HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2014 February 01.

Published in final edited form as:

Nat Methods. 2013 August ; 10(8): 747–750. doi:10.1038/nmeth.2522.

Genome-Wide Profiling of Cap-Independent Translation Enhancing Elements in the Human Genome

Brian P. Wellensiek[1], **Andrew C. Larsen**[1], **Bret Stephens**[1], **Kim Kukurba**[1,3], **Karl Waern**[5], **Natalia Briones**[1], **Li Liu**[1], **Michael Snyder**[5], **Bertram L. Jacobs**[2,4], **Sudhir Kumar**[1,4], and **John C. Chaput**[1,3,*]

[1]Center for Evolutionary Medicine and Informatics, Arizona State University, Tempe, AZ 85287

[2]Center for Infectious Diseases and Vaccinology, The Biodesign Institute, Arizona State University, Tempe, AZ 85287

[3]Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287

[4]School of Life Sciences, Arizona State University, Tempe, AZ 85287

[5]Center for Genomics and Personalized Medicine, Department of Genetics, Stanford University, Stanford, CA 94305

Abstract

We report an *in vitro* selection strategy to identify RNA sequences that mediate cap-independent translation initiation. This method entails the mRNA display of trillions of genomic fragments, selection for translation initiation, and high-throughput deep sequencing. We identified >12,000 translation enhancing elements (TEEs) in the human genome, generated a high-resolution map of human TEE bearing regions (TBRs), and validated the function of a subset of sequences *in vitro* and in cells.

Keywords

Human Genome; Internal Ribosomal Entry Sites (IRES); mRNA Display

Eukaryotic translation initiation usually follows a cap-dependent (CD) mechanism where the 43S ribosomal pre-initiation complex is recruited to a 7-methylguanosine cap located at the 5' end of the mRNA strand via recognition of the cap-binding complex eIF4F[1,2]. While progress over the years has provided a detailed structural and mechanistic understanding of each step in the CD process[1,2], very little is known about the molecular basis of cap-independent (CI) translation initiation[3]. CI translation occurs during normal cellular

*To whom correspondence should be addressed. john.chaput@asu.edu..

AUTHOR CONTRIBUTIONS. J.C. conceived the project. J.C, B.W., B.J., S.K., and L.L. designed the experiments. B.W., B.S., K.W., A.L., K.K., N.B. performed the experiments. J.C. B.W., A.L., K.K., L.L., S.K. and M.S. analyzed the data. J.C. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS. The authors declare competing financial interests: B.J., S.K. and J.C. are named inventors on a patent application on the technology described in this manuscript.

processes (e.g., mitosis and apoptosis) or when the CD translation machinery is compromised by viral infection or disease[4,5]. To address this critical gap in our understanding of protein translation, we developed an *in vitro* selection strategy to identify sequences in the human genome that mediate CI translation initiation.

Our selection strategy relies on mRNA display, which is a cell-free method for covalently linking newly translated proteins to their encoding RNA message[6]. In this approach (Fig. 1a), a genomic library is inserted into the 5' untranslated region (UTR) of a DNA construct containing the genetic information necessary for mRNA display. The library is *in vitro* transcribed into a pool of uncapped single-stranded mRNA that is photo-ligated at the 3' end to a DNA linker containing a 3' puromycin residue. When translated *in vitro*, RNA sequences that mediate CI translation initiation become covalently linked to a peptide affinity tag encoded in the open reading frame. Chemical bond formation between newly translated peptides and their encoding mRNA occurs via the natural peptidyl transferase activity of the ribosome, which recognizes puromycin as a tyrosyl-tRNA analogue (Fig. 1b). Functional RNAs are then isolated, reverse-transcribed, and amplified by PCR to regenerate the pool of DNA for another selection cycle.

We began the selection with a library of ~10^{13} RNA-DNA-puromycin molecules containing a random region of genomic fragments (~150 nts) that were derived from total human DNA[7]. The library was translated for 1 hour at 30°C and fusion formation was promoted by incubating the translation mixture overnight at −20°C under high salt conditions. mRNA-peptide fusions were isolated from the crude lysate by oligo(dT) affinity purification, reverse-transcribed, and sequences displaying a His-6 affinity tag were immobilized on Ni-NTA agarose beads. The beads were thoroughly washed to remove RNA molecules that did not form mRNA-peptide fusions or did not translate in the correct reading frame. mRNA-peptide fusions that remained bound were selectively eluted with imidazole, exchanged into buffer, and amplified by PCR to reinitiate another selection cycle.

The abundance of mRNA-peptide fusions plateaued after six rounds of mRNA display, indicating that the library had become dominated by sequences that could enhance CI translation initiation (Fig. 1c). To assess the level of sequence diversity that remained in the pool, we cloned and sequenced members from the selection output. A total of 636 unique sequences were identified, 225 of which showed 100% identity to the human reference genome (hg18) (Supplementary Table 1). The remaining 411 sequences have high homology (85–99% identity), but contain small degrees of sequence variation that include single nucleotide polymorphisms in addition to small insertions and deletions (Supplementary Table 2). This level of variation is expected for individuals in a population, and it is known that functionally relevant sequences can differ between individual genomes[8,9].

To test our selected sequences for functional activity in human cells, we constructed two luciferase reporter vectors (Fig. 2a)[10]. The first vector contained an unstructured 5' UTR designed to quantify the activity of the translation enhancing elements (TEEs). The second vector contained a stable stem-loop structure ($\Delta G = -58$ kcal/mol) upstream of the insert, which blocks translation in the absence of an internal ribosomal entry site (IRES).

Translation of both mRNA templates containing a no-insert 13-nt control sequence confirmed that the stem-loop structure inhibits translation (~99% inhibition) *in vitro* and in cells (Fig. 2b). Quantitative realtime PCR (qRT-PCR) confirmed that the translational differences were not caused by differences in RNA expression.

Because cryptic splicing activity is a common cause of IRES misinterpretation[11], we used a cytoplasmic expression system that bypasses nuclear expression[12]. In this system, mammalian cells transfected with an expression vector carrying a vaccinia virus (VACV)-specific promoter are immediately infected with VACV. The virus produces its own RNA polymerase that recognizes the viral promoter and mediates RNA expression in the cytoplasm. We confirmed that nuclear expression did not contribute to translation by measuring the luciferase activity of transfected cells that were not infected with VACV. These cells yielded luciferase values equivalent to untreated control cells (data not shown).

Next, we tested the set of perfectly matched sequences for TEE and IRES function in human cells. Using the unstructured vector, we found that the selected sequences produce up to 100-fold more luciferase than the no insert control (Fig. 2c), demonstrating that our *in vitro* selection strategy successfully enriched for sequences that enhance translation. Approximately 20% of our TEEs remain functional when tested in the 5' hairpin construct (Fig. 2c), suggesting that a large number of TEEs are capable of internal ribosomal initiation. To ensure that the observed IRES activity was not due to a cryptic promoter[13], we screened 20 high activity sequences in HeLa cells using a stem-loop vector lacking the VACV promoter. This assay identified 8 sequences with modest to high cryptic promoter activity (Supplementary Fig. 1). We labeled the remaining 12 sequences human IRESs, as they do not exhibit cryptic promoter activity and are not an artifact of RNA splicing.

We then compared the set of 12 human IRESs to a set of 12 randomly chosen sequences from the starting library in the structured and unstructured luciferase reporter systems, both in HeLa cells and in HeLa cell lysate. Strong concordance was observed for the unstructured luciferase reporter vector, which showed ~100-fold greater translation enhancing activity in HeLa cells and in HeLa cell lysate as compared to the set of unselected sequences (Fig. 2d, Supplementary Table 3). A similar trend was observed for the 5' hairpin reporter, which showed that the selected sequences exhibit up to ~400-fold higher activity in cells and up to ~100-fold higher activity *in vitro* than the unselected sequences (Fig. 2d, Supplementary Table 3). Collectively, these results establish the ability of our *in vitro* selection strategy to identify RNA sequences from the human genome that function as efficient translation enhancing elements, a subset of which function as IRESs.

One caveat of our HeLa cell assay is that the mRNA transcripts likely contain a 5' cap due to the strong capping enzymes encoded in the VACV genome[12]. This is not a concern for the hairpin construct as the stem-loop structure was shown to block CD translation initiation (Fig. 2b). However, in the case of the unstructured templates, where a 5' cap could aid translation initiation, further experiments are needed to define the activity of the TEE. We therefore selected 26 sequences that exhibited a range of TEE activity, but had no observable IRES activity (Fig. 2c). We then measured their luciferase activity under CI conditions relative to the no insert control. Consistent with the functional constraints of our

in vitro selection, the selected TEEs maintain their activity in the absence of a 5' cap (Supplementary Fig. 2). In some cases, activity increased significantly when the 5' cap was missing, suggesting that certain TEEs prefer CI translation initiation pathways. This observation provides new insight into the mechanism of translation initiation where the 5'-cap is thought to inhibit alternative pathways[14].

Since only a small number of human TEEs are known[15], we decided to perform Illumina deep sequencing on the starting library (R_0) and the selection output (R_6). Sequence analysis revealed that only 2% of the R_0 sequences remain in the pool after six rounds of selection. We aligned the sequences to the reference human genome (hg19) and identified 12,278 unique regions that were enriched by at least 10-fold (see Methods, Supplementary Fig. 3, Supplementary Table 4). The *in vitro* selected TEE-bearing regions (TBRs) map to ~2 million base pairs (Mb). A vast majority of TBRs were shorter than 250 bps (99.5%) and widely dispersed across all 24 chromosomes (Fig. 3a, Supplementary Fig. 4). Of these, 12% (1,532) mapped to genomic regions containing known genes, even though genic regions (introns and exons) account for ~40% of the human genome (Fig. 3b)[16]. This under-abundance in genic regions may be a result of negative selection against TEEs aimed at avoiding disruptive translation in nature, which would be consistent with our results of TEE activity *in vitro* and in cells (Fig. 2). Moreover, the TBRs are preferentially found in the 5'-UTR regions of genes (3-fold over-representation), which would suggest potential functional roles for these elements. We also observed a small, but statistically significant, enrichment of TBRs in long non-coding RNA (lncRNA) regions as compared to the entire human genome (12.2% vs. 11.5%, binomial test, P=0.003), which could lead to the production of novel proteins since these sites are located in intragenic regions of the genome.

GeneOntology analysis revealed that many TBRs associate with genes involved in signal transduction, cell communication, and neurological system development pathways (Supplementary Fig. 5). These functional categories are frequently reported for genes that have undergone adaptive evolution[17,18]. One such example are genes encoding glutamate receptors, which are important for neural communication, memory formation, learning, and regulation[19]. Among the 21 human genes encoding glutamate receptors, 8 harbor TBRs in their introns. Of these, two were enriched by more than 1000-fold after *in vitro* selection using mRNA display. Some of these sequences are flanked by regions that are highly conserved among species and show transcriptional activity in cells (Fig. 3c), indicating a possible role for TBRs in the translation of proteins involved in important developmental pathways.

In summary, we present an *in vitro* selection strategy that makes it possible to search entire genomes for RNA sequences that enhance cap-independent translation initiation. Using this technique, we identified >12,000 TEEs in the human genome, generated a high-resolution map of human TEE bearing regions, and validated the function of a subset of sequences *in vitro* and in cells. Our approach is time and cost effective, cell-line independent, and scalable, making it an effective tool for studying translation mechanisms in other genomes.

Online Methods

Library assembly and mRNA display selection

The pool of fragmented human genomic DNA was previously constructed with conserved sequences flanking the random region[7]. The library was modified by overlap PCR to add all necessary sequence information required for mRNA display. This included a T7 RNA polymerase promoter site upstream of the random region and an open reading frame and photo-crosslinking site downstream of the random region. The open reading frame included a canonical AUG start site followed by a nucleotide sequence encoding a flexible linker and His-6 protein affinity tag. The library was amplified using the forward primer (5' TTCTAATACGACTCACTATAGGGGATCCAAGCTTCAGACGTGCCTCACTACG) and reverse primer (5' ATAGCCGGTGTCCACTTCCATGATGATGGTGATGGTGGGCCATG GCTGAGCTTGACGCTTTGC). For each round of selection, 120 pmol of the dsDNA library was transcribed with T7 RNA polymerase into single-stranded RNA and purified by 10% denaturing urea-PAGE gel. Purified RNA was photo-ligated to a psoralen-DNA-puromycin linker (5'-psoralen-TAGCCGGTG-$(PEG_9)_2$-A_{15}-ACC-puromycin) by irradiating at 366 nm for 15 minutes. The RNA-DNA-puromycin product was ethanol precipitated and the cross-linked RNA (400 pmol) was translated *in vitro* by incubating the library with micrococcal nuclease-treated rabbit reticulocyte lysate and ^{35}S-methionine for 1 hour at 30°C. The mixture was then incubated overnight at −20°C in the presence of KCl (600 mM) and $MgCl_2$ (75 mM) to promote fusion formation. The mRNA-peptide fusion molecules were purified from the crude lysate using oligo (dT)-cellulose beads (NEB) and reverse transcribed with SuperScriptII (Invitrogen) by extending the DNA primer (5' TTTTTTTTTTTTTTTATCC ACTTCCATGATGATGGT) with dNTPs. Fusion molecules containing the correctly translated His-6 tag were isolated on Ni-NTA agarose beads (Qiagen). Functional sequences were recovered by eluting the column with 500 mM imidazole, dialyzing the sample into water, and amplifying the cDNA by PCR using previously described overlap PCR primers to add back the necessary sequences for mRNA display. The selection progress was monitored by measuring the fraction of S^{35}-labeled mRNA-peptide fusions that bound to and eluted from the oligo (dT) and Ni-NTA affinity columns. After 6 rounds of selection and amplification, the dsDNA library was cloned into a pJET plasmid (Fermentas), and individual isolates were sequenced at the ASU core DNA sequencing facility.

Luciferase reporter plasmids

A monocistronic luciferase reporter vector with an unstructured 5' UTR, that contains both a T7 RNA polymerase promoter and a vaccinia virus synthetic late promoter (slp), was constructed from a pT3_R-luc<IRES>F-luc$(pA)_{62}$ luciferase reporter plasmid[10]. The vector was first modified using PCR to exchange the T3 promoter with a T7 promoter (forward primer 5' GATCCCGGGATTAATAACGACTCACTATAGGGAACAAAAGCTGGGTACCGG and reverse primer 5' GATCCCGGGTGCGCGCTTGGCGTAATCATGG). The resulting PCR product was cut with *Sma*I restriction endonuclease, and recircularized using T4 DNA ligase. A synthetic double-stranded DNA molecule containing the slp promoter was inserted

immediately downstream of the T7 promoter using *Kpn*I and *Xho*I restriction sites. Finally, the *renilla luciferase* gene was removed by PCR using forward primer 5' ACTAGGATCCGCTTCTGTTGGGAAATGC and reverse primer 5' CGCGGATCCAAGCTTATCGATACCGTCGAC. The PCR product was cut with *Bam*HI restriction endonuclease and recircularized using T4 DNA ligase. To assay for IRES activity, two additional luciferase reporter vectors were used, both of which contain a stable stem-loop structure in the 5' UTR. The first vector was the pT7-stem_F-luc$(pA)_{62}$ luciferase reporter plasmid described previously[2]. This plasmid contains a T7 RNA polymerase promoter upstream of the stem-loop. The second vector was constructed by removing the stem-loop structure from pT7-stem_F-luc$(pA)_{62}$ using *Stu*I and *Xho*I restriction sites and reciprocally inserting it into the unstructured vector, immediately downstream of the slp promoter. Plasmids to assay for cryptic promoter activity were generated by removing the T7 and slp promoters from the unstructured vector using *Sma*I and *Bam*HI restriction sites. T4 DNA ligase was then used to insert a 22-nucleotide spacer (5' ATAGCGCCACCGAGATATCTGG 3') in place of the promoters. To insert the human genomic sequences into the luciferase reporter vectors, the genomic fragments were amplified by PCR (forward primer 5' TAGGGGGATCCCAGAC GTGCCTCACTACGT and reverse primer 5' TGGGCCATGGCTGAGCTTGACGCTTTGCT) to add *Bam*HI and *Nco*I restriction sites to the 5' and 3' ends respectively. The PCR products were then reciprocally inserted into the vectors immediately upstream of the luciferase-coding region by restriction endonuclease digestion.

Cell culture

HeLa cells, obtained from American Type Culture Collection, were maintained in DMEM (Invitrogen) supplemented with 5% fetal bovine serum (HyClone) and 5 μg/mL gentamicin (Invitrogen). Cells were kept at 37°C in a humidified atmosphere containing 5% CO_2. The cells were free of mycoplasma contamination, as determined by PCR during routine monitoring of cell lysates.

Luciferase reporter assay

HeLa cells were seeded at a density of 15,000 cells per well in white 96-well plates 18 hours prior to transfection. Cells were transfected with a complex of the luciferase reporter plasmid (200 ng) and Lipofectamine 2000 (0.5 μl) in Opti-MEM (Invitrogen), and immediately infected with the Copenhagen strain (VC-2) of wild-type vaccinia virus at a multiplicity of infection (m.o.i) of 5 PFU/cell. Cells were lysed (6 hours post-infection) in the 96-well plates and luciferase activity was measured using the Promega Luciferase Assay System with a Glomax microplate luminometer (Promega). Cell-free characterization of the top translation enhancing sequences was performed using a Human *In vitro* Protein Expression Kit (Pierce). Luciferase expression was achieved following manufacturer's protocols using 300 ng of linear template for a two-hour transcription at 32°C followed by a 90 min translation at 30°C.

RNA characterization

A portion of the cells used in the luciferase reporter transfection studies were separately lysed to evaluate the quality of the cellular RNA. RNA isolation was performed using the PerfectPure RNA cultured cell kit (5 Prime) according to manufacturer's protocol. Isolated RNA was reverse transcribed with an oligo (dT) primer and Superscript II (Invitrogen). Realtime PCR (iQ™ SYBR® Green Supermix, Bio-Rad) was used to determine the mRNA levels of luciferase (forward primer 5' GCTGGGCGTTAATCAGAGAG and reverse primer 5' GTGTTCGTCTTCGTCCCAGT) as well as the housekeeping gene hypoxanthine-guanine phosphoribosyltransferase (HPRT, forward primer 5' TGCTGAGGATTTGGAAAGGGTG and reverse primer 5' CCTTGAGCACACAGAGGGCTAC). Using the ΔΔCt method, the amount of luciferase mRNA was normalized to HPRT mRNA levels. Luminescence values were then adjusted according to the normalized luciferase mRNA levels.

Sequence analysis

An in-house pipeline was used to process Illumina HiSeq sequences. First, base-calling and quality control were performed using the Illumina HiSeq2000 according to the manufacturer's instructions (Supplementary Table 4a). The average length of reads was 80 base pairs. To detect and trim the PCR primers at both ends of each Illumina read, we used the “cutadapt” program (http://code.google.com/p/cutadapt/) allowing a maximum of 2 mismatches. Both primers were detected in a vast majority of the reads (85% in R_0 and 98% in R_6). However, multiple primers were found to be concatenated in some reads, which is common for HiSeq data. For these reads, we used “cutadapt” iteratively until all primer sequences were trimmed. Finally, reads shorter than 35 bps or longer than 75 bps were discarded, because they contained too many or no copies of the primers (Supplementary Table 4b). To ensure correct orientation for all reads, sequences were reverse complemented if the 5' primer was present at the 3' end or the 3' primer was present at the 5' end.

All trimmed reads were aligned to the human reference genome build 19 (hg19) using iterative execution of “bowtie” alignment and end trimmings[20]. Sequentially, with one base at a time, 16 bps from the 3' end, 5 bps from the 5' end, and another 15 bps from the 3' end were trimmed from unaligned reads, which is done to ensure low-quality base calls do not interfere with sequence alignment. In all iterations, “bowtie” was executed in “-n” mode with “-n 2 –e 70” setting. Reads uniquely mapped to exactly one location, 2 – 10 locations, and more than 10 locations in the hg19 genome were denoted as “single-copy”, “low-copy” and “high-copy” reads, respectively (Supplementary Table 4c).

Based on reads mapped to the human genome, we used the command-line version of the CisGenome[21] to call peaks where R_6 served as the positive sample and R_0 served as the negative control sample; parameters were set as “-c 1 -m 10 -w 60 -s 20 -p 0.009948 -br 0 -ssf 0”. Because TEEs are directional, we applied single-strand filtering and labeled a peak as “forward” or “reverse” depending on which strand of the genome it resided on. To further reduce spurious peaks, we required a peak to have a strand-specific global false discovery rate less than 10%, total number of reads > 10 and at least 1 read present in the R_0 library (Supplementary Table 4d). The CisGenome program compared the normalized number of R_6 reads with the normalized number of R_0 reads in a peak, which represented the fold

enrichment level (Supplementary Table 4e). Because repetitive elements can complicate downstream analysis, we focused on peaks derived from single-copy reads. Furthermore, single-copy peaks containing low-complexity sequences were detected using RepeatMasker (www.repeatmasker.org) with parameters "-noint –species human –q". Peaks with no repeat masked and with more than 10-fold enrichment were called putative TEE-bearing regions (TBRs) (Supplementary Table 4f). Chromosomal distributions of TBRs were converted into ideograms using the Idiographica website[22].

We performed bionomical tests for evaluating the null hypothesis that TBRs are randomly distributed in the human genome. In this case, the random probability of a base to belong to a genomic category was first estimated using the RefSeq database. This was equal to 0.43, 0.005, 0.005, and 0.57, for genes (all exons and introns), 5'-UTRs, 3'-UTRs, and inter-genic regions, respectively. We also conducted GeneOntology enrichment analyses to identify functional categories that were over-represented in the collection of genes found to harbor TBRs (Supplementary Fig. 5). We used GeneOntology classifications from the PANTHER[23] website and applied Bonferroni correction for multiple testing, using a cutoff *p*-value of 10^{-3}. Enriched biological processes were reported (Supplementary Fig. 5). Because the naïve library was generated by randomly sampling the genome, longer genes were sampled more often than shorter genes. To account for this gene length effect, we constructed a background sample from the human genome that matched the length distribution of TBR genes, and redid the GeneOntology enrichment analysis. This process was repeated ten times. The Bonferroni corrected p-values from each analysis were combined using Fisher's method. Biological processes that have significant p-values (<0.01) in at least one of these ten gene-length adjusted analyses or have significant combined p-values (χ^2 p-value <0.05) were highlighted.

Illumina library construction and generation

The Illumina sequencing libraries were generated according to Illumina DNA Sample Kit Instructions (Illumina Part # 0801– 0303). The protocol was modified such that enzymes were obtained from other suppliers, as previously described[24]. Briefly, DNA from the output of round 6 was end-repaired and phosphorylated using the `End-It' kit (Epicentre). The blunt, phosphorylated ends were treated with Klenow fragment (3' to 5' exo minus; NEB) and dATP to yield a 3' A overhang for ligation of Illumina's adapters. Following adapter ligation (LigaFast, Promega) DNA was PCR amplified with Illumina genomic DNA primers 1.1 and 2.1. The final libraries were band-isolated (150–300 bp) from an agarose gel to remove residual primers and adapters. Purified library DNA was captured on an Illumina flowcell for cluster generation and sequenced on an Illumina HiSeq 2000 following the manufacturer's protocols.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank J. Szostak and J. Doudna for providing the human genome library and luciferase vectors, respectively; N. Pakulis, M. Leon, J. Flores, and K. Fenton for technical assistance; G. McInnes for initial bioinformatics analysis; and members of the Chaput lab for helpful discussions and comments on the manuscript. This work was supported by NIH to J.C. (Eureka Award; GM085530) and S.K. (HG002096-11).

References

1. Sonenberg N, Hinnebusch AG. Cell. 2009; 136:731–745. [PubMed: 19239892]
2. Jackson RJ, Hellen CUT, Pestova TV. Nat. Rev. Mol. Cell Biol. 2010; 10:113–127. [PubMed: 20094052]
3. Shatsky IN, Dmitriev SE, Terenin IM, Andreev DE. Molecular Cells. 2010; 30:285–293.
4. Johannes G, Carter MS, Eisen MB, Brown PO, Sarnow P. Proc. Natl. Acad. Sci. USA. 1999; 96:13118–13123. [PubMed: 10557283]
5. Spriggs KA, Stoneley M, Bushell M, Willis AE. Biol. Cell. 2008; 100:27–38. [PubMed: 18072942]
6. Roberts RW, Szostak JW. Proc. Natl. Acad. Sci. USA. 1997; 94:12297–13302. [PubMed: 9356443]
7. Salehi-Ashtiani K, Luptak A, Litovchick A, Szostak JW. Science. 2006; 313:1788–1792. [PubMed: 16990549]
8. Korbel JO, et al. Science. 2007; 318:420–426. [PubMed: 17901297]
9. Kasowski M, et al. Science. 2010; 328:232–235. [PubMed: 20299548]
10. Gilbert WV, Zhou KH, Butler TK, Doudna JA. Science. 2007; 317:1224–1227. [PubMed: 17761883]
11. Baranick BT, et al. Proc. Natl. Acad. Sci. USA. 2008; 105:4733–4738. [PubMed: 18326627]
12. Moss B. Science. 1991; 252:1662–1667. [PubMed: 2047875]
13. Van Eden ME, Byrd MP, Sherrill KW, Lloyd RE. RNA. 2004; 10:720–730. [PubMed: 15037781]
14. Mitchell SF, et al. Mol. Cell. 2010; 39:950–962. [PubMed: 20864040]
15. Mokrejs M, et al. Nuc. Acids Res. 2010; 38:D131–D136.
16. Sakharkar MK, Chow VT, Kangueane P. In Silico Biol. 2004; 4:387–397. [PubMed: 15217358]
17. Akey JM. Genome Res. 2009; 19:711–722. [PubMed: 19411596]
18. Sabeti PC, et al. Science. 2006; 312:1614–1620. [PubMed: 16778047]
19. Traynelis SF, et al. Pharmacol. Rev. 2010; 62.3:405–496. [PubMed: 20716669]
20. Langmead B, Trapnell C, Pop M, Salzberg SL. Genome Biol. 2009; 10:R25. [PubMed: 19261174]
21. Ji HK, et al. Nat. Biotechnol. 2008; 26:1293–1300. [PubMed: 18978777]
22. Kin T, Ono Y. Method Biochem. Anal. 2007; 23:2945–2946.
23. Thomas PD, et al. Genome Res. 2003; 13:2129–2141. [PubMed: 12952881]
24. Auerbach RK, et al. Proc. Natl. Acad. Sci. U.S.A. 2009; 106:1492–1493.

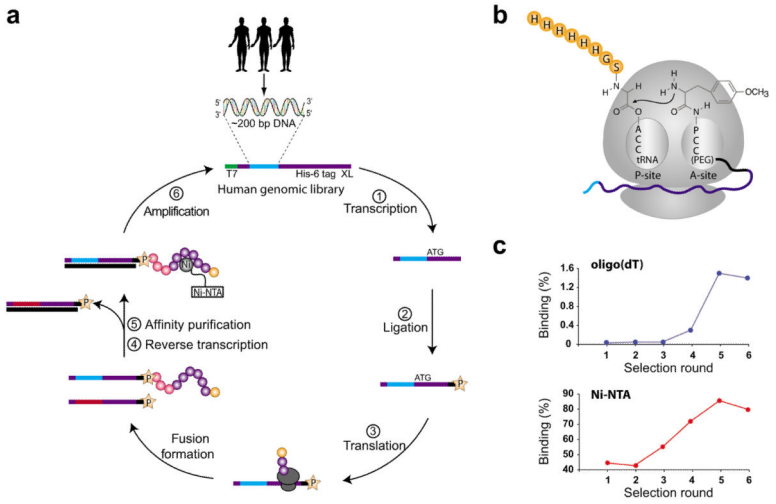

Figure 1. ***In vitro*** **selection of RNA elements that mediate cap-independent (CI) translation**

(**a**) A library of human genomic DNA fragments was inserted into a DNA cassette containing all of the sequence information necessary for mRNA display. For each selection round, the dsDNA pool was *in vitro* transcribed into ssRNA, conjugated to a DNA-puromycin linker, and translated *in vitro*. Uncapped mRNA sequences that initiate translation of an intact ORF become covalently linked to a His-6 protein affinity tag encoded in the RNA message. Functional molecules are recovered, reverse transcribed, and amplified by PCR to generate the DNA for the next selection cycle. (**b**) RNA-protein fusion molecules are generated via the natural peptidyl transferase activity of the ribosome. (**c**) Selection progress was monitored by measuring the fraction of S^{35}-labeled fusion molecules recovered from the oligo-dT and Ni-NTA affinity columns.

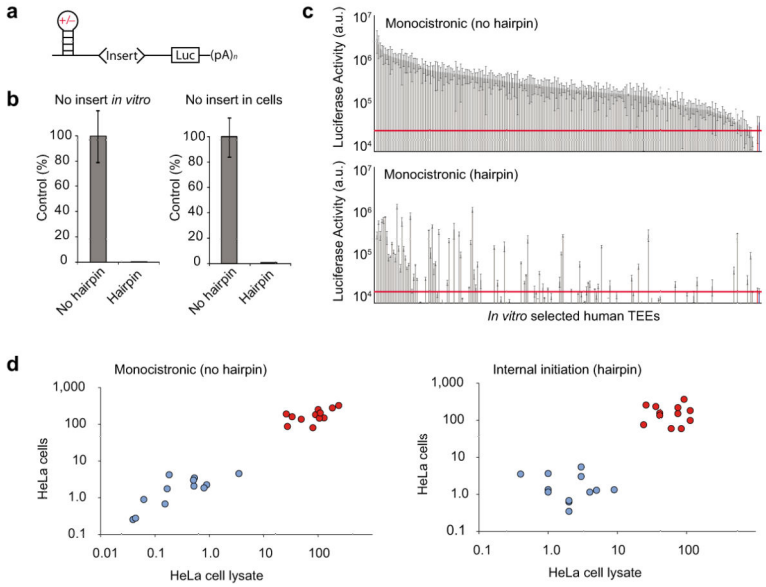

Figure 2. Functional analysis of selected TEEs in human cells and *in vitro*

(**a**) Firefly luciferase reporter with or without a stable stem-loop structure in the 5' UTR. (**b**) Translation efficiency of a no-insert control in the absence and presence of the stem loop structure assayed in HeLa cell lysate and in HeLa cells. (**c**) Translation enhancing activity of 225 representative sequences after six rounds of *in vitro* selection. Sequences were assayed in the absence (top) or presence (bottom) of the stem-loop structure in HeLa cells. Results were compared to an unstructured 13-nucleotide insert (red bar), which defined the basal level of bioluminescence activity for the reporter plasmid. (**d**) A set of twelve high activity sequences (red) were compared to an equal number of unselected sequences from the starting library (blue) in the absence (top) and presence (bottom) of the stem-loop structure in HeLa cells and in HeLa cell lysate. Fold enhancement of translation was measured relative to a no insert reporter containing 13-nt unstructured sequence in place of the TEE. Luciferase values were normalized to luciferase mRNA levels for cell-based experiments in (b) and (d), however not in (c).

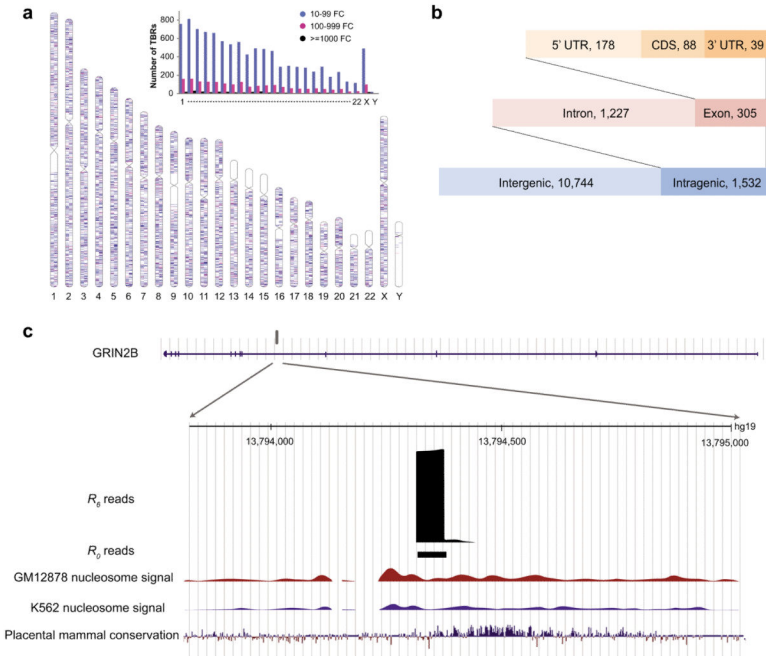

Figure 3. Genomic landscape of human TEEs

(**a**) Chromosomal ideogram of TBRs with different levels of fold-change (FC) between the starting (R_0) and the selected library (R_6): Low (10–99 FC; blue), Medium (100–999 FC; green), and High (>=1,000 FC; red). The blank regions in the chromosome correspond to the unsequenced regions in the reference genome (hg19). The total number of TBRs per chromosome is displayed in the inset and is sorted by enrichment levels. (**b**) Genetic distribution of TBRs revealed under-representation in intragenic and exonic regions (binomial test, both $p < 10^{-16}$), and over-representation in 5'-UTRs (binomial test, $p < 10^{-16}$). (**c**) Genomic context of a TBR residing in an intron of the *GRIN2B* gene. This TBR was enriched by over 1,000 times. It overlaps with active nucleosome binding sites in the ENCODE cell lines GM12878 and K562, and is upstream of a highly conserved region among placental mammals. Population polymophisms were found upstream, but not within or downstream, of this TBR.