Award Number: W81XWH-07-1-0044

TITLE: Identification of the Transformational Properties and Transcriptional Targets of the Oncogenic SRY Transcription Factor SOX4

PRINCIPAL INVESTIGATOR: Christopher Scharer

CONTRACTING ORGANIZATION: Emory University
Atlanta, GA 30322

REPORT DATE: January 2009

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: (Check one)

✓ Approved for public release; distribution unlimited

Distribution limited to U.S. Government agencies only; report contains proprietary information

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.
SOX4 is a critical developmental transcription factor in vertebrates and is required for precise proliferation and differentiation in multiple tissues. SOX4 has also been implicated in tumorigenesis of multiple tumor types and has been shown by our lab to be upregulated in prostate cancer. However, the exact molecular role and the genes SOX4 transcriptionally influences remain unknown. Using chromatin immunoprecipitation coupled to DNA microarrays (ChIP-chip) we have identified 282 high confidence SOX4 target genes and 3600 genomic binding sites. We have also used a unique protein binding, double-stranded DNA microarray to determine a novel SOX4 position-weight matrix for in silico binding site searches. Direct targets of SOX4 include key cellular regulators such as EGFR, Tenascin C, Frizzled-5 and Patched-1. In addition, SOX4 impacts other transcriptional networks through regulation of 23 transcription factors. We also show that SOX4 impacts the microRNA-processing pathway through direct regulation of Dicer and Argonaute 1. For the first time, these data provide a snapshot of the global transcriptional regulatory network of SOX4 in prostate cancer.
Introduction:

SOX4 is a critical developmental transcription factor and is required for precise differentiation and proliferation in multiple tissues. SOX4 is a 47-kDa protein that is encoded by a single exon and contains a conserved high mobility group (HMG) DNA-binding domain (DBD) related to the TCF/LEF family of transcription factors. Our lab has previously shown SOX4 mRNA and protein to be overexpressed in prostate cancer, and this expression is correlated with increasing Gleason score. Other labs have shown SOX4 mRNA to be overexpressed in other tumors such as leukemia, melanoma, glioblastoma and bladder carcinomas. However, despite this knowledge little is known of the direct transcriptional targets of SOX4, and how misregulation of these networks affects human cancers and development. The goal of this research is to determine the transcriptional target genes of SOX4 and to determine SOX4’s role in murine prostate development. To determine the direct transcriptional targets on a global scale we performed chromatin immunoprecipitation coupled to DNA microarrays. We used human promoter arrays from NimbleGen, Inc. that tiled roughly 5 kb of promoter and intronic sequence for 25,000 known transcripts. In total the array tiled 110 Mb of DNA. Using this technique we were able to determine the genes with SOX4 bound at their promoter in living prostate cancer cells. Furthermore, expression profiling of prostate cancer cells overexpressing either SOX4 or a control vector identified those genes that are transcriptionally regulated by SOX4. We have also obtained a SOX4 floxed mouse that will enable the prostate specific deletion of SOX4 in mice. This information will determine if SOX4 is required for the development of a functional prostate. Determining the transcriptional targets and in vivo functions of SOX4 will contribute critical knowledge to the SOX4 field and further our understanding of SOX4’s role in development and carcinogenesis.
AIM 1: Determine the Direct Transcriptional Targets of SOX4 on a Global Scale using a ChIP-chip and microarray approach.

Chromatin immunoprecipitation (ChIP) relies on high quality, specific antibodies which can immunoprecipitated the protein of interest with little background. While commercial antibodies that recognize SOX4 in immunoblotting applications exist, none have shown activity in immunoprecipitations in our hands. Therefore, an HA epitope tag was introduced onto the N-terminus of SOX4 and cloned into a lentiviral vector for stable infection of mammalian cells (Fig. 1A). The lentivirus contains an eYFP gene to enable the purification by Flourescent Activated Cell Sorting (FACS) of stably infected cells. For both the LNCaP and RWPE-1 prostate cancer cell lines, both a control eYFP only and an HA-SOX4-eYFP cell line were created and infected cells FACS purified (Fig. 1B). Both cell lines were tested for transgene expression and to ensure HA-SOX4 could be immunoprecipitated using our 12CA5, anti-HA monoclonal antibody (Fig. 1C). ChIP assays were performed from the LNCaP-HA-SOX4 cells in triplicate and in duplicate from the LNCaP-YFP cells. DNA was extracted and purified according to published protocols, and amplified using a Ligation-mediated PCR approach (8). 4 ug of immunoprecipitated and total input DNA was sent to NimbleGen and hybridized to their 25K dual chip Promoter array. The array tiles roughly 5 Kb of promoter and intronic sequence for 25,000 known transcripts with a total coverage of 110 Mb.

Signal intensities were Z-score normalized, log2 transformed and ratios of immunoprecipitated to total input signal calculated for each probe set. ChIPOTle software (2) was used with a 500 bp sliding window to look for sets of neighboring probes that are enriched together. Peaks that overlapped in 2 of the three data sets and were not present in the control data set were identified and called SOX4 binding sites (Fig. 2A). This analysis identified 3,600 binding sites in the promoters of 3,470 different genes. 28 peaks were chosen and 10 verified by ChIP-quantitative Real-Time PCR (qPCR) and 18 by traditional ChIP-PCR. 24 of 28 sites (86%) chosen were specifically enriched in the LNCaP-HA-SOX4 cells and not in the control LNCaP-YFP cells (Fig. 2B and 2C). All validated peaks were also validated in the RWPE-1 cell line except ANKRD15, further validating the data set (Fig. 2C).

While SOX4 binds to the promoters of 3,470 different genes we do not know how SOX4 influences transcription of each gene. To identify genes whose expression changes when SOX4 levels are altered we performed whole genome expression profiling of LNCaP cells transfected with either control vector or HA-SOX4. In order to enrich for direct SOX4 targets, total RNA was harvested 24 hours post-transfection and profiled using an Illumina Human 6-v2 whole genome array. Compared to vector control, 1,766 genes were altered at least 1.5 fold when SOX4 was overexpressed in LNCaP cells (Fig. 3A). Ten of these genes were confirmed by qPCR (Fig. 3B) and SOX4’s induction of DICER1 was confirmed at the protein level (Fig. 3C). Previous expression profiling of LNCaP cells transfected with either control siRNA or SOX4 siRNA identified 465 down stream target genes for SOX4 (7). Combining these three data sets we identified 282 genes that had SOX4 bound to their promoter regions and were transcriptionally altered when SOX4 levels were perturbed (Fig. 3D). 9 genes overlapped in all three data sets (PIK4CA, DHX9, BTN3A3, CDK2, MVK, ADAM10, RYK, ISG20, and DBI). Although only 10% of the significant differentially expressed genes overlapped with the ChIP-chip data, this is likely a conservative estimate because the NimbleGen 25K promoter array only queries proximal promoter sequences. Thus, more of the 1,900 genes that responded to changes in SOX4 mRNA levels (but were not detected by ChIP-chip) could still be direct targets. Excellent candidates would be the 40 genes that responded to SOX4 on both microarray platforms, such as the IL6 receptor, SOX12, and NME1. Alternate methods such as ChIP-SEQ would provide a truly unbiased, genomic picture of SOX4 binding. Nevertheless, this is the first
global study of SOX4 binding and provides a foundation for understanding the SOX4 transcriptional network in prostate cancer.

HMG domain transcription factors bind AT rich DNA in the minor groove and two previous reports identify a 7mer SOX4 binding motif (15, 16). While this knowledge can aid in the search for putative binding sites it does not take into account the role of alternate bases at various positions. A SOX4 specific position-weight matrix is required to fully utilize the power of bioinformatic searches. Apart from the consensus core SOX family binding site WWCAAW, where W represents either A or T, little is known about what preferences SOX4 exhibits at each base position during binding (6). In order to facilitate bioinformatic searches for SOX4 DNA binding sites we sought to determine a SOX4 specific position-weight matrix (PWM) using a unique, protein-binding, double stranded DNA microarray (1). The array allows recombinant protein to interact with and bind, every possible 10mer, thus allowing in vitro binding site specificities to be calculated. We generated an N-terminal, GST-SOX4-DBD fusion protein, and expressed and purified it from E. coli (Fig. 4B). To ensure the purified recombinant fusion protein was functional we performed an electromobility shift assay (EMSA) using a published SOX4 binding site of AACAAAG (15). Increasing concentrations of GST-SOX4-DBD was incubated with radiolabeled specific probe alone, with a cold specific competitor or a cold non-specific competitor. GST-SOX4-DBD was able to bind the probe and cause a shift that was abolished when cold specific competitor probe, but not when cold non-specific probe was added (Fig. 4A). These data show that the truncated GST-SOX4-DBD fusion protein is functionally active in vitro. The GST-SOX4-DBD was incubated with the protein binding microarray and a novel PWM (RWYAAWRV) (R – A or G, Y – C or T, and V – G, A or C) was calculated according to published protocols (Fig. 4C) (1). Two groups have previously reported similar binding site sequences for SOX4: AACAAAG (15) and AACAAT (16). Our PWM confirms both of the previous known binding sites and adds new information on the binding preferences in the 8th position as well as alternate bases at the 6th and 7th positions.

Using our newly determined PWM we sought to establish if the peaks in the promoters of our SOX4 target genes are enriched for SOX4 binding sites. We applied CONFAC software (5) and analyzed the peaks in our 282 high-confidence target genes as well as 10 sets of random control promoter sequence. Control peaks of equal size were selected from at random from promoter sequences covered on the NimbleGen array and each control set represents equal sequence coverage as our 282 high-confidence peaks. With stringent criteria (core similarity: >0.85; matrix similarity: >0.75) we find that 60% of our high-confidence peaks contain SOX4 binding sites. SOX4 sites were significantly enriched compared to our 10 random control sets by Mann-Whitney U test with Benjamini correction for multiple hypothesis testing (q < 0.0019). To further characterize the data set we searched each of the 3,600 SOX4 binding sites and 10 sets of control peaks (assembled in the same manner as above) for the presence of Protein-binding Microarray (PBM) bound k-mers. These k-mers are the individual, ungapped 8mer sequences SOX4 bound on the PBM. The specificity of PBM k-mers can be defined by the enrichment score (ES), which ranges from -0.5 to 0.5 (10). We analyzed the enrichment of PBM k-mers with 0.45 > ES > 0.40 (moderate) and ES > 0.45 (stringent). Both SOX4-bound peaks and control peaks contained stringent and moderate k-mers, SOX4 bound peaks contained significantly more stringent (p = 0.0002) and moderate (p = 1.08 x 10^-5) k-mers by two-tailed Mann-Whitney test. SOX transcription factors have been reported to mediate their transcriptional activity through interactions with other transcription factors such as the SOX2-OCT3/4 pair (6). We applied CONFAC software to search for the presence of co-occurring binding sites enriched in our SOX4 peaks. Interestingly, the E2F family was the most frequently co-occurring motif (Table 1) and Ingenuity Pathway Assist1 identified cell-cycle as a functionally enriched process in the 3,470 SOX4 target genes. This suggests that part of SOX4’s function is to regulate genes involved in the cell-cycle progression. CONFAC also identified co-occurring binding sites for transcription factors involved in the TGFβ, WNT, and NF-κB pathways (Table 1). The presence of

1 http://www.ingenuity.com
WNT pathway transcription factors was particularly interesting considering a previous report that SOX4 co-operates with TCF4 and β-catenin to alter transcription (14). We confirmed this finding in LNCaP cells and found that SOX4 co-operates with β-catenin to increase transcription of a WNT reporter construct (Fig. 5).

In order to determine the biological processes and pathways enriched in SOX4 target genes we performed GO ontology analysis using DAVID software (3). As expected, the top annotated process was transcription \((p = 3.17 \times 10^{-18})\) but surprisingly we also find transmembrane \((5.59 \times 10^{-10})\) and protein phosphorylation/dephosphorylation \((3.5 \times 10^{-18}/6.6 \times 10^{-7})\) as enriched functions. DAVID software also identified 23 transcription factors as direct SOX4 target genes (Table 2). These data suggest that SOX4 modulates signaling networks at all three cellular levels: at the membrane, in the cytoplasm and inside the nucleus. IPA analysis identified biological pathways enriched in both the 3,470 direct target genes and the 1,766 genes altered when SOX4 is overexpressed. As expected the top annotations were cancer, cell-cycle and tissue development and SOX4 target genes were found to influence wide variety of developmental signaling pathways such as WNT, NOTCH, WNT-β-catenin, PI3K-AKT and the EGFR signaling network. Interestingly, microRNA processing enzymes DICER, AGO1 and the RNA helicase DHX9 were both direct target genes and showed expression changes when SOX4 was overexpressed. For the first time we report a link between a SOX family member and the microRNA processing pathway. Key SOX4 target genes and their cellular localization are illustrated in Figure 6A and 6B.

For a detailed discussion of these results see Appendix II and (13).

AIM2: Determine the effects of Loss or Overexpression in vivo

SOX4 is required for the development and differentiation of multiple murine tissues (4, 9, 11, 12, 17). We hypothesize that deletion of SOX4, specifically in the prostate, will affect normal murine prostate development. Dr. Neal Copeland has provided us with mice that contain the endogenous SOX4 allele flanked by LOXP sites to facilitate CRE mediated deletion of SOX4. Here at Emory we already have a colony of mice containing the CRE transgene driven by the prostate specific Probasin promoter. Probasin is initially expressed at the onset of puberty (roughly two weeks of age) in all lobes of the prostate and seminal vesicles and mostly epithelial cells (19). We initially obtained SOX4\(^{+/+}\) heterozygote mice and these mice are being bred to homozygosity as well as being crossed to the Probasin-CRE (Pb-CRE) mice to obtain homozygous SOX4 floxed males (SOX4\(^{+/+}\)) who are Pb-CRE positive. Currently we have obtained Pb-CRE negative, SOX4\(^{+/+}\) males and females as well as SOX4\(^{+/+}\), Pb-CRE positive males and females. Once male, prostate specific SOX4 knockout mice are obtained we will dissect out the prostate and harvest RNA and protein to assess SOX4 expression levels. This will provide a unique opportunity to investigate the expression status of direct SOX4 target genes predicted by our ChIP-chip analysis. Tissue sections will also be H&E stained for morphology analysis as well as immunohistochemical staining to determine the status of different prostate cell types.
Key Research Accomplishments:

- Expanded the known SOX4 target genes in the prostate to 282
- Identified 3,600 SOX4 binding sites in the proximal promoter of 3,470 different genes
- Developed a novel PBM k-mer based SOX4 binding site search algorithm in the perl programming language
- Identified biological pathways and processes SOX4 influences
- Significantly advanced the breeding of prostate specific SOX4 knockout mice

Reportable Outcomes:

- Manuscripts: The research presented in Aim 1 has been accepted for publication in Cancer Research and will be published on January 15, 2009 (13).
- Abstracts: The research in Aim 1 was presented as a poster at the 2008 Keystone meeting: Signaling Pathways in Cancer and Development.
- Presentations: All research presented in Aim I is presented as annually as an oral lecture as a requirement of my graduate program (Genetics and Molecular Biology).
- Database: All ChIP-chip and Expression profiling data has been deposited in the GEO database as required for publication under the Accession number: GEO11915
- Funding Application: All research presented in this report is part of an NIH Competitive Renewal application, applied for by my Principle Investigator Dr. Carlos Moreno.
- Training: As a student of the Genetics and Molecular Biology program I attend research seminars twice weekly and have taken 8 hours of course work comprising two classes: 1- a comprehensive Cancer Biology course, and 2- a introductory Bioinformatics course. My mentor and principle investigator, Dr. Carlos Moreno, has informally instructed me in the Perl Programming language as well as intensive direction in the analysis and data mining of microarray data from different platforms. In the next year I plan on writing and defending a dissertation consisting of the work presented in this report.

Conclusion:

In recent years various labs have utilized expression microarray data mining to identify a handful of SOX4 target genes. This report, for the first time, identifies the SOX4 target genes on a truly global scale. Interestingly, this data has highlighted a previously unknown function of SOX4. The vast array of transcription factor targets suggests SOX4 has a role in modulating other transcriptional programs towards a common goal. In vivo experiments presented in Aim 2 will aid our understanding of SOX4’s role in prostate development and the consequences of prostate specific ablation of SOX4 will be studied and linked to our transcriptional target data.

One draw back from our ChIP-chip approach was that our NimbleGen chip only contained proximal promoter sequences. SOX4 has been reported to bind at least one enhancer in T-cells (18) and most likely affects other enhancers in our prostate model. Performing either ChIP-SEQ or ChIP-chip using a whole genome tiling array would lend more insight and truly define a global SOX4 regulatory network. Of particular interest to our lab is SOX4’s role in WNT signaling. Our lab will explore the details of SOX4’s interaction with β-catenin and how this affects the target genes SOX4 affects.

SOX4 has been shown to be overexpressed in prostate cancer as well as many other types of human cancers such as melanoma, medulloblastomas, glioblastomas and leukemias. Identifying the transcriptional programs SOX4 controls is a first step in elucidating how SOX4 promotes
carcinogenesis and evaluating SOX4 as a potential drug target in prostate cancer and other malignancies.

References:

Appendices:

I. Curriculum Vitae

CHRISTOPHER SCHARER

Genetics and Molecular Biology
Graduate School of Arts and Sciences, GDBBS
Emory University
Atlanta, GA 30322
(404)712-2808

Education

Emory University, Atlanta, Georgia

- **Ph.D.** in Biomedical and Biological Sciences,
  - Program: Genetics and Molecular Biology - May, 2009
  - Dissertation: “Global Identification of Transcriptional Targets for SOX4 in Prostate Cancer.”
  - Advisor: Dr. Carlos S. Moreno
  - GPA: 4.0

Emory University, Atlanta, Georgia

- **B.S.** in Biology – May, 2004
  - GPA: 3.4

Academic Awards and Fellowships

**Department of Defense Predoctoral Training Grant in Prostate Cancer Research** – 2006 – 2009
GDBBS Student Symposium, 2nd Place Poster Award - 2008
GDBBS Excellence in Teaching Award – 2007
**NIH Predoctoral Training Grant** – GMB, 2005 - 2006
Thomas Aliberti Scholar/Athlete Award – 2004

Research Experience

**Doctoral Research:**
Genetics and Molecular Biology, GDBBS, Emory University, Atlanta, Georgia, 2004-2009
(Advisor: Dr. Carlos Moreno - cmoreno@emory.edu)

- Analysis of the transcriptional targets for the oncogenic transcription factor SOX4 using Chromatin Immunoprecipitation (ChIP), followed by DNA microarray and analysis with computational software developed by our lab.
- Investigation into the role of SOX4 in prostate cancer formation using both a prostate specific over-expression and knockout mouse model.
- Improve treatment options for recurrent ovarian cancer by investigating whether an Aurora kinase family inhibitor can overcome Paclitaxel resistance in ovarian cancer cell lines.

**Undergraduate Research:**
Investigation into localized transcription in neuronal axons grown both in culture and purified from mice.

Analysis of the function of the chimeric, mutant gene \textit{Wlds} and its role in slow Wallerian degeneration in neurons.

**Teaching Experience**

**Teaching Assistant:**
Undergraduate Cancer Biology, Emory University, Spring 2006
(Professor: Dr. Gregg Orloff – gregg.orloff@emory.edu)

- Taught one lecture, assisted in student presentations, writing and grading tests, as well as tutoring students.

**Undergraduate Tutoring:**
- Served as a mentor and tutor for several undergraduates enrolled in Biology classes at Emory University – 100 hours

**Additional Activities and Honors**

- **Varsity Soccer**, Emory University – 2000-2003
  - Captain – 2003
  - UAA All-Conference Honorable Mention - 2002, 2003
  - Thomas-Aliberti Scholar/Athlete Award - 2004
- **Sigma Chi**, Beta Chi Chapter
- **USLlive Broadcaster** for the Atlanta Silverbacks – 2007-present

**Peer Reviewed Publications**


**Published Abstracts**


II. Aim 1 Publication: Scharer et al. *Cancer Research, in press*
Genome-wide Promoter Analysis of the SOX4 Transcriptional Network in Prostate Cancer Cells

Christopher D. Scharer, Colleen D. McCabe, Mohamed Ali-Seyed, Michael F. Berger, Martha L. Bulyk, and Carlos S. Moreno

Abstract

SOX4 is a critical developmental transcription factor in vertebrates and is required for precise differentiation and proliferation in multiple tissues. In addition, SOX4 is overexpressed in many human malignancies, but the exact role of SOX4 in cancer progression is not well understood. Here, we have identified the direct transcriptional targets of SOX4 using a combination of genome-wide localization chromatin immunoprecipitation–chip analysis and transient overexpression followed by expression profiling in a prostate cancer model cell line. We have also used protein-binding microarrays to derive a novel SOX4-specific position-weight matrix and determined that SOX4 binding sites are enriched in SOX4-bound promoter regions. Direct transcriptional targets of SOX4 include several key cellular regulators, such as EGR1, HSP70, Tenascin C, Frizzled-5, Patched-1, and Delta-like 1. We also show that SOX4 targets 23 transcription factors, such as MLL, FOXA1, ZNF281, and NKK3-1. In addition, SOX4 directly regulates expression of three components of the RNA-induced silencing complex, namely Dicer, Argonaute 1, and RNA Helicase A. These data provide new insights into how SOX4 affects developmental signaling pathways and how these changes may influence cancer progression via regulation of gene networks involved in microRNA processing, transcriptional regulation, the TGFβ, Wnt, Hedgehog, and Notch pathways, growth factor signaling, and tumor metastasis.

[Research Article]

Introduction

The sex determining region Y-box 4 (SOX4) gene is a developmental transcription factor important for progenitor cell development and Wnt signaling (1, 2). SOX4 is a 47-kDa protein that is encoded by a single exon and contains a conserved high-mobility group DNA-binding domain (DBD) related to the TCF/LEF family of transcription factors that mediate transcriptional responses to Wnt signals. SOX4 directly interacts with β-catenin, but its precise role in the Wnt pathway is unknown (2). In adult mice, SOX4 is expressed in the gonads, thymus, T-lymphocyte and pro-B-lymphocyte lineages, and to a lesser extent in the lungs, lymph nodes, and heart (1). Embryonic knockout of SOX4 is lethal around day E14 due to cardiac failure, and these mice also showed impaired lymphocyte development (3). Tissue-specific knockout of SOX4 in the pancreas results in failure of normal development of pancreatic islets (4). SOX4 heterozygous mice have impaired bone development (5), whereas prolonged expression of SOX4 inhibits correct neuronal differentiation (6). These studies suggest a critical role for SOX4 in cell fate decisions and differentiation.

Whereas SOX2 is critical for maintenance of stem cells (7), SOX4 may specify transit-amplifying progenitor cells that are the immediate daughters of adult stem cells and have been proposed to be the population that gives rise to cancer stem cells. In humans, SOX4 is expressed in the developing breast and osteoblasts and up-regulated in response to progestins (8). SOX4 is up-regulated at the mRNA and protein level in prostate cancer cell lines and patient samples, and this up-regulation is correlated with Gleason score or tumor grade (9). In addition, SOX4 is overexpressed in many other types of human cancers, including leukemias, melanomas, glioblastomas, medulloblastomas (10), and cancers of the bladder (11) and lung (12). A meta-analysis examining the transcriptional profiles of human cancers found SOX4 to be 1 of 64 genes up-regulated as a general cancer signature (12), suggesting that SOX4 has a role in many malignancies. Furthermore, SOX4 cooperates with Evi1 in mouse models of myeloid leukemogenesis (13). Recently, we showed that SOX4 can induce anchorage-independent growth in prostate cancer cells (9). Consistent with the concept that SOX4 is an oncogene, three independent studies searching for oncogenes have found SOX4 to be one of the most common retroviral integration sites, resulting in increased mRNA (14–16).

Despite these findings, the role that SOX4 plays in carcinogenesis remains poorly defined. Whereas the transactivational properties of SOX4 have been characterized (17), genuine transcriptional targets remain elusive. To date, three studies have used expression profiling of cells after either small interfering RNA (siRNA) knockdown or overexpression of SOX4 to identify downstream target genes (9, 11, 18). Very recently, 31 SOX4 target genes were confirmed by chromatin immunoprecipitation (ChIP) in a hepatocellular carcinoma cell line (19). Although interesting, this study was limited by the fact that it focused on a specific tumor stage transition and did not use a genome-wide localization approach.

Here, we have performed a genome-wide localization analysis using a ChIP-chip approach to identify those genes that have SOX4 bound at their proximal promoters in human prostate cancer cells. We have identified 282 genes that are high-confidence direct SOX4 targets, including many genes involved in microRNA (miRNA) processing, transcriptional regulation, developmental pathways, growth factor signaling, and tumor metastasis. We have also used

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Requests for reprints: Carlos S. Moreno, Department of Pathology and Laboratory Medicine, Winship Cancer Institute, Emory University, Whitehead Research Building, Room 105, 615 Michael Street, Atlanta, GA 30322. Phone: 404-712-2809; Fax: 404-727-8538; E-mail: csmoreno@emory.edu.

©2009 American Association for Cancer Research.

doi:10.1158/0008-5472.CAN-08-3415
unique protein-binding DNA microarrays (PBM; refs. 20–22) to query the binding of recombinant SOX4 to every possible 8-mer. The PBM-derived SOX4 DNA binding data will further facilitate computational analyses of genomic SOX4 binding sites. These data provide new insights into how SOX4 affects key growth factor and developmental pathways and how these changes may influence cancer progression.

Materials and Methods

Cell culture and stable cell line construction. All cell lines were cultured, as described by American Type Culture Collection except LNCaP cells, which were cultured with T-Medium (Invitrogen). HA-tagged SOX4 was cloned into the pHR-UBQ-IRES-eYFP-U3 lentiviral vector (gift from Dr. Hihn Ly, Emory University), and stable cells were isolated, as previously described (23).

ChIP. Two 90% confluent P150s of both LNCaP-YFP and LNCaP-YFP/HA-SOX4 or RWPE-1-YFP and RWPE-1-YFP/HA-SOX4 cells were formaldehyde fixed and sonicated, and ChIP assay was performed, as described previously (23). Anti-HA 12CA5 or mouse IgG was used to immunoprecipitate protein-DNA complexes overnight at 4°C and collected using DynaM280 sheep anti-mouse IgG beads for 2 h. Dyna beads were washed, protein-DNA complexes were eluted, and DNA was purified, as described previously (24). A detailed description of the ChIP-chip protocol can be found in Supplementary Methods. Anti-HA 12CA5, anti–Flag-M2 (Sigma-Aldrich), or mouse IgG was used to immunoprecipitate protein-DNA complexes overnight at 4°C. All PCR primers used in ChIP-PCR can be found in Supplementary Table S7.

ChIP-chip analysis. To determine the direct SOX4 target genes on a global scale, we performed ChIP assays in triplicate from the LNCaP cell line stably expressing SOX4 and in duplicate from a control cell line that expressed YFP alone. Immunoprecipitated and input DNA were subjected to whole genome amplification, Cy3/Cy5 fluorescent labeling, and hybridization to the NimbleGen 25K human promoter array set. Input and immunoprecipitated DNA isolated from LNCaP-YFP and LNCaP-YFP/HA-SOX4 cells were amplified using linker-mediated PCR as described previously (25). Amplified DNA was labeled and hybridized in triplicate by NimbleGen Systems, Inc., to their human 25K promoter array. This set consists of two microarrays that tile 4 kb of upstream promoter sequence and 750 bp of downstream intronic sequence on average, with a total genomic coverage of 110 Mb. Raw hybridization data were Z-score normalized, and ratios of immunoprecipitation to input DNA were determined for each sample. ChIPOTle software was used to determine enriched peaks using a 500-bp sliding window every 50 bp, as previously described (23). NimbleGen microarray data are available from the GEO database accession number GEO11915.

Luciferase assays. PCR fragments representing the binding sites in the EGFR, ERBB2, and TLE1 genes were cloned in front of the pGL3-promoter luciferase construct (Promega). Primers sequences used can be found in Supplementary Table S7. LNCaP cells were transfected with 100 ng of TK-Resilla construct, 500 ng of pGL3-promoter vector alone and with cloned inserts, and 500 ng of either a SOX4 or vector expression construct. Dual luciferase assays were performed 48 h posttransfection, according to the manufacturer’s guidelines (Promega). All assays were performed in triplicate on separate days.

Quantitative real-time PCR. LNCaP cells were plated in six-well culture dishes and grown to 90% confluency before transfection with 1 μg of SOX4 plasmid or vector control using Lipofectamine 2000 (Invitrogen). At 24 h posttransfection, total RNA was harvested using the RNeasy kit (Qiagen), and reverse transcription was performed using Superscript III reverse transcriptase (Invitrogen). Quantitative real-time PCR (qPCR) was performed using SYBR Green I (Invitrogen) on a Bio-Rad iCycler using 18s or y-actin as a control, and data were analyzed using the 2ΔΔCt method (26). All primers used in this study are listed in Supplementary Table S7.

Microarray analysis. Total RNA was isolated from three independent experiments of either vector control or SOX4-transfected LNCaP cells, as
described above. Each transfection was performed in triplicate, and each sample was hybridized in duplicate, creating six data points for each condition. Total RNA was submitted to the Winship Cancer Institute DNA Microarray Core facility.8

FN1 All samples showed RNA integrity of 8.3 or greater using an Agilent 2100 Bioanalyzer. RNA was hybridized to the Illumina Human6 v2 Expression Beadchip that query roughly 47,000 transcripts with 48,701 probes, and after normalization, significantly changed probes were calculated using significance analysis of microarrays (SAM) software (27). Settings for SAM were two-class unpaired (versus vector control) imputation engine (10 nearest neighbor), permutations (500), RNG seed (1234567), Delta (1.316), fold change (1.5), and false discovery rate (0.749%). Microarray data are available in the GEO database accession number GEO11915.

Immunoblotting. Cells were lysed in lysis buffer [0.137 mol/L NaCl, 0.02 mol/L TRIS (pH 8.0), 10% glycerol, and 1% NP40], and 50 μg total lysate were separated by SDS-PAGE electrophoresis and transferred to nitrocellulose for immunoblotting. Immunoblots were probed with polyclonal rabbit SOX4 antisera described previously (9) and DICER (Santa Cruz). To control for equal loading, immunoblots were also probed with a mouse monoclonal antibody to protein phosphatase 2A (PP2A) catalytic subunit (BD Biosciences).

Results

SOX4 transcriptionally activates EGFR. Using expression profiling to determine the genes whose mRNA levels change when SOX4 is either overexpressed or eliminated using siRNA (9), we identified EGFR as a candidate SOX4 transcriptional target (Fig. 1A). Analysis of the promoter and first intron of EGFR and other family members with CONFAC software (28) revealed the presence of potential SOX4 binding sites within the first intron of EGFR and ERBB2 (Fig. 1B). CONFAC functions by identifying the conserved sequences in the 3-kb proximal promoter region and first intron of human-mouse orthologue gene pairs and then identifying transcription factor binding sites (TFBS), defined by position weight matrices from the MATCH software (29), which are conserved between the two species (28).

Whereas limited commercial antibodies exist for SOX4 and show activity in immunoblots, in our hands, none of them have been

Figure 2. A, graph showing enrichment in the three HA-SOX4 lanes over the average of the two YFP replicates for the SOX4 target gene FM04. Y axis is the signal intensity across the genomic coordinates on the X axis. B, qPCR ChIP analysis of 10 randomly selected genes verified in both the RWPE-1 and LNCaP cell lines. Graph shows fold enrichment of the HA-SOX4 immunoprecipitation over the YFP negative control immunoprecipitation. Numbers above the bars represent the mean log2 of fold enrichment of ChIP-chip signal for the probes contained in the peak relative to YFP. Bars, SD (n = 3 technical replicates). C and D, genes that were verified by conventional ChIP assay. HA-SOX4 and YFP cells were subjected to conventional ChIP followed by PCR in both the LNCaP (C) and RWPE-1 (D) prostate cell lines. Six genes verified in the LNCaP cell lines and five in the RWPE-1 cell lines.
useful in a ChIP assay. Therefore, we used epitope-tagged SOX4, as described in other SOX4 ChIP studies (9, 19). Although the FLAG epitope tag was not tested directly for activity, a glutathione S-transferase (GST)-SOX4 construct showed binding to a known SOX4 motif and not a control motif (Supplementary Fig. S2B), validating that the epitope tag does not interfere with SOX4 binding. To determine if SOX4 directly bound the EGFR and ERBB2 enhancers, we performed ChIP analysis on RWPE-1 prostate cancer cells stably infected with FLAG-SOX4 or a control lentiviral vector. DNA representing the predicted SOX4 sites was specifically amplified from the FLAG-SOX4 cell line and not from the control cell line, indicating that SOX4 binds to intronic sequence of EGFR and ERBB2 (Fig. 1C). EGFR is expressed in RWPE-1 cells, but not in LNCaP cells, and SOX4 did not bind to these sequences in LNCaP cells (data not shown).

To characterize the transcriptional effect of SOX4 levels on the regions bound by SOX4 in ChIP assays, the amplified ChIP fragments were cloned in front of a minimal promoter luciferase reporter plasmid and tested in transient transfections in LNCaP cells. Compared with a vector control, SOX4 significantly increased transcription of the EGFR fragment 3-fold and the TLE1-positive control fragment roughly 4-fold. Although not found significant, ERBB2 was activated 1.5-fold compared with the vector control (Fig. 1D). Consistent with microarray data, SOX4 transcriptionally activates the EGFR enhancer.

**Genome-wide localization analysis.** To determine the direct SOX4 target genes on a global scale, we performed ChIP assays in triplicate from the LNCaP HA-SOX4 stable cell line and in duplicate from the control LNCaP-YFP cell line. Peaks (P < 0.001) that overlapped in at least two of the three data sets and were not

---

**Figure 3.** A, heat map (top) illustrating Illumina expression data of the 1,766 significant genes, as determined by SAM analysis. Red, overexpressed genes; green, underexpressed genes. Venn diagram (bottom) depicts the overlap between 3,470 ChIP-chip SOX4 direct target genes, the Illumina expression data set of 1,766 genes, and the Affymetrix expression data set of 465 genes. B, qPCR expression analysis of SOX4 direct target genes after SOX4 overexpression in LNCaP cells. All 10 genes were up-regulated over a vector control transfection, similar to values determined by the illumina array with a P value of <0.005 by Student’s t test. Bars, SD (n = 3 independent biological replicates performed on separate days). C, DICER protein expression is up-regulated by SOX4. HA-SOX4 or vector control was transfected into LNCaP cells, and immunoblots were probed for DICER, SOX4, and PP2Ac as a loading control. D, PBM-derived 8-mer PWM for SOX4 displayed both graphically and numerically for each base position derived from incubation of recombinant GST-SOX4-DBD with a universal “all 8-mer” double-stranded DNA protein-binding microarray. With stringent criteria (core similarity, >0.85; matrix similarity, <0.75) we find 60% of the peaks in the 282 high-confidence promoters contain SOX4 binding sites.
present in the LNCaP-YFP cell line were called significant (Fig. 2A). Based on these variables, we classified 3,600 significant, overlapping peaks as SOX4 target sequences. Because some transcription start sites (TSS) are quite close to each other (<3 kb), it was not always possible to assign a unique gene to every peak. In addition, many genes had multiple peaks in their promoters, and thus, we mapped the 3,600 peaks to 3,470 different genes (Supplementary Table S1).

To verify the set of 3,600 SOX4 peaks, 28 candidate SOX4 target sites representing a range of P values in promoters of genes of biological interest were chosen, primers were designed around the peaks and enrichment was verified by conventional ChIP. Ten of these 28 candidates were analyzed by ChIP qPCR and 18 by ChIP-PCR. Overall, 24 of 28 (86%) of the candidate targets were confirmed, validating our data set. All 10 of the peaks chosen to validate by qPCR were reproducibly enriched over the YFP control in both the LNCaP-HA-SOX4 cell line and the RWPE-1 cell line (Fig. 2B). Of the target sites validated by conventional PCR, 14 of 18 genes were confirmed in both the LNCaP and RWPE-1 cell lines, whereas a mock, control PCR was negative (Fig. 2C; D: data not shown). The only exception was ANKRDI15, which was enriched only in the LNCaP cell line and not in the RWPE-1 line.

**Target gene expression analysis.** To determine whether SOX4 binding affects transcription of the 3,470 genes that have SOX4 bound at their promoters, we performed whole genome expression analysis on LNCaP cells after transfection with SOX4 or a control vector. To increase the likelihood of identifying direct SOX4 targets, total RNA was isolated at a relatively early time point (24 hours posttransfection) and hybridized to Illumina Human 6-v2 whole genome arrays. A total of 1,766 genes were changed at least 1.5-fold with a false discovery rate of 0.749% (Fig. 3A; Supplementary Table S2). Of those 1,766 genes, 244 were also directly SOX4 targets by ChIP-chip analysis (Fig. 3A; Supplementary Table S3). Seven of these genes were confirmed by qPCR (Fig. 3B).

Our previous expression profiling of LNCaP cells after SOX4 siRNA knockdown (9) identified 465 downstream targets, and we confirmed that SOX4 regulates the expression of DICER, DLL1, and HES2 in LNCaP cells by qPCR (Fig. 3B). We further confirmed SOX4 regulation of DICER at the protein level (Fig. 3C). Out of those 465 candidate targets, 47 genes overlapped with the 3,470 ChIP-chip targets, increasing the number of direct SOX4 targets to 282 genes (Fig. 3A; Supplementary Table S3). We classified these 282 genes bound by SOX4 in ChIP-chip and significantly changed by expression profiling as high confidence direct SOX4 target genes. Nine genes (PKH4CA, DDX9, BTNL3A3, CDK2, MVK, ADAM10, RYK, ISG20, and DBI) overlapped in all three data sets. The transcription factor SOX4 and purine biosynthetic enzyme GART, two genes on chromosome 21 that are transcribed in opposite directions and regulated by a bidirectional promoter, were affected in opposite ways. SOX4 was activated by SOX4 1.8-fold, as detected by SOX4 overexpression, whereas GART was increased almost 3-fold as determined by SOX4 siRNA knockdown, suggesting that SOX4 regulates the directionality of this promoter.

We next analyzed the P values of the peaks in our ChIP-chip data set, comparing the P values of the genes that were altered by transient overexpression of SOX4 with those that were not (Supplementary Fig. S2). We found no difference in the distributions of the ChIP-chip P values for those genes that were changed in expression profiling experiments and those that were not. Thus, based on our ChIP-chip validation experiments and the similar P-value distributions, we conclude that SOX4 is genuinely bound at the promoters of the 3,188 genes that did not change but that SOX4 by itself is not limiting or sufficient to generate changes in transcription without corresponding changes in the cellular context, such as activation of cofactors or signaling pathways.

**Novel SOX4 position weight matrix.** To facilitate computational analyses of SOX4 DNA binding sites, we sought to determine the DNA binding preferences of SOX4 using universal PBMs (20). This universal PBM array allows recombinant SOX4 protein to interact with and bind every possible 8-mer, thus allowing in vitro binding site specificities to be calculated.

We generated an NH4 terminal, GST-SOX4-DBD fusion protein, expressed and purified it from E. coli, and tested for activity (Supplementary Fig. S3). The GST-SOX4-DBD was incubated with the protein binding microarray and a novel position weight matrix (PWM; RWYAWRWV) was calculated from the PBM data (Supplementary Table S4) using the Seed-and-Wobble algorithm (Fig. 3D; ref. 20). Three groups have previously reported similar binding site sequences for SOX4: AACAAAG (30), ACAAT (31), and WWCAAWG (19). Our PWM confirms the SOX4 core binding sequence of the previously known binding sites but there are some differences in the specificity at the 1st and 7th positions and we find a bias toward A, C, and G at the 8th position. These differences could be due to the fact that earlier reports used no more than 31 sequences to develop the binding motif, whereas our study queried every possible 8-mer.

**SOX4 peaks contain SOX4 binding sites.** Using our newly derived PWM, we applied CONFAC software (28) to analyze the enriched sequences for the presence of SOX4 binding sites. We analyzed the sequences of the peaks in the promoters of our 282 high confidence genes against 10 sets of control promoter sequences to see if SOX4 sites were enriched in our target gene set. Control promoter peaks of equal size to SOX4 peaks were chosen randomly from sequences covered by the NimbleGen array, and each control set contained equal total sequence coverage as our 282 high confidence peaks. With stringent criteria (core similarity, >0.85; matrix similarity, >0.75), we find that 60% of the peaks contain SOX4 binding sites. SOX4 sites were significantly enriched relative to 10 sets of random promoter sequence by Mann-Whitney U test using Benjamini correction for multiple hypothesis testing (q < 0.0019).

To further characterize the SOX4 binding sites, we searched the entire set of 3,600 SOX4 peaks and 10 equal sets of random promoter sequence for the presence of PBM-bound k-mers (here, ungapped 8-mers). The specificity of PBM k-mers can be quantified by the enrichment score (ES), which ranges from 0.5 to 0.5 (32). We analyzed the specificity of PBM k-mers with 4.45 > ES > 0.40 (moderate) and ES > 0.45 (stringent). Whereas both SOX4-bound peaks and random promoter sequence contained moderate and stringent k-mers, SOX4 peaks contained significantly more stringent (P = 0.0002) and moderate (P = 1.08 x 10^-5) k-mers by two-tailed Mann-Whitney test (Supplementary Fig. S4).

To investigate interaction with protein partners that may increase SOX4 affinity for poor matching sites in vivo, we searched for enrichment of cooccurring TFBS in the SOX4 peaks. We applied CONFAC software to search the sequences for the presence of cooccurring transcription factor binding sites within the same peak (Table 1). Using the same criteria as above, we determined that the E2F family had the most frequently co-occurring motif (similar to TTTCGCGC, q = 1.78 x 10^-8). Interestingly, ingenuity pathway analysis (IPA) identified cell cycle as a functionally enriched process in the 3,470 SOX4 target genes (P = 0.00916), suggesting...
that part of SOX4’s function is to control the expression of genes involved in cell cycle progression. CONFAC analysis identified other significant TFBS motifs enriched in the SOX4 peaks (Table 1), including those for transcription factors in the TGFβ, Wnt, and NF-kB pathways. SOX4 modulates Wnt signaling via interaction with β-catenin and the TCF4 transcription factor (2), suggesting a possible role for SOX4 in transcriptionally modulating Wnt signals. We confirmed the recent report that SOX4 cooperates with constitutively active β-catenin to activate TOP-Flash luciferase reporters (2) and found that SOX4 synergistically induces activation of these constructs, further highlighting a role for SOX4 in the Wnt pathway (Supplementary Fig. S5).

### SOX4 target genes
To determine the biological processes and functions of the SOX4 targets, we performed a gene ontology analysis using DAVID software (33) on the 282 high confidence SOX4 targets. Among the SOX4 targets were 23 transcription factors (Table 2), and DAVID analysis determined that the top annotations were transcription (P = 3.7 × 10⁻¹⁸), transmembrane (P = 5.59 × 10⁻¹⁰), and protein phosphorylation/dephosphorylation (P = 3.5 × 10⁻¹⁸/6.6 × 10⁻⁷). These findings are paralleled by expression profiling of SOX4 overexpression in HU609 bladder carcinoma cells where top annotated functions were signal transduction and protein phosphorylation (11).

### FN2
Commercial IPA software⁹ identified biological pathways and functions that are enriched in our 282 high confidence targets, 1,766 significant genes identified by SAM analysis, and the 3,470 unique genes that had SOX4 bound at their promoters in ChIP-chip. As anticipated, among the most significant annotations were cell cycle, cancer, and tissue development. In the significant expression data set of 1,766 genes, we observed an up-regulation of cell cycle, cancer, and tissue development. In the significant expression data set of 1,766 genes, we observed an up-regulation of cell cycle, cancer, and tissue development.

### Discussion
Whereas many studies have identified SOX4 as a crucial developmental transcription factor that is often overexpressed in many types of malignancies, little is known of how SOX4 regulates in cancer cells. We have used a ChIP-chip approach to report the first genome-wide localization analysis of SOX4 and mapped 3,600 binding peaks that represent 3,470 unique genes possibly under the transcriptional control of SOX4. We have also identified 1,766 genes that respond to increased SOX4 levels by whole genome expression profiling. Integration of these data sets mapped 282 high-confidence direct targets in the SOX4 transcriptional network. In addition, we have used protein-binding microarrays

<table>
<thead>
<tr>
<th>Transcription factor</th>
<th>Family</th>
<th>Benjamini corrected q value</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2F4</td>
<td>E2F</td>
<td>1.78E−11</td>
</tr>
<tr>
<td>E2F1</td>
<td>E2F</td>
<td>3.06E−11</td>
</tr>
<tr>
<td>PAX5</td>
<td>Paired box</td>
<td>2.07E−10</td>
</tr>
<tr>
<td>WHN</td>
<td>Forkhead</td>
<td>2.94E−10</td>
</tr>
<tr>
<td>SMAD3</td>
<td>SMAD</td>
<td>1.82E−09</td>
</tr>
<tr>
<td>SMAD4</td>
<td>SMAD</td>
<td>3.33E−09</td>
</tr>
<tr>
<td>MYC</td>
<td>MYC</td>
<td>6.25E−09</td>
</tr>
<tr>
<td>NFKAPPAB</td>
<td>NF-κB</td>
<td>2.95E−08</td>
</tr>
<tr>
<td>LEF1/TCF1</td>
<td>LEF</td>
<td>1.12E−06</td>
</tr>
</tbody>
</table>

### Table 2. DAVID analysis identified 23 transcription factors present in our high confidence SOX4 target genes

<table>
<thead>
<tr>
<th>Entrez ID</th>
<th>Symbol</th>
<th>Microarray fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td>196528</td>
<td>ARID2</td>
<td>1.99</td>
</tr>
<tr>
<td>2001</td>
<td>ELF5</td>
<td>−2.65</td>
</tr>
<tr>
<td>3169</td>
<td>FOXA1</td>
<td>−2.47</td>
</tr>
<tr>
<td>2976</td>
<td>GTF3C2</td>
<td>−3.12</td>
</tr>
<tr>
<td>64412</td>
<td>GZF1</td>
<td>2.42</td>
</tr>
<tr>
<td>84558</td>
<td>LCOR</td>
<td>2.41</td>
</tr>
<tr>
<td>4173</td>
<td>MCM4</td>
<td>1.55</td>
</tr>
<tr>
<td>58508</td>
<td>MLL3</td>
<td>2.06</td>
</tr>
<tr>
<td>10953</td>
<td>MORF4L1</td>
<td>2.07</td>
</tr>
<tr>
<td>8031</td>
<td>NCOA4</td>
<td>2.64</td>
</tr>
<tr>
<td>4784</td>
<td>NFIX</td>
<td>−2.83</td>
</tr>
<tr>
<td>4824</td>
<td>NKX3-1</td>
<td>−4.53</td>
</tr>
<tr>
<td>7799</td>
<td>PRDM2</td>
<td>2.48</td>
</tr>
<tr>
<td>5933</td>
<td>RBL1</td>
<td>1.80</td>
</tr>
<tr>
<td>55309</td>
<td>SNFT</td>
<td>−2.32</td>
</tr>
<tr>
<td>6722</td>
<td>SRF</td>
<td>−2.03</td>
</tr>
<tr>
<td>54816</td>
<td>SUHVW4</td>
<td>−1.93</td>
</tr>
<tr>
<td>9412</td>
<td>SURB7</td>
<td>−2.24</td>
</tr>
<tr>
<td>9238</td>
<td>TCEAL1</td>
<td>−1.57</td>
</tr>
<tr>
<td>7718</td>
<td>ZNF165</td>
<td>1.53</td>
</tr>
<tr>
<td>7738</td>
<td>ZNF184</td>
<td>1.66</td>
</tr>
<tr>
<td>23528</td>
<td>ZNF281</td>
<td>1.71</td>
</tr>
<tr>
<td>30834</td>
<td>ZNRD1</td>
<td>−1.63</td>
</tr>
</tbody>
</table>

NOTE: Gene ontology term: transcription, DNA dependent (P = 3.7 × 10⁻¹⁸).

⁹ http://www.ingenuity.com
to determine a novel PWM specific for SOX4 and show that our ChIP-chip predicted peaks are significantly enriched for SOX4 binding sites. These data provide several new insights into the roles that SOX4 plays in the cell.

**SOX4 direct target genes.** Although only 10% of the significant differentially expressed genes overlapped with the ChIP-chip data, this is likely a conservative estimate because the NimbleGen 25k promoter array only queries proximal promoter sequences and not more than 1 kb downstream of the TSS. We found that SOX4 binds EGFR and ERBB2 in the first intron over 20 kb downstream of the TSS (Fig. 1D), and unsurprisingly, we did not detect EGFR or ERBB2 in our ChIP-chip experiment. Thus, more of the 1,900 genes that responded to changes in SOX4 mRNA levels (but were not detected by ChIP-chip) could still be direct targets. Excellent candidates would be the 40 genes that responded to SOX4 on both microarray platforms, such as the IL6 receptor, SOX12, and NME1 (Supplementary Table S6). Whereas 3,600 is a fairly large number of SOX4 bound regions, some background can be expected. Nevertheless, we were able to validate 24 of 28 (86%) candidate binding sites chosen, adding confidence to our data set. In fact, an even higher number of over 4,200 genomic binding sites had been previously observed for c-Myc in ChIP–positron emission tomography whole genome studies (39). Whole genome tiling arrays or ChIP-seq could provide additional binding sites that may show more overlap with the Illumina expression data set.

Conversely, many of the bound genes may not respond to changes in SOX4 mRNA levels alone but to multiprotein activator complexes of which SOX4 is only one component. Furthermore, the stability of SOX4 bound to a promoter could be greater than unbound SOX4, limiting the effects observed by siRNA knockdown. In different cell types or cellular contexts, SOX4 may activate a different subset of these genes. Of the 31 SOX4 target genes reported by Liao and colleagues (19), only six are represented in our NimbleGen data set and three found to be changed in our Illumina expression profiling data set. The small overlap could be due to the fact that those genes were identified in hepatocellular carcinomas, whereas we have examined prostate cancer cells. Interestingly, DKK was one of the six genes that overlapped in both data sets, further implicating SOX4 in the Wnt pathway. Because SOX4 is known to interact with β-catenin and other coactivators, it may be poised at many of these promoters to enable responses to developmental signals from the Wnt or TGFβ pathway.

**Receptor and signaling regulation.** Our data suggest that SOX4 regulates cellular differentiation through a variety of transcription factors and receptors. SOX4 is up-regulated in response to numerous external ligands ranging from TGFβ (38)
and BMP-6 (40) to parathyroid hormone and progesterone (8). Previous work has shown that SOX4 directly signals from IL-SR (41), and here, we have shown that SOX4 directly regulates EGFR (Fig. 1). Membrane receptors in the SOX4 transcriptional network also include Frizzled family members FZD3, FZD5, FZD8; the Hedgehog receptor PTCH-1; the Notch ligand DLL1; TRAIL decoy receptor TNFRSF10D; and other growth factor receptors, such as FGFR1 and JGF2R. DAVID analysis also revealed protein phosphorylation/dephosphorylation \((P = 3.5 \times 10^{-18} / 6.6 \times 10^{-7})\) and transcription \((P = 3.7 \times 10^{-18})\) are enriched annotations, identifying 23 transcription factors that are direct targets of SOX4. This evidence suggests that SOX4 regulates signaling events both at the external input level and the internal output or transcription level. This regulation could be direct, as with IL-SR, or through the transcriptional targets SOX4 activates.

**Transcription factors and SOX4.** Here, we have reported DNA binding specificity data for SOX4, which will improve computational analyses for SOX4 specific binding sites. Our data confirm the known SOX family core-binding motif and add new specificity at the 1st, 7th, and 8th positions. Whereas crystal structure of the known SOX family core-binding motif and add new specificity SOX4.

**SOX4 and cancer.** Based on the target genes we identified, SOX4 seems to influence cancer progression in several ways. First, it plays a key role in the activation of and response to developmental pathways, such as Wnt, Notch, Hedgehog, and TGFβ. Second, SOX4 inhibits differentiation via repression of transcription factors, such as NKK3.1, and activation of MLL and MLL3, two histone H3 K4 methyltransferases that induce activation of HOX gene expression (42). MLL methyltransferase complexes also facilitate E2F activation of S-phase promoters, facilitating cell cycle progression. Activation of MLL also suggests a mechanism for the role of SOX4 in myeloid leukemogenesis, because MLL is a critical oncogene that is often translocated or amplified in this disease (43). Thirdly, SOX4 targets growth factor receptors, such as EGFR, FGFR1, and JGF2R, enhancing proliferative signals in tumors and potentially activating the PI3K-akt pathway. Mice heterozygous for NKK3.1 and PTEN in the prostate develop prostate adenocarcinomas and metastases to the lymph node (44). Thus, our data suggest that SOX4 may promote prostate cancer progression directly through NKK3.1 repression and indirectly through PI3K-akt activation. Finally, SOX4 seems to promote metastasis via up-regulation of tenascin C. Recently, both SOX4 and tenascin C were shown to enhance metastasis of breast cancer cells to the lung (45), as has the TGFβ pathway, which activates their expression (46). Other metastasis-associated SOX4 target genes include integrin α7, and Rac1. Rac1 was recently shown to control nuclear localization of β-catenin in response to Wnt signals (47).

**SOX4 regulates components of the RISC complex and small RNA pathway.** miRNAs are small noncoding RNA species that regulate the translation and stability of mRNA messages for hundreds of downstream target genes via partial complementarity to short sequences in the 3’ untranslated regions of mRNAs. The RISC, which is composed of AGO1 or AGO2, TBBP, and Dicer processes miRNAs from precursors (pre-miRNA) to their mature form, cleaves target mRNAs, and participates in translational inhibition. RNA Helicase A (RNA/DxH9) interacts with the RISC complex and participates in loading of small RNAs into the RISC complex (48). We observed that three components of the RISC complex, Dicer, AGO1, and RNA/DxH9, are high-confidence direct targets of SOX4 (Supplementary Table S3), and we confirmed these data by qPCR (Fig. 3B). Dicer has been independently observed to be overexpressed in prostate cancers (49).

In addition, we observed that Toll-like receptor 3 (TLR3), which binds to double-stranded RNAs, induces gene silencing, and can induce apoptosis (50), was induced 2.8-fold upon overexpression of SOX4. This induction may be indirect because TLR3 was not detected by ChIP-chip, but we cannot exclude the possibility that SOX4 may directly regulate TLR3 from a distal or intronic enhancer.

Our observation that SOX4 targets three genes important in small RNA processing is of particular interest in light of the role of SOX4 in development and cancer progression. miRNAs have been implicated in numerous physiologic processes from development to oncogenesis. miRNAs can also act as suppressors of breast cancer metastasis via targeting of tenascin C and SOX4 (45) and as promoters of breast cancer metastasis (51). The finding that SOX4 can affect expression of multiple components of the RISC complex also provides insight into why long-term loss of SOX4 induces widespread apoptosis (9, 18). In summary, these data shed light on the mechanisms and pathways through which SOX4 may exert its effects during development and cancer progression. Further studies are necessary to elucidate the precise role of SOX4 in the functioning of these pathways.

**Disclosure of Potential Conflicts of Interest**

No potential conflicts of interest were disclosed.

**Acknowledgments**

Received 9/3/2008; revised 10/8/2008; accepted 11/3/2008.

**Grant support:** National Cancer Institute R01 CA106026 (C.S. Moreno), NIH/ NHGRI grant R01 HG009985 (M.F. Berger and M.L. Bulyk), DOD CDMPR Prostate Cancer Predoctoral Training Fellowship PC060145 (C.D. Scharer), and Postdoctoral Training Fellowship PC060114 (C.D. McCabe).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank Dr. Maja Orlandic-Kodani at Winship Cancer Institute Microarray Core Facility for performing Illumina microarray labeling and hybridization, Robert Karaffa at the Emory University FlowCore for cell sorting, Dr. Hinh Ly for IRES-eYFP lentiviral vector, and Dr. Anita Corbett for pGEX-4T-1 plasmid.

**References**


Figures:

**Figure 1:** (A) Schematic diagram of the lentiviral constructs used to stably infect LNCaP and RWPE-1 prostate cancer cells showing the locations of LTRs and promoters. The top figure represents the control, eYFP only construct, and the lower figure represents the HA-SOX4 construct. (B) Histogram charts showing the control uninfected, pre-sorted and post-sorted cell populations. Lower axis displays YFP signal intensity. (C) Immunoblot showing that HA-SOX4 is expressed and specifically immunoprecipitated from the LNCaP-HA-SOX4 cell line and not the control LNCaP-YFP cell line.

**Figure 2:** (A) Graph showing enrichment in the three HA-SOX4 lanes over the average of the two YFP replicates for the gene FMO4. (B) QRT-PCR analysis of 10 randomly selected genes verified in both the RWPE-1 and LNCaP cell lines. Graph shows fold enrichment of the HA-SOX4 IP over the YFP control IP. (C) Genes that were verified by conventional ChIP assay. LNCaP-HA-SOX4 and LNCaP-YFP cells were subjected to conventional ChIP followed by PCR in both the LNCaP and RWPE-1 prostate cell lines.
Figure 3: (A) Heat map illustrating Illumina expression data of the 1,766 significant genes as determined by SAM analysis. Red indicates overexpressed and green denotes underexpressed genes. (B) qPCR data of SOX4 direct target genes after SOX4 overexpression in LNCaP cells. All ten genes were upregulated over a vector control transfection, similar to values determined by the Illumina array with a p-value less than 0.005 by students T-test. Error bars indicate 1 SD (n = 3 independent biological replicates performed on separate days). (C) DICER is regulated by SOX4 at the protein level. Empty vector or one expressing HA-SOX4 was transfected into LNCaP cells and immunoblotting performed. DICER is upregulated specifically by SOX4 and not in the control transfection. (D) Venn diagram depicts the overlap between 3,470 ChIP-chip SOX4 direct target genes, the Illumina expression data set of 1,766 genes, and the Affymetrix expression dataset of 465 genes.
Figure 4: (A) EMSA assay of recombinant GST-SOX4-DBD binding to a known SOX4 binding motif of a 35mer oligo. NP – No protein, SP – specific probe, SC – Specific cold competitor, NSC – non-specific cold competitor. (B) SDS-PAGE gel of GST-SOX4-DBD from an IPTG uninduced (U) or induced (I) cell line. (C) Novel 8mer PWM for SOX4 displayed both graphically and numerically for each base position.

Figure 5: Luciferase assay of LNCaP cells transfected with either a vector control or 100, 200, or 300 ng of a SOX4 expression vector. LNCaP cells were also co-transfected with either a vector control or the β-catenin S33Y constitutively active mutant. All cells were transected with the TOP flash luciferase reporter and luciferase activity was measured 24 hrs post-transfection. SOX4 does not function alone but instead cooperates with β-catenin to activate the TOP flash reporter in a dose dependent manner.
Figure 6: (A) IPA analysis of direct target genes graphically illustrating the cellular location of the SOX4 transcriptional target genes. SOX4 regulates a host of nuclear and membrane localized proteins as well as multiple components of the RISC complex. Red indicates target genes upregulated by SOX4, green denotes downregulated genes and white represents genes for which no expression change was detected. (B) IPA analysis of Illumina expression genes changed at least 2-fold by SAM analysis. SOX4 regulatory targets include a host of membrane and nuclear proteins. Red indicates genes upregulated by SOX4 overexpression and green denotes downregulated genes.
### Tables:

**Table 1:** Benjamini corrected q-values for co-occurring transcription factor binding sites.

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>Family</th>
<th>Benjamini Corrected q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2F4</td>
<td>E2F</td>
<td>1.78E-11</td>
</tr>
<tr>
<td>E2F1</td>
<td>E2F</td>
<td>3.06E-11</td>
</tr>
<tr>
<td>PAX5</td>
<td>Paired Box</td>
<td>2.07E-10</td>
</tr>
<tr>
<td>WHN</td>
<td>Forkhead</td>
<td>2.94E-10</td>
</tr>
<tr>
<td>SMAD3</td>
<td>SMAD</td>
<td>1.82E-09</td>
</tr>
<tr>
<td>SMAD4</td>
<td>SMAD</td>
<td>3.33E-09</td>
</tr>
<tr>
<td>MYC</td>
<td>MYC</td>
<td>6.25E-09</td>
</tr>
<tr>
<td>NFKAPPAB</td>
<td>NF-κB</td>
<td>2.95E-08</td>
</tr>
<tr>
<td>LEF1/TCF1</td>
<td>LEF</td>
<td>1.12E-06</td>
</tr>
</tbody>
</table>

Table 2: DAVID analysis identified 23 transcription factors present in our high confidence SOX4 target genes.

<table>
<thead>
<tr>
<th>Entrez ID</th>
<th>Symbol</th>
<th>Microarray Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>196528</td>
<td>ARID2</td>
<td>1.99</td>
</tr>
<tr>
<td>2001</td>
<td>ELF5</td>
<td>-2.65</td>
</tr>
<tr>
<td>3169</td>
<td>FOXA1</td>
<td>-2.47</td>
</tr>
<tr>
<td>2976</td>
<td>GTF3C2</td>
<td>-3.12</td>
</tr>
<tr>
<td>64121</td>
<td>GZF1</td>
<td>2.42</td>
</tr>
<tr>
<td>84458</td>
<td>LCOR</td>
<td>2.41</td>
</tr>
<tr>
<td>4173</td>
<td>MCM4</td>
<td>1.55</td>
</tr>
<tr>
<td>58508</td>
<td>MLL3</td>
<td>2.06</td>
</tr>
<tr>
<td>10933</td>
<td>MORF4L1</td>
<td>2.07</td>
</tr>
<tr>
<td>8031</td>
<td>NCOA4</td>
<td>2.64</td>
</tr>
<tr>
<td>4784</td>
<td>NFIX</td>
<td>-2.83</td>
</tr>
<tr>
<td>4824</td>
<td>NKKX3-1</td>
<td>-4.53</td>
</tr>
<tr>
<td>7799</td>
<td>PRDM2</td>
<td>2.48</td>
</tr>
<tr>
<td>5933</td>
<td>RBL1</td>
<td>1.80</td>
</tr>
<tr>
<td>55509</td>
<td>SNFT</td>
<td>-2.32</td>
</tr>
<tr>
<td>6722</td>
<td>SRF</td>
<td>-2.03</td>
</tr>
<tr>
<td>54816</td>
<td>SUHW4</td>
<td>-1.93</td>
</tr>
<tr>
<td>9412</td>
<td>SURB7</td>
<td>-2.24</td>
</tr>
<tr>
<td>9338</td>
<td>TCEAL1</td>
<td>-1.57</td>
</tr>
<tr>
<td>7718</td>
<td>ZNF165</td>
<td>1.53</td>
</tr>
<tr>
<td>7738</td>
<td>ZNF184</td>
<td>1.66</td>
</tr>
<tr>
<td>23528</td>
<td>ZNF281</td>
<td>1.71</td>
</tr>
<tr>
<td>30834</td>
<td>ZNRD1</td>
<td>-1.63</td>
</tr>
</tbody>
</table>

GO Term: transcription, DNA dependent (p = 3.7x10^{-18}).