Genome-wide Identification of PAX3-FKHR Binding Sites in Rhabdomyosarcoma Reveals Candidate Target Genes Important for Development and Cancer

Liang Cao[1], Yunkai Yu[1, 2], Sven Bilke[1], Robert L. Walker[1], Linnia H. Mayeenuddin[1, 2], David O. Azorsa[4], Fan Yang[1], Marbin Pineda[1], Lee J. Helman[3], Paul S. Meltzer[1]

[1]Genetics Branch, Center for Cancer Research , National Cancer Institute, Bethesda, MD 20892, USA

[2]Laboratroy of Proteomics and Analytical Technologies, SAIC-Frederick Inc, National Cancer Institute-Fredetrick, Frederick, MD 21703, USA

[3]Pediatric Oncology Branch, Center for Cancer Research , National Cancer Institute, Bethesda, MD 20892, USA

[4]Cancer Genetics Branch, National Human Genome Research Institute (Present address: Translational Genomics Research Institute, Phoenix AZ 85004)

Corresponding author: Liang Cao, Genetics Branch, Center for Cancer Research , National Cancer Institute, Bethesda, MD 20892, USA. Phone: (301) 435-9039; Fax: (301) 402-3241; Email: caoli@mail.nih.gov

Supplementary Methods

Generation of PFM2 antibody

Monoclonal antibodies were prepared as previously described (Azorsa and Meltzer, 1999) with the following modifications: peptide PF corresponding to the PAX3-FKHR translocation region (C**TIGNGLSPQNSIRHNLSL**) was synthesized by Genosys (The Woodlands, TX) with the additional cysteine at the *N*-terminus for coupling. The peptide was coupled to KHL and ovalbumin using a maleamide conjugation kit (Pierce, Rockford, IL). Balb/c mice were injected i.p. with 100 µg of KLH peptides in Hunter's adjuvant (Sigma) followed by additional i.p. injections of 100 µg of purified KLH peptide in Hunter's adjuvant at two weeks intervals. Two weeks after the third injection, one mouse received an i.p. injection of 75 µg of purified KLH peptide in PBS and i.v. injections of 30 µg of purified KLH peptide in PBS for three consecutive days. A day later, the spleen of the mouse was removed and fused with the myeloma cell line P3x63Ag8.653 as previously described using 50% PEG with 5% DMSO. Fused cells were resuspended in HY media supplemented with 20% FBS, HAT, and Nutridoma-CS and seeded in eight flat bottom 96-well plates. Hybridoma colonies were screened for secretion of mAbs that bound to ovalbumin couple peptide by ELISA and by western blot analysis as previously described (Azorsa and Meltzer, 1999). Hybridomas secreting mAbs of interest were subcloned twice by limiting dilution. Final hybridoma clones were isotyped using an isotyping kit (Boehringer Mannheim). Tissue culture supernatants from the final clones were collected, treated with 0.02% sodium azide, and stored at 4°C. For chromatin immunoprecipitation, antibody was purified from ascites. Western blot analysis demonstrated a high specificity for PAX3-FKHR (Fig. 1).

Chromatin immunoprecipitation

Chromatin immunoprecipitation was performed with reagents from Active Motif, (Carlsbad, CA) using the manufacture protocol. Briefly, cross-linking was performed with $5X10^7$ Rh4 and RD cells in medium with 1% formaldehyde for 15 min. Cells were collected, lysed in lysis buffer and sonicated to DNA fragment of 200-300 bp in length. Cell lysates were added to ChIP buffer, precleared for 1 hr with protein G-Sepharose and incubated overnight at 4°C with 20 µg anti-PAX3-FKHR antibody PFM2. Protein G-Sepharose was then added for 1 hr at 4°C. After sequential washes, DNA was eluted from protein G-beads, reversed cross-linked and removed from RNA with RNase A digestion. Proteins were then digested with Proteinase K, and DNA was purified with mini-columns. Total recovered DNA was determined with picogreen and fold enrichment was determined with qPCR using primers (Table S6) specific for MYF5 enhancer for PAX3-FKHR and GAPDH promoter for Pol-II control.

DNA library construction and sequencing

Immunoprecipitated DNA samples were processed and analyzed using on a GA1 sequencer (Illumina). Briefly, DNA ends were repaired using a 1:5 mixture of T4 and Klenow DNA polymerases following the manufacturer's instructions. After the addition of a single adenine base to the DNA using Klenow exo$^-$ enzyme, adapters were ligated to the ends of the adenine-tailed purified DNA. DNA was then size-selected at around 300 bp on a 12% PAGE gel. Adapter-modified DNA fragments were enriched by PCR using Phusion polymerase and PCR primers 1.1 and 2.1 (Illumina) following the manufacturer's

instruction. Cluster generation was done on one channel of the Illumina cell for each sample, and 27 cycles of sequencing were performed on the Illumina cluster station and 1G analyzer.

Processing sequence data

Sequence reads were mapped to the human reference genome (NCBIv36, hg18) using Illumina's extended Eland alignment program. Stacked reads, oriented reads starting at identical positions, as well as low quality reads with more than two deviations from the reference or an alignment score less than 25 were removed from the resulting datasets. Local read densities then estimated by counting coverage of read-events for each nucleotide in the genome, where the oriented reads were extended to the insert length (100bp) size-selected during library preparation.

P-values *p* used to identify significantly increased read densities were estimated based on the cumulative Poisson distribution. The local emission coefficient $\lambda_{(x)}$ was estimated from input (non-IP) data using the average read densities of windows centered around *x* of sizes 1bp, 100bp, 1000bp, respectively. Of those, the most conservative (largest) estimate max $\lambda_{i(x)}$ was used in order to minimize the false discovery rate. Discrete enriched regions were identified using the following heuristic: a continuous stretch of DNA was called significantly enriched if the following conditions were met simultaneously, $p < 10^{-9}$ anywhere within that region, and $p \leq 10^{-6}$ everywhere else. Subsequently, distinct significant regions were merged into a single region if they were closer than ½ fragment size (50bp) apart. Finally, regions determined in this way smaller than half the median fragment length (100bp/2= 50 bp), were rejected.

The bed files for PFM2 IP-seq can be down-loaded from GEO (no. GSE19063). The resulting read-density data obtained from the PFM2 IP experiment can be visualized under pFM2 using UCSC version hg18 at http://genome.ucsc.edu/cgi-bin/hgTracks?org=human&position=chr22&hgt.customText=http://watson.nci.nih.gov/~sven/pFM2.wig.

The discrete regions of enrichment were analyzed for conservation by reporting the maximum phast-score (vertebrate, 44-way conservation scores downloaded from UCSC within the discrete regions of enriched read density.

Identification of putative PAX3-FKHR regulatory target genes

Putative regulatory targets of PAX3-FKHR binding sites were identified based on the closest gene heuristic, this is, for each binding site, the closest RefSeq transcript with a unique Entrez gene was identified. In cases where multiple RefSeq transcripts mapped to the same Entrez-GI, the longest transcript was selected. Gene ontology analysis of the genes was made with Ingenuity (Redwood City, CA).

DNA Motif Analyses

Coordinates of sequences similar to known transcription factor binding site motifs were identified using a position weighted matrix based approach (Quandt et al., 1995), the frequency of known motifs (Transfac 11.4 database) present in regions of enriched read density was counted and compared to their respective frequency in (a) the entire genome or (b) regions selected randomly from the genome. For (b), sequencing data for non-selected (input) DNA was used to generate the random location distribution. P-values for

over-representation were derived using (for a) Fishers exact test or (for b) by counting the number of random iterations, where the frequency of a given motif in the random set was larger or equal to the frequency observed in the ChIP-Seq dataset.

Co-enrichment of other transcription factors around PAX3 motifs were estimated by enlarging a hit for PAX3 with +/- 100 bases from that site. The numbers of co-occurrences were counted and comparing with their respective frequency in the entire genome. Fisher's exact test was used to calculate the P-values.

De novo sequence discovery for novel motifs was performed with DME (Smith et al., 2005). The output sequences were further investigated for similarities using STAMP (http://www.benoslab.pitt.edu/stamp/index.php) using Transfac database.

References

Azorsa, D. O., and Meltzer, P. S. (1999). Production and characterization of monoclonal antibodies to the steroid receptor coactivator AIB1. Hybridoma *18*, 281-287.
Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic acids research *23*, 4878-4884.
Smith, A. D., Sumazin, P., and Zhang, M. Q. (2005). Identifying tissue-selective transcription factor binding sites in vertebrate promoters. Proceedings of the National Academy of Sciences of the United States of America *102*, 1560-1565.