LongHorn application for transcription (TF)-target, miRNA-target, and RNA-binding proteins (RBP)-target interacitons

TF-target predictions. We collected a total of 6,566 non-redundant and experimentally-verified human TF-target interactions for 557 TFs and 2528 targets from 3 sources; of these 388 have characterized motifs. Interactions were collected from the following sources:

- HTRIdb (Bovolenta et al., 2012) build dating 03/20/2014: 2209 interactions involving 277 TFs and 1381 targets that were verified by small and mid-scale techniques. These excluded interactions detected by ChIP-chip or ChIP-seq due to their lower confidence.
- Table 3 of Whitfield et al. (Wang et al., 2012; Whitfield et al., 2012) which included 63 interactions between 7 TFs and 54 target genes.
- TRANSFAC Professional,(Matys et al., 2006) from February 2013, 4,888 interactions between 501 TFs and 1669 targets. We excluded interactions involving more than one TF per target to avoid non-specific binding by co-factors.

We used ENCODE (Encode, 2012) data to predict TF targets based on ChIP-Seq, including 108 TFs that were profiled in 37 cell lines, with the majority of assays performed in replicates. ChIP-seq data were downloaded from the UCSC genome browser, using hg19 annotation. Transcription factor binding sites in proximal promoters were selected as sequence-based targets and used in the subsequent expression-based analysis.

In total, we collected 1634 position weight matrices (PWMs) for 642 human TFs with expression in TCGA RNASeqV2 from 5 sources. To avoid matrix entries of value 0, a pseudo count 1 was added to each entry before calculating the relative occurrence frequencies (%) of nucleotides at each position. We used this frequency table to scan TF binding sites from the proximal promoters and lncRNA transcript sequences. Sources include the following:

- JASPAR (Sandelin et al., 2004) version: 5.0_ALPHA: 104 PWMs for 100 TFs.
- SwissRegulon (Pachkov et al., 2007) downloaded on 03/18/2014: 353 PWMs for 340 TFs.
- HumanTF (Jolma et al., 2013), downloaded from Table S3 in their paper: 661 PWMs for 365 TFs. Only higher-confidence motifs were included (motifs indicated in orange or green were not included).
- HOCOMOCO (Kulakovskiy et al., 2013) version: 9.0: 430 PWMs for 402 TFs. Only motifs of quality A, B, C, or D were extracted.
- Factorbook (Wang et al., 2012), downloaded from Table S2 in their paper: 86 PWMs for 76 TFs. These excluded unannotated motifs in their publication.

PWMs were used to predict TFBS in proximal promoters, 5'-flanking regions, and lncRNA transcripts. TF-target predictions were based on combining evidence from verified interactions, ChIP-Seq assays, sequence-based motif analysis and co-expression networks (Lefebvre et al., 2010; Margolin et al., 2009; Pankowicz et al., 2016; Smith et al., 2005a; Smith et al., 2006; Smith et al., 2005b; Zhou et al., 2010). Predictions are given in Supplementary Table S5. We used ENCODE ChIP-Seq data sets to select candidate TF-target interactions based on significant peaks (Q value<1E-10) in proximal promoters of coding genes. In addition, 1634 PWMs for 642 TFs were used to infer binding sites on proximal promoters and their corresponding 5'-flanking regions. Only significant binding sites ($p < 1E-5$, when compared to flanking regions) were included. As a TF could have multiple binding sites, with different binding strength on multiple promoters for the same gene, we integrated binding strength and relative position to the TSS of all sites for the same TF-promoter pair into a single weighted score S using the following formula to estimate the binding likelihood for this TF-promoter pair. Here, d_i is the distance between the TSS and the mid-point of binding site, L is the length of promoter, i.e., 2000 bps, M is the total number of binding sites associated with this TF-promoter pair, P_{min} is the minimal attainable p-value genome-wide, and P_i is the binding significance for site i. (Sikora-Wohlfeld et al., 2013).

$$S = \sum_{i=1}^{M} \frac{-log10(P_i)}{-log10(P_{min})} * \{1 - [d_i/(L/2)]\}$$

If an expressed TF-target pair of was either (1) experimentally verified, (2) had significant (q < 1E-10) evidence for binding on the promoter of any transcript from at least one ENCODE ChIP-Seq data, or (3) had a nonzero S score, as predicted by at least one PWM on either forward or reverse strand of the promoter of any transcript, we tested its significance of correlation using dCor (described in the following section) to reverse-engineer tumor type-specific TF-target interactomes. Spearman's correlation was also calculated to determine the sign of the correlation, which indicates whether the TF is activating or repressing the target. TF-target pairs with evidence for sequence binding and significant expression measured by dCor (p < 1E-3) were included in the transcriptional interactome (Zhou et al., 2010).

MiRNA-target predictions. Verfieid miRNA-target interactions were compiled from miRecords, TarBase, TRANSFAC, and miRTarBase (v4.5 in 11/01/2013). Only human miRNA-target gene interactions with strong experimental evidence, i.e., reporter assay or western blot, were selected. In addition, we included validated targets from the Table S2 of Grosswendt et al. (Grosswendt et al., 2014), which included interactions between 359 miRNAs and 2463 genes. In total, these 4,696 interactions were used to train classifiers and predict miRNA-target interactions genome wide. All miRNA targets—in 3' UTRs and lncRNAs—were inferred using Cupid (step 2 and without computing step 3) with standard parameters (Chiu et al., 2015). All predicted interactions are included in the Supplementary Table S7. Note that both RBP-targets and miRNA-targets form the post-transcriptional network.

RBP-target predcitions. We used ENCODE (Encode, 2012) data to predict RBP targets based on eCLIP-profiled targets for 96 RBPs, with each assay performed in duplicates. RBP-binding sites on 3'-UTRs of protein-coding genes and lncRNA transcripts were inferred based on ENCODE eCLIP data sets exclusively, and using a q < 1E-10 cutoff. If multiple peaks are mapped to the same 3' UTR/lncRNA transcript, the best q value is assigned to determine the strength of association. Similar to TF-target prediction, we required p<1E-3 for the significance of dCor between RBP and target. We predicted RBP-targets via integrating both sequence binding and co-expression evidence.

lncRNA-target predcitions. LongHorn predicts modulation of TFs, RBPs and miRNAs (aggregately referred to as effectors) by lncRNAs. We model lncRNAs as Decoys, Co-factors, Guides and Switches. In all cases, lncRNAs are described as modulators and affect the activity of effectors. Predictions rely on building blocks that include inferred transcriptional and post-transcriptional networks, predicted lncRNAs binding and delta distance correlation. However, unlike other 3 types of predictions, the methodology for predicting switch lncRNAs did not require evidence for direct binding. For detail, please see LoongHorn manuscript, which was submitted as a part of the revised deliverables.

Cross-species conservation estimates by phastCons (Siepel et al., 2005) was used for predicting TF and miRNA binding sites. Both complete hg19 human genome and genome-wide phastCons46way conservation scores for vertebrate were downloaded from UCSC Genome Browser annotation. All scores were normalized between 0 and 1.
LINCS assays were used to verify predictions by Cupid and LongHorn (Chiu et al., BMC Genomics, 2017).