2018-11-01

Uncovering human transcription factor interactions associated with genetic variants, novel DNA moti...

Uncovering human transcription factor interactions associated with genetic variants, novel DNA motifs, and repetitive elements using enhanced yeast one-hybrid assays

Jared Allan Sewell[1]#, Shaleen Shrestha[1]#, Clarissa Stephanie Santoso[1], Elena Forchielli[1], Sebastian Carrasco Pro[2], Melissa Martinez[1], Juan Ignacio Fuxman Bass[1,2]

[1] Department of Biology, Boston University, Boston, MA, USA.

[2] Bioinformatics Program, Boston University, Boston, MA, USA.

Correspondence: Dr. Juan Fuxman Bass (fuxman@bu.edu)

# these authors contributed equally to this work

ABSTRACT

Identifying transcription factor (TF) binding to noncoding variants, uncharacterized DNA motifs, and repetitive genomic elements has been difficult due to technical and computational challenges. Indeed, current experimental methods such as chromatin immunoprecipitation are capable of only testing one TF at a time and motif prediction algorithms often lead to false positive and false negative predictions. Here, we address these limitations by developing two approaches based on enhanced yeast one-hybrid assays. The first approach allows to interrogate the binding of >1,000 human TFs to single nucleotide variant alleles, short insertions and deletions (indels), and novel DNA motifs; while the second approach allows for the identification of TFs that bind to repetitive DNA elements. Using the former approach, we identified gain of TF interactions to a GG→AA mutation in the TERT promoter and an 18 bp indel in the TAL1 super-enhancer, both of which are associated with cancer, and identified the TFs that bind to three uncharacterized DNA motifs identified by the ENCODE Project in footprinting assays. Using the latter approach, we detected the binding of 75 TFs to the highly repetitive Alu elements. We anticipate that these approaches will expand our capabilities to study genetic variation and under-characterized genomic regions.

INTRODUCTION

The study of transcription factor (TF) binding to different genomic regions, and how this binding is affected by noncoding variants, is critical to understand the mechanisms by which gene expression is controlled in normal and pathogenic conditions (1). Chromatin-immunoprecipitation followed by next-generation sequencing (ChIP-seq) has been instrumental in identifying the genomic regions occupied by a TF and studying TF function (2,3). However, it has been challenging to employ ChIP-seq to address many central human functional genomics problems such as determining whether disease-associated single nucleotide variants (SNVs) and short insertions/deletions (indels) in noncoding regions alter TF binding, as well as identifying the TFs that bind to specific DNA motifs and repetitive genomic DNA elements.

Experimentally determining whether TF binding is altered by genomic variants associated with genetic diseases or cancer has been challenging as this cannot be performed using ChIP-seq without *a priori* TF candidates. This is because all ~1,500 human TFs would have to be evaluated individually, and because samples from the appropriate tissues and conditions from healthy and sick individuals need to be obtained (1,4). Thus, the most widely used approach to prioritize TFs consists of using known DNA binding specificities (available for ~50% of human TFs) and motif search algorithms such as FIMO, BEEML-PWM or TFM-pvalue to compare predicted TF binding between the different noncoding alleles (5-9). However, this approach often results in multiple false positive and false negative predictions given that: 1) DNA motifs are missing for nearly half of the known human TFs, 2) it is estimated that only ~1% of the predicted DNA motifs in the genome are actually occupied by the TF *in vivo*, 3)

multiple genomic regions occupied by TFs do not contain the corresponding TF binding sites, and 4) sequence preferences for naked DNA may be different than for nucleosomal DNA (5,10-12).

Many genomic regions identified by DNase I or ATAC-seq footprinting studies are occupied by unidentified proteins, in many cases in a sequence specific manner (13,14). Indeed, using genome-wide DNase I footprinting, the ENCODE Project identified 683 de novo motifs, 289 of which could not be matched to any TF based on known DNA binding specificities (13). Due to the lack of TF candidates, it is nearly impossible to use ChIP to identify the TFs interacting with these genomic sites, as hundreds of TFs would need to be tested in each of the cell lines where the footprints were found. Thus, to this day, most of the novel DNA motifs derived from DNase I footprinting remain orphan.

More than half of the human genome is comprised of tandem or interspaced repetitive DNA elements, many located within promoter, enhancer, and silencer sequences (15). TF binding to these repetitive genomic elements is challenging to study by ChIP-seq, not only because hundreds of TFs need to be assayed, but also because repetitive DNA sequences are difficult to map to the reference genome and are thus often filtered out in most bioinformatics analysis pipelines (16,17). This is particularly true for highly repetitive elements such as the Alu short interspaced nuclear elements, which are present in more than one million copies in the human genome (18).

Enhanced yeast one-hybrid (eY1H) assays provide a complementary approach to ChIP-seq, where physical interactions between TFs and DNA regions are tested in the milieu of the yeast nucleus using reporter genes (19,20). eY1H assays involve two

components: a 'DNA-bait' (*e.g.*, a genomic variant, a novel DNA motif, or a repetitive element) and a 'TF-prey'. DNA-baits are cloned upstream of two reporter genes (HIS3 and LacZ) and integrated into the yeast genome. The DNA-bait yeast strains are then mated with strains that express TFs fused to the yeast Gal4 activation domain (AD) to generate diploid yeast containing both bait and prey. If the TF binds to the DNA-bait, the AD will induce reporter expression which can be measured by the conversion of the colorless X-gal to a blue compound (by the β–galactosidase enzyme encoded by LacZ), and by the ability of the yeast to grow on media lacking histidine even in the presence of 3-amino-triazole (a competitive inhibitor of the His3 enzyme).

Given that eY1H assays can be parallelized to study >1,000 TFs in a single experiment, this framework is particularly well suited to identify the sets of TFs that bind to a DNA region of interest (rather than the sets of DNA regions bound by a TFs as in ChIP-seq). In particular, eY1H assays (and other variations of the assay) have been used to identify the repertoire of TFs that bind to gene promoters and enhancers in humans, mice, nematodes, flies, and plants (1,21-26). Further, using eY1H assays, we determined altered TF binding to 109 SNVs associated with different genetic diseases including immune disorders, developmental malformations, cancer, and neurological disorders (1). This pipeline was based on PCR from human genomic DNA using wild type and mutated primers to generate the DNA-baits. As a consequence, the approach used was not suitable to clone and study indels (unless patient DNA samples are used) or novel DNA motifs from footprinting studies. In addition, the eY1H pipeline was not previously adapted to evaluate repetitive DNA elements given that the cloning steps were optimized for unique genomic DNA regions.

Here, we present two novel cloning approaches, one to study short DNA sequences (SNVs, indels, and novel DNA motifs), and another to study repetitive DNA elements. Using the former approach, we identified gain of TF binding to a two nucleotide substitution in the TERT promoter found in a patient with melanoma and to an 18 bp indel in a TAL1 super-enhancer found in a patient with T-cell acute lymphoblastic leukemia (27,28). In addition, we uncovered TFs that potentially bind to three novel motifs derived from genome-wide DNase I footprinting by the ENCODE Project that were previously uncharacterized (13). Finally, using the later eY1H approach to study repetitive DNA, we found 75 TFs that can bind to different Alu sequences present in the human genome. Overall, these eY1H-based approaches provide a novel toolkit to address genomic questions that have been challenging to address using current experimental and computational approaches.

MATERIALS AND METHODS

eY1H assays

DNA-baits were generated using different approaches depending on the type of sequence cloned (**Figure 1**). Genomic variants (SNVs and indels) and novel DNA motifs were synthesized as oligonucleotides (Thermofisher) flanked by the attB4 and attB1R sequences for cloning using the Gateway recombination system (**Figure 1** and **Supplementary Table S1**). Double-stranded oligonucleotides were generated by primer extension using Taq polymerase (Thermofisher) and a primer complementary to the attB1R (**Supplementary Table S1**) site using an initial denaturation step of 3 min at

95°C, ten cycles of 30 seconds at 55°C and 30 seconds at 72°C, followed by one cycle of 5 min at 72°C. The double-stranded oligonucleotides were then cloned into the pDONR-P4P1R by Gateway cloning using the BP Clonase II (ThermoFisher), and then transferred to the pMW#2 and pMW#3 vectors (Addgene) using the LR Clonase II (ThermoFisher), upstream of two reporter genes (HIS3 and LacZ). Both reporter constructs were integrated into the Y1HaS2 yeast strain genome by site-specific recombination to generate chromatinized DNA-bait strains as previously described (29,30).

DNA-baits corresponding to repetitive DNA elements were generated by PCR using human genomic DNA (Clontech) as a template, Platinum Hifi Taq polymerase (ThermoFisher), and degenerate primers complementary to different family members of Alu elements (Alu-Fw and Alu-Rv, **Supplementary Table S1**) or different variations of Alphoid DNA (Alphoid-Fw and Alphoid-Rv, **Supplementary Table S1**) (**Figure 1**). These primers include the attB4 and attB1R sequences for Gateway cloning. The PCR cycle involved an initial denaturation step of 2 min at 94°C, 35 cycles of 30 sec at 94°C, 15 sec at 58°C, and 75 sec at 72°C, followed by a final extension for 7 min at 72°C. The random libraries containing the Alu sequences or the Alphoid DNA were cloned into the pDONR-P4P1R vector by Gateway cloning and transformed into DH5α bacteria. Individual colonies were picked and sequenced to identify the sequences cloned (**Supplementary Table S2**). Each sequence was then cloned using the Gateway system into the HIS3 and LacZ reporter vectors and integrated into the Y1HaS2 yeast strain genome. DNA-bait yeast strains were then sequenced to verify the identity of the yeast integrants.

DNA-bait strains were mated with an array of yeast strains expressing 1,086 human TF-preys using a Singer RoToR robotic platform, as previously described (1,19). Each interaction was tested in quadruplicate, and only interactions detected with at least two colonies were considered positive (**Figure 1**). As previously observed, ~90% of interactions were detected by all four replicates (1,19,23).

Transient transfections and luciferase assays

TF interactions with noncoding alleles were validated by luciferase assays in HEK293T cells. Given that testing the noncoding alleles in the short sequence context (20-40 bp) used in eY1H assays led to barely above background levels of luciferase activity (not shown), DNA-bait luciferase reporter clones were generated corresponding to the 500 bp genomic sequence surrounding the noncoding alleles (**Supplementary Table S3**). Wild-type sequences were generated by PCR using human genomic DNA (Clontech) as a template, Platinum Hifi or SuperFi Taq polymerases (ThermoFisher), and primers surrounding the noncoding alleles. Mutant sequences were generated from the wild-type sequence by PCR stitching using primers that contain the mutated nucleotide. DNA-bait luciferase reporter clones were generated by cloning the noncoding regions upstream of the firefly luciferase into a Gateway compatible vector generated from pGL4.23[luc2/minP] (1). TF-prey clones were generated by Gateway cloning the TF coding sequence into the pEZY3-VP160 vector (31).

HEK293T cells were plated in 96-well opaque plates (~1 x 10^4 cells/well) 24 hours prior to transfection in 100 µl DMEM + 10% FBS + 1% Antibiotic-Antimycotic 100X. Cells were transfected with Lipofectamine 3000 (Invitrogen) according to the

manufacturer's protocol using 20 ng of the DNA-bait pGL4.23[luc2/minP] luciferase reporter vector, 80 ng of the TF-pEZY3-VP160 vector, and 10 ng of renilla luciferase control vector. The empty pEZY3-VP160 vector co-transfected with the recombinant firefly luciferase plasmid was used as a normalization control. 48 hours after transfection, firefly and renilla luciferase activities were measured using the Dual-Glo Luciferase Assay System (Promega) according to the manufacturer's protocol. Non-transfected cells were used to subtract background luciferase activities, followed by normalizing firefly luciferase activity to renilla luciferase activity.

ChIP-seq analysis

ChIP-seq peaks for GATA4 (ENCSR590CNM) and ZBTB26 (ENCSR229DYF) were obtained from the ENCODE Project (32). A peak was assigned to a gene promoter if the midpoint of the peak was located within the -2kb to +250bp from the transcription start site (according to Gencode v19). In addition, promoter positions overlapping with coding regions were excluded from this analysis. The midpoint of peaks overlapping with promoter regions was calculated using the intersect option from BEDTools (33). The fraction of genes with GATA4 and ZBTB26 peaks in their promoters was calculated as a function of the number of UW.Motif.0118 and UW.Motif.0146 motifs, respectively.

Prediction of DNA binding motif for zinc finger proteins

DNA motif prediction for the Cys2-His2 zinc finger TFs ZBTB26 and ZNF646 was performed using an expanded support vector machine model available at

http://zf.princeton.edu/fingerSelect.php (34). For ZBTB26 zinc fingers 1-3 were used for predictions. For ZNF646 zinc fingers 1-3, 4-6, 11-13, 14-16, 20-22, and 23-25 were used for predictions.

Co-expression between TFs and potential targets

For each of the three DNA motifs derived from DNase I footprints tested by eY1H assays, potential target genes were determined by the presence of the DNA motifs within their promoter sequences. Promoter regions were considered between 2 kb upstream and 250 bp downstream of the transcription start sites. Alternative promoters were also considered. Co-expression between the TFs found to bind to the DNA motifs by eY1H assays and the genes containing different the motifs in their promoters was determined by calculating the Pearson correlation coefficient (PCC) between their expression levels across 32 tissues from the Expression Atlas (https://www.ebi.ac.uk/gxa/home, E-MTAB-2836).

Gene ontology enrichment

Biological Process Gene Ontology enrichment was determined using the Gene Ontology Consortium enrichment analysis tool using all human genes as background (http://geneontology.org/page/go-enrichment-analysis) for genes with at least two motifs in their promoter regions, for each of the DNA motifs derived from DNase I footprints. For each cluster of terms, only the term with the highest fold enrichment (>2) was considered. For each of the TFs found to bind to the DNA motifs in eY1H assays

(except ZBTB26 and ZNF646 for which no or few publications are available), the association between the TF and the gene ontology terms were annotated (**Supplementary Table S4**).

RESULTS

Identifying altered TF binding to noncoding SNVs and indels

In our previous cloning strategy used to generate DNA-baits for identifying altered TF binding to noncoding variants by eY1H assays, DNA sequences were generated by PCR using human genomic DNA as a template and primers containing wild type or mutant sequences to introduce the allele variants (1). As a consequence, this cloning strategy presented several limitations for our current study: 1) the requirement of human DNA samples, 2) indels could not be successfully evaluated (unless genomic DNA was obtained from patient samples) as primers containing indels would fail to anneal to the wild type DNA template, and 3) the introduction of unwanted mutations during PCR even when using high fidelity polymerases, thus reducing the efficiency to generate DNA-baits without spurious mutations. To tackle these limitations, we modified our protocol to generate DNA-baits by using synthesized oligonucleotides containing a short sequence of interest (21 bp for SNVs, and 20 bp + indel length for indels) flanked by the attB4 and attB1R gateway cloning sites, followed by second strand synthesis using a universal primer complementary to the attB1R site (**Figure 1**).

To determine whether this modified cloning approach coupled with eY1H assays can identify altered TF binding caused by variants associated with human disease, we

screened two noncoding somatic mutations found in cancer patients: a two-nucleotide substitution in the TERT promoter and a complex 18 bp indel in a super-enhancer of TAL1 (27,28). First, we tested the two-nucleotide substitution (GG→AA) in the TERT promoter at position -138/-139. This particular mutation, derived from a patient with melanoma, was predicted to create an ETS/TCF binding site (28,35). Using eY1H assays against an array of 1,086 human TFs, we identified the binding of several ETS factors including GABPA, ELK1, ELK4, ETV1, ELF2, and ERF to the mutant AA allele but not to the wild-type GG allele, consistent with previous *in silico* predictions (**Figure 2A**). Indeed, the -138/-139 GG→AA mutation creates binding sites for these TFs (**Figure 2B**) but not for other ETS factors such as SPI1 and ETV7 that have slightly different DNA binding specificities (not shown). To validate the differential interactions observed, we performed luciferase assays by co-transfecting HEK293T cells with a plasmid carrying a 500 bp wild-type or mutant GG→AA containing TERT promoter sequence driving luciferase expression and a plasmid expressing the indicated TFs fused to the VP160 (10 copies of VP16) activation domain. We found that GABPA, ELK1, ELK4, ETV1, and ERF lead to a stronger activation of the mutant promoter, whereas ELF2 did not induce reporter expression (**Figure 2C**). Interestingly, a previous study identified a gain of interaction with GABPA to two G→A TERT promoter mutations at positions -124 and -146 (36). This suggests that binding sites for GABPA (and other ETS factors) created by single or di-nucleotide mutations at different positions in the TERT promoter may be responsible for TERT reactivation in tumors derived from different cancer patients.

In addition, we evaluated altered TF binding caused by an 18 bp insertion in a TAL1 super-enhancer found in a patient with T-cell acute lymphoblastic leukemia (27). We detected interactions involving ELK1, GABPA, ELF2, ELF3, and MYB with the insertion allele, but not with the wild type sequence or a sequence replacing the insertion with an $(AT)_9$ repeat (**Figure 3A**). The interaction between MYB and the insertion allele was confirmed in human cells by luciferase assays (**Figure 3B**), consistent with a previous study that found that this particular insertion creates a binding site for MYB, further validating our approach (27). However, the eY1H interactions between the insertion allele and the ETS factors ELK1, GABPA, ELF2, and ELF3 could not be confirmed by luciferase assays (**Figure 3B**), even though these TFs are predicted to bind outside the indel (**Figure 3C**). This difference between eY1H assays (and motif predictions) and luciferase assays may be related to differences in chromatin context (*i.e.*, eY1H assays test interactions within chromatinized DNA *versus* luciferase assays in episomal vectors) or differences in cellular context (*i.e.*, interactions tested in yeast for eY1H assays *versus* human cells in luciferase assays). Additional experiments in the endogenous locus using genome edited cell lines will ultimately determine whether ETS factors affect gene expression caused by the insertion. Regardless of whether the interactions with ETS factors ultimately validate, it is important to note that eY1H assays narrowed down the follow-up studies to five candidate TFs compared to *in silico* analyses using default settings in CIS-BP (5) that led to predicting 30 losses and 98 gains of interactions with the TAL1 super-enhancer insertion (not shown). This is particularly important, given that validation methods such as reporter assays, ChIP, and TF knockdowns are generally low-throughput.

Identifying TFs binding to novel DNA motifs

Different experimental methods, including protein-binding microarrays, SELEX, bacterial one-hybrid assays, and ChIP-seq have identified DNA binding motifs for hundreds of human TFs (5,37,38). However, 289 (out of 683) DNA motifs identified by genome-wide footprinting using DNase I by the ENCODE Project (13) remain orphan (*i.e.*, no TF has been predicted to bind these motifs). This can stem from the lack of DNA binding motifs for many human TFs (~50%), differences between DNA motifs determined *in vitro* and those occupied *in vivo*, from motif quality, and limitations in prediction algorithms. To determine whether eY1H assays can identify the TFs that bind to these orphan DNA motifs, we tested three of the DNA motifs identified by DNase I footprints by the ENCODE Project (13) that could not be matched to any human TF.

Using eY1H assays, each DNA motif was tested using three tandem repeats and a mutated version of the motif as control (**Figure 4A**). For UW.Motif.0118 (GCTGATAA) we determined that GATA4, GATA5, and DMBX1 bind to the wild type but not the mutant DNA motif (**Figure 4B**). Indeed UW.Motif.0118 matches the DNA binding motifs for GATA4 and GATA5, which were reported by later publications (**Figure 4C**) (37,39). More importantly, we found that gene promoters that contain one or more instances of UW.Motif.0118 are enriched in ChIP-seq peaks for GATA4 (**Figure 4H**). DMBX1 can be discarded as a candidate to bind UW.Motif.0118 as its motif matches the junction between two motif copies in the tandem repeat rather than the motif itself (not shown). We also determined that UW.Motif.0146 (ATTTCTGG), but not the mutated motif, binds ZBTB26 (**Figure 4D**). The DNA binding motif for ZBTB26 has not yet been determined.

Using a position weight matrix prediction algorithm designed for cys2-his2 zinc finger TFs, we predicted a likely recognition motif based on the amino acid sequence of zinc fingers 1-3, which closely resembles UW.Motif.0146 (**Figure 4E**). Further, we found that gene promoters that contain one or more instances of UW.Motif.0146 are enriched in ChIP-seq peaks for ZBTB26, further validating our approach (**Figure 4H**). Finally, for UW.Motif.0167 (ACAAAAGA) we found that multiple SOX TFs and ZNF646 bind to the wild type but not the mutant motif in eY1H assays (**Figure 4F**). This DNA motif, partially matches SOX motifs (**Figure 4G**), while motifs are not available for ZNF646, a protein with 31 zinc fingers according to Uniprot. For several clusters of three ZNF646 zinc fingers, we predicted a CAAAA binding preference, a sequence present in UW.Motif.0167 (**Supplementary Figure S1**).

To determine whether the TFs that bind to the orphan DNA motifs are functionally related to their potential respective target genes, we determined TF-target co-expression across tissues. We found that for motifs UW.Motif.0146 and UW.Motif.0167, the higher the occurrence of the motifs in target gene promoters the higher the correlation between the expression levels of the target genes and the TFs predicted to bind to those DNA motifs (**Figure 4I**). More importantly, we found that the Biological Process Gene Ontology terms associated with potential target genes that contain more than one instance of the motif in their promoter, are related to known functions of the TFs that bind the DNA motifs in eY1H assays (**Supplementary Table S4**). For example, genes that contain more than one instance of the UW.Motif.0118 in their promoters are associated with cardiovascular and neuronal development. This is consistent with the role of GATA4 and GATA5 in heart development and angiogenesis,

and the role of GATA4 in neuronal development and function (40-43). Similarly, genes that contain more than one instance of the UW.Motif.0167 in their promoters are associated with dendritic spine development, among other biological processes, as are SOX2, SOX5, and SOX11 (44-46). Altogether, this shows that our approach can identify the TFs that bind to uncharacterized motifs, including zinc fingers which are generally difficult to study, and that the TFs identified are functionally related to their potential target genes.

TF binding to repetitive DNA elements

The binding of TFs to repetitive DNA elements has been challenging to study experimentally given the difficulty of mapping sequencing reads derived from these elements to genomic locations (16,17). Indeed, most ChIP-seq pipelines currently remove a large fraction of the reads corresponding to repetitive elements (16). To illustrate the power of eY1H assays to identify TFs that bind to repetitive elements, we evaluated Alu sequences, a type of short interspaced nuclear element present in more than one million copies in the human genome. Studying TF binding to these sequences is particularly important given that Alu sequences are embedded within gene promoters, enhancers, and introns, and have been shown to play significant roles in gene regulation (18,47). In addition, these sequences are often silenced by mechanisms that are not fully understood which could in part be mediated by transcriptional repressors (48,49).

To evaluate TF binding to these repetitive elements, we cloned 20 Alu sequences into our eY1H pipeline (**Supplementary Table S2**). This was performed using degenerate primers complementary to the 5' and 3' ends of Alu sequences, which allowed us to obtain clones belonging to different Alu families, including the ancestral AluJ, the derived AluS, and the younger AluY elements. Using eY1H assays, we identified 75 TFs that bind to at least one Alu sequence, and 34 TFs that bind to at least 20% of the 20 Alu sequences tested (**Figure 5A** and **Supplementary Table S5**). Interestingly, Alu sequences are enriched in binding to TFs belonging to the nuclear hormone receptor (NHR), zinc finger DHHC (ZF-DHHC), ETS, and regulatory factor X (RFX) families compared to the array of TFs tested (**Figure 5B**). Of note, we did not detect interactions with NHR or ZF-DHHC TFs for the two AluJ and a subset of AluS sequences tested, suggesting that binding sites for these TFs may have been acquired sometime during AluS divergence. Other than this, differences in TF binding between Alu sequences do not seem to cluster by Alu family, likely because of differences in deletions and truncations within family members.

The widespread TF binding to Alu sequences observed is not a general feature of repetitive elements or of eY1H assays, as screening alphoid DNA (*i.e.*, centromeric DNA sequences) only led to marginal TF binding (**Figure 5C**). Indeed, we did not identify any TFs that bound more than 20% of the 12 alphoid sequences tested (not shown), compared to 34 TFs for the Alu sequences. This is expected, as alphoid DNA is known to recruit multiple centromeric and heterochromatin proteins but not TFs (50). Altogether, we show that eY1H assays can identify TF binding to highly repetitive genomic sequences such as Alu sequences. Whether these TFs globally affect the

function of Alu sequences or the function of Alu sequences at specific loci remains to be determined.

DISCUSSION

In this study we describe two approaches to evaluate TF binding to short DNA sequences (*e.g.*, SNVs, indels, and novel DNA motifs) and repetitive elements. Testing these types of sequences has previously been challenging due to limitations in ChIP-seq and DNA motif analyses. Indeed, motif analyses using CIS-BP (5) led to predicting 128 and 56 differential TF interactions with the TAL1 super-enhancer insertion and the TERT -138/-139 GG→AA mutation, respectively. Our approach greatly reduced the number of differential TFs that would need to validated, which is particularly important given that reporter assays, ChIP, and TF knockdowns followed by RT-qPCR are generally low-throughput. More importantly, it allows the identification of TF-DNA binding even in the absence of *a priori* TF candidates or human DNA template as in the case of novel DNA motifs identified by DNase I footprinting assays. The eY1H approach can also be applied to DNA motifs enriched in the regulatory regions of functionally related genes, in particular when these DNA motifs cannot be assigned to any TF.

Using eY1H assays we also evaluated 20 sequences belonging to different Alu families and identified 34 TFs that bind to at least 20% of the sequences tested. These TFs may be involved in regulating the expression of nearby genes or silencing Alu sequences. Alternatively, Alu sequences may also act as sinks for some TFs reducing their effective nuclear concentration. However, we did not detect a significantly higher

number of sequencing reads matching Alu sequences in ChIP-seq datasets from the ENCODE Project corresponding to Alu binding *versus* non-Alu binding TFs (not shown). This could be associated with epigenetic silencing of many Alu sequences which prevents TF binding in human cells in most tissues and conditions. Indeed, most Alu sequences have been found to be enriched in the H3K9me mark and to be actively silenced in somatic tissues (48,51). Nonetheless, thousands of Alu sequences in the human genome contain active histone marks and may be permissive for TF binding which could contribute to the transcriptional control of nearby genes in specific cells and conditions (18,52). For example, a recent study found that *de novo* ChIP-seq peaks for the H3K4me1 mark in macrophages infected with *Mycobacterium tuberculosis* contain Alu sequences enriched for binding sites of several TFs including ETS and NHR factors, consistent with our findings by eY1H assays (52). We anticipate that the approach we developed to study TF binding to Alu and alphoid sequences will shed light into the role of other repetitive elements in gene regulation, silencing, and establishment of heterochromatin.

Overall, the eY1H approaches described here unlock the possibility of characterizing altered TF binding to different types of genomic variants and studying the role of TFs in regulating the function of repetitive genomic elements.

ACKNOWLEDGMENTS

We thank Dr. Trevor Siggers for critically reviewing the manuscript.

FUNDING

This work was supported by the National Institutes of Health [R00-GM114296 and R35-GM128625 to J.I.F.B.; and 5T32HL007501-34 to J.A.S.] and the National Science Foundation [NSF-REU BIO-1659605 to M.M.].

AUTHORS CONTRIBUTIONS

J.I.F.B., J.A.S., S.S., E.F., and M.M. performed eY1H assays. S.S. and C.S.S. performed luciferase assays. J.I.F.B. and S.C.P. performed the data analyses. J.I.F.B. conceived the project and wrote the manuscript. All authors approved the content of the manuscript.

Competing interests

The authors declare no competing interests.

REFERENCES

1. Fuxman Bass, J.I., Sahni, N., Shrestha, S., Garcia-Gonzalez, A., Mori, A., Bhat, N., Yi, S., Hill, D.E., Vidal, M. and Walhout, A.J. (2015) Human gene-centered transcription factor networks for enhancers and disease variants. *Cell*, **161**, 661-673.
2. Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91-100.
3. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, **4**, 651-657.
4. Gan, K.A., Carrasco Pro, S., Sewell, J.A. and Fuxman Bass, J.I. (2018) Identification of Single Nucleotide Non-coding Driver Mutations in Cancer. *Front Genet*, **9**, 16.
5. Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K. *et al.* (2014)

Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431-1443.

6. Grant, C.E., Bailey, T.L. and Noble, W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017-1018.
7. Rheinbay, E., Parasuraman, P., Grimsby, J., Tiao, G., Engreitz, J.M., Kim, J., Lawrence, M.S., Taylor-Weiner, A., Rodriguez-Cuevas, S., Rosenberg, M. *et al.* (2017) Recurrent and functional regulatory mutations in breast cancer. *Nature*.
8. Touzet, H. and Varre, J.S. (2007) Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms Mol Biol*, **2**, 15.
9. Zhao, Y. and Stormo, G.D. (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat Biotechnol*, **29**, 480-483.
10. Zia, A. and Moses, A.M. (2012) Towards a theoretical understanding of false positives in DNA motif finding. *BMC Bioinformatics*, **13**, 151.
11. Talebzadeh, M. and Zare-Mirakabad, F. (2014) Transcription factor binding sites prediction based on modified nucleosomes. *PLoS One*, **9**, e89226.
12. Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S.O., Nitta, K.R., Morgunova, E., Taipale, M. *et al.* (2018) The interaction landscape between transcription factors and the nucleosome. *Nature*, **562**, 76-81.
13. Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83-90.
14. Ramirez, R.N., El-Ali, N.C., Mager, M.A., Wyman, D., Conesa, A. and Mortazavi, A. (2017) Dynamic Gene Regulatory Networks of Human Myeloid Differentiation. *Cell Syst*, **4**, 416-429 e413.
15. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.
16. Rozowsky, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M. and Gerstein, M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol*, **27**, 66-75.
17. Chung, D., Kuan, P.F., Li, B., Sanalkumar, R., Liang, K., Bresnick, E.H., Dewey, C. and Keles, S. (2011) Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. *PLoS Comput Biol*, **7**, e1002111.
18. Deininger, P. (2011) Alu elements: know the SINEs. *Genome Biol*, **12**, 236.
19. Reece-Hoyes, J.S., Diallo, A., Lajoie, B., Kent, A., Shrestha, S., Kadreppa, S., Pesyna, C., Dekker, J., Myers, C.L. and Walhout, A.J. (2011) Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. *Nat Methods*, **8**, 1059-1064.
20. Sewell, J.A. and Fuxman Bass, J.I. (2018) Options and Considerations When Using a Yeast One-Hybrid System. *Methods Mol Biol*, **1794**, 119-130.
21. Burdo, B., Gray, J., Goetting-Minesky, M.P., Wittler, B., Hunt, M., Li, T., Velliquette, D., Thomas, J., Gentzel, I., dos Santos Brito, M. *et al.* (2014) The

Maize TFome--development of a transcription factor open reading frame collection for functional genomics. *Plant J*, **80**, 356-366.

22. Gubelmann, C., Waszak, S.M., Isakova, A., Holcombe, W., Hens, K., Iagovitina, A., Feuz, J.D., Raghav, S.K., Simicevic, J. and Deplancke, B. (2013) A yeast one-hybrid and microfluidics-based pipeline to map mammalian gene regulatory networks. *Mol Syst Biol*, **9**, 682.
23. Fuxman Bass, J.I., Pons, C., Kozlowski, L., Reece-Hoyes, J.S., Shrestha, S., Holdorf, A.D., Mori, A., Myers, C.L. and Walhout, A.J. (2016) A gene-centered C. elegans protein-DNA interaction network provides a framework for functional predictions. *Mol Syst Biol*, **12**, 884.
24. Reece-Hoyes, J.S., Pons, C., Diallo, A., Mori, A., Shrestha, S., Kadreppa, S., Nelson, J., Diprima, S., Dricot, A., Lajoie, B.R. *et al.* (2013) Extensive rewiring and complex evolutionary dynamics in a C. elegans multiparameter transcription factor network. *Mol Cell*, **51**, 116-127.
25. Hens, K., Feuz, J.D., Isakova, A., Iagovitina, A., Massouras, A., Bryois, J., Callaerts, P., Celniker, S.E. and Deplancke, B. (2011) Automated protein-DNA interaction screening of Drosophila regulatory elements. *Nat Methods*, **8**, 1065-1070.
26. Brady, S.M., Zhang, L., Megraw, M., Martinez, N.J., Jiang, E., Yi, C.S., Liu, W., Zeng, A., Taylor-Teeples, M., Kim, D. *et al.* (2011) A stele-enriched gene regulatory network in the Arabidopsis root. *Mol Syst Biol*, **7**, 459.
27. Mansour, M.R., Abraham, B.J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A.D., Etchin, J., Lawton, L., Sallan, S.E., Silverman, L.B. *et al.* (2014) Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science*, **346**, 1373-1377.
28. Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K. *et al.* (2013) TERT promoter mutations in familial and sporadic melanoma. *Science*, **339**, 959-961.
29. Fuxman Bass, J.I., Reece-Hoyes, J.S. and Walhout, A.J. (2016) Gene-Centered Yeast One-Hybrid Assays. *Cold Spring Harb Protoc*, **2016**, pdb top077669.
30. Fuxman Bass, J.I., Reece-Hoyes, J.S. and Walhout, A.J. (2016) Generating Bait Strains for Yeast One-Hybrid Assays. *Cold Spring Harb Protoc*, **2016**, pdb prot088948.
31. Carrasco Pro, S., Dafonte Imedio, A., Santoso, C.S., Gan, K.A., Sewell, J.A., Martinez, M., Sereda, R., Mehta, S. and Fuxman Bass, J.I. (2018) Global landscape of mouse and human cytokine transcriptional regulation. *Nucleic Acids Res*.
32. Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.
33. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
34. Persikov, A.V. and Singh, M. (2014) De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Res*, **42**, 97-108.
35. Vallarelli, A.F., Rachakonda, P.S., Andre, J., Heidenreich, B., Riffaud, L., Bensussan, A., Kumar, R. and Dumaz, N. (2016) TERT promoter mutations in

melanoma render TERT expression dependent on MAPK pathway activation. *Oncotarget*, **7**, 53127-53136.

36. Bell, R.J., Rube, H.T., Kreig, A., Mancini, A., Fouse, S.D., Nagarajan, R.P., Choi, S., Hong, C., He, D., Pekmezci, M. *et al.* (2015) Cancer. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. *Science*, **348**, 1036-1039.
37. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327-339.
38. Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H. and Wolfe, S.A. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, **133**, 1277-1289.
39. Kulakovskiy, I.V., Medvedeva, Y.A., Schaefer, U., Kasianov, A.S., Vorontsov, I.E., Bajic, V.B. and Makeev, V.J. (2013) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res*, **41**, D195-202.
40. Walsh, K. and Shiojima, I. (2007) Cardiac growth and angiogenesis coordinated by intertissue interactions. *J Clin Invest*, **117**, 3176-3179.
41. Lawson, M.A. and Mellon, P.L. (1998) Expression of GATA-4 in migrating gonadotropin-releasing neurons of the developing mouse. *Mol Cell Endocrinol*, **140**, 157-161.
42. Holtzinger, A. and Evans, T. (2007) Gata5 and Gata6 are functionally redundant in zebrafish for specification of cardiomyocytes. *Dev Biol*, **312**, 613-622.
43. Ang, Y.S., Rivas, R.N., Ribeiro, A.J.S., Srivas, R., Rivera, J., Stone, N.R., Pratt, K., Mohamed, T.M.A., Fu, J.D., Spencer, C.I. *et al.* (2016) Disease Model of GATA4 Mutation Reveals Transcription Factor Cooperativity in Human Cardiogenesis. *Cell*, **167**, 1734-1749 e1722.
44. Whitney, I.E., Keeley, P.W., St John, A.J., Kautzman, A.G., Kay, J.N. and Reese, B.E. (2014) Sox2 regulates cholinergic amacrine cell positioning and dendritic stratification in the retina. *J Neurosci*, **34**, 10109-10121.
45. Naudet, N., Moutal, A., Vu, H.N., Chounlamountri, N., Watrin, C., Cavagna, S., Malleval, C., Benetollo, C., Bardel, C., Dronne, M.A. *et al.* (2018) Transcriptional regulation of CRMP5 controls neurite outgrowth through Sox5. *Cell Mol Life Sci*, **75**, 67-79.
46. Hoshiba, Y., Toda, T., Ebisu, H., Wakimoto, M., Yanagi, S. and Kawasaki, H. (2016) Sox11 Balances Dendritic Morphogenesis with Neuronal Migration in the Developing Cerebral Cortex. *J Neurosci*, **36**, 5775-5784.
47. Batzer, M.A. and Deininger, P.L. (2002) Alu repeats and human genomic diversity. *Nat Rev Genet*, **3**, 370-379.
48. Kondo, Y. and Issa, J.P. (2003) Enrichment for histone H3 lysine 9 methylation at Alu repeats in human cells. *J Biol Chem*, **278**, 27658-27662.
49. Humphrey, G.W., Englander, E.W. and Howard, B.H. (1996) Specific binding sites for a pol III transcriptional repressor and pol II transcription factor YY1 within the internucleosomal spacer region in primate Alu repetitive elements. *Gene Expr*, **6**, 151-168.

50. Buxton, K.E., Kennedy-Darling, J., Shortreed, M.R., Zaidan, N.Z., Olivier, M., Scalf, M., Sridharan, R. and Smith, L.M. (2017) Elucidating Protein-DNA Interactions in Human Alphoid Chromatin via Hybridization Capture and Mass Spectrometry. *J Proteome Res*, **16**, 3433-3442.
51. Ward, M.C., Wilson, M.D., Barbosa-Morais, N.L., Schmidt, D., Stark, R., Pan, Q., Schwalie, P.C., Menon, S., Lukk, M., Watt, S. *et al.* (2013) Latent regulatory potential of human-specific repetitive elements. *Mol Cell*, **49**, 262-272.
52. Bouttier, M., Laperriere, D., Memari, B., Mangiapane, J., Fiore, A., Mitchell, E., Verway, M., Behr, M.A., Sladek, R., Barreiro, L.B. *et al.* (2016) Alu repeats as transcriptional regulatory platforms in macrophage responses to M. tuberculosis infection. *Nucleic Acids Res*, **44**, 10571-10587.

FIGURE LEGENDS

Figure 1. eY1H approaches to test short DNA sequences and repetitive DNA elements. Short DNA sequences are synthesized flanked by the attB4 and attB1R Gateway sites. The reverse strand is synthesized by primer extension (attB1R). The double stranded DNA generated is cloned into the pDONR-P4P1R vector by Gateway BP reaction. Repetitive DNA sequences are amplified from genomic DNA using degenerate primers flanked by the attB4 (forward) or the attB1R (reverse) sites. The repetitive element DNA library generated is cloned *en masse* into the pDONR-P4P1R vector and individual sequences are selected after bacterial transformation and picking of individual colonies. Both short DNA sequences and repetitive DNA are then transferred into eY1H reporter vectors (HIS3 and LacZ) and integrated into the yeast genome to generate DNA-baits strains. DNA-bait strains are tested for interactions against an array of 1,086 human TF-preys (TFs fused to the yeast Gal4 activation domain) by mating. Interactions are identified by the ability to grow in the absence of histidine and in the presence of the his3p inhibitor 3-Amino-1,2,4-triazole and turn blue in the presence of X-gal. Interactions are tested in quadruplicate.

Figure 2. Identification of altered TF binding to a two-nucleotide substitution in the TERT promoter. (**A**) eY1H screen for wild type and a -138/-139 GG→AA mutation in the TERT promoter associated with cancer. Each interaction was tested in quadruplicate. AD – empty vector control. (**B**) Motifs obtained from CIS-BP that match the differential TFs identified by eY1H assays. (**C**) Luciferase assays to validate the differential TF interactions with the TERT promoter alleles. HEK293T cells were co-transfected with reporter plasmids containing the wild type or mutant TERT promoter region cloned upstream of the firefly luciferase reporter gene, and expression vectors for the indicated TFs (fused to the activation domain VP160). After 48 h, cells were harvested and luciferase assays were performed. Relative luciferase activity is plotted as fold change compared to cells co-transfected with the wild type TERT promoter construct and the VP160 vector control. A representative experiment of three is shown. The average of three replicates is indicated by the black line. *$p<0.05$ by one-tailed log-transformed Student's t-test with Benjamini-Hochberg correction.

Figure 3. Identification of altered TF binding to a TAL1 super-enhancer insertion. (**A**) eY1H screen for an 18bp insertion in a TAL1 super-enhancer associated with T-cell acute lymphoblastic leukemia. Wild type and an $(AT)_9$ sequence were screened as controls. Each interaction was tested in quadruplicate. AD – empty vector control. (**B**) Motifs obtained from CIS-BP that match the differential TFs identified by eY1H assays. (**C**) Luciferase assays to validate the differential TF interactions with the TAL1 super-enhancer wild type and insertion alleles. HEK293T cells were co-transfected with

reporter plasmids containing the wild type or insertion TAL1 super-enhancer region cloned upstream of the firefly luciferase reporter gene, and expression vectors for the indicated TFs (fused to the activation domain VP160). After 48 h, cells were harvested and luciferase assays were performed. Relative luciferase activity is plotted as fold change compared to cells co-transfected with the wild type TAL1 super-enhancer construct and the VP160 vector control. A representative experiment of three is shown. The average of three replicates is indicated by the black line. *p<0.05 by one-tailed log-transformed Student's t-test with Benjamini-Hochberg correction.

Figure 4. Identification of TFs that bind to novel DNA motifs. (**A**) Motifs were tested by eY1H assays as three tandem copies. Motifs with two mutated bases were tested as controls. (**B, D, F**) eY1H screen of three motifs identified by DNase I footprinting by the ENCODE Project. Tested sequences are indicated. Each interaction was tested in quadruplicate. AD – empty vector control. (**C, G**) Alignment of motif logos from CIS-BP to the tested sequences. (**E**) Predicted ZBTB26 motif based on the amino acid sequence of zinc fingers 1-3. (**H**) Fraction of genes with ChIP-seq peaks for GATA4 and ZBTB26 in their promoters as a function of the number of UW.Motif.0118 and UW.Motif.0146 motifs, respectively. *p<0.001 vs absence of motif by proportion comparison test. (**I**) Correlation between TF and target gene expression levels across 32 tissues based on the occurrence of each motif in the target gene promoter. *p<0.001 vs absence of motif by Mann-Whitey's U test.

Figure 5. Identification of TFs that interact with Alu sequences. (**A**) TFs that interact with 20% or more of the Alu sequences tested by eY1H assays. TFs are ordered from top to bottom based on the number of Alu sequences they bind to. (**B**) Distribution by family for TFs that interact with at least 20% of the Alu sequences tested, compared to the distribution of TFs in the eY1H array. NHR – nuclear hormone receptor, ZF-DHHC – zinc finger DHHC, ZF-C2H2 – zinc finger cys2his2, HD – homeodomain, ETS – E26 transformation specific, HMG – high mobility group, RFX – regulatory factor X, bHLH – basic helix-loop-helix, bZIP – basic leucine zipper domain. *$p<0.05$ by proportion comparison test after Bonferroni correction. (**C**) Comparison between the number of PDIs detected for Alu and Alphoid sequences. Statistical significance determined by two-tailed Mann-Whitney's U test.

Figure 1

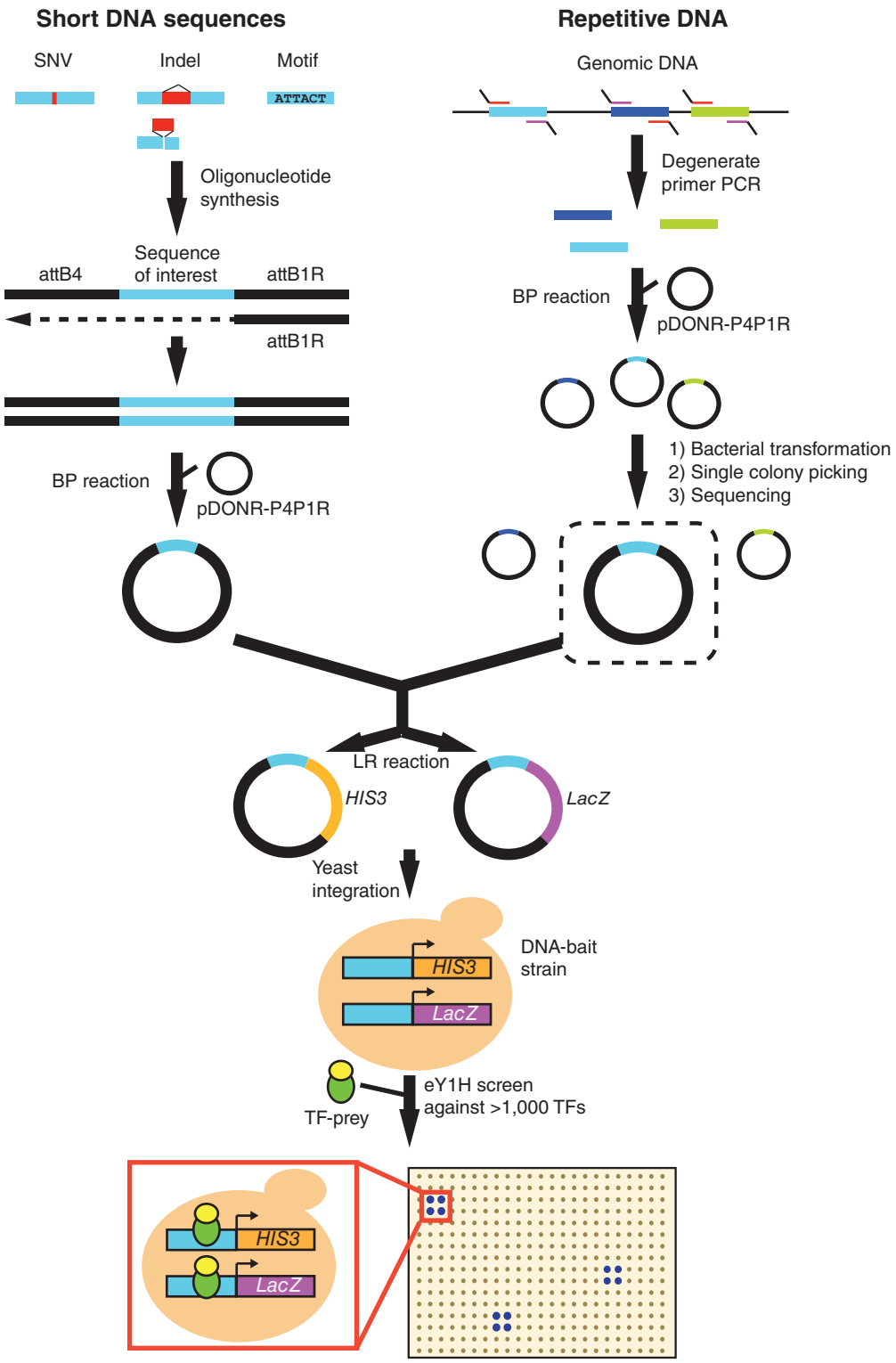
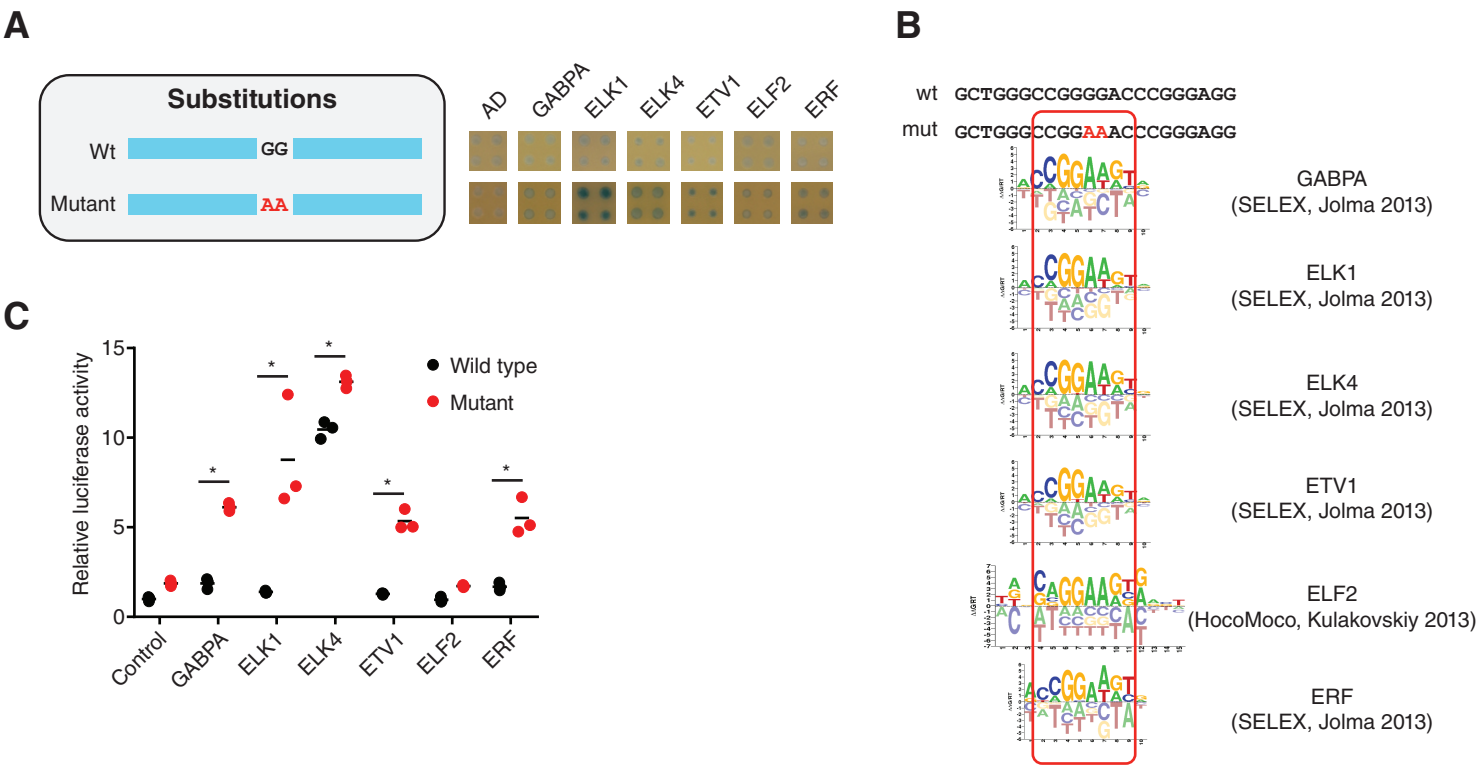


Figure 2



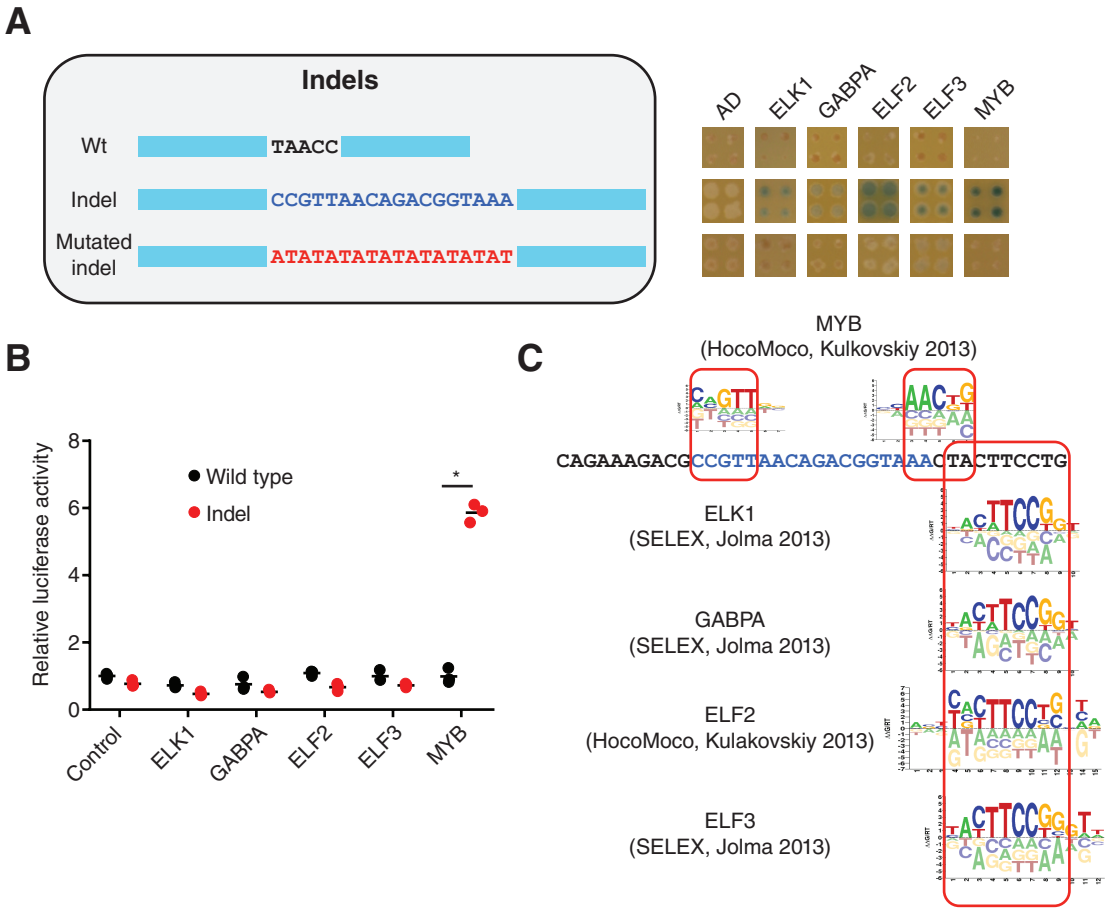


Figure 4

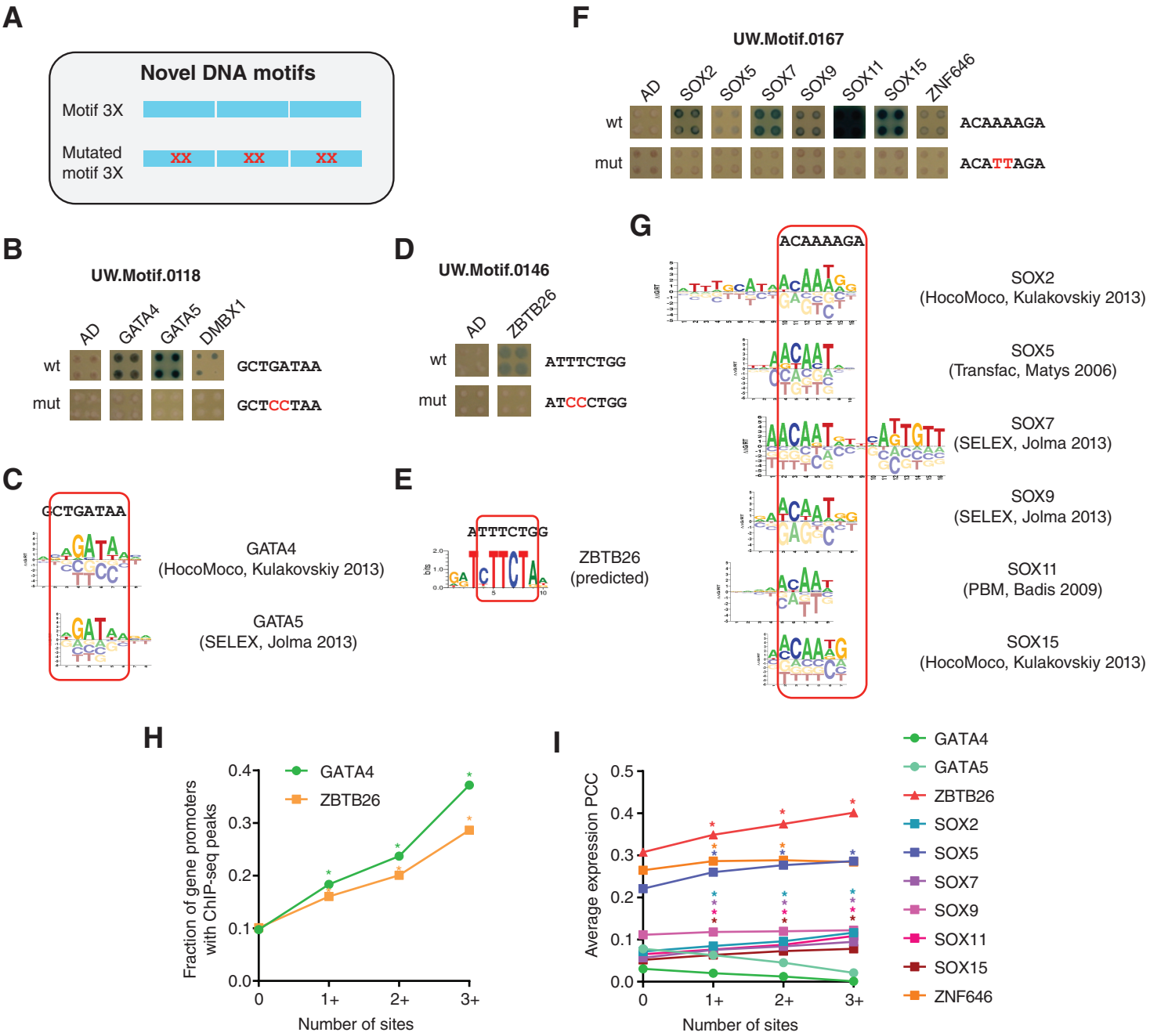


Figure 5