Supplemental Material for Systematic Interrogation of Human Promoters by Weingarten-Gabbay et al.

Index

Content of supplemental tables

Supplemental Table S1. Characterized promoters from the literature and reported TSSs

Supplemental Table S2. Expression measurements of oligos from full-length promoters.

Supplemental Table S3. Expression measurements of 508 Pre-Initiation Complex (PIC) binding sequences from the human genome.

Supplemental Table S4. Core promoters with bidirectional activity.

Supplemental Table S5. Tested core promoter elements.

Supplemental Table S6. Tested core promoter backgrounds.

Supplemental Table S7. Expression measurements of synthetic configurations of core promoter elements.

Supplemental Table S8. Expression measurements of 1875 native core promoters from the human genome.

Supplemental Table S9. Expression measurements of various distances between TATA and Inr elements.

Supplemental Table S10. Expression measurements of 133 TF binding-sites.

Supplemental Table S11. Expression measurements of multiple homotypic sites for four TFs.

Additional methods

Synthetic library design

Dissection of reported promoters

The full sequences of 11 characterized promoters (*AdML*, *EF1alpha*, *UBC*, *HIV*, *MMLV LTR*, *hFerL*, *HBV*, *mPGK1*, *hPGK1*, *CMV* and *SV40*) and the reported TSSs were collected from the literature (Van Beveren et al. 1982; Hansen and Sharp 1983; Adachi et al. 1986; Hennighausen and Fleckenstein 1986; Adra et al. 1987; Eisenstein and Munro 1990; Moriyama et al. 1994; Nenoi et al. 1996; Lay et al. 2000; Wang et al. 2008). Detailed information on the sequences, TSS positions and references appears in Supplemental Table S1. Promoters sequences were dissected into fragments of 153nt with overlap of 103nt between oligos (i.e., 50nt distance between the start positions of two sequential oligos).

Pre-initiation complex (PIC) binding sequences

PIC binding sequences were analyzed from ChIP-exo measurements in K562 cells (Supplemental Tables S1 and S2 from Venters *et al*. (Venters and Pugh 2013)). Coordinates of hg18 were converted to hg19 using USCS "Lift Genome Annotations" tool. Binding sequences were selected to represent different chromatin states (promoters, enhancers and heterochromatin) and different binding intensities according to ChIP-exo measurements. For each of the 508 binding sequence two oligos were designed from -103 to +50 relative to the binding sequences midpoint on either the plus or minus strand of the DNA.

Endogenous core promoters from the human genome

We selected 1875 native core promoters from the human genome to represent a wide range of promoters:

(a) Promoters that span various endogenous expression levels – we analyzed CAGE-seq measurements from the ENCODE Consortium(The ENCODE Project Consortium 2012) in K562, the cell line in which we constructed the library, to represent promoters with expression levels between 0 RPM (no expression in K562 cells) to 10,000 RPM (high expression levels).
(b) Constitutive and induced promoters – in addition to K562, we also analyzed CAGE-seq measurements in 11 additional cell lines: A459, HepG2, HeLa S3, MCF-7, H1-hESC, GM12878, HUVEC, SK-N-SH, BJ, AG04450 and NHEK. We included in the design promoters that were expressed in at least 10 of the 12 cell lines ("constitutive promoters") and promoters that were expressed in less than 6 cell lines ("induced promoters").
(c) Promoters with long-range physical interactions – we analyzed chromosome conformation capture carbon copy (5C) measurements in K562 cells(Sanyal et al. 2012) to include TSS for which interaction with distal DNA was observed.
(d) TATA-containing promoters - our design also includes native promoters with a perfect match to either 'TATAAA' or 'TATATA'.

Promoters backgrounds

(a) Background sequences for core promoter elements: Five core promoter were selected to represent human and viral sequences, with and without endogenous TATA box: *ACTB*, *CMV*, *HIV*, *RPLP0* and *RTRAF*. For each promoter a sequence in the length 153nt was extracted from positions -103 to +50 relative to the TSS and existing core promoter elements

were mutated. A description of the wildtype sequences, the mutated elements and the resulting backgrounds sequences appears in Supplemental Table S6.

(b) Background sequences for TF binding-sites: Two promoters that contain endogenous TATA were selected: the human *ACTB* and the *CMV* promoters. For each promoter a sequence in the length 153nt was extracted from positions -148 to +5 relative to the TSS. Existing TF binding-sites were mutated as detailed in Supplemental Table S6. Core promoter elements were maintained in the original sequences and served as a "minimal core promoter".

Core promoter elements

The sequence of the TATA (TATATAAG), Inr (TCAGATC), MTE (CGAGCCGAGC) and DPE (GGTTGT) were extracted from the rationally designed "super core promoter", which yields high expression levels in human cells(Juven-Gershon et al. 2006). The sequences of the BREu (CGGCGCC) and BREd (GTGTGGG) elements were derived from the reported consensuses sequences (Lagrange et al. 1998; Deng and Roberts 2005). A detailed description of the sequences used and references appears in Supplemental Table S5.

TF binding-sites

(a) TFs activity screen: TF bindings-sites were analyzed from a high-throughput SELEX experiment performed on a large collection of human TFs(Jolma et al. 2013). We extracted the binding sequences of the TFs that are abundant in the proximal-promoter region SP1, ETS1, YY1, CREB, USF1, NRF1 and SRF, as well as TFs that were part of the 119 factors tested during the ENCODE Project(Wang et al. 2012).

(b) TF binding-sites multiplicity: We selected four TF binding-site with different representation in homotypic clusters in the human genome according to bioinformatic analysis (Supplemental Table S2 from Gotea *et al*. (Gotea et al. 2010)) – SP1 (CCCCGCCCC), ETS1 (ACCGGAAGT), YY1 (GCCGCCATTTTG) and CREB (TGACGTCA). Each motif was placed in 1-7 copies in the following predefined positions on the oligo sequence: 15, 30, 45, 60, 75, 90, and 105 on the oligo's sequence. To control for the effects of position within the oligo, distance between elements and flanking sequences, for each number of sites oligos with all possible positions were designed. For example: for two sites oligos were designed with the motif's sequences in positions 15,30; 15,45; 15,60; 15,75; 15, 90; 15,105; 30,45; 30,60 etc.

Synthetic library production and amplification

The production and amplification steps were carried out essentially as in (Sharon et al. 2012). Agilent Oligo Library Synthesis (OLS) technology was used to produce a pool of 55,000 different fully designed single-stranded 200-mers, a subset of 15,753 of which comprised the synthetic library presented in this study. Each designed oligo contains subset specific priming sites, leaving 164nt for the variable region. The library was synthesized using Agilent's on-array synthesis technology (Cleary et al. 2004; LeProust et al. 2010) and then provided to us as an oligo pool in a single tube (10pmol). The pool of oligos was dissolved in 200µl TE. 5.5ng of the library (1:50 dilution) were divided into 16 tubes, and each tube was amplified using PCR. The primers used for amplification of the library included sites for the restriction enzymes AscI and RsrII, for cloning into the library master plasmid. The oligonucleotides were amplified using constant primers in the length of 51nt, which are complementary to the subset primer (underlined) and adds the restriction sites (bold) and a tail of approx. 30nt to allow identification

of products that were not properly cut by restriction enzymes in the next step. Primers sequences: upstream primer – 5' – TTGTTCCGCCGCTTCGCTGACTGTG**GGCGCGCC**CGCGTCGCCGTGAGGAGG -3', downstream primer 5' – TCAGTCGCCGCTGCCAGATCGCGGT**CGGTCCG**AGCCCCACGGAGGTGCCAC – 3'. Each PCR reaction contained 24µl of water with 0.323ng DNA, 10µl of 5× Herculase II reaction buffer, 5µl 2.5mM dNTPs each, 2.5µl 20uM Fw primer, 2.5µl 20uM Rv primer, and 1µl Herculase II Fusion DNA Polymerase (Agilent Technologies, Santa Clara, California). The parameters for PCR were 95^0C for 1 min, 14 cycles of 95^0C for 20 sec and 68^0C for 1 min, each, and finally one cycle of 68^0C for 4 min. The PCR products from all 16 tubes were joined and concentrated using Amicon Ultra, 0.5ml 30K centrifugal filters (Merck Millipore) for DNA Purification and Concentration. The concentrated DNA was then purified using a PCR MinElute purification kit (QIAGEN) according to the manufacturer's instructions.

Synthetic library cloning into the master plasmid

Library cloning into the master plasmid was adapted from a protocol that was previously described for a lenti-virus based library(Weingarten-Gabbay et al. 2016). Purified library DNA (720ng total) was cut with the unique restriction enzymes AscI and RsrII (Fermentas FastDigest) for 2 hours at 37^0C in four 40 µl reactions containing 4µl FD buffer, 1µl of AscI enzyme, 2.5µl of RsrII enzyme, 0.8µl DTT, and 18µl DNA, followed by a heat inactivation step of 20 min at 65^0C. Digested DNA was separated from smaller fragments and uncut PCR products by electrophoresis on a 2.5% agarose gel stained with GelStar (Cambrex Bio Science Rockland). Fragments in the size of 200bp were cut from the gel and eluted using electroelution Midi GeBAflex tubes (GeBA, Kfar Hanagid, Israel)). Eluted DNA was precipitated using standard NaAcetate\Isopropanol protocol. The master plasmid was cut with AscI and RsrII (Fermentas FastDigest) for 2.5 hours at 37^0C in a reaction mixture containing 9µl FD buffer, 3µl of each enzyme, 3µl Alkaline Phosphatase (Fermentas), and 4.5µg of the plasmid in a total volume of 90µl, followed by a heat inactivation step of 20 min at 65^0C. Digested DNA was purified using a PCR purification kit (QIAGEN). The digested plasmids and DNA library were ligated for 0.5 hr at room temperature in two 10µl reactions, each containing 150ng plasmid and the library in a molar ratio of 1:1, 1µl CloneDirect 10× Ligation Buffer, and 1µl CloneSmart DNA Ligase (Lucigen Corporation) followed by a heat inactivation step of 15 min at 70^0C. 14µl ligated DNA was transformed into a tube of *E. cloni*® 10G CLASSIC Electrocompetent Cells (Lucigen) divided to 7 aliquots (25µl each) which were then plated on 28 LB agar (200mg/ml amp) 15cm plates. To ensure that the ligation products only contain a single insert we performed colony PCR on 93 random colonies. The volume of each PCR reaction was 30µl; each reaction contained a random colony picked from a LB plate, 3µl of 10× DreamTaq buffer, 3µl 2mM dNTPs mix, 1.2µl 10µM 5' primer, 1.2µl 10µM 3' primer, 0.3µl DreamTaq Polymerase (Thermo scientific). The parameters for PCR were 95^0C for 5 min, 30 cycles of 95^0C for 30s, 68^0C for 30s, and 72^0C for 40s, each, and finally one cycle of 72^0C for 5 min. The primers used for colony PCR were taken from the *ACTB* promoter (5' – CTCTTCCTCAATCTCGCTCTCGCTC – 3') and the chimeric intron (5' – GACCAATAGGTGCCTATCAGAAACGC – 3'). Out of the 93 colonies evaluated, only 3 had multiple inserts. To ensure that all ~15,000 oligos are represented we collected over $2{\cdot}10^6$ colonies sixteen hours after transformation, by scraping the plates into LB medium. Library pooled plasmids were purified using a NucleoBond Xtra maxi kit (Macherey Nagel). Following the purification, the library plasmids were extracted from a 0.8% agarose gel,

in order to clean them from free library DNA that presented a toxic effect on library nucleofected cells.

***in-vitro* transcription of ZFN mRNA**

ZFN mRNA was *in-vitro* transcribed from pZFN1 and pZFN2 plasmids (Sigma) according to the manufacturer's protocol, using MessageMAX T7 ARCA-Capped Message Transcription Kit and Poly(A) Polymerase Tailing Kit (CellScript). The RNA was then purified using MEGAclear kit (Ambion), the concentration was measured and integrity and polyadenylation were verified by high-sensitivity RNA TapeStation (Agilent). Small aliquots (5-10µl) containing 600ng/µl of each of the two ZFN mRNAs were stored in -80^{0}C.

Preparing samples for sequencing

In order to maintain the complexity of the library amplified from gDNA, PCR reactions were carried out on gDNA amount calculated to contain an average of 200 copies of each oligo included in the sample. For each of the 16 bins, 20µg of gDNA were used as template in a two-step nested PCR in two tubes (to include the required amount of gDNA), each containing 100µl (in both steps); In the 1st step each reaction contained 10µg gDNA, 50µl of Kapa Hifi ready mix ×2 (KAPA biosystems), 5µl 10µM 5' primer, and 5µl 10µM 3' primer. The parameters for the first PCR were 95°C for 5 min, 18 cycles of 94°C for 30s, 65°C for 30s, and 72°C for 40s, each, and finally one cycle of 72°C for 5 min. Primers used for the first reaction were from the *ACTB* promoter (5'-CTCTTCCTCAATCTCGCTCTCGCTC-3') and the chimeric intron (5'-GACCAATAGGTGCCTATCAGAAACGC-3'). In the 2nd PCR step each reaction contained 5µl of the first PCR product (uncleaned), 50µl of Kapa Hifi ready mix ×2 (KAPA biosystems), 5µl 10µM 5' primer, and 5µl 10µM 3' primer. The PCR program was similar to the first step, using 24 cycles. For the second reaction the 5' primer was comprised of a random 5nt sequence to increase complexity, followed by an 8nt barcode (one of three for each bin, underlined) and a library specific sequence (5'-HNHNH<u>XXXXXXXX</u>CGCGTCGCCGTGAGGAGG -3'). The common 3' primer was (5'- HNHNHNHNGCCCCACGGAGGTGCCAC -3'. In both, the 'N's represent random nucleotides, and 'H' is A,C or T, in order to avoid synthesis of stretches of G that can affect initial clusters definition in NextSeq runs. The concentration of the PCR samples was measured using Quant-iT dsDNA assay kit (ThermoFisher) in a monochromator (Tecan i-control), and the samples were mixed in ratios corresponding to their ratio in the population. The library was separated from unspecific fragments by electrophoresis on a 2% agarose gel stained by EtBr, cut from the gel, and cleaned in 2 steps: gel extraction kit (QIAGEN), and SPRI beads (Agencourt AMPure XP). The sample was assessed for size and purity at the TapeStation, using high sensitivity D1K screenTape (Agilent Technologies, Santa Clara, California). 10ng library DNA were used for library preparation for NGS; specific Illumina adaptors were added, and DNA was amplified using 14 amplification cycles, protocol adapted from Blecher-Gonen et al(Blecher-Gonen et al. 2013). The sample was reanalyzed using TapeStation.

Isolated clones measurements

Thirty isolated clones, at least one from each of the 16 expression bins were grown from single cells that were sorted into 96-wells plate. The clones were chosen based on their verified emergence from a single cell. After 28 days cell populations were analyzed in Flow Cytometry for eGFP expression and genomic DNA (gDNA) was purified. DreamTaq DNA polymerase (Thermo scientific) was used to amplify the library from 200ng gDNA, with same conditions

and primers as in the library colony PCR. The PCR product was Sanger sequenced from the PCR Fw primer.

Retroviruses production and infection

Phoenix virus packaging cells were used for retroviruses production as described before(Sigal et al. 2007). 5×10^5 cells were plated on 6cm plates 24hr prior to transfection. Cells were transfected with pPRIGp mChHA, a Moloney Murine Leukemia Virus (MMLV) retroviral plasmid expressing a bicistronic transcript encoding mCherry and eGFP separated by the EMCV Internal Ribosome Entry Site (IRES)(Albagli-Curiel et al. 2007). Each transfection included: 100µl DMEM with no serum or antibiotics, 12µl of FuGENE 6 transfection reagent (Promega) and 4µg of the retroviral plasmid. Transfection was performed according to the manufacturer's instructions. After 24hr medium was replaced with fresh DMEM and H1299-EcoR cells were plated on 10cm plates for infection. After additional 24hr (48hr past transfection) 4ml of viruses-containing media were collected from Phoenix cells and centrifuged for 5 minutes in 1,500rpm. 3.5ml of viruses-containing media were added to 1.5ml RPMI media in each H1299 plate (total volume of 5ml) in addition to 5µl of Polybrene (AL-118, Sigma). After 24hr cells were washed 3 times with PBS, and fresh RPMI complete medium was added.

Computing promoter activity threshold using empty vector measurements

In order to determine the activity threshold of core promoters we constructed and measured K562 cells for which we integrated an "empty vector" plasmids containing a linker sequence of 25bp between the AscI and RsrII restriction sites. We then measured the fluorophore levels of cells expressing the empty vector in flow cytometry using the same lasers intensities and settings as in the library sorting. We computed the normal distribution of the GFP/mCherry and extracted the mean and standard deviation (std). We set a threshold of 2 stds from the mean. Oligos with expression levels above this threshold (>=1.58) were considered as positive core promoters.

Statistical analyses

To assess the difference between expression levels of two groups that are distributed normally (e.g., native core promoters with and without TATA elements) we used a two-sample *t*-test. When expression levels were not distributed normally, such as in the case of the pre-initiation complex (PIC) binding sequences, we performed non-parametric Wilcoxon rank-sum test (for two samples) or Kruskal-Wallis test (for >2 samples). To compare the expression of pairs of sequences (e.g., adding a poly(dA:dT) tract to a sequence with two TF binding-sites) we performed Wilcoxon signed rank test. To evaluate differences in the variance resulting from the ZFN and retroviruses systems two-sample *F*-test for equal variances was performed. To examine significant differences between the proportions of positive core promoters in two groups (e.g., promoters and enhancers regions) we performed a two-proportion z-test. All correlations reported in the manuscript and the corresponding p-values were computed using Pearson correlation.

Intersecting PIC binding-regions from ChIP-exo measurements with TSSs from GRO-cap measurements

We used BEDTools(Quinlan and Hall 2010; Dale et al. 2011) to intersect the genomic coordinates of the 508 PIC binding-regions that were derived from the ChIP-exo data from Venters *et al*. (Venters and Pugh 2013) (defined as -100 to +100 relative to the reported TFIIB

midpoint) with the 128,471 TSSs identified by GRO-cap measurements in K562 cells by Core *et al.*(Core et al. 2014), (Supplementary Data, file name: "tss_all_k562.bed"). To evaluate the probability of observing 160 mutual sites by chance we computed the hypergeometric distribution of genomic sites that are captured by the two assays with total sample size of 15,000,000 possible genomic regions (3×10^9 divided by 200bp). To test the robustness of our results we reduced the overall population size to 5% of the human genome (20-fold reduction).

Fitting a logistic function

To examine the relationship between the number of binding-sites and promoter activity we fitted a logistic function with three parameters: maximal expression levels (L), the steepness of the curve (k), and the number of binding-sites at the sigmoid's midpoint (X_0).

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Oligos with expression levels below the activity threshold were filtered out. To test the agreement between the data and the fitted function we computed for each binding-site in each background the correlation and p-value between the measured expression levels and the fitted values.

References

Adachi A, Gendelman HE, Koenig S, Folks T, Willey R, Rabson A, Martin MA. 1986. Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. *Journal of virology* **59**: 284-291.

Adra CN, Boer PH, McBurney MW. 1987. Cloning and expression of the mouse pgk-1 gene and the nucleotide sequence of its promoter. *Gene* **60**: 65-74.

Albagli-Curiel O, Lecluse Y, Pognonec P, Boulukos KE, Martin P. 2007. A new generation of pPRIG-based retroviral vectors. *BMC Biotechnol* **7**: 85.

Blecher-Gonen R, Barnett-Itzhaki Z, Jaitin D, Amann-Zalcenstein D, Lara-Astiaso D, Amit I. 2013. High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. *Nature protocols* **8**: 539-554.

Cleary MA, Kilian K, Wang Y, Bradshaw J, Cavet G, Ge W, Kulkarni A, Paddison PJ, Chang K, Sheth N et al. 2004. Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis. *Nature methods* **1**: 241-248.

Core LJ, Martins, A.L., Danko, C.G., Waters, C., Siepel A. & Lis, J.T. 2014. Analysis of transcription start sites from nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature genetics*.

Dale RK, Pedersen BS, Quinlan AR. 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**: 3423-3424.

Deng W, Roberts SG. 2005. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes & development* **19**: 2418-2423.

Eisenstein RS, Munro HN. 1990. Translational regulation of ferritin synthesis by iron. *Enzyme* **44**: 42-58.

Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome research* **20**: 565-577.

Hansen U, Sharp PA. 1983. Sequences controlling in vitro transcription of SV40 promoters. *The EMBO journal* **2**: 2293-2303.

Hennighausen L, Fleckenstein B. 1986. Nuclear factor 1 interacts with five DNA elements in the promoter region of the human cytomegalovirus major immediate early gene. *The EMBO journal* **5**: 1367-1371.

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**: 327-339.

Juven-Gershon T, Cheng S, Kadonaga JT. 2006. Rational design of a super core promoter that enhances gene expression. *Nature methods* **3**: 917-922.

Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH. 1998. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes & development* **12**: 34-44.

Lay AJ, Jiang XM, Kisker O, Flynn E, Underwood A, Condron R, Hogg PJ. 2000. Phosphoglycerate kinase acts in tumour angiogenesis as a disulphide reductase. *Nature* **408**: 869-873.

LeProust EM, Peck BJ, Spirin K, McCuen HB, Moore B, Namsaraev E, Caruthers MH. 2010. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic acids research* **38**: 2522-2540.

Moriyama K, Takada T, Tsutsumi Y, Fukada K, Ishibashi H, Niho Y, Maeda Y. 1994. Mutations in the transcriptional regulatory region of the precore and core/pregenome of a hepatitis B virus with defective HBeAg production. *Fukuoka Igaku Zasshi* **85**: 314-322.

Nenoi M, Mita K, Ichimura S, Cartwright IL, Takahashi E, Yamauchi M, Tsuji H. 1996. Heterogeneous structure of the polyubiquitin gene UbC of HeLa S3 cells. *Gene* **175**: 179-185.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. *Nature* **489**: 109-113.

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature biotechnology* **30**: 521-530.

Sigal A, Danon T, Cohen A, Milo R, Geva-Zatorsky N, Lustig G, Liron Y, Alon U, Perzov N. 2007. Generation of a fluorescently labeled endogenous protein library in living human cells. *Nature protocols* **2**: 1515-1527.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Van Beveren C, Rands E, Chattopadhyay SK, Lowy DR, Verma IM. 1982. Long terminal repeat of murine retroviral DNAs: sequence analysis, host-proviral junctions, and preintegration site. *Journal of virology* **41**: 542-556.

Venters BJ, Pugh BF. 2013. Genomic organization of human transcription initiation complexes. *Nature* **502**: 53-58.

Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research* **22**: 1798-1812.

Wang R, Liang J, Jiang H, Qin LJ, Yang HT. 2008. Promoter-dependent EGFP expression during embryonic stem cell propagation and differentiation. *Stem Cells Dev* **17**: 279-289.

Weingarten-Gabbay S, Elias-Kirma S, Nir R, Gritsenko AA, Stern-Ginossar N, Yakhini Z, Weinberger A, Segal E. 2016. Comparative genetics. Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science* **351**.