown as “Max Change” in the figure. These genes are then sorted by descending order of this MLFC to create a master distribution for each treatment (e.g. Mock vs. EGFP). A subdistribution, subL, of each master distribution is then created. This is done using an initial MLFC cutoff equal to one. Starting with the Mock vs. EGFP list, the first gene who’s MLFC is below one is added to the subL being generated from EGFP (subL Mock vs. EGFP). The MLFC of this first gene is then set as the new, lower cutoff for the next gene to be drawn. This lower cutoff will then be used to select the next lower MLFC gene from the Mock vs. TCERG1-C distribution to be added to the subL being generated from Mock vs. TCERG1-C (subL Mock vs. TCERG1-C). In this way genes are draw alternatively from either distribution, selecting a lower MLFC each time. In this way two subLs are generated, which are matched for maximum log fold changes. Dots within the original distributions indicate multiple genes in a row and are not shown for the sake of space, and indicate that the original Mock vs. TCERG1-C distribution has overall higher maximum log fold changes compared to the Mock vs. EGFP distribution. C. Survival plot of odds scores for subL EGFP & subL TCERG1-B. D. Survival plot of odds scores for subL EGFP & subL TCERG1-C.
TABLE 1. Analysis summary.

|                      | Down Set ($S_{dn}$) | Up Set ($S_{up}$) |
|----------------------|---------------------|-------------------|
| Total probes         | 22283               | 22283             |
| EGFP 1.2 fold        | 582                 | 649               |
| EGFP corrected Total | 21701               | 21634             |
| TCERG1 siRNA-B       | 1205                | 1208              |
| TCERG1 siRNA-C       | 913                 | 821               |
| TCERG1 Common B&C    | 554 $^{Eq1}$        | 485 $^{Eq2}$      |

Eq1

\[
[(M_{(n=3)} \geq 1.2 \times TCERG1 - B_{(n=3)}) \cap (M_{(n=3)} \geq 1.2 \times TCERG1 - C_{(n=3)})] \cap \{all \text{ genes} - (M_{(n=3)} \geq 1.2 \times EGFP_{(n=3)})\}
\]

Eq2

\[
[(1.2 \times M_{(n=3)} \leq TCERG1 - B_{(n=3)}) \cap (1.2 \times M_{(n=3)} \leq TCERG1 - C_{(n=3)})] \cap \{all \text{ genes} - (1.2 \times M_{(n=3)} \leq EGFP_{(n=3)})\}
\]
### TABLE 2. Up and Down regulated “High confidence” Targets

#### Top 35 Up-regulated “High confidence” Targets (Fold change derived from HEK293T KD)

| Probe ID | Fold Change | Common Genbank | Gene Title |
|----------|-------------|----------------|------------|
| 214073_at | 4.88 | CTTN (EMS1) | BG475299 | ems1 sequence (mammary tumor and squamous cell carcinoma-associated (p80/85 src substrate) |
| 200799_at | 2.21 | HSPA1A | NM_005345 | heat shock 70kDa protein 1A |
| 212834_at | 1.79 | DDX52 | BE963238 | DEAD (Asp-Glu-Ala-Asp) box polypeptide 52 |
| 218566_s_at | 1.78 | CHORDC1 | NM_012124 | cysteine and histidine-rich domain (CHORD)-containing, zinc binding protein 1 |
| 214157_at | 1.64 | GNAS | AA401492 | GNAS complex locus |
| 206335_at | 1.62 | GALNS | NM_000512 | galactosamine (N-acetyl)-6-sulfatase (Morquio syndrome, mucopolysaccharidosis type IVA) |
| 204034_at | 1.60 | ETHE1 | NM_014297 | ethylmalonic encephalopathy 1 |
| 213637_at | 1.59 | BE503392 | Homo sapiens transcribed sequence with weak similarity to protein ref:NP_060265.1 (H.sapiens) hypothetical protein FLJ20378 [Homo sapiens] |
| 203157_s_at | 1.57 | GLS | AB020645 | glutaminase |
| 204423_at | 1.57 | MKLN1 | NM_013255 | muskelin 1, intracellular mediator containing kelch motifs |
| 200962_at | 1.56 | RPL31 | A348010 | ribosomal protein L31 |
| 219499_at | 1.52 | SEC61A2 | NM_018144 | Sec61 alpha 2 subunit (S. cerevisiae) |
| 210508_s_at | 1.51 | KCNQ2 | D82346 | potassium voltage-gated channel, KQT-like subfamily, member 2 |
| 213459_at | 1.50 | RPL37A | AU155515 | ribosomal protein L37a |
| 204526_s_at | 1.50 | TBC1D8 | NM_007063 | TBC1 domain family, member 8 (with GRAM domain) |
| 220153_at | 1.50 | ENTPD7 | NM_020354 | lysosomal aminopeptidase-like protein 1 |
| 203992_s_at | 1.49 | UTX | AF000992 | ubiquitously transcribed tetratricopeptide repeat gene, X chromosome |
| 202800_at | 1.49 | SLC1A3 | NM_004172 | solute carrier family 1 (glial high affinity glutamate transporter), member 3 |
| 209179_s_at | 1.49 | LENG4 | BC003164 | leukocyte receptor cluster (LRC) member 4 |
| 219484_at | 1.48 | HCF-2 | NM_013320 | host cell factor 2 |
| 222309_at | 1.48 | AK2 | AW972292 | chromosome 6 open reading frame 62 |
| 212173_at | 1.47 | SPATA5L1 | BE90973 | hypothetical protein MGC5347 |
| 220607_x_at | 1.46 | TH1L | NM_016397 | TH1-like (Drosophila) |
| 214169_at | 1.46 | UNC84A | BE615699 | unc-84 homolog A (C. elegans) |
| 215223_s_at | 1.45 | SOD2 | W46388 | superoxide dismutase 2, mitochondrial |
| 214056_at | 1.45 | MCL1 | BF981280 | myeloid cell leukemia sequence 1 (BCL2-related) |
| 200815_s_at | 1.44 | PAFAH1B1 | L13386 | platelet-activating factor acetylhydrolase, isoform lb, alpha subunit 45kDa |
| 209626_s_at | 1.41 | OSBPL3 | AY009837 | oxysterol binding protein-like 3 |
| 201025_at | 1.41 | EIF5B | AB018284 | translation initiation factor IF2 |
| 203938_s_at | 1.41 | TAF1C | NM_005679 | TATA box binding protein (TBP)-associated factor, RNA polymerase I, C, 110kDa |
| 214857_at | 1.41 | AL050035 | Homo sapiens mRNA; cDNA DKFZp566H0124 (from clone DKFZp566H0124) |
| 202067_s_at | 1.40 | LDLR | AI861942 | low density lipoprotein receptor (familial hypercholesterolemia) |
| 209282_at | 1.40 | PRKD2 | AF309082 | protein kinase D2 |
| 202861_at | 1.40 | PER1 | NM_002616 | period homolog 1 (Drosophila) |
| Probe ID     | Fold Change | Common       | Genbank      | Gene Title                                                                 |
|-------------|-------------|--------------|--------------|----------------------------------------------------------------------------|
| 208319_s_at | 2.10        | RBM3         | NM_006743    | RNA binding motif protein 3                                               |
| 205238_at  | 2.03        | FLJ12687     | NM_024917    | hypothetical protein FLJ12687                                           |
| 218431_at  | 1.97        | C14orf133    | NM_022067    | chromosome 14 open reading frame 133                                    |
| 212222_at  | 1.85        | PSME4        | AU143855     | proteasome (prosome, macropain) activator subunit 4                      |
| 207076_s_at | 1.85        | ASS          | NM_000050    | argininosuccinate synthetase                                              |
| 204143_s_at | 1.83        | HSRTSBETA    | NM_017512    | rTS beta protein                                                          |
| 217886_at  | 1.80        | EPS15        | BF213575     | epidermal growth factor pathway substrate 15                             |
| 207761_s_at | 1.80        | DKFZP586A0522| NM_014033    | DKFZP586A0522 protein                                                     |
| 212061_at  | 1.76        | SR140        | AB002330     | U2-associated SR140 protein                                                |
| 203227_s_at | 1.75        | SAS          | AL514076     | sarcoma amplified sequence                                                |
| 204142_at  | 1.74        | HSRTSBETA    | NM_017512    | rTS beta protein                                                          |
| 201456_s_at | 1.74        | BUB3         | NM_004725    | BUB3 budding uninhibited by benimidazoles 3 homolog (yeast)               |
| 202396_at  | 1.73        | TCERG1       | NM_006706    | transcription elongation regulator 1 (CA150)                              |
| 218966_at  | 1.71        | MYO5C        | NM_018728    | myosin VC                                                                 |
| 218961_s_at | 1.69        | PNKP         | NM_007254    | polynucleotide kinase 3'-phosphatase                                      |
| 203962_s_at | 1.68        | NEBL         | NM_006393    | nebulette                                                                  |
| 209894_at  | 1.68        | LEPR         | U50748       | leptin receptor                                                            |
| 216837_at  | 1.68        | IMPACT       | NM_018439    | hypothetical protein IMPACT                                               |
| 204333_s_at| 1.66        | AGA          | NM_000027    | aspartylglucosaminidase                                                   |
| 202447_at  | 1.66        | DECR1        | NM_001359    | 2,4-dienoyl CoA reductase 1, mitochondrial                                |
| 203226_s_at| 1.66        | SAS          | AL514076     | sarcoma amplified sequence                                                |
| 202561_at  | 1.65        | TNKS         | AF070613     | tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase         |
| 210980_s_at| 1.64        | ASAH1        | U47674       | N-acylsphingosine amidohydrolase (acid ceramidase) 1                      |
| 218341_at  | 1.64        | FLJ1838      | NM_024664    | hypothetical protein FLJ20972                                            |
| 219785_s_at| 1.63        | FBXO31       | NM_024735    | MGC15419 protein                                                          |
| 212631_at  | 1.63        | SAS          | AL514076     | sarcoma amplified sequence                                                |
| 209817_at  | 1.62        | PPP3CB       | M29550       | protein phosphatase 3 (formerly 2B), catalytic subunit, beta isoform       |
| 205321_at  | 1.62        | EIF2S3       | NM_001415    | eukaryotic translation initiation factor 2, subunit 3 gamma, 52kDa        |
| 219469_at  | 1.61        | DNHCH2       | NM_024606    | dynein, cytoplasmic, heavy polypeptide 2                                 |
| 213704_at  | 1.61        | RABGGTB      | AA129753     | Rab geranylgeranyltransferase, beta subunit                               |
| 212062_at  | 1.60        | ATP9A        | AB014511     | ATPase, Class II, type 9A                                                |
| 210425_x_at| 1.60        | GOLGIN-67    | AF164622     | golgin-67                                                                 |
| 218620_s_at| 1.59        | HEMK         | NM_016173    | HEMK homolog 7kb                                                          |
| 215735_s_at| 1.58        | TSC2         | AC005600     | tuberous sclerosis 2                                                      |
| 212091_s_at| 1.58        | COL6A1       | AI141603     | collagen, type VI, alpha 1                                                |
TABLE 3: Statistical and experimental validation of SPLICER AV.

| Rank | Gene     | RT-PCR Evidence | Odds score | P-Value   |
|------|----------|-----------------|------------|-----------|
| 1    | CTTN     | VALID           | 99.745     | 1.5E-04*  |
| 2    | BUB3     | VALID           | 4.908      | 1.1E-08*  |
| 3    | GNAS     | ND              | 3.468      | 1.2E-01   |
| 4    | SYNCRIP  | VALID           | 2.521      | 5.4E-06*  |
| 5    | MAP2K5** | ND              | 2.441      | 1.5E-08*  |
| 6    | ACACA    | VALID           | 2.265      | 2.3E-03*  |
| 7    | MTCPI    | ND              | 2.244      | 4.0E-04*  |
| 8    | RBM3**   | ND              | 2.22       | 7.4E-06*  |
| 9    | ASAH1    | invalid         | 2.168      | 5.7E-11*  |
| 10   | PAFAH1B  | invalid         | 2.148      | 2.2E-02   |
| 11   | APPBP2   | invalid         | 2.112      | 1.1E-04*  |
| 12   | PPP3CB   | VALID           | 2.005      | 3.7E-04*  |
| 43   | RABGGTB  | VALID           | 1.378      | 3.6E-03*  |

* Significant at the p<0.01 level
** No testable hypothesis.
Table 4. Gene Set* Enrichment in TCERG1(-)_293 (n=6) phenotype

| Motifs Gene Set Name | SIZE | ES   | NES   | NOM p-val | FDR q-val |
|----------------------|------|------|-------|-----------|-----------|
| ATGCAGT,MIR-217      | 81   | -0.455 | -1.795 | 0.000     | 0.141     |
| ATTACAT,MIR-380-3P   | 74   | -0.449 | -1.758 | 0.002     | 0.118     |
| CTCTATG,MIR-368      | 31   | -0.549 | -1.752 | 0.009     | 0.085     |
| AAGGGAT,MIR-188      | 57   | -0.463 | -1.750 | 0.000     | 0.065     |
| CATTCTA,MIR-203      | 215  | -0.382 | -1.746 | 0.000     | 0.055     |
| V$SHNF4ALPHA_Q6      | 182  | -0.372 | -1.664 | 0.000     | 0.111     |
| AACATTC,MIR-409-3P   | 113  | -0.392 | -1.628 | 0.000     | 0.137     |
| ATCATGA,MIR-433      | 84   | -0.405 | -1.619 | 0.004     | 0.130     |
| CTCAGAGA,MIR-526B    | 50   | -0.449 | -1.610 | 0.009     | 0.128     |
| AGTCTTA,MIR-499      | 56   | -0.426 | -1.580 | 0.006     | 0.158     |
| TTCTNRGNNNTTC_V$HSF_Q6 | 110 | -0.370 | -1.547 | 0.007     | 0.189     |
| ATGCTGG,MIR-338      | 82   | -0.387 | -1.532 | 0.009     | 0.203     |
| CATGTAA,MIR-496      | 145  | -0.344 | -1.516 | 0.002     | 0.209     |
| V$SHNF4_01_B         | 181  | -0.336 | -1.509 | 0.000     | 0.187     |
| V$HIF1_Q3            | 158  | -0.347 | -1.507 | 0.003     | 0.180     |
| ATATGCA,MIR-448      | 157  | -0.342 | -1.498 | 0.002     | 0.170     |
| TTCYRGAA_UNKNOWN     | 219  | -0.330 | -1.496 | 0.000     | 0.166     |
| ATGTACA,MIR-493      | 242  | -0.324 | -1.480 | 0.005     | 0.187     |
| V$MYC_Q2             | 127  | -0.345 | -1.479 | 0.005     | 0.181     |
| TTCRRNRTGC_UNKNOWN   | 130  | -0.350 | -1.479 | 0.009     | 0.175     |
| TTGCACT,MIR-130A,MIR-301,MIR-130B | 306 | -0.311 | -1.465 | 0.000     | 0.184     |
| ACCATT,T,MIR-522     | 129  | -0.339 | -1.463 | 0.007     | 0.181     |
| V$SMYCMAX_02         | 194  | -0.323 | -1.441 | 0.000     | 0.203     |
| GTGCAAA,MIR-507      | 101  | -0.347 | -1.438 | 0.009     | 0.202     |
| V$HIF1_Q5            | 164  | -0.323 | -1.428 | 0.010     | 0.216     |
| AAGCACT,MIR-520F     | 168  | -0.321 | -1.423 | 0.007     | 0.214     |
| V$SYY1_Q6            | 162  | -0.325 | -1.422 | 0.009     | 0.211     |
| V$SYY1_02            | 172  | -0.319 | -1.421 | 0.003     | 0.202     |
| TTTTGTG,MIR-373      | 178  | -0.322 | -1.418 | 0.009     | 0.204     |
| ATTCCTT,MIR-186      | 206  | -0.306 | -1.404 | 0.007     | 0.219     |
| CTNTGT,MIR-524       | 334  | -0.294 | -1.388 | 0.003     | 0.235     |
| V$USF_01             | 187  | -0.305 | -1.387 | 0.007     | 0.234     |
| AATGTGA,MIR-23A,MIR-23B | 329 | -0.287 | -1.378 | 0.003     | 0.242     |

*Gene Set: Broad c3.v2.symbols.gmt [motif]
Figure 1

A.  

|                | 24 hrs. | 48 hrs. | 72 hrs. |
|----------------|---------|---------|---------|
| Mock           | EGFP    | TCERG1-A| TCERG1-B| TCERG1-C|
| EGFP           |         |         |         |         |
| TCRG1-A        |         |         |         |         |
| TCRG1-B        |         |         |         |         |
| TCRG1-C        |         |         |         |         |

72 hrs. FACS

B.  

Exp #1  

|    | 48 hrs. | 72 hrs. |
|----|---------|---------|
| LUC|         |         |
| TCERG1-A|       |         |

Exp #2  

|    | 48 hrs. | 72 hrs. |
|----|---------|---------|
| LUC|         |         |
| TCERG1-A|       |         |

-siRNA

-TCERG1

-PTB
A. 

\[ L = \text{TCERG1}(+)_{293T} \text{ vs. TCERG1}(-)_{293T} \]
\[ S = 485 \text{ up} \]

\[ L = \text{TCERG1}(+)_{293T} \text{ vs. TCERG1}(-)_{293T} \]
\[ S = 554 \text{ down} \]

B. 

\[ L = \text{TCERG1}(+)_{\text{HeLa}} \text{ vs. TCERG1}(-)_{\text{HeLa}} \]
\[ S = 485 \text{ up} \]

\[ L = \text{TCERG1}(+)_{\text{HeLa}} \text{ vs. TCERG1}(-)_{\text{HeLa}} \]
\[ S = 554 \text{ down} \]
**Figure 6**

A. Odds Distribution of Genes with Two or More Probes

L = Mock vs. EGFP
L = Mock vs. TCERG1-B
L = Mock vs. TCERG1-C

B. Original Mock vs. EGFP Gene Name Max Change Original Mock vs. TCERG-C Gene Name Max Change

SAT 0.939 SAT 0.939
TMEG5 0.652 TMEG5 0.652
PRPF4B 0.614 PRPF4B 0.614

Original Mock vs. EGFP Gene Name Max Change Original Mock vs. TCERG-C Gene Name Max Change

AXL 0.0020 AXL 0.0020
CSDE1 0.0017 CSDE1 0.0017

C. Probe Score Normalized TCERG1-B Odds Distribution

SubL = Mock vs. EGFP
SubL = Mock vs TCERG1-B

D. Probe Score Normalized TCERG1-C Odds Distribution

SubL = Mock vs. EGFP
SubL = Mock vs TCERG1-C
Identification of the cellular targets of the transcription factor TCERG1 reveals a prevalent role in mRNA processing

James L. Pearson, Timothy J. Robinson, Manuel J. Muñoz, Alberto R. Kornblihtt and Mariano A. García-Blanco

J. Biol. Chem. published online January 10, 2008

Access the most updated version of this article at doi: 10.1074/jbc.M709402200

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

Supplemental material:
http://www.jbc.org/content/suppl/2008/01/10/M709402200.DC1