Genomics Data 2 (2014) 192–194

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Identifying direct targets of transcription factor Rfx2 that coordinate ciliogenesis and cell movement

Taejoon Kwon [a,1], Mei-I Chung [a,1], Rakhi Gupta [b], Julie C. Baker [b], John B. Wallingford [a,c,d], Edward M. Marcotte [a,c,*]

[a] Department of Molecular Biosciences, University of Texas at Austin, United States
[b] Department of Genetics, Stanford University, United States
[c] Center for Systems & Synthetic Biology, Institute for Cellular & Molecular Biology, University of Texas at Austin, United States
[d] Howard Hughes Medical Institute, United States

ARTICLE INFO

Article history:
Received 6 June 2014
Accepted 17 June 2014
Available online 26 June 2014

Keywords:
Xenopus laevis
Rfx2
Ciliogenesis

ABSTRACT

Recently, using the frog *Xenopus laevis* as a model system, we showed that the transcription factor Rfx2 coordinates many genes involved in ciliogenesis and cell movement in multiciliated cells (Chung et al., 2014). To our knowledge, it was the first paper to utilize the genomic resources, including genome sequences and interim gene annotations, from the ongoing *X. laevis* genome project. For researchers who are interested in the application of genomics and systems biology approaches in *Xenopus* studies, here we provide additional details about our dataset (NCBI GEO accession number GSE50593) and describe how we analyzed RNA-seq and ChIP-seq data to identify direct targets of Rfx2.

Specifications	
Organism/cell line/tissue	*Xenopus laevis* animal caps (dissected ectoderm) or whole embryos
Sex	Not specified
Sequencer or array type	Illumina HiSeq2000
Data format	FASTQ (raw); tab-delimited text files (processed)
Experimental factors	RNA-seq: wild-type control vs Rfx2 morphants (100 animal caps at developmental stage 20) ChIP-seq: GFP vs Rfx2-GFP (600 whole embryos at developmental stage 20)
Experimental features	Very brief experimental description
Consent	All raw sequencing data are free to use. Genome and gene annotation data is free to use for high-throughput experiment data analysis, such as RNA-seq, ChIP-seq, or proteomics. Otherwise, please contact us (Edward Marcotte marcotte@icmb.utexas.edu) or a member of the International Xenopus Genome Consortium (Daniel Rokhsar dsrokhsar@lbl.gov or Masanori Taira m_taira@biol.s.u-tokyo.ac.jp).
Sample source location	N/A

Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50593.

Experimental design, materials and methods

RNA-seq experiments

The detailed procedure for our sample preparation has been previously reported [1]. Briefly, we injected 12 ng of morpholino into 4-cell stage *X. laevis* embryos to knock down Rfx2 expression; the morpholino sequence has been previously reported [1]. We then prepared 100 animal caps (ectodermal explants of stage 10 *X. laevis* embryos, dissected with forceps), both for control samples and Rfx2 morphants, and cultured them until stage 20. The stage of animal caps was estimated by comparison against embryos from the same clutch. Total RNA was collected using the Trizol method, and then processed using a non-strand-specific Illumina RNA-seq library preparation kit with poly-A enrichment (TruSeq v2). We sequenced these libraries in a 2 × 50 bp paired-end configuration using an Illumina HiSeq 2000.

RNA-seq analysis

The *X. laevis* genome project was ongoing when we collected these data, so for this study we used a draft genome sequence (JGI

[*] Corresponding author at: 2500 Speedway MBB 3.148, University of Texas, Austin, TX 78712, USA. Tel.: +1 512 471 5435; fax: +1 512 232 3472.
E-mail address: marcotte@icmb.utexas.edu (E.M. Marcotte).
[1] These authors contributed equally.

http://dx.doi.org/10.1016/j.gdata.2014.06.015

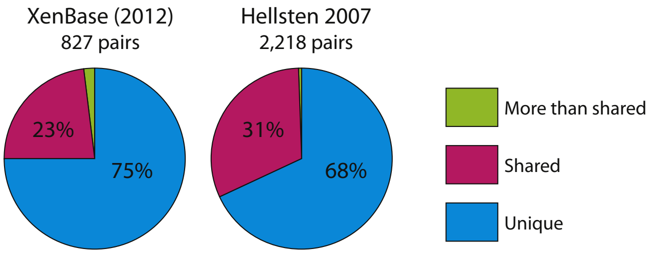

Fig. 1. Mapping results of RNA-seq reads on *X. laevis* homoeologs."Unique", "Shared", and "More than shared" indicate reads that map only once, twice, or more than twice to the set of *X. laevis* homoeologs, respectively.

version 6.0 genome scaffolds; available at ftp://ftp.xenbase.org/pub/Genomics/JGI/Xenla6.0/) and annotation ('Oktoberfest' version of putative transcripts, mainly derived from RNA-seq de novo assembly and then confirmed against JGI version 6.0 genome scaffolds; see http://www.marcottelab.org/index.php/XENLA_Oktoberfest for more details). All scaffolds and transcripts are available at XenBase (ftp://xenbaseturbofrog.org/sequence_information/UTA/) and our supplementary website (http://www.marcottelab.org/index.php/ChungKwon2013_RFX2). Because it is easier for gene-level expression analysis, we conducted RNA-seq mapping against putative transcripts rather than the whole genome. Using bowtie1 (version 0.12.7) [4], we mapped our RNA-seq reads to the Oktoberfest models (which contain 25,537 putative transcripts for each gene) using the longest transcript model for each locus. Then we used edgeR [6] to identify differentially expressed genes, focusing on genes with greater than 2-fold difference and a false discovery rate less than 0.05.

One of the challenges in *X. laevis* RNA-seq analysis is the presence of homoeologs, i.e. duplicated genes that arise as a result of allotetraploidy. Using an allowance of 2 mismatches within a 50-bp read (the '-v 2' option in bowtie1), we evaluated how many reads were mapped interchangeably between homoeologs. We used two datasets for this test: (1) 827 gene pairs previously identified by a variety of labs and curated at XenBase using an '-a/-b' gene name suffix [3], and (2) 2218 assembled EST pairs identified as involved in a trio relationship with *Xenopus tropicalis* [2]. As shown in Fig. 1, 68–75% of reads were uniquely mapped and only 23–31% of reads were mapped to both duplicated genes. We were particularly interested in the differential expression between wild-type embryos and Rfx2 morphants. Thus, in order to maximize the expression signals in our analysis, we allowed for all possible hits in mapping with the '-a' option (i.e. interchangeably mapped reads would be counted twice), and then conducted differential expression analysis. We also tested (1) randomly assigning multi-hit reads to a 'best target' and (2) using only uniquely mapped reads. Ultimately we found no major differences in differential analysis between these approaches (data not shown).

Out of 24,089 *X. laevis* transcripts detected in our RNA-seq experiments, 3209 transcripts were down-regulated in the Rfx2 knockdown condition, and 1523 transcripts were up-regulated. To perform functional network analysis using HumanNet [5], we converted these gene lists to human orthologs (based on EnsEMBL version 69). Note that initial orthology assignments are already captured by the *X. laevis* Oktoberfest transcript gene names, because as part of the transcript set construction, all *X. laevis* protein sequence candidates were compared to the reference proteome of five different species (human,

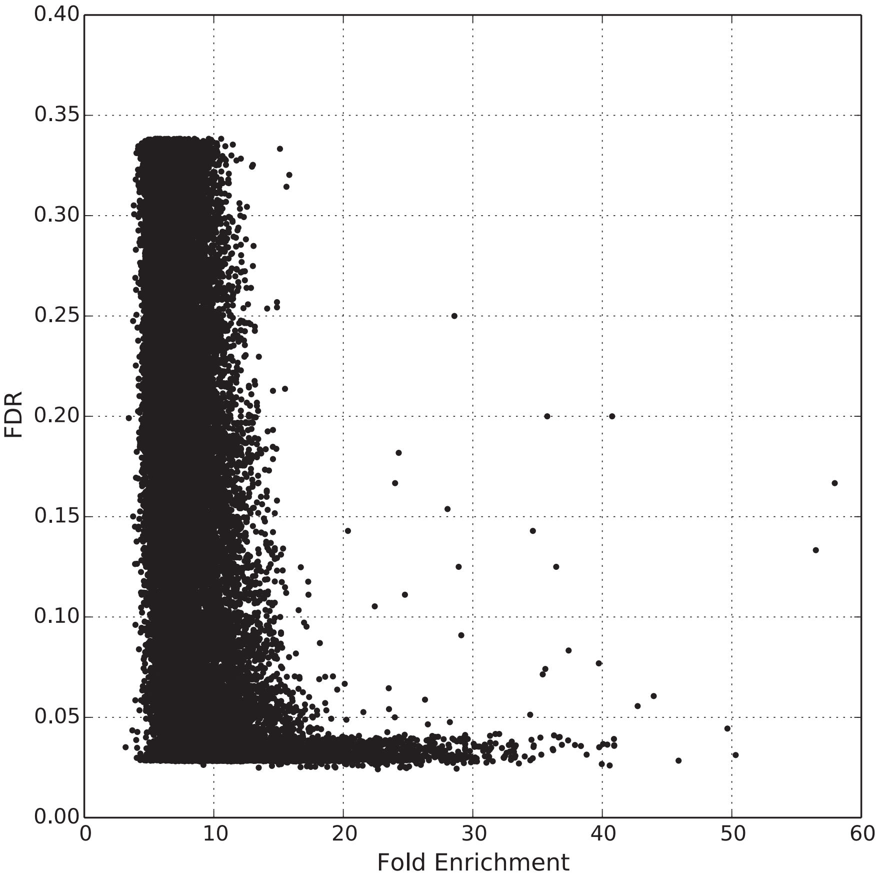

Fig. 2. Distribution of fold enrichment and false discovery rate (FDR) in ChIP-seq peak calling. Most peaks with greater than 20 fold enrichment exhibit an FDR less than 0.05. Based on this observation, we included peaks with FDR greater than 0.05 in successive analyses if they exhibited a fold enrichment greater than 20.

mouse, zebrafish, chicken, and *X. tropicalis*) for the purpose of assigning gene names consistent with the human orthologs. For homoeologs, if only one of two duplicated genes was determined to be significantly differentially expressed but not the other, we still assigned the corresponding human gene as being differentially expressed. After converting all *X. laevis* genes into human orthologs, we identified 2750 human candidate genes transcriptionally regulated by Rfx2.

ChIP-seq experiments

The detailed description of our ChIP-seq sample preparation has been previously reported [1]. Briefly, we injected mRNA encoding GFP-tagged Rfx2 into 4-cell stage *X. laevis* embryos and then pulled down the tagged protein with α-GFP antibody (ab290) from 600 whole embryos (stage 20). Before immunoprecipitation, we cross-linked Rfx2-genomic DNA complexes with 1% formaldehyde and fragmented them with a Branson 450 Sonifier (expected fragment size was from 200 to 500 bp). As a control, we injected GFP messenger RNAs alone and conducted the same immunoprecipitation procedure. DNA fragments were extracted with phenol–chloroform and purified with a QIAquick PCR purification kit (Qiagen). Sequencing libraries were prepared with a standard Illumina genomic library construction kit (TruSeq) and sequenced with an Illumina HiSeq 2000 in 1 × 50 bp configuration.

ChIP-seq analysis

Similar to the RNA-seq data analysis, we conducted ChIP-seq analysis to discriminate between homoeolog genes. We applied a more stringent criteria for ChIP-seq read mapping, requiring uniquely mapped reads to the genome scaffold (JGI version 6.0) and a maximum of 2 mismatches within the seed sequence (i.e. the '-m 1 -n 2' options in bowtie1 [4]). For peak calling, we used MACS (version 1.4.2) with default options [7].

We initially determined significant Rfx2-bound peaks by using a false discovery rate (FDR) cutoff (<0.05) reported by MACS. However, as shown in Fig. 2, only a few peaks demonstrated an FDR above 0.05 if the fold enrichment of the peak was greater than 20, so we included these peaks as well in our further analysis. For each peak, we assigned the closest protein-coding gene as its target gene, so long as it was within 10 kb. As shown in Fig. 3, most of these peaks were located less than 1000 bp from the transcript start site of their assigned gene, suggesting that, if anything, our criteria for associating ChIP-seq peaks to target genes were over-generous.

Out of 29,448 peaks identified in total, 6646 peaks were selected for further study that exhibited either an FDR < 5% or a fold-enrichment >20, and 5024 of those peaks were assigned to their neighboring genes. As with our RNA-seq data analysis, we converted the 5024 *X. laevis* target gene IDs to human genes, collapsing duplicated genes into a single human ortholog based on their names. This analysis resulted in a final set of 911 putative directly bound Rfx2 target genes that also showed significantly differential gene expression after Rfx2 knockdown [1]. A list of all 911 genes is available in Supplemental File 1 in our previous report [1].

Acknowledgments

This work was supported by grants to J.B.W. from the NIH/NIGMS and NIH/NHLBI; to J.C.B. from the NIH/NIGMS; and to E.M.M. from the NIH/NIGMS, the NSF, the Cancer Prevention Research Institute of Texas, the U.S. Army (58343-MA), and the Welch Foundation (F1515). J.B.W. is an Early Career Scientist of the Howard Hughes Medical Institute.

References

[1] M.-I. Chung, T. Kwon, F. Tu, E.R. Brooks, R. Gupta, M. Meyer, J.C. Baker, E.M. Marcotte, J.B. Wallingford, Coordinated genomic control of ciliogenesis and cell movement by RFX2. Elife 3 (2014) e01439.

[2] U. Hellsten, M.K. Khokha, T.C. Grammer, R.M. Harland, P. Richardson, D.S. Rokhsar, Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog *Xenopus laevis*. BMC Biol. 5 (2007) 31.

[3] C. James-Zorn, V.G. Ponferrada, C.J. Jarabek, K.A. Burns, E.J. Segerdell, J. Lee, K. Snyder, B. Bhattacharyya, J.B. Karpinka, J. Fortriede, et al., Xenbase: expansion and updates of the *Xenopus* model organism database. Nucleic Acids Res. 41 (2013) D865–D870.

[4] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10 (2009) R25.

[5] I. Lee, U.M. Blom, P.I. Wang, J.E. Shim, E.M. Marcotte, Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 21 (2011) 1109–1121.

[6] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 (2010) 139–140.

[7] Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nusbaum, R.M. Myers, M. Brown, W. Li, et al., Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9 (2008) R137.

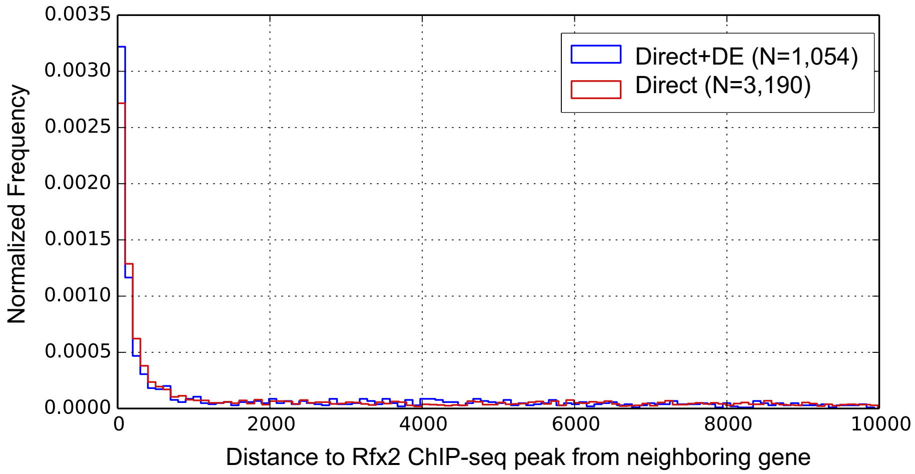

Fig. 3. Distance between ChIP-seq-identified Rfx2 binding sites and nearby genes. "Direct + DE" represents genes that have an Rfx2 binding peak and a significantly differentially expressed pattern in the Rfx2 knockdown condition. "Direct" represents genes that have an Rfx2 binding peak but lack significant differential expression in Rfx2 knockdown. In both cases, however, most peaks are located less than 1000 bp away from annotated genes.