

## Washington University School of Medicine Digital Commons@Becker

---

### Open Access Publications

---

2008

# Genome wide screens in yeast to identify potential binding sites and target genes of DNA-binding

Jue Zeng

*University of California - Irvine*

Jizhou Yan

*National Institutes of Health*

Ting Wang

*Washington University School of Medicine in St. Louis*

Deborah Mosbrook-Davis

*National Institutes of Health*

Kyle T. Dolan

*National Institutes of Health*

*See next page for additional authors*

Follow this and additional works at: [http://digitalcommons.wustl.edu/open\\_access\\_pubs](http://digitalcommons.wustl.edu/open_access_pubs)

---

### Recommended Citation

Zeng, Jue; Yan, Jizhou; Wang, Ting; Mosbrook-Davis, Deborah; Dolan, Kyle T.; Christensen, Ryan; Stormo, Gary D.; Haussler, David; Lathrop, Richard H.; Brachmann, Rainer K.; and Burgess, Shawn M., "Genome wide screens in yeast to identify potential binding sites and target genes of DNA-binding." *Nucleic Acids Research*.36,1. e8. (2008).  
[http://digitalcommons.wustl.edu/open\\_access\\_pubs/113](http://digitalcommons.wustl.edu/open_access_pubs/113)

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact [engeszer@wustl.edu](mailto:engeszer@wustl.edu).

---

**Authors**

Jue Zeng, Jizhou Yan, Ting Wang, Deborah Mosbrook-Davis, Kyle T. Dolan, Ryan Christensen, Gary D. Stormo, David Haussler, Richard H. Lathrop, Rainer K. Brachmann, and Shawn M. Burgess

# Genome wide screens in yeast to identify potential binding sites and target genes of DNA-binding proteins

Jue Zeng<sup>1</sup>, Jizhou Yan<sup>2,3</sup>, Ting Wang<sup>3,4</sup>, Deborah Mosbrook-Davis<sup>2</sup>, Kyle T. Dolan<sup>2</sup>, Ryan Christensen<sup>4</sup>, Gary D. Stormo<sup>4</sup>, David Haussler<sup>3</sup>, Richard H. Lathrop<sup>5,6</sup>, Rainer K. Brachmann<sup>1,6,7</sup> and Shawn M. Burgess<sup>2,\*</sup>

<sup>1</sup>Division of Hematology/Oncology, Department of Medicine, University of California, Irvine, CA,

<sup>2</sup>Genome Technology Branch, National Human Genome Research Institute, NIH, Bethesda, MD, <sup>3</sup>Howard Hughes Medical Institute, Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA,

<sup>4</sup>Department of Genetics, Washington University School of Medicine, Saint Louis, MO, <sup>5</sup>Departments of Computer Science and Biomedical Engineering, <sup>6</sup>Institute for Genomics and Bioinformatics and <sup>7</sup>Departments of Biological Chemistry and Pathology & Laboratory Medicine, University of California, Irvine, CA, USA

Received October 24, 2007; Revised November 28, 2007; Accepted November 29, 2007

## ABSTRACT

Knowledge of all binding sites for transcriptional activators and repressors is essential for computationally aided identification of transcriptional networks. The techniques developed for defining the binding sites of transcription factors tend to be cumbersome and not adaptable to high throughput. We refined a versatile yeast strategy to rapidly and efficiently identify genomic targets of DNA-binding proteins. Yeast expressing a transcription factor is mated to yeast containing a library of genomic fragments cloned upstream of the reporter gene *URA3*. DNA fragments with target-binding sites are identified by growth of yeast clones in media lacking uracil. The experimental approach was validated with the tumor suppressor protein p53 and the forkhead protein Foxl1 using genomic libraries for zebrafish and mouse generated by shotgun cloning of short genomic fragments. Computational analysis of the genomic fragments recapitulated the published consensus-binding site for each protein. Identified fragments were mapped to identify the genomic context of each binding site. Our yeast screening strategy, combined with bioinformatics

approaches, will allow both detailed and high-throughput characterization of transcription factors, scalable to the analysis of all putative DNA-binding proteins.

## INTRODUCTION

Identification of DNA-binding sites for transcription factors can be a difficult task, in particular because they often consist of short, degenerate DNA sequences. Traditionally, DNA-binding sites have been identified by electrophoretic mobility shift assays (EMSA) (1), SELEX enrichment (2,3) and DNase I footprinting assays (4). More recently, microarray technology has been adapted to this task (5), but the identified *in vitro* binding sites do not necessarily address the *in vivo* activity of a particular binding site. The human genome encodes 2000–3000 transcription factors (6–9), but at present only 123 transcription factors (from any species) have experimentally determined binding sites in the JASPAR database ([http://mordor.cgb.ki.se/cgi-bin/jaspar2005/jaspar\\_db.pl](http://mordor.cgb.ki.se/cgi-bin/jaspar2005/jaspar_db.pl)). This indicates a strong need for additional approaches that are more efficient.

Experimental strategies in the yeast *Saccharomyces cerevisiae* can be fast, cheap and almost universally accessible, as has been demonstrated by the broad use

\*To whom correspondence should be addressed: Tel: +301 594 8224; Fax: +301 496 0474; Email: [burgess@mail.nih.gov](mailto:burgess@mail.nih.gov)

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Present addresses:

Rainer K. Brachmann, Genentech, Inc. 1 DNA Way, MS443B South San Francisco, CA, USA

Jizhou Yan, Abramson Family Cancer Research Institute, Department of Cancer Biology, University of Pennsylvania, Philadelphia, PA, USA

© 2007 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

of the two-hybrid assay (10) and a wide array of additional experimental strategies in yeast for high-throughput screens (11–18). More recently, Deplancke and colleagues (19) studied 72 digestive tract promoters of *Caenorhabditis elegans* in high-throughput yeast assays with 117 proteins and found a large number of previously unknown protein–DNA interaction networks. Recent techniques involving high-throughput sequencing called ChIP-Seq (20), provides a promising new approach, but requires high-quality antibodies to the transcription factor of interest and sufficient starting material for chromatin immunoprecipitation. Thus the technique is not readily scalable to hundreds or thousands of transcription factors.

We have created the necessary tools and protocols to perform yeast screens that identify both the sequences of DNA-binding targets of transcription factors and biologically active sites of binding. This article demonstrates the utility using libraries made from the zebrafish and mouse genomes. We have validated the approach for both libraries using two transcription factors: FoxI1 and p53. For both transcription factors, yeast screens generate accurate consensus DNA-binding sites and potential target genes. The techniques are readily scalable to new, high-throughput sequencing approaches for comprehensive binding data on large numbers of transcription factors.

## METHODS

### Transcription factor expression plasmid

A cDNA expression vector, pYoh-1, was constructed by inserting a double-stranded oligo containing multiple restriction enzyme sites (forward, 5'-TCG AGC TCA GTC GAC TGG TAC CGA TAT CGA ATT CGG ATC CCC GGG GCC TC-3' and reverse, 5'-CAT GGA GGC CCC GGG GAT CCG AAT TCG ATA TCG GTA CCA TGC GAC TGA GC-3') into pACT2 (Clontech) at *NcoI*/*XhoI* sites and replacing *LEU2* with *ADE2* (Supplementary Data). The *ADE2* gene was amplified from yeast genomic DNA using the primers 5'-AAT GCA ATC GAT TAA CGC CGT ATC GTG ATT AAC-3' and 5'-ACG TAA GCG GCC GCC GCT ATC CTC GGT TCT GC-3'. Zebrafish *FoxI1* cDNA coding region was subcloned into pYoh-1 at the multiple cloning site and fused with the Gal4 activation domain and the HA epitope tag [originally described in Ref. (21)]. Yeast expression plasmids for human p53 have been described previously (22,23).

Transcription factor plasmids were introduced into *MAT $\alpha$*  yeast strains W303 (kindly provided by Carl Wu's laboratory; genotype *ade2-1, can1-100, his3-11,15, leu2-3,112, trp1-1, ura3-1*) or BY4735 (*ade2 $\Delta$ ::hisG his3 $\Delta$ 200 leu2 $\Delta$ 0 met15 $\Delta$ 0 trp1 $\Delta$ 63 ura3 $\Delta$ 0*) (24). Standard yeast manipulations were performed as described (25). Fusion protein expression of FoxI1 was confirmed by western blot using anti-HA tag antibody.

### Yeast genomic libraries

To generate the genomic libraries, the *URA3* reporter plasmid pHQ366 (26) was modified by replacing the

*PstI*-p53-binding site-EcoRI linker with a new *PstI*-EcoRI linker (5'-GTA TCT CGA GG-3' and 5'-AAT TCC TCG AGA TAC TGC A-3'), yielding plasmid pYoh366. Zebrafish or mouse genomic DNA was partially *Tsp509I*-digested, gel-purified and cloned into pYoh366 at the EcoRI site. Ligation reactions were ethanol precipitated and electroporated into *Escherichia coli* ElectroMAX DH5 $\alpha$ -E cells (Invitrogen). The library complexity was assessed by counting a serial dilution of transformants on LB-ampicillin plates. The remainder was plated on large LB-ampicillin plates, allowed to grow and washed into a flask containing LB media. Plasmids were isolated using standard procedures. Yeast strain BY404 (*MAT $\alpha$  ade2 $\Delta$ ::hisG his3 $\Delta$ 200 leu2 $\Delta$ 0 trp1 $\Delta$ 63 ura3 $\Delta$ 0*) (24) was transformed with genomic fragment-containing pYoh366 and frozen in aliquots ( $5 \times 10^8$  cells/ml).

### Yeast mating

We used the method from Ref. (27) as an optimized mating protocol, originally designed for yeast two-hybrid screening. Mating of two haploid yeast strains of opposite mating type, each harboring one of the respective plasmids, results in the formation of doubly transformed diploid zygotes. An aliquot of BY404 with the genomic library in pYoh366 was combined with a freshly raised culture of p53-expressing or FoxI1-expressing yeast (*MAT $\alpha$* ) and then subjected to the standard mating procedures. Mating efficiency was calculated by the number of diploid colonies on plates selecting for diploid yeast per total colony number on YPAD or YPD plates. When *ADE2* plasmids were used (i.e. the FoxI1 screening), yeast expressing the transcription factor also contained an empty *HIS3* plasmid. This allowed for selection of diploid yeast on the actual screening plates (-his, -trp, -ura) while using the *ADE2*-based color assay for detection of the transcription factor plasmid.

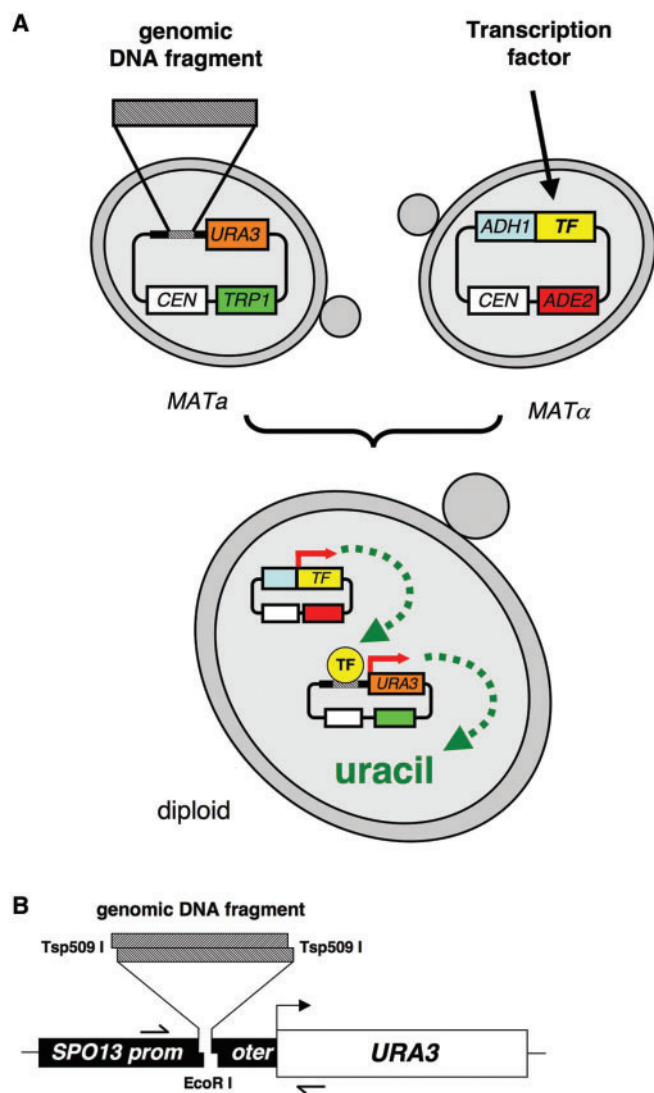
### Sequencing of positive clones

Library plasmids were rescued from positive yeast clones as described (28) and sequenced using Forward366 (5'-GCG CTT TAA GAG AAA ATA TTT GTC CTG-3') and Reverse366 (5'-CGG CTA TTT CTC AAT ATA CTC CTA ATT AAT AC-3'). Alternatively, the genomic library fragments were PCR amplified directly from yeast using Forward366 and ReverseUra (5'-GTA GCA GCA CGT TCC TTA TAT GTA GC-3'). ReverseUra targets the plasmid in a region outside the *SPO13* promoter. Before sequencing, 20  $\mu$ l of PCR product were treated with 0.3 U shrimp alkaline phosphatase (Amersham) and 3 U EXO I (USB) in 20 mM Tris–HCl, pH 8.0 and 10 mM MgCl<sub>2</sub> for 1 h at 37°C and 80°C for 15 min. The fragments were diluted with 95  $\mu$ l of ddH<sub>2</sub>O and sequenced with nested primer Reverse366.

## RESULTS

### General strategy in yeast for genome-wide screening for binding sites of DNA-binding proteins

We devised an improved assay system to perform whole genome screens for transcription factor-binding sites.



**Figure 1.** (A) Strategy of genomic library screens in yeast to identify binding sites for transcription factors. The transcription factor of interest is expressed in haploid yeast of mating type *MAT* $\alpha$ . A library of genomic fragments cloned upstream of *URA3* in the reporter plasmid is maintained in *MAT**a* yeast. Yeast mating combines the two plasmids resulting in *URA3* activation and growth on SC-Ura plates if the transcription factor binds to the genomic DNA fragment. (B) Details of the cloning site for genomic fragments. The *SPO13* promoter represses basal transcription and the *Tsp509I* fragments are cloned into an engineered *EcoRI* site. Primers for amplification of the genomic fragments directly from the yeast colonies are indicated with half arrows. By placing one primer in the *URA3* gene and the other in the *SPO13* promoter, it prevents amplification of the endogenous sequences.

The screens in yeast yield minimal background and quickly eliminate false-positives (Figure 1). In essence, a transcription factor is tested against random genomic fragments to isolate DNA that can be directly bound by the protein of interest. By fusing the transcription factor to the activating domain of the yeast transcription factor GAL4 (GAL4AD), the protein will become a transcriptional activator, regardless of the normal role it plays *in vivo* or any co-factors that it would normally need to activate transcription. Binding of the transcription factor

to the genomic DNA fragment results in activation of the *URA3* reporter gene and growth of yeast on plates lacking uracil.

The screening efficiency relies on several key aspects: (i) high-quality libraries of genomic DNA with potentially several-fold coverage of the zebrafish and mouse genomes; (ii) tight repression of the *SPO13* promoter upstream of the *URA3* reporter gene; (iii) negative selection of genomic library yeast in 5-FOA containing media prior to screens resulting in few false-positives; (iv) *ADE2*-based expression plasmids for both auxotrophic selection and a visual color 'sectoring' assay; (v) maintenance of libraries and transcription factors in haploid yeast of opposite mating type for fast and easy execution of library screens and (vi) PCR amplification of the genomic fragments directly from yeast for rapid sequencing and reanalysis of the positive genomic fragments.

### Construction of whole genome libraries for zebrafish and mouse

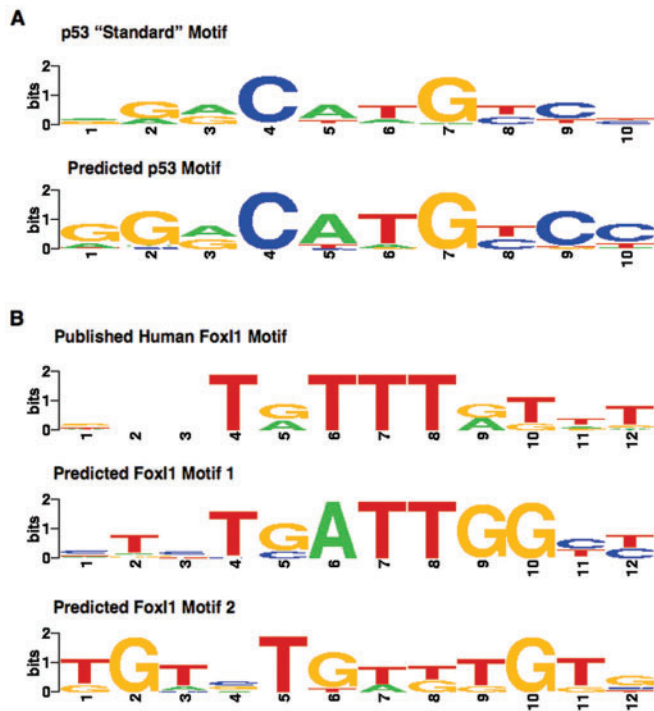
Libraries were constructed in plasmid pYoh366 that contains the *URA3* reporter gene downstream of the *SPO13* promoter (28) in a *CEN* plasmid (one copy per yeast cell) with the selective marker gene *TRP1* (Figure 1) (29). Genomic DNA from zebrafish and mouse was partially digested with *Tsp509I*. Genomic fragments of ~500 bp in length were gel-purified and cloned into the *EcoRI* site of pYoh366. A total of  $3 \times 10^7$  independent clones with an average size of 300 bp were obtained for the zebrafish library, an approximate 4- to 6-fold coverage of the zebrafish genome. A total of  $1.7 \times 10^7$  independent clones with an average size of 700 bp were obtained for the mouse library, representing an approximately 3- to 4-fold coverage of the mouse genome. For both libraries, >90% of plasmids had inserts. The libraries were transformed into yeast strain BY404 (*MAT**a*) (24), and  $>10^8$  independent colonies from each library were pooled and frozen in aliquots.

### Yeast expression plasmids for DNA-binding proteins

Several marker genes (such as *HIS3* or *LEU2*) are available for yeast expression plasmids. These markers allow for screens that involve more than one transcription factor. The marker gene *ADE2* is particularly advantageous because loss of the plasmid can be monitored by accumulation of a red adenine precursor on plates with limiting amounts of adenine (30). A colony that grows on plates lacking uracil yet shows red sectors or is completely red indicates that part or all of the colony has lost the transcription factor plasmid, easily identifying it as a false-positive clone.

### Whole-genome screens for DNA-binding proteins in yeast

Screens were performed according to the outline in Figure 2. A total of  $2 \times 10^8$  haploid yeast of mating type *MAT**a* containing the genomic library were mated on non-selective YPD plates with an equal number of haploid yeast (*MAT* $\alpha$ ) containing the expression plasmid for a transcription factor. This approach is significantly more efficient than transforming the library into yeast



**Figure 2.** Definition of p53 and FoxI1 consensus DNA-binding sites based on genomic fragments isolated in yeast screens. **(A)** A 'standard' p53 PSSM (position-specific scoring matrix) model was obtained based on alignment of 162 experimentally verified p53 DNA-binding sites. A motif was predicted by applying the CONSENSUS program to sequences of 94 library clones and was compared to the standard p53 DNA-binding site model. The logos of the standard motif and predicted motif are shown. **(B)** The published human FoxI1-binding site compared to the CONSENSUS predicted motifs of zebrafish FoxI1. Motif 1 is similar except for a noticeable difference at position 6 (T to A). Motif 2 matches the published consensus but with slightly weaker preferences (core binding sequence is shifted from 4–8 to 5–9 in graph).

expressing the target transcription factor, because (i) maintaining the library in yeast makes it an easily renewable resource and (ii) mating assays are more convenient and efficient than DNA transformation. The typical mating efficiency was 5–10%, yielding  $>10^7$  diploid yeast per screen. After 3 days of growth at 30°C, Ura<sup>+</sup> clones (non-sectored white colonies in the case of *ADE2* plasmids) were single-colony purified and checked for plasmid-dependency of the Ura<sup>+</sup> phenotype. The library plasmids were rescued, transformed into BY404, retested for the Ura<sup>+</sup> phenotype in mating assays for the transcription factor, sequenced and analyzed computationally.

### Characterization of the mouse genomic DNA library using p53

p53 is a tumor suppressor protein that is activated in response to cellular stresses. It has well-established DNA-binding characteristics (2,31). All experiments were performed with the complete mouse library of  $1.7 \times 10^7$  independent genomic fragments. Growing the yeast containing the genomic library in 5-FOA-containing media prior to the screen eliminated background from

'self-activating' fragments almost entirely (18 Ura<sup>+</sup> clones per  $1 \times 10^6$  yeast).

We performed a screen with p53 under fully optimized conditions. Yeast with an empty control plasmid was processed in parallel to determine the background of false-positives. We obtained 330 Ura<sup>+</sup> colonies per  $1 \times 10^6$  diploid yeast for p53, whereas yeast with an empty plasmid yielded only two colonies per  $1 \times 10^6$ .

The very low level of false-positives in our assay allowed us to streamline the screening procedure significantly. Once Ura<sup>+</sup> clones emerged after 2–5 days, a PCR reaction was used to amplify the genomic insert directly from yeast. One primer hybridized to the upstream *SPO13* sequence and one to the downstream *URA3* sequence so that only plasmid, but not genomic yeast, DNA was amplified. The PCR product was reintroduced into haploid yeast by co-transformation with a partially overlapping gapped plasmid, and the reporter gene plasmid was generated in yeast by homologous recombination (32). The phenotype of the genomic fragment was confirmed in yeast after mating to haploid yeast with the transcription factor. In parallel, the PCR product was purified and sequenced. This approach reduces manual processing of yeast clones, speeds up screens and is adaptable to high-throughput screening.

### Characterization of the zebrafish genomic DNA library using FoxI1

FoxI1 is a forkhead class transcription factor involved in proper organization of the zebrafish and mouse otic vesicle and zebrafish cranio-facial cartilage during early embryonic development (33–36). Because FoxI1 can be either an activator or a repressor *in vivo* (21), we fused the open reading frame of FoxI1 to the Gal4 activating domain (to ensure transcriptional activation) and screened the zebrafish library (to ensure transcriptional activation) and screened the zebrafish library ( $3 \times 10^7$  unique genomic inserts, 4- to 6-fold coverage of the zebrafish genome). In parallel, we performed a control-mating assay for haploid yeast with an empty plasmid. For this particular screen, we cloned the *FoxI1* open reading frame into the pYOH-1 (modified from the Clontech plasmid pACT-1) plasmid marked with *ADE2*. False-positives will not need to maintain the *FoxI1* plasmid to grow on SC-Ura plates, so the colonies would take on a 'sectored' or all red appearance. True-positives should remain completely white.

Using the approach of Figure 1 and described for p53,  $2.5 \times 10^7$  diploid yeast were screened on plates lacking uracil for 5 days at 30°C. The negative control resulted in 14 false-positive Ura<sup>+</sup> colonies per  $1 \times 10^6$  diploid yeast screened. In contrast, FoxI1 resulted in 132 Ura<sup>+</sup> colonies per  $1 \times 10^6$ . A total of 710 white colonies were single colony purified. We PCR-amplified the genomic inserts directly from yeast and sequenced the fragments with nested PCR primers. The fragments were simultaneously re-tested in yeast as described before. All 710 fragments were again Ura<sup>+</sup> and dependent on FoxI1 for growth.

### Computational analysis of the genomic fragments for consensus DNA-binding sites

*p53*. Of the 140 starting clones, 37 were duplicates and 3 had no insert. Of the remaining 100, 90 clones could be fully sequenced using primers up- and downstream of the genomic fragments, and 10 clones had an interior sequence gap. Sequences from the reference mouse genome were retrieved to fill the gap. The remaining six clones were chimerical and could not be mapped precisely because the sequencing gap included the region where the two distinct genomic fragments join. The February 2006 mouse genome assembly and BLAT at <http://genome.ucsc.edu/> were used to map the 94 complete clones. Thirty-five clones could not be mapped because they contained exclusively repetitive sequences. Of the remaining 59 clones, 33 were mapped to a single and 26 split clones to more than one genomic locus (24 with 2, 1 with 3 and 1 with 4 fragments). Chimerical clones were a result of the library construction because the genomic fragments had compatible ends. In total, because of some chimerism, 59 clones mapped to 81 unique loci. Typically, chimeric clones had a predicted binding site in only one of the two (or more) genomic fragments in the clone. In summary, ~70% of library plasmids (94/140) were useful for the p53 DNA-binding site analysis and ~40% could be uniquely mapped to the genome.

The consensus p53 DNA-binding site is two half-sites of RRRCWWGYYY separated by 0–13 bp (2). We built a new matrix from a larger data set of 162 p53 DNA-binding sites. According to the site predicted by our 162 known p53-binding sites, the spacing of 0 between the p53 half-sites is much preferred over any other spacing. This result was confirmed in yeast assays (26). A logo of this half-site model is shown in Figure 2A.

We applied the CONSENSUS program (37) to 94 isolated clones with complete sequence data. The best motif is a 10 bp approximate palindrome that matches our standard p53 DNA-binding site model perfectly with a *P*-value of  $1.73 \times 10^{-6}$  (Figure 2A).

The p53 motif signal was very strong when analyzing the complete set of 94 clones. To test whether we can obtain an accurate motif for p53 with fewer sequences, we generated random subsets of the 94 sequences ranging from 5 to 50 and applied CONSENSUS to these data sets. With 20 or more sequences, the top predictions matched perfectly with our p53 DNA-binding site model at least 70% of the time. We then added two additional motif-finding programs, Gibbs sampler (38) and Projection (39). The combined sensitivity of these three programs was 100% on 10 random subsets of 20 sequences each indicating that just 20 genomic fragments per transcription factor may be sufficient to characterize the majority of putative transcription factors.

We mapped the likely p53 DNA-binding site in the library sequences. Out of the 94 complete library clones, 81 contained at least one perfect p53 DNA-binding site. Some clones contained as many as five predicted sites. Eleven of the remaining clones contained a good match to the p53 consensus DNA-binding site, but scored below

our very stringent cutoff value. Two clones contained only one good p53 half-site.

*FoxI1*. A total of 154 unique sequences were obtained from the 710 total clones. One hundred and thirty-four of the 154 unique sequences could be mapped to the zebrafish genome (zebrafish build Zv6). Occasionally a sequence mapped to more than one location with 100% identity, these occurrences can often be explained by errors in the current zebrafish genomic build. In general, mapping was less ‘robust’ because of high polymorphism rates in the zebrafish genome and the less ‘complete’ nature of the assembly when compared to mouse. Approximately 50% of the clones were chimeric consisting of two different fragments fused together (Supplementary Data).

Human FoxI1 (HFH-3, FREAC6, HNF-3, FKHL10) binds to the motif TRTTTRKDD as determined by SELEX enrichment (40). The CONSENSUS program identified two common motifs in the isolated fragments. The less frequent motif was identical to the published binding site (Figure 2B), however, the more commonly represented binding site, TSATTGGYY, while similar, had some obvious differences (Figure 2B), particularly the presence of an A in position 6 that is a T in the published consensus. This could either reflect slight differences in the preference of FoxI1 binding between the human and zebrafish forms, or differences in binding preference in the context of histone packaging *in vivo*. Forkhead class transcription factors have been shown to bind more stably to DNA in the context of histones (41), this may change the recognition site preferences when compared to *in vitro* determinations. Thirteen fragments did not contain either predicted motif. The discovery of another putative binding site for FoxI1 as well as the confirmation of the previously published site reinforces the value of the technique. It is of particular note that motif 1 contains the NF-Y core sequence CCAAT, common in most eukaryotic promoters (42). It is unclear what the presence of this sequence signifies, but NF-Y has been shown to be involved in DNA compaction (43), similar to the role described for FoxI1 (21), and like forkhead proteins, has a structure that is similar to histones (44,45). Further study is needed to determine the significance of the CCAAT motif.

### Identification of putative target genes

*p53*. Fifty-nine unique clones identified 81 fragments with unique genomic locations because several clones were chimeric. Of the 56 fragments with a good to excellent p53 DNA-binding site, 10 mapped to introns. The remaining fragments were found several thousand base pairs up- or downstream of genes, similar to the findings of ChIP studies for p53 (46,47) (Supplementary Data). These 46 fragments would not have been sampled by typical ‘promoter’ arrays. From the list of genes neighboring the isolated fragments with p53 DNA-binding sites, six genes (*Sec61a2*, *Ass1*, *Aldh2*, *Kit1*, *Ela3b* and *Rprm*) are known to be up-regulated in a p53-dependent fashion, but their p53 DNA-binding sites have not been reported (48–51).

*FoxI1*. Because of chimerism, 154 unique sequences were mapped to 134 genomic loci. The remaining fragments were unmappable (Supplementary Data). There were only 20 instances where the fragment mapped within 10kb of the transcriptional start site. This is consistent with a proposed second role for FoxI1 as a global chromatin remodeling protein (21). There are not many genes known to be directly regulated by FoxI1, but within the small set of genes that have FoxI1 sites <10kb away, two genes, *Lhx3* and *Cldna*, have roles in ear development (52,53). This makes them candidates for regulation by FoxI1 since FoxI1 has a known role in ear development (34–36,54).

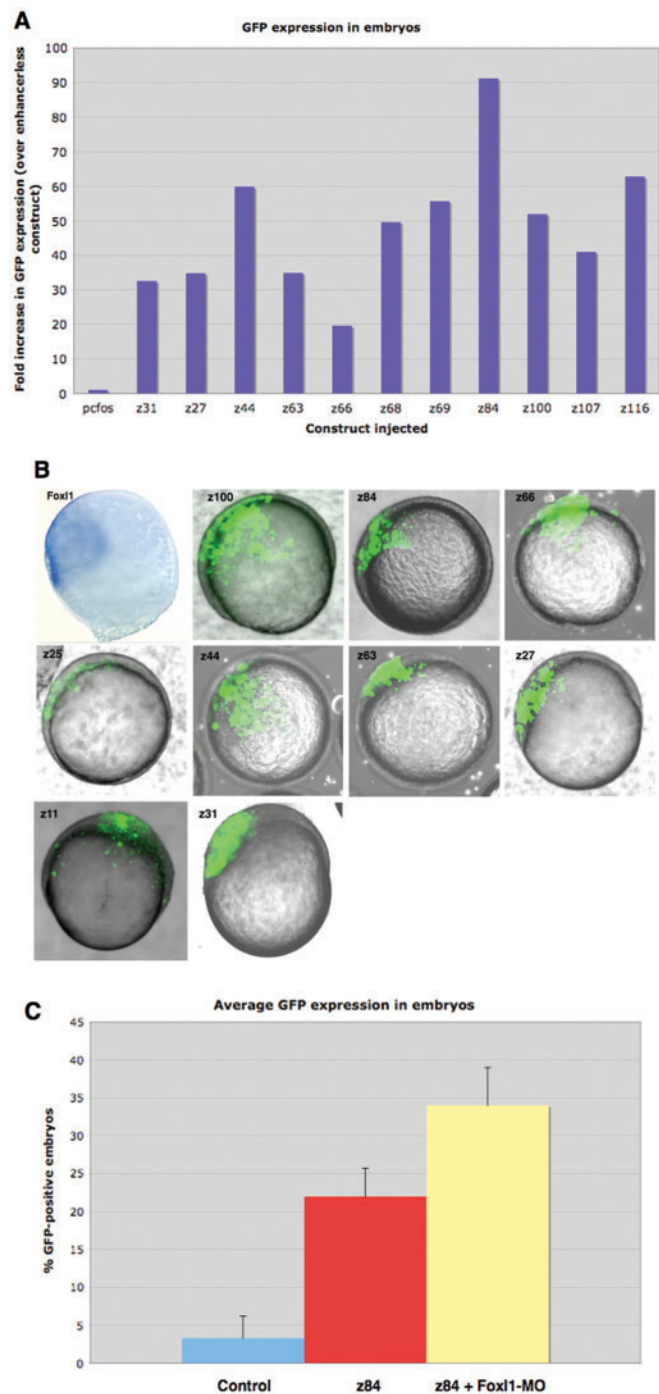
### *In vivo* testing of zebrafish fragments

Recently developed techniques allow researchers to use zebrafish embryos as a rapid readout for testing *in vivo* activity of putative *cis*-regulatory elements (55,56). We selected 13 fragments and subcloned them into a TOL2-based transposon vector containing Gateway™ compatible cloning sites upstream of a minimal promoter and *GFP* (gift of A. McCallion). All 13 fragments showed a significant increase in GFP expression compared to empty vector alone (Figure 3A). This demonstrated a specific transcriptional response dependent on the inserted genomic fragment since previous research has shown that random genomic fragments do not consistently activate GFP expression (55,56). Typically, expression of GFP was similar to the known expression pattern of FoxI1 (Figure 3B). In order to demonstrate that transactivation was dependent on FoxI1, we selected two fragments, z84 and z11, and tested them in the presence or absence of morpholino oligonucleotides that inhibit FoxI1 expression (OpenBiosystems). Fragment z11 showed no significant differences in the presence or absence of embryonic FoxI1. However, fragment z84 showed a significant increase in GFP expression when FoxI1 protein was inhibited (Figure 3C,  $P < 0.001$ ). We have previously demonstrated that FoxI1 is capable of remodeling chromatin higher order structure and only a few genes are activated or inhibited by FoxI1 expression (21). In this study, we further found that FoxI1 has an inhibitory role on GFP expression in the context of certain genomic fragments. It is likely that the FoxI1 binding is crucial to recruit other transcription factor to the sites on those fragments that are acting as positive and/or negative regulators.

In summary, the fragments isolated using the yeast technique showed strong enrichment for *cis*-regulatory elements (13/13 tested) and we have demonstrated in one instance (of two tested) that FoxI1 is responsible for transcriptional regulation from the fragment.

## DISCUSSION

We present an approach in yeast that can rapidly identify the target binding sequence of nearly any transcription factor of interest, and simultaneously identify target sites that are biologically active and relevant to the function of that transcription factor. This redesigned and optimized yeast approach combines three important elements: (i) yeast mating (instead of transformation) for efficient



**Figure 3.** *In vivo* testing of isolated FoxI1 fragments. (A) Relative GFP activation compared to the empty TOL2 construct (pcfoss). (B) Comparison of GFP activation from several of the fragments compared to the *in situ* expression pattern of FoxI1 (upper left). Some genomic fragments mimic FoxI1 expression, e.g. z100, while others appear to have a different pattern, (z84, z11). (C) Testing of fragment z84 in the presence or absence of FoxI1 expression. z84 has clear activation over the control empty vector, but when co-injected with *FoxI1* targeting morpholinos, the expression increases significantly ( $P < 0.0001$ ).

screening of an entire genome in one pass; (ii) red/white sectoring of the *ADE2* gene for visual identification of false-positives and (iii) use of *URA3* as the reporter gene for expansion of libraries in 5-FOA to eliminate yeast with



'self-activating' fragments and reduce false-positives to near zero. In two test cases, p53 and FoxI1, a yeast-based screening approach proved to be a powerful and efficient way to study DNA-binding proteins. Recent studies using our p53-binding site data as a starting point, have identified endogenous retroviral (ERV) LTR's carrying functionally active p53-binding sites specific to the primate lineage, indicating the potential utility of this approach both to identify binding sites that are active *in vivo*, and to make novel discoveries about *cis*-regulatory regions in general (57). The genomic fragments can be directly PCR-amplified from yeast for both sequencing and immediate retesting in yeast using 'gap repair'. The approach is readily extended to hetero-dimeric or -multimeric transcription factor complexes.

This technique provides a low-cost, potentially high-throughput and rapid assessment of DNA-binding targets for transcription factors that nicely complements the current technologies, such as ChIP-Seq. We have shown that DNA-binding sites can be predicted in as few as 20 isolated sequences and the identified fragments have a very high likelihood of being biologically active.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Carl Wu for providing reagents. Support for T.W. is from the Helen Hay Whitney Foundation. G.S. is supported through funding from NIH grant HG00249. This research was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (S.B.). Funding to pay the Open Access publication charges for this article was provided by the National Human Genome Research Institute.

*Conflict of interest statement.* None declared.

## REFERENCES

- Hogbom, E., Magnusson, A.C. and Leanderson, T. (1991) 'E-boxes' as promoter elements in B cell lines and untransformed B lymphocytes. *Scand. J. Immunol.*, **33**, 615–620.
- el-Deiry, W.S., Kern, S.E., Pietenpol, J.A., Kinzler, K.W. and Vogelstein, B. (1992) Definition of a consensus binding site for p53. *Nat. Genet.*, **1**, 45–49.
- Klug, S.J. and Famulok, M. (1994) All you wanted to know about SELEX. *Mol. Biol. Rep.*, **20**, 97–107.
- Kim, S., Ip, H.S., Lu, M.M., Clendenin, C. and Parmacek, M.S. (1997) A serum response factor-dependent transcriptional regulatory program identifies distinct smooth muscle cell sublineages. *Mol. Cell. Biol.*, **17**, 2266–2278.
- Mintseris, J. and Eisen, M.B. (2006) Design of a combinatorial DNA microarray for protein-DNA interaction studies. *BMC Bioinformatics*, **7**, 429.
- Kummerfeld, S.K. and Teichmann, S.A. (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res.*, **34**, D74–D81.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Messina, D.N., Glasscock, J., Gish, W. and Lovett, M. (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.*, **14**, 2041–2047.
- Tupler, R., Perini, G. and Green, M.R. (2001) Expressing the human genome. *Nature*, **409**, 832–833.
- Fields, S. and Song, O. (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245–246.
- Brachmann, R.K. and Boeke, J.D. (1997) Tag games in yeast: the two-hybrid system and beyond. *Curr. Opin. Biotechnol.*, **8**, 561–568.
- Parsons, A.B., Geyer, R., Hughes, T.R. and Boone, C. (2003) Yeast genomics and proteomics in drug discovery and target validation. *Prog. Cell. Cycle Res.*, **5**, 159–166.
- Simon, J.A. and Bedalov, A. (2004) Yeast as a model system for anticancer drug discovery. *Nat. Rev. Cancer*, **4**, 481–492.
- Wilson, T.E., Fahrner, T.J., Johnston, M. and Milbrandt, J. (1991) Identification of the DNA binding site for NGFI-B by genetic selection in yeast. *Science*, **252**, 1296–1300.
- Li, J.J. and Herskowitz, I. (1993) Isolation of ORC6, a component of the yeast origin recognition complex by a one-hybrid system. *Science*, **262**, 1870–1874.
- Wang, M.M. and Reed, R.R. (1993) Molecular cloning of the olfactory neuronal transcription factor Olf-1 by genetic selection in yeast. *Nature*, **364**, 121–126.
- Tokino, T., Thiagalingam, S., el-Deiry, W.S., Waldman, T., Kinzler, K.W. and Vogelstein, B. (1994) p53 tagged sites from human genomic DNA. *Hum. Mol. Genet.*, **3**, 1537–1542.
- Mastick, G.S., McKay, R., Oligino, T., Donovan, K. and Lopez, A.J. (1995) Identification of target genes regulated by homeotic proteins in *Drosophila melanogaster* through genetic selection of Ultrabithorax protein-binding sites in yeast. *Genetics*, **139**, 349–363.
- Deplancke, B., Mukhopadhyay, A., Ao, W., Elewa, A.M., Grove, C.A., Martinez, N.J., Sequerra, R., Doucette-Stamm, L., Reece-Hoyes, J.S. *et al.* (2006) A gene-centered *C. elegans* protein-DNA interaction network. *Cell*, **125**, 1193–1205.
- Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
- Yan, J., Xu, L., Crawford, G., Wang, Z. and Burgess, S.M. (2006) The forkhead transcription factor FoxI1 remains bound to condensed mitotic chromosomes and stably remodels chromatin structure. *Mol. Cell. Biol.*, **26**, 155–168.
- Baroni, T.E., Wang, T., Qian, H., Dearth, L.R., Truong, L.N., Zeng, J., Denes, A.E., Chen, S.W. and Brachmann, R.K. (2004) A global suppressor motif for p53 cancer mutants. *Proc. Natl Acad. Sci. USA*, **101**, 4930–4935.
- Dearth, L.R., Qian, H., Wang, T., Baroni, T.E., Zeng, J., Chen, S.W., Yi, S.Y. and Brachmann, R.K. (2006) Inactive full-length p53 mutants lacking dominant wild-type p53 inhibition highlight loss-of-heterozygosity as an important aspect of p53 status in human cancers. *Carcinogenesis*, **28**, 289–298.
- Brachmann, C.B., Davies, A., Cost, G.J., Caputo, E., Li, J., Hieter, P. and Boeke, J.D. (1998) Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast*, **14**, 115–132.
- Sherman, F. (1991) Getting started with yeast. *Methods Enzymol.*, **194**, 3–21.
- Qian, H., Wang, T., Naumovski, L., Lopez, C.D. and Brachmann, R.K. (2002) Groups of p53 target genes involved in specific p53 downstream effects cluster into different classes of DNA binding sites. *Oncogene*, **21**, 7901–7911.
- Soellick, T.R. and Uhrig, J.F. (2001) Development of an optimized interaction-mating protocol for large-scale yeast two-hybrid analyses. *Genome Biol.*, **2**, research0052.1–0052.7.
- Vidal, M., Brachmann, R.K., Fattaey, A., Harlow, E. and Boeke, J.D. (1996) Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and DNA-protein interactions. *Proc Natl Acad. Sci. USA*, **93**, 10315–10320.

29. Sikorski, R.S. and Hieter, P. (1989) A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics*, **122**, 19–27.
30. Hieter, P., Mann, C., Snyder, M. and Davis, R.W. (1985) Mitotic stability of yeast chromosomes: a colony color assay that measures nondisjunction and chromosome loss. *Cell*, **40**, 381–392.
31. Funk, W.D., Pak, D.T., Karas, R.H., Wright, W.E. and Shay, J.W. (1992) A transcriptionally active DNA-binding site for human p53 protein complexes. *Mol. Cell. Biol.*, **12**, 2866–2871.
32. Ma, H., Kunes, S., Schatz, P.J. and Botstein, D. (1987) Plasmid construction by homologous recombination in yeast. *Gene*, **58**, 201–216.
33. Hulander, M., Wurst, W., Carlsson, P. and Enerback, S. (1998) The winged helix transcription factor Fkh10 is required for normal development of the inner ear. *Nat. Genet.*, **20**, 374–376.
34. Lee, S.A., Shen, E.L., Fiser, A., Sali, A. and Guo, S. (2003) The zebrafish forkhead transcription factor Foxi1 specifies epibranchial placode-derived sensory neurons. *Development*, **130**, 2669–2679.
35. Nissen, R.M., Yan, J., Amsterdam, A., Hopkins, N. and Burgess, S.M. (2003) Zebrafish foxi one modulates cellular responses to Fgf signaling required for the integrity of ear and jaw patterning. *Development*, **130**, 2543–2554.
36. Solomon, K.S., Kudoh, T., Dawid, I.B. and Fritz, A. (2003) Zebrafish foxi1 mediates otic placode formation and jaw development. *Development*, **130**, 929–940.
37. Hertz, G.Z. and Stormo, G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
38. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
39. Buhler, J. and Tompa, M. (2002) Finding motifs using random projections. *J. Comput. Biol.*, **9**, 225–242.
40. Overdier, D.G., Ye, H., Peterson, R.S., Clevidence, D.E. and Costa, R.H. (1997) The winged helix transcriptional activator HFH-3 is expressed in the distal tubules of embryonic and adult mouse kidney. *J. Biol. Chem.*, **272**, 13725–13730.
41. Cirillo, L.A. and Zaret, K.S. (1999) An early developmental transcription factor complex that is more stable on nucleosome core particles than on free DNA. *Molecular Cell*, **4**, 961–969.
42. Mantovani, R. (1998) A survey of 178 NF-Y binding CCAAT boxes. *Nucleic Acids Res.*, **26**, 1135–1143.
43. Guerra, R.F., Imperadori, L., Mantovani, R., Dunlap, D.D. and Finzi, L. (2007) DNA compaction by the nuclear factor-Y. *Biophys. J.*, **93**, 176–182.
44. Clark, K.L., Halay, E.D., Lai, E. and Burley, S.K. (1993) Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature*, **364**, 412–420.
45. Romier, C., Cocchiarella, F., Mantovani, R. and Moras, D. (2003) The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y. *J. Biol. Chem.*, **278**, 1336–1345.
46. Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J. *et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.
47. Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y. *et al.* (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell*, **124**, 207–219.
48. Burns, T.F. and El-Deiry, W.S. (2003) Microarray analysis of p53 target gene expression patterns in the spleen and thymus in response to ionizing radiation. *Cancer Biol. Ther.*, **2**, 431–443.
49. Lin, S.W., Cook, M., Finnberg, N., Bernhard, E., Baldwin, D. and El-Deiry, W.S. (2004) p53-dependent induction of serine proteases in irradiated mouse colon. *Cancer Biol. Ther.*, **3**, 1290–1297.
50. Ohki, R., Nemoto, J., Murasawa, H., Oda, E., Inazawa, J., Tanaka, N. and Taniguchi, T. (2000) Reprimo, a new candidate mediator of the p53-mediated cell cycle arrest at the G2 phase. *J. Biol. Chem.*, **275**, 22627–22630.
51. Sax, J.K., Stoddard, A., Murphy, M.E., Chodosh, L. and El-Deiry, W.S. (2003) Microarray expression profiling of p53-dependent transcriptional changes in an immortalized mouse embryo fibroblast cell line. *Cancer Biol. Ther.*, **2**, 416–430.
52. Kollmar, R., Nakamura, S.K., Kappler, J.A. and Hudspeth, A.J. (2001) Expression and phylogeny of claudins in vertebrate primordia. *Proc. Natl Acad. Sci. USA*, **98**, 10196–10201.
53. Hertzano, R., Dror, A.A., Montcouquiol, M., Ahmed, Z.M., Ellsworth, B., Camper, S., Friedman, T.B., Kelley, M.W. and Avraham, K.B. (2007) Lhx3, a LIM domain transcription factor, is regulated by Pou4f3 in the auditory but not in the vestibular system. *Eur. J. Neurosci.*, **25**, 999–1005.
54. Hulander, M., Kiernan, A.E., Blomqvist, S.R., Carlsson, P., Samuelsson, E.J., Johansson, B.R., Steel, K.P. and Enerback, S. (2003) Lack of pendrin expression leads to deafness and expansion of the endolymphatic compartment in inner ears of Foxi1 null mutant mice. *Development*, **130**, 2013–2025.
55. Allende, M.L., Manzanares, M., Tena, J.J., Feijoo, C.G. and Gomez-Skarmeta, J.L. (2006) Cracking the genome's second code: enhancer detection by combined phylogenetic footprinting and transgenic fish and frog embryos. *Methods*, **39**, 212–219.
56. Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L. and McCallion, A.S. (2006) Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science*, **312**, 276–279.
57. Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess, S.M., Brachmann, R.K. and Haussler, D. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl Acad. Sci. USA*, **104**, 18613–18618.