Supplementary Material

Massively parallel functional annotation of 3' untranslated regions

Wenxue Zhao, Joshua L. Pollack, Denitza P. Blagev, Noah Zaitlen, Michael T. McManus, David J. Erle

Supplementary Figures 1-7.
Supplementary Tables 1-4.
Supplementary Notes 1 and 2.
List of Supplementary Data 1-7.
Supplementary Figure 1. The CXCL2 and CXCL3 3' UTRs destabilize reporter transgene mRNAs. Cells were transduced with BTV reporter with (a) no 3' UTR test sequence (BTV) or with BTV containing (b) full-length CXCL2 or (c) full-length CXCL3 3' UTRs and harvested at indicated times after addition of doxycycline. GFP mRNA remaining at each time point was quantified with real-time PCR. Values are mean ± standard error from triplicate measurements normalized to the amount of the same mRNA present before doxycycline addition (time 0). Half-life ($t_{1/2}$) was calculated by fitting an exponential decay curve.
Supplementary Figure 2. Flow cytometric analysis demonstrates effects of 3' UTRs on reporter transgene expression. Cells were transduced with (a) BTV lentivirus lacking the GFP transgene, (b) BTV with GFP but no 3' UTR test sequence, or BTV with 3' UTR sequences from (c) the full-length CXCL2 3' UTR, (d) a proximal segment of the CXCL2 3' UTR containing an ARE (ARE1), (e) the full-length CXCL3 3' UTR, (f) an ARE from GM-CSF (CSF2), (g) a synthetic miR-21 target, or (h) a scrambled miR-21 target. Flow cytometry was used to measure reporter GFP and reference LNGFR transgene products. Cells with high LNGFR expression (to the right of the vertical dotted line) were used to calculate expression levels. Expression levels (shown as percentages) were quantified by measuring GFP/LNGFR ratios and normalized relative to cells transduced with BTV lacking GFP (defined as 0%) and BTV lacking a test 3' UTR sequence (defined as 100%).
Supplementary Figure 3. Analysis of conserved cis-regulatory elements in CXCL2 and CXCL3 3’ UTRs. (a) Conserved 3’ UTR cis-regulatory elements in CXCL2 and CXCL3. Sequences of four highly active cis-regulatory regions in CXCL2 aligned with the homologous regions of CXCL3. Values for median ΔmRNA represent the largest effects of mutations within 8-nt windows within each of these regions. *P values were calculated using the Wilcoxon rank sum test. (b) Elements identified by fast-UTR make non-redundant contributions to the activity of the complete CXCL3 3’ UTR. Flow cytometry was used to measure effects of wild type (WT) and mutant CXCL3 3’ UTRs on protein production. Mutants had deletions (Δ) of a control sequence (C) found to have minimal activity by fast-UTR or one of the 4 highly active sequences shown in panel a. Values are means of triplicate assays and error bars represent SEM; *, p < 0.05 versus both WT and ΔC control by Tukey’s HSD test.
Supplementary Figure 4. Flow cytometric sorting enriched for 3' UTR sequences that alter protein production. Reporter protein levels before and after flow cytometric sorting. BTV: empty vector (no 3' UTR test sequence, defined as 100%), UNS: unsorted, HI: high sort gate, MID: intermediate sort gate, LO: low sort gate. Values represent mean ± standard error from triplicate flow cytometry analyses.
Supplementary Figure 5. Motifs enriched in segments with differential activity in 3 cell types. The effects of conserved 3' UTR segments on steady state mRNA level in three cell types were determined by fast-UTR (see Fig. 2h-j). Segments that had differential effects on mRNA levels in pairs of cells (e.g., relatively low in BEAS-2B versus Jurkat) were used for discriminative motif discovery (DREME software, see Online Methods). P-values and E-values (used to adjust for multiple comparisons) were computed by DREME. Cumulative distribution plots illustrate steady state mRNA levels for segments with or without motifs identified using DREME in the two cell types of interest. For example, panel a shows that in BEAS-2B cells segments with the DUAUUUAW motif are associated with lower levels of mRNA than segments lacking this motif, whereas the opposite is true in Jurkat cells.
Supplementary Figure 6. Mutation of AU-rich sequences increases mRNA stability. We measured mutation effects for all 8-nt long intervals within the conserved segments. (a) Mutations in intervals containing only A and/or U (AU content = 8) were more likely to stabilize mRNA ($p < 0.0001$ versus all other AU content groups by Dunn multiple comparisons procedure). (b) This difference is attributable to intervals that contain or overlap the core AU-rich element motif AUUUA.
Supplementary Figure 7. Fast-UTR analysis of Pumilio motifs. (a) An active sequence containing a UGUACAG motif. This motif is related to the previously described human Pumilio motif and matches a motif recognized by Drosophila Pumilio (see text). (b) An active sequence containing a previously identified motif bound by human Pumilio 1 and 2 (hPUM). (c) Another hPUM motif with no detectable activity. Black bars depict positions of the Pumilio motifs and sequences matching the motifs are shown in parenthesis. ΔStability values represent median differences between mutant and wild type clones for 8-mer sliding windows. P values were calculated by comparing clones with mutations in Pumilio motifs to clones without Pumilio motif mutations using the Wilcoxon rank sum test. Horizontal axis values represent nucleotide positions within each 160 nt segment. Active elements shown here and all other elements identified by fast-UTR are listed in Supplementary Data 6, which includes the numbers of mutant and wild type clones used to produce statistics for each active element. Statistics for the inactive predicted hPUM site in AKAP2 were computed using data from 58 clones with mutations and 470 wild type clones.
Supplementary Table 1. Functional effects of proximal CXCL2 3' UTR SNPs found in human populations

| Genome position | dbSNP ID     | CXCL2 3' UTR position | substitution | ΔmRNA $^1$ | $P^2$    |
|-----------------|--------------|------------------------|--------------|------------|----------|
| chr4:74963329   | rs199500030  | 141                    | C>T          | -16%       | 1.0 x 10^-4 |
| chr4:74963327   | rs201539067  | 143                    | A>C          | 20%        | 8.1 x 10^-4 |
| chr4:74963311   | rs200614340  | 159                    | A>C          | 74%        | 3.6 x 10^-10|
| chr4:74963311   | rs200614340  | 159                    | A>G          | 52%        | 1.4 x 10^-10|
| chr4:74963307   | rs189458219  | 163                    | A>G          | 6%         | ns        |
| chr4:74963303   | rs200621501  | 167                    | G>A          | 5%         | ns        |
| chr4:74963299   | rs200223707  | 171                    | G>A          | 9%         | ns        |
| chr4:74963299   | rs200223707  | 171                    | G>C          | 5%         | ns        |
| chr4:74963288   | rs201877060  | 182                    | T>G          | -10%       | ns        |
| chr4:74963278   | rs200919020  | 192                    | A>G          | 1%         | ns        |

$^1$ Steady state mRNA levels for 3' UTR sequences with SNPs compared to the reference 3' UTR sequence as determined by fast-UTR (see Fig. 1b).

$^2$ $P$ values for comparisons between 3' UTR sequences with SNPs and the reference 3' UTR sequence (Wilcoxon rank sum test).
**Supplementary Table 2. Mutation-sensitive elements affecting mRNA stability**

| Destabilizing elements¹ | Number of Elements |
|-------------------------|--------------------|
| Element with known destabilizing motifs² | 47 |
| miRNA targets (TargetScan) | 31 |
| Motifs recognized by destabilizing mRNA binding proteins | 22 |
| ARE motifs (AUUUA and/or ELAVL1 motifs) | 10 |
| CDE stem-loop motifs | 3 |
| PUF motifs | 9 |
| Human PUM1/PUM2 motifs (UGUANUA) | 4 |
| Drosophila Pum motifs (e.g., UGUACAG) | 5 |
| Elements with other RNA binding protein recognition motifs³ | 9 |
| No miRNA target or RNA binding protein recognition motif | 50 |

| Stabilizing elements⁴ | 44 |
|-----------------------|----|
| Element with known stabilizing motifs | 4 |
| HNRNP motifs (including the CU-rich element motif) | 3 |
| BRUNOL4/BRUNOL5 motif | 1 |
| Elements with other RNA binding protein recognition motifs³ | 7 |
| No known RNA binding protein recognition motif | 33 |

¹ Mutations significantly increased mRNA stability (FDR < 5%): \( \Delta \text{mRNA stability} = 48\% \ (29\% \text{ to } 73\%) \ [\text{median} \ (\text{interquartile range})] \).

² Includes 6 elements that overlap with both an miRNA target and an RNA binding protein motif.

³ From a set of 297 motifs associated with RNA binding proteins from *Homo sapiens* and other species and included in the CISBP-RNA database.

⁴ Mutations significantly decreased mRNA stability (FDR < 5%); \( \Delta \text{mRNA stability} = -29\% \ (-22\% \text{ to } -37\%) \ [\text{median} \ (\text{interquartile range})] \).
Supplementary Table 3. Correlation of fast-UTR and endogenous mRNA stability

| Comparison                      | Correlation (ρ) | P value | Number of genes |
|--------------------------------|-----------------|---------|-----------------|
| Fast-UTR–Yang et al.            | 0.09            | 0.01    | 665             |
| Fast-UTR–Goodarzi et al.        | 0.09            | 0.0004  | 1394            |
| Yang et al.–Goodarzi et al.     | 0.05            | 0.0005  | 4055            |

1 Pairwise comparisons involving three datasets: 1) fast-UTR stability measures in BEAS-2B cells (this study), 2) endogenous mRNA stability in HepG2 hepatocellular carcinoma (Yang et al.), and 3) endogenous mRNA stability in MDA-MB-231 breast cancer cells (Goodarzi et al.) See Online Methods for details.

2 Number of genes with stability measurements in both datasets.
**Supplementary Table 4. Oligonucleotides**

**PCR primers for full-length chemokine 3’ UTR sequences**

| Chemokine | Forward primer (5’→3’) | Reverse primer (5’→3’) |
|-----------|-------------------------|-----------------------|
| CXCL2 full length | CCGAAGAAGAGAAGAAGA | TCTCCTGGTTTTTGAGAAAGAT |
| CXCL3 full length | CAGCCTATCTTGACACTTCC | TATCTATACCTTGGTCAAAATTTAAC |

**Oligonucleotides annealed to create short 3’ segments for testing individually**

| Segment | RNA strand (5’→3’) | Reverse complement of RNA strand (5’→3’) |
|---------|-------------------|----------------------------------------|
| miR-21 sensor | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |
| Scrambled sensor | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |
| A26C1B | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |
| A26C1B mutant | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |
| C3orf58 | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |
| C3orf58 mutant | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |
| MPP6 | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |
| MPP6 mutant | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |
| SETD1B | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |
| SETD1B mutant | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |
| SMAD7 | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |
| SMAD7 mutant | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |
| SMAP1 | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |
| SMAP1 mutant | CCGGTTCCACATCGTCGAGTTACGTTAAGTTCCTGTTCACCGGACTGCCGAGCGAAGCGGCTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | TATATCAGTTATCATCAGGATGAGCTTCGCCTGCAGGGTCACG |

**PCR primers for fast-UTR library construction and analysis**

**Proximal CXCL2 SNP library**

| Oligonucleotide pool amplification1 | Forward primer (5’→3’) | Reverse primer (5’→3’) |
|-----------------------------------|-------------------------|-----------------------|
| CTTGACCTCTCGAGAAGCTGACG | CGATACGCGATGTTTGGAG | AGCTGACTGGAGTTCAGACG |

**RNA (cDNA), DNA amplification**

| Forward primer (5’→3’) | Reverse primer (5’→3’) |
|-------------------------|-----------------------|
| CTTGACCTCTCGAGAAGCTGACG | AGCTGACTGGAGTTCAGACG |

**Full-length CXCL2 & CXCL3 library**

| Oligonucleotide pool amplification1 | Forward primer (5’→3’) | Reverse primer (5’→3’) |
|-----------------------------------|-------------------------|-----------------------|
| GCTCCTACGGCGGTGTGGAG | GTGAGCTTTCGCGAGGTTG |

**RNA (cDNA), DNA amplification**

| Forward primer (5’→3’) | Reverse primer (5’→3’) |
|-------------------------|-----------------------|
| GCTCCTACGGCGGTGTGGAG | GTGAGCTTTCGCGAGGTTG |

**Conserved library (high fidelity)**

| Oligonucleotide pool amplification1 | Forward primer (5’→3’) | Reverse primer (5’→3’) |
|-----------------------------------|-------------------------|-----------------------|
| ACAATCGGATGGTTTTCGACG | CCGGTGCAAACTGTTGCTGCCC |

**RNA (cDNA), DNA amplification**

| Forward primer (5’→3’) | Reverse primer (5’→3’) |
|-------------------------|-----------------------|
| ACAATCGGATGGTTTTCGACG | CCGGTGCAAACTGTTGCTGCCC |

**Conserved library (error-prone PCR)**

| Oligonucleotide pool amplification1 | Forward primer (5’→3’) | Reverse primer (5’→3’) |
|-----------------------------------|-------------------------|-----------------------|
| ACAATCGGATGGTTTTCGACG | CCGGTGCAAACTGTTGCTGCCC |

**RNA (cDNA), DNA amplification**

| Forward primer (5’→3’) | Reverse primer (5’→3’) |
|-------------------------|-----------------------|
| ACAATCGGATGGTTTTCGACG | CCGGTGCAAACTGTTGCTGCCC |

---

1 The underlined sequences match primer recognition sequences at the 5’ and 3’ ends of the 200-mer oligonucleotides in each pool.

2 We used a set of oligonucleotides with different 6 or 7 nt sample indexes (XXXXXX or XXXXXXX). We amplified each mRNA (cDNA) or genomic DNA sample using a primer with a different index, combined the PCR products for multiplexed sequencing, and used sequencing read data to de-multiplex.

3 During PCR primer synthesis, we incorporated a mixture of nucleotides at 8 consecutive positions (NNNNNNNN) to generate random clone indexes.
Supplementary Note 1

We used two fast-UTR libraries to analyze cis-regulatory effects of the conserved 3' UTR segments. We generated both fast-UTR libraries from the same pool of 2828 oligonucleotides containing a total of 3000 conserved 3' UTR elements (see Online Methods). We generated the first conserved element library using high fidelity, low bias PCR. We sequenced reporter DNA and RNA from cells transduced with this library and identified a total of 151,093 clones representing 98.3% (2781/2828) of the designed 3' UTR segments. We used this library for analyzing steady state mRNA levels and for analyzing protein production. We generated the second conserved element library using error-prone PCR in order to generate a large number of mutations. This library contained a total of 392,252 clones representing 91.2% (2580/2828) of the designed 3' UTR segments. These clones contained a total of 924,201 mutations (mean 1.7 substitutions and 0.7 deleted nt per clone). We used this library to analyze mRNA stability and identify mutation-sensitive elements that altered stability.

We estimate that the probability of incorrectly assigning a sequencing read to the wrong clone is small. Reads were first mapped to a given segment by analyzing the portion of the paired end reads corresponding to the 160 nt 3' UTR segment and then assigned to a clone by analyzing the 8 nt index sequence. Based upon reported error rates for the polymerase we used to amplify genomic DNA or cDNA\(^1\) (KAPA HiFi, 2.8 \times 10^{-7} \text{ errors/nt}) and Illumina HiSeq sequencing\(^2\) (~2 \times 10^{-3} \text{ errors/nt}), the probability of a single error in the 8 nt index sequence is 2% or less. Furthermore, most index read errors were not included in downstream analysis, since clone indexes found in fewer than 10 reads were filtered out. Only errors that produced an index read that matched a different clone from the same segment were likely to be included in our measurements. There are \(8 \times 3 = 24\) index sequences that differ by a single nt from any given index; these represent \(24/(4^8)\) or 0.04% of all possible 8-nt indexes. For a segment with an average number of clones (152 clones/segment for the conserved 3' UTR library used for analyzing mutation effects), the probability that a single nucleotide error would yield a read that matched another clone can be estimated as \(151 \times 0.04\%\), or \(~6\%\). Therefore we estimate the overall probability of incorrectly assigning a read to the wrong clone is \(~2\% \text{ (sequencing error rate) } \times 6\% \text{ (probability that an error will result in incorrect}\)
assignment), or ~0.1%. This probability increases for segments represented by more clones, but we expect that any increase in errors due to incorrect clone assignment would be more than offset by power gained from analyzing the additional clones.

Supplementary Note 2
We examined correlations between endogenous mRNA stability and the effects of 3' UTR segments from the same mRNAs on stability as measured by fast-UTR. We used published data from microarray-based analyses of endogenous mRNA stability in HepG2 hepatocellular carcinoma cells (the "Yang et al." dataset) and MDA-MB-231 breast cancer cells (the "Goodarzi et al." dataset) and compared these with fast-UTR stability measurements made in BEAS-2B cells (see Online Methods). For each of the pairwise comparisons between fast-UTR and endogenous mRNAs, we found statistically significant positive correlations (Supplementary Table 3). These results support the relevance of fast-UTR results to endogenous mRNA stability. The correlations are weak, suggesting that activities of 3' UTR segments measured by fast-UTR explain only a portion of the half-life of an endogenous transcript (measured in different cells). This is expected, since fast-UTR measurements of segments of 3' UTRs cannot account for the activity of the remainder of the 3' UTR and other regions of the endogenous mRNA, and since the activity of 3' UTR active elements and other determinants of mRNA stability vary depending upon cellular context. The strikingly low correlation between the two different endogenous mRNA datasets can be attributed to differences in cellular context and to technical differences between those two studies.

Supplementary Note References
1. Fazekas, A., Steeves, R. & Newmaster, S. Improving sequencing quality from PCR products containing long mononucleotide repeats. BioTechniques 48, 277-285 (2010).
2. Minoche, A.E., Dohm, J.C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biology 12, R112 (2011).
Supplementary Data 1. CXCL2 fast-UTR genome browser tracks
The accompanying file contains fast-UTR data for the experiments shown in Fig. 1b and c. Data are presented in Wiggle format and the file can be loaded directly into genome browsers to generate tracks for visualizing fast-UTR data. There are eight tracks representing the data in Fig. 1b (CXCL2 proximal 3' UTR SNP analysis). For each possible nucleotide substitution (i.e., change of the RNA sequence to A, C, G, or U), one track reports the difference in steady-state mRNA levels between CXCL2 segments with variant and reference sequence and a second track reports the $-\log_{10} p$-value for this difference. Two tracks represent the data in Fig. 1c (analysis of effects of spontaneous mutations in segments spanning the full-length CXCL2 3' UTR). One track reports the effect of mutations in 8 nt regions (centered at the indicated position) on steady-state mRNA levels and the other track reports the $-\log_{10} p$-value for this difference. Since CXCL2 is encoded by the negative strand of chromosome 4, fast-UTR results are more easily visualized by reversing the genome browser display. All coordinates are based on the GRCh37/hg19 human genome assembly.

Supplementary Data 2. Conserved 3' UTR segment sequences
See accompanying file.

Supplementary Data 3. 3' UTR effects on mRNA levels and stability
See accompanying file.

Supplementary Data 4. 3' UTR sequences enriched in sorted cells with high or low reporter protein levels
See accompanying file.

Supplementary Data 5. miRNA profiling of BEAS-2B cells
See accompanying file.

Supplementary Data 6. Cis-regulatory elements identified using fast-UTR
See accompanying file. If Supplementary Data 7 has been uploaded to the UCSC Genome Browser (http://genome.ucsc.edu/), links in Supplementary Data 6 will allow visualization of fast-UTR data in the browser.

Supplementary Data 7. Conserved 3' UTR segment fast-UTR genome browser tracks
The accompanying file contains fast-UTR data for the elements listed in Supplementary Data 6. Data are presented in Wiggle format and the file can be loaded directly into genome browsers to generate tracks for visualizing fast-UTR data. One track reports the effect of mutations in 8 nt regions (centered at the indicated position) on RNA stability and the other track reports the $-\log_{10} p$-value for this difference. If the Supplementary Data 7 file has been loaded into the UCSC Genome Browser (http://genome.ucsc.edu/), the links in Supplementary Data 6 can be used to navigate to 3' UTR segments containing elements identified by fast-UTR. All coordinates are based on the GRCh37/hg19 human genome assembly.