# In Vivo Validation of a Computationally Predicted Conserved Ath5 Target Gene Set

Filippo Del Bene[1,2,3,4☯], Laurence Ettwiller[1,5,6☯], Dorota Skowronska-Krawczyk[7], Herwig Baier[2,3,4], Jean-Marc Matter[7], Ewan Birney[5*], Joachim Wittbrodt[1*]

1 Developmental Biology Programme, European Molecular Biology Laboratory, Heidelberg, Heidelberg, Germany, 2 Department of Physiology, University of California San Francisco, San Francisco, California, United States of America, 3 Programs in Neuroscience, Genetics, and Developmental Biology, University of California San Francisco, San Francisco, California, United States of America, 4 Center for Human Genetics, University of California San Francisco, San Francisco, California, United States of America, 5 European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, 6 Neurobiologie et diversité cellulaire, Ecole Supérieure de Physique et de Chimie Industrielles, CNRS, UMR7637, Paris, France, 7 University of Lausanne, Eye Hospital Jules Gonin, Lausanne, Switzerland

**So far, the computational identification of transcription factor binding sites is hampered by the complexity of vertebrate genomes. Here we present an in silico procedure to predict target sites of a transcription factor in complex genomes using its binding site. In a first step sequence, comparison of closely related genomes identifies the binding sites in conserved *cis*-regulatory regions (phylogenetic footprinting). Subsequently, more remote genomes are introduced into the comparison to identify highly conserved and therefore putatively functional binding sites (phylogenetic filtering). When applied to the binding site of atonal homolog 5 (Ath5 or ATOH7), this procedure efficiently filters evolutionarily conserved binding sites out of more than 300,000 instances in a vertebrate genome. We validate a selection of the linked target genes by showing coexpression with and transcriptional regulation by Ath5. Finally, chromatin immunoprecipitation demonstrates the occupancy of the target gene promoters by Ath5. Thus, our procedure, applied to whole genomes, is a fast and predictive tool to in silico filter the target genes of a given transcription factor with defined binding site.**

Citation: Del Bene F, Ettwiller L, Skowronska-Krawczyk D, Baier H, Matter JM, et al. (2007) In vivo validation of a computationally predicted conserved *Ath5* target gene set. PLoS Genet 3(9): e159. doi:10.1371/journal.pgen.0030159

## Introduction

To understand regulatory networks, it is important to unravel the direct interactions of its transcriptional regulators. For this, the corresponding transcription factor binding sites in the upstream region of the respective target genes have to be identified. However, available approaches have not been able to overcome problems related to the fact that the transcription factor binding sites are short (6–20 bp) and consequently are found very frequently, spread all over the genome. These motifs are functional in only a small fraction of their instances [1]. It has been suggested that epigenetic processes, in particular histone modifications, permit or prevent the access to chromatin [2].

Cooperative binding of multiple transcription factors to combinations of motifs also account for the high selectivity in vivo. Combinations of transcription factor binding sites have therefore been used to computationally predict regulatory modules [3–6]. Comparative genomic approaches applied methods commonly termed "phylogenetic footprinting" [7]. These techniques are based on the fact that functional genomic regions are under selective pressure, resulting in the evolutionary conservation of the respective sequences. Phylogenetic footprinting identifies conserved stretches of noncoding DNA in sequence alignments of related species with limited complexity. To apply this approach to complex genomes, the complexity can be reduced by focusing on the sequences flanking identified genes.

In closely related species, neutrally evolving sequences, as well as functionally relevant and therefore conserved sequences, result in an alignment. Consequently, functional motifs are masked by the high degree of overall sequence similarity. On the other hand, if the genomes are too diverged, sequence comparison may fail to detect short conserved functional motifs due to the lack of significant alignment. Thus, the evolutionary distance of the genomes analyzed has to be considered.

To overcome these problems, we developed a novel evolutionary filtering approach that takes advantage of the increasing number of sequenced vertebrate genomes. In a first step, we limited the complexity of closely related genomes by restricting the analysis to the upstream region of annotated genes. Considering only those genes that contain a transcription factor binding site in this region, we subsequently performed alignments with their orthologs from closely related genomes. In the second step, the regions of their orthologs in more diverged genomes were scanned for the presence of the motif. This evolutionary double

**Editor:** Stuart K. Kim, Stanford University Medical Center, United States of America

**Received** March 15, 2007; **Accepted** August 1, 2007; **Published** September 21, 2007

A previous version of this article appeared as an Early Online Release on August 1, 2007 (doi:10.1371/journal.pgen.0030159.eor).

**Abbreviations:** bHLH, basic helix loop helix; ChIP, chromatin immunoprecipitation; EMSA, electrophoresis mobility shift assay; GO, gene ontology; PWM, position weight matrix; RGC, retinal ganglion cell

* To whom correspondence should be addressed. E-mail: Jochen.Wittbrodt@embl.de (JW); (birney@ebi.ac.uk (EB)

☯ These authors contributed equally to this work.

## Author Summary

To establish regulatory gene networks that drive key biological processes is of crucial importance to identify the genes that are directly controlled by transcriptional regulators. Ideally, this can be accomplished by identifying the direct transcription factor binding site in the *cis*-regulatory regions of the respective target genes. However, problems related to the fact that the motifs recognized and bound by transcription factors are short (6–20 bp) and consequently found very frequently and spread all over the genome, have limited this approach. The transcription factor Ath5 is involved in the specification and differentiation of retinal ganglion cells in the developing vertebrate eye. We show that Ath5 directly regulates its own expression by binding to a small region of its proximal promoter that contains two identical motifs. Using this motif description, together with conservation across large evolutionary distances, we then searched in the genome for other target genes of Ath5 and predicted 166 direct target genes. We then validated a subset of these predictions both in vitro and in vivo. Our analysis therefore provides an example of computation prediction of transcriptional target genes. At the same time, the genes identified represent the most comprehensive list of effectors mediating the role of Ath5 during eye development.

filtering allowed to identify—in the large number of occurrences of a short motif—the small number of evolutionarily conserved transcription factor binding sites.

We benchmarked this procedure using the available dataset for the transcription factor E2F by comparing the results with the existing chromatin immunoprecipitation (ChIP) on chip [8]. Eighty-five percent of our in silico predicted targets contained in the ChIP on chip dataset were experimentally validated. This demonstrates the predictive power of the procedure in the context of the complex human genome.

We next used our procedure to de novo identify of a set of Ath5 target genes. The basic helix loop helix (bHLH) transcription factor Ath5 is a key regulator of vertebrate retinal development. Ath5 is required for the differentiation of retinal ganglion cells (RGCs), which provide the axonal link of the retina to the respective visual centers [9–11]. Loss of *ath5* function results in the absence of RGC formation in vertebrates [12–14]. Conversely, gain of *ath5* function by overexpression in the retina promotes RGC formation [15,16]. So far, only a few Ath5 target genes have been identified, including Ath5 itself [17,18] and its binding site is only poorly defined.

We show that Ath5 interacts with its own promoter and autoregulates its own expression via binding to an extended E-box motif (CCACCTG) containing the consensus site recognized by bHLH transcription factors [19]. Using this motif, we predict by phylogenetic double filtering a conserved set of target genes and experimentally validate a number of those targets in vivo.

## Results

### Identification of the Ath5 Binding Site

We first experimentally defined an Ath5 binding site to be used as a signature for the computational prediction of its conserved target genes. Ath5 had been shown to control its own expression in a conserved positive regulatory feedback loop [17,18]. Since our aim is to identify conserved target genes, we also searched for motifs within the Ath5 regulatory region that are conserved throughout vertebrates. In a comparative approach using promoterwise (http://www.ebi.ac.uk/~birney/wise2/) we identified two evolutionarily conserved (from teleosts to mammals) extended E-box motifs (CCACCTG) within 2 kb of upstream sequences that in medaka fish embryos faithfully recapitulate *ath5* expression in a reporter construct (Figure 1A–1C).

To test the interaction of Ath5 with these conserved CCACCTG motifs, electrophoresis mobility shift assays (EMSAs) were performed with oligos containing the two wild-type motifs or different variants in which the motif was altered with or without affecting the E-box consensus (see Materials and Methods). We found that the presence of at least one E-box was sufficient to allow binding of Ath5. Binding was only abolished if the consensus E-box in both motifs was changed (Figure S1A). Furthermore, only those oligos in which one of the E-boxes was preserved competed with the wild-type probe when added in excess (Figure S1B). Those results confirm the specificity of the interaction and indicate a high affinity of Ath5 for the conserved CCACCTG motif.

To investigate the ability of Ath5 to activate its own promoter, we used cos7 cells in a luciferase transcription assay. As previously demonstrated for chick Ath5 [17], the medaka 2-kb Ath5 promoter is also strongly activated by Ath5 in a dose-dependent manner (Figure 1D). Our mutational analysis revealed that changing one of the motifs while preserving the E-box consensus results in reduced transcriptional activation (2-fold versus 6.5-fold of the wild-type promoter; Figure 1D). No activation was observed in all the other variants tested. Furthermore, embryos injected with corresponding GFP reporter constructs, in which the E-box consensus in the two conserved motifs is disrupted, failed to express GFP in the endogenous domain (unpublished data; see also Materials and Methods). This indicates that only the identified CCACCTG motifs are efficiently recognized and bound by Ath5.

### Computational Prediction of the Gene Set Regulated by Ath5

To identify functional target sites, and, consequently a conserved target gene set of a given transcription factor in a genome-wide manner, we devised a multistep procedure that relies on the evolutionary conservation of functionally relevant transcription factor binding sites.

First, we reduce the complexity by limiting the search space to the region upstream of annotated human genes. We subsequently search for the presence of the motif corresponding to the transcription factor binding site in a conserved region with rodents (see Materials and Methods for the definition of conservation). In a last step, we scan orthologous regions in more diverged species for the presence of the motif. This additional filtering step is independent of any alignment, i.e., the motif does not have to lie in a conserved stretch. All the genes with an upstream region that passes the last filter are defined as the predicted target genes of the analyzed transcription factor (see Text S1 for details).

To assess the performance of our in silico procedure, we benchmarked it using the binding site of the transcription factors E2F (Transfac, Jaspar [20,21]) by comparing our dataset with that obtained by ChIP [8]. The details of the

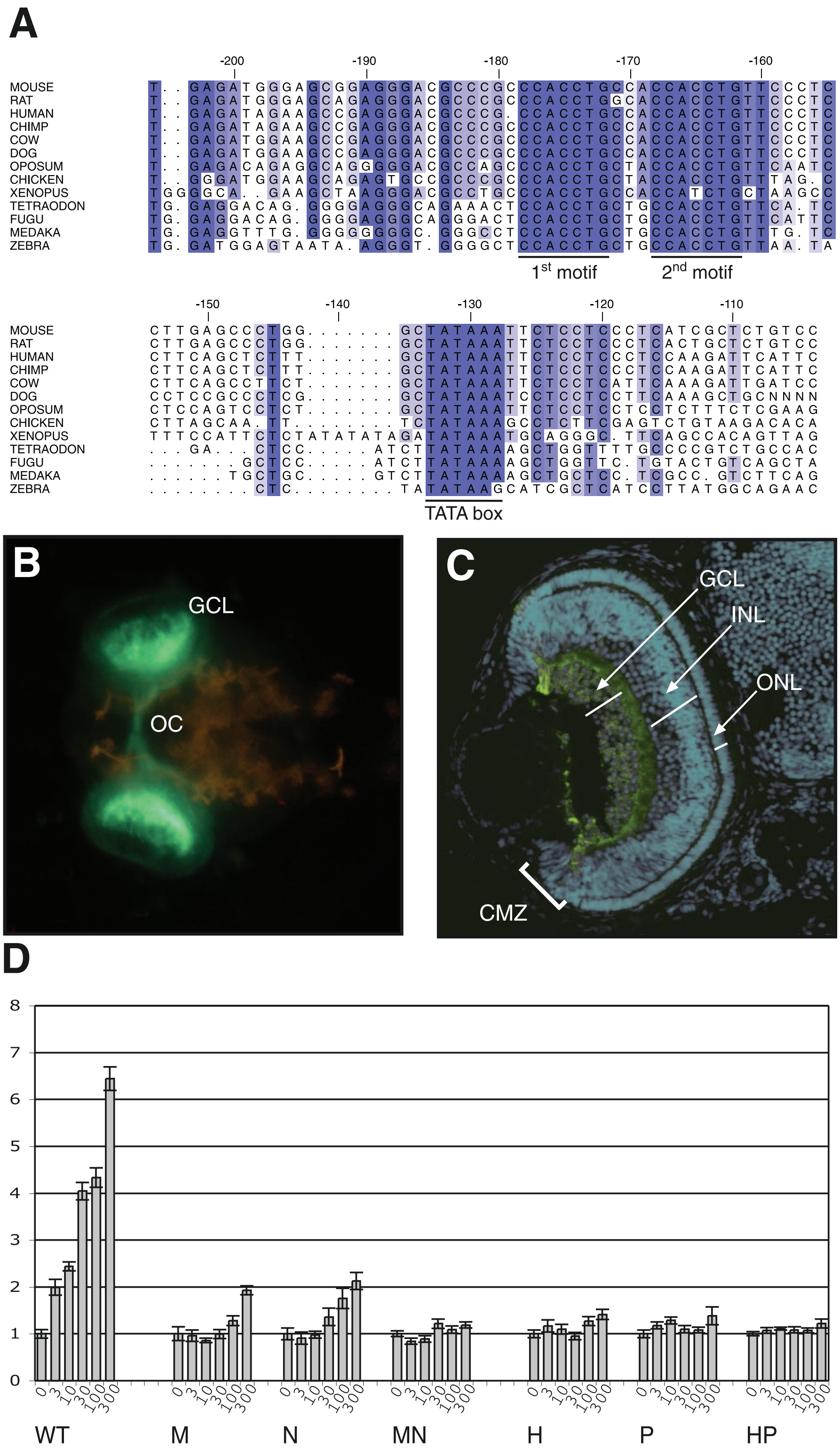

**Figure 1.** Ath5 Promoter Alignment and Characterization

(A) Proximal promoter region alignment of *ath5* in various vertebrate species. Numbering indicates bases 5′ of the start codon.

(B, C) Transgenic medaka embryos at day four of development. (C) Cryosection and anti-GFP antibody staining of a transgenic embryo representing the full Ath5 expression pattern. Note that the reporter is weakly expressed all over the mature retina and specifically enriched in the GCL and in a few other cells indicated by arrows. Complete absence of expression is observed in the distal part of the ciliary marginal zone as well as outside the eye. Nuclei are visualized by DAPI counterstaining. CMZ, ciliary marginal zone; GCL, ganglion cell layer; INL, inner nuclear layer; ONL outer nuclear layer; and OC, optic chiasm.

(D) Transactivation of Ath5 on its own promoter assayed by luciferase reporter transcription activity. Two-kilobase Ath5 promoter or its mutant forms generated by PCR introducing base changes into the two conserved binding motifs were used (see Table S3 for oligo sequences). Values on the *x*-axis are the quantity (in ng) of Ath5pCS2+ -expressing vector transfected. WT corresponds to the wild-type promoter, M and N are mutations of the first and

second conserved e-boxes, respectively, without disrupting the e-box consensus sequence. H and P are mutations of the first and second conserved e-boxes, respectively, with disruption of the e-box consensus sequence. MN and HP are mutations in both e-boxes without and with disruption of both e-boxes, respectively (see Materials and Methods for details)
doi:10.1371/journal.pgen.0030159.g001

benchmarking procedure are described in Text S1. Of the 1,342 genes with Ensembl identification numbers that were tested by Ren et al. [8], we predict 14 to be bound by E2F, of which 12 (85.7 %) are correct. This is a significant improvement over a control where genes are randomly sampled (*p*-value $< 0.00001$). We note, however, that our stringent conservation requirement misses 89% of the bound genes. Low sensitivity is, at this point, an unavoidable consequence of comparative studies that aim at high specificity using evolutionarily distant species.

Using the defined Ath5 binding site, we applied our evolutionary double filtering procedure to identify conserved Ath5 binding sites and, by this, potential Ath5 target genes. In previous studies, the majority of conserved regulatory regions had been found within 5 kb upstream of genes [22]. Therefore, in our search for the Ath5 binding site, we concentrated on the 5-kb upstream sequence of all annotated genes in the vertebrate genomes analyzed. Candidate genes were thus identified by the presence of the conserved CCACCTG motif or its corresponding reverse complement within this region. Our procedure (Text S1) filtered the number of occurrences of the 7-bp Ath5 binding site from about 324,000 instances in the entire human genome (Ensembl v42, repeat masked sequences) to 166 evolutionarily conserved sites and the corresponding genes (Table S1). We noted that the majority of these sites are found within the first 2 kb upstream of the annotated transcriptional start site (Figure S2). This is in contrast to the random distribution of Ath5 motifs present in the 5 kb upstream sequences of all annotated human genes and further confirms previous

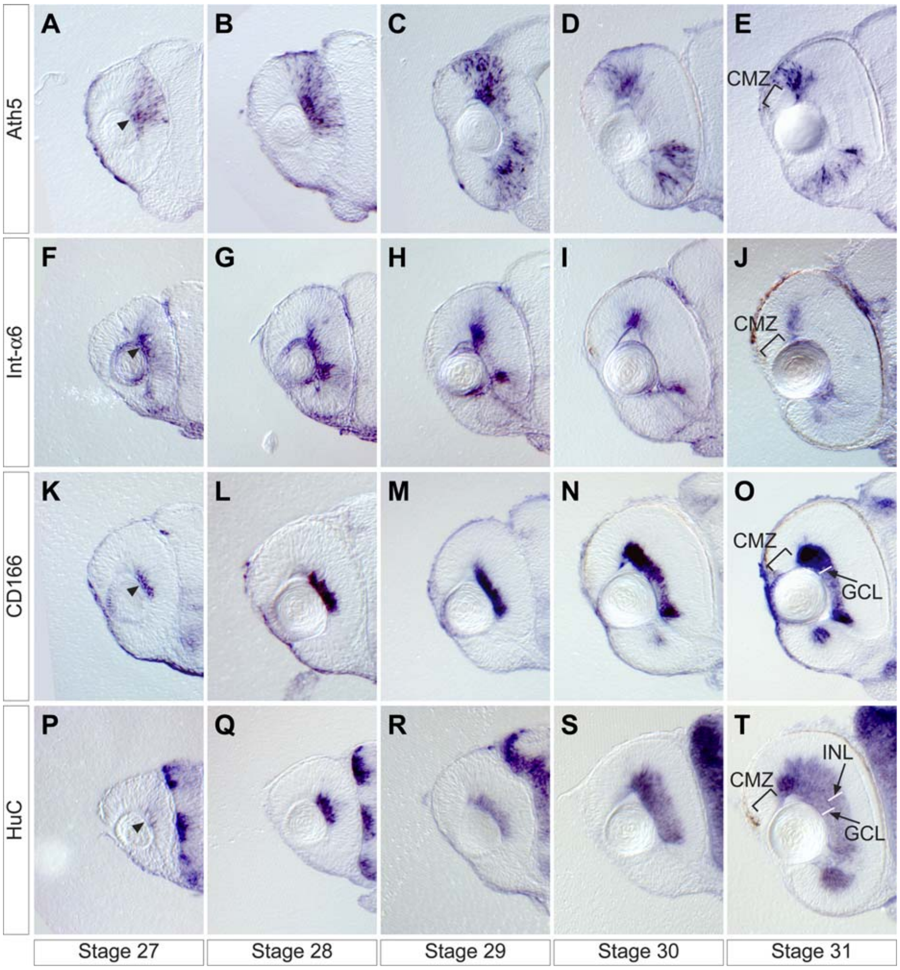

**Figure 2.** Developmental Time-Course of the Expression of *ath5* (A–E) and Three Examples Representing the Major Groups of Target Genes

Embryos are stained by whole mount in situ hybridization and sectioned transversally at the level of the optic nerve. *int-α6* overlaps with *ath5* expression and follows it (F–J). *CD166* is continuous with the *ath5* expression domain and then remains in the cells that had expressed *ath5* in the ganglion cell layer (K–O). *HuC* is expressed in the ganglion cell layer and part of the inner nuclear layer (P–T), during medaka retina development from stage 27 to stage 31. Arrowheads indicate the simultaneous onset of the expression in the central retina of the target genes within *ath5* expression domain at stage 27. CMZ, ciliary marginal zone; GCL, ganglion cell layer; and INL, inner nuclear layer.
doi:10.1371/journal.pgen.0030159.g002

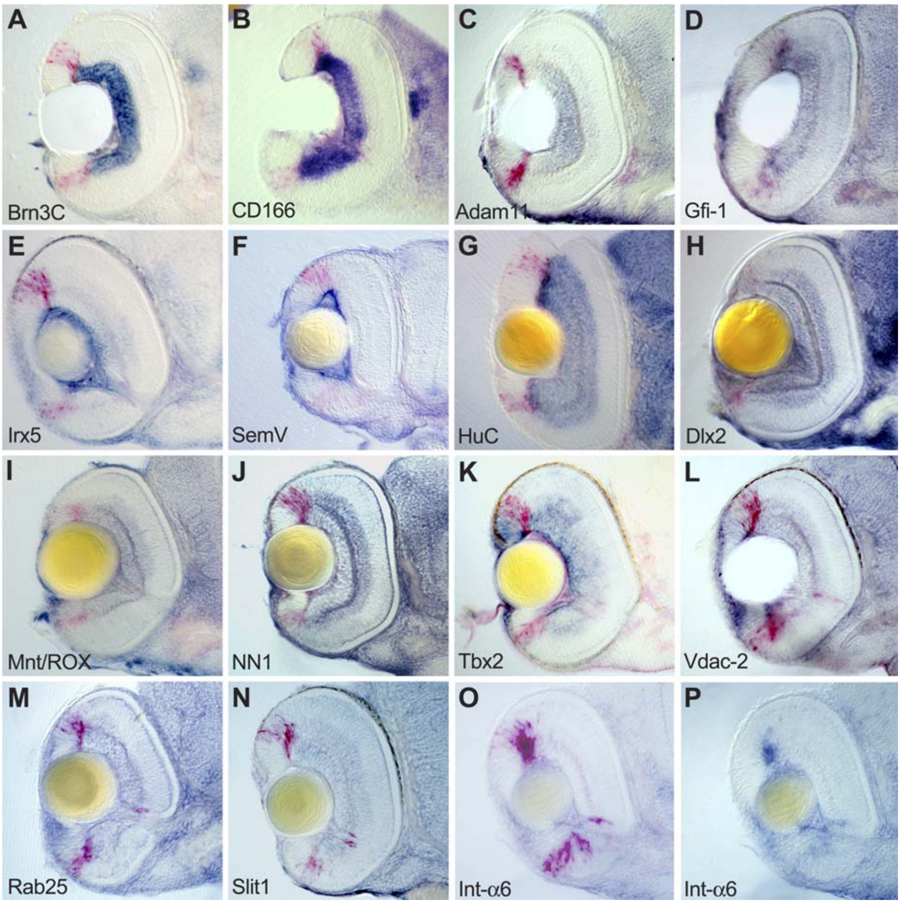

**Figure 3.** Double In Situ for *ath5* (Red in All Pictures) and Target Gene (Blue) Expression in Stage 32 (Day 4) Medaka Retinae

(A–P) Embryos are sectioned transversally at the level of the optic nerve. Dorsal is to the top. Abbreviations of target genes are indicated in Table S1.

doi:10.1371/journal.pgen.0030159.g003

studies on the position of relevant regulatory elements [23] relative to the gene start.

## Computational Assessment and Experimental Validation of the Predicted Target Genes

We compared the gene ontology annotation (GO) [24] of the identified gene set to that of the entire annotated human genome (Figure S2). We found an enrichment of the cellular component "nucleus" ($p$ = 1.2e–04), the biological process "transcription factor activity" ($p$ = 1.40e–08), and the biological function "development" ($p$ = 7.02e–12). We organized the set of predicted target genes of Ath5 into functional categories: transcription factor, neuronal function, axon guidance and growth, cell cycle and signaling, development, and others ($n$ = 166; Table S1 and references therein).

We analyzed the expression pattern of thirty predicted target genes within relevant categories (see Table S1) in the medaka fish retina in comparison to *ath5* expression (Figures 2 and 3; unpublished data) by whole mount in situ hybridization and found retinal expression for 19 of them (Figures 2 and 3; unpublished data).

At the onset of retinal differentiation (stage 27) [25], all the target genes expressed in the retina show an expression overlapping with that of *ath5* in the central retina (Figure 2; unpublished data). At subsequent retinal differentiation stages, the expression patterns of the different target genes can be classified into three major groups. In the first group, the expression pattern remains entirely overlapping with that of *ath5* (Figure 2A–2J). The second group is composed of genes expressed late in mature RGCs in the central retina, abutting the *ath5* expression domain (Figure 2K–2O). In the third group, in addition to the GCL, late expression is also found in neurons of the inner nuclear layer (Figure 2P–2T).

Analyzing the predicted target genes with respect to the GO categories we found Ath5 among the transcription factors, in agreement with its autoregulatory function, as well as a number of factors that have been implicated to function in RGC differentiation, including Brn3C (POU4F3, Figure 3A), Gfi-1 (GFI1, Figure 3D), Irx5 (IRX5, Figure 3E), Dlx2 (DLX2, Figure 3H), Dlx1 (DLX1), and Tbx2 (TBX2 Figure 3K) [26]. In some cases, their involvement in differentiation and/or survival of RGCs has been well documented, such as for Brn3C and Dlx1/Dlx2 (Figure 3A and 3H) [16,27,28]. The majority of the genes in the category "neuronal function" are ion channels such as the voltage dependent anion channel Vdac-2 (Figure 3L). This category also contains the RNA binding protein ELAVL3 (HuC, ElavC), which has been shown to function in RGC development (Figure 3G). The category "axon guidance" contains the cell adhesion molecules CD166 (ALCAM, Figure 3B), MCAM, Slit-1 (SLIT1, Figure 3N) and integrin alpha-6 (Int-α6, ITGA6, Figure 3O

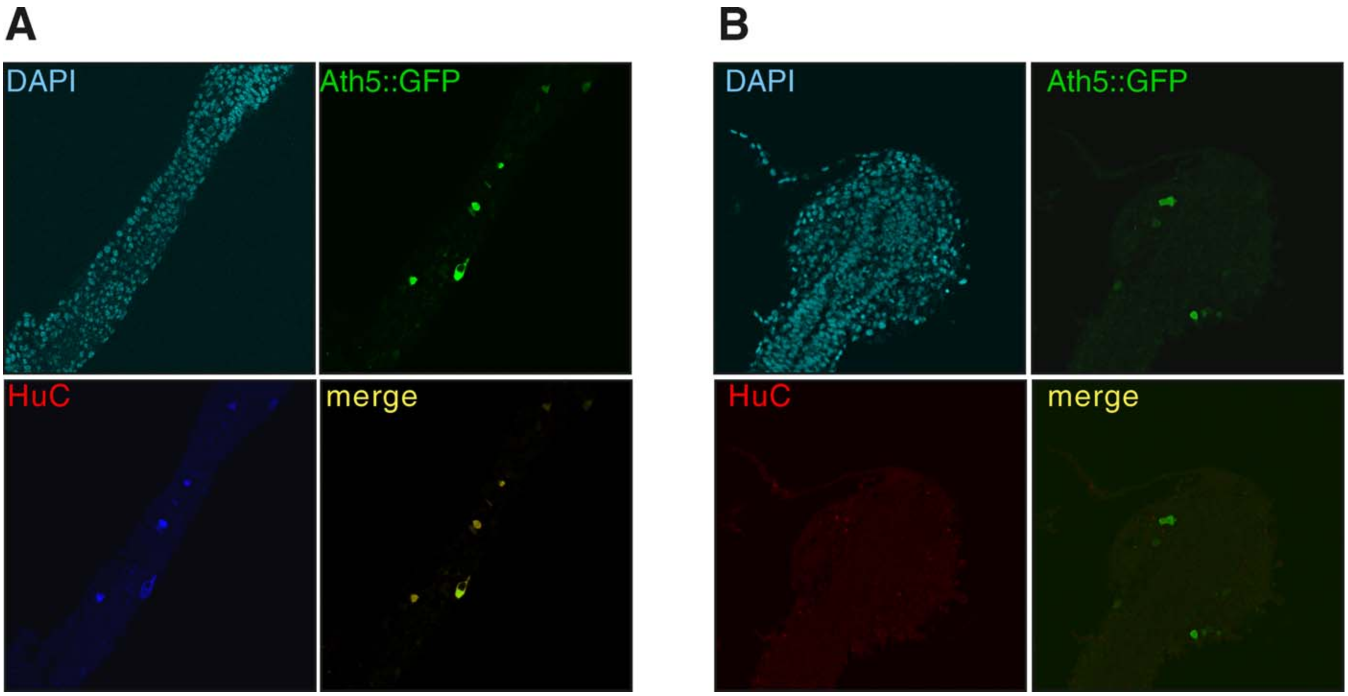

**Figure 4.** Ectopic Activation of Target Genes in Response to Ectopic Ath5 In Vivo

(A, B) Mosaic expression of Ath5 leads to activation of target genes in those cells expressing Ath5. Upper left, DAPI nuclear staining; lower left, HuC fluorescent in situ; upper right, GFP reporter indicating Ath5 expression; and lower right, merge of GFP and in situ signals. (A) Medaka embryo injected with 2-kb Ath5 promoter driving GFP expression in response to Ath5 and Ath5 pCS2+ expression vector. Note the complete overlap of Ath5 activity and expression of its target HuC. (B) Medaka embryo control injected with 2-kb Ath5 promoter driving GFP expression and empty pCS2+ vector.

doi:10.1371/journal.pgen.0030159.g004

and 3P) that play a role in axonal guidance (see Table S1). Furthermore, this category contains genes that were not previously shown to be expressed in RGCs. Our analysis confirmed expression in RGCs for ADAM11 (Figure 3C) and NN1 (NAV1, Figure 3J). The last category includes genes involved in cell cycle regulation and cell signaling (RAB25, Figure 3M and MNT/ROX, Figure 3I). Some of those genes, e.g., NDRG1 and NDRG2, play a role in cell differentiation, whereas others, e.g., CABLES1 and CABLES2 stimulate neurite outgrowth [29].

In conclusion, we analyzed 30 putative target genes by whole mount in situ hybridization in medaka fish and found retinal expression for 19 of them (Figure 3; unpublished data). The remaining 11 genes either showed no expression or a pattern that was not consistent with regulation by Ath5. Furthermore, retinal expression had already been shown in other species for five additional predicted target genes (Table S1). Thus, out of these 35 genes analyzed, 24 (63%) are expressed in a pattern consistent with their regulation by Ath5 (Figures 2 and 3).

## Functional Validation of Predicted Target Genes

We used ectopic Ath5 expression in the developing medaka embryo to examine the transcriptional regulation of the target genes. To monitor ectopic Ath5 expression, a plasmid expressing Ath5 under the control of a strong and ubiquitous promoter was injected into one-cell stage embryos together with the 2kb Ath5::GFP reporter. This results in a mosaic distribution of the cosegregating plasmids in the injected embryo [30]. Cells expressing Ath5 (as visualized by GFP) also ectopically express the putative Ath5 target gene *HuC,* as visualized by fluorescent in situ hybridization (Figure 4A). Similar results were obtained for other target genes such as *Brn3C* and *CD166* (unpublished data). Control embryos coinjected with the empty expression vector and the Ath5::GFP reporter did not show any colocalization of ectopic GFP with any of the target genes analyzed (Figure 4B). Ectopic overexpression of the related bHLH transcription factors Xath3(Xenopus NeuroM, Neurod4) or Xash1 (Xenopus Ash1) did not result in ectopic activation of these Ath5 targets genes (Figure S3; Table S2; Text S1; unpublished data). Taken together, these experiments show that the expression of *HuC*, *Brn3C,* and *CD166* is specifically activated by Ath5.

We next analyzed whether Ath5 binds to the promoters of the predicted target genes using ChIP on chick retinal chromatin preparations [18]. We concentrated on the chick orthologs of the target genes *Dlx2, HuC, Nn1,* and Int-α6 (Table S1) [31]. Ath5 in vivo occupancy of target sequences was found in all cases tested (Figure 5). As a negative control, in the same extracts we found no Ath5 occupancy of the *neuroM* promoter, a gene also expressed in the retina but not activated by Ath5 [18]. In addition, no occupancy of the Ath5 target sequences was detected in extracts from the optic tectum, where *ath5* is not expressed (Figure 5).

Our results show that our procedure efficiently identifies novel transcriptional targets of Ath5. Out of 35 predicted genes analyzed, 24 are expressed in a pattern consistent with regulation by Ath5. When tested for ectopic induction by Ath5, in fish embryos three out of three tested genes were directly activated by ectopic Ath5. Finally, ChIP showed the

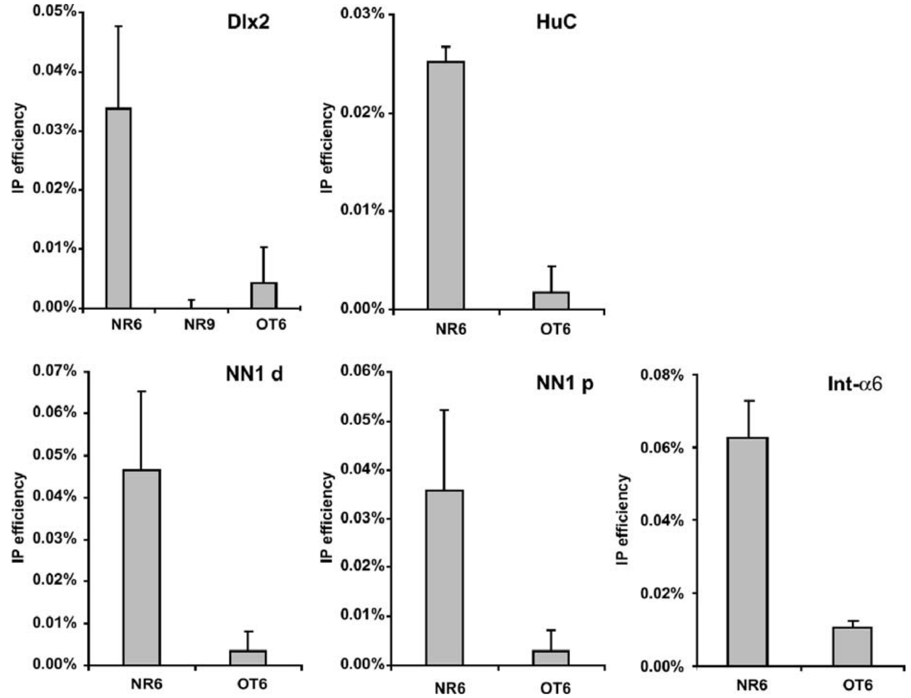

**Figure 5.** In Vivo Occupancy of Four New Target Promoters by Ath5

Antibody directed against Ath5 was used to immunoprecipitate crosslinked chromatin fragments prepared from chicken neuroretina and optic tectum. Occupation was measured by real time PCR with primers specific for the predicted target genes (*Dlx2, HuC, NN-1,* and *Int-α6*). In one case, NN-1, the two predicted binding sites within the promoter region are at a distance sufficient to discriminate between the different occupancies of each of them with the resolution of the ChIP technique (about 700 bp). Both sites are in vivo occupied at different levels. In the *y*-axis, the absolute ChIP efficiency is indicated. NR6, neuroretina at day 6 post fertilization; NR9, neuroretina at day 9 post fertilization; OT6, optic tectum at day 6 post fertilization; d, distal; and p, proximal.

doi:10.1371/journal.pgen.0030159.g005

occupancy by Ath5 of all four (out of four) target loci tested. Some of these target genes have been implicated to function in RGC differentiation. We demonstrate that Ath5 regulates the transcription of these genes and furthermore is bound to their promoter during retinogenesis.

## Discussion

In the work presented here we describe an approach for the identification of relevant target genes that relies on a novel computational procedure. This in silico procedure provides predictions for functionally relevant instances of transcription factor binding sites. This is achieved by a phylogenetic double filtering process that relies on the use of evolutionarily diverged genomes, reducing the large number of spurious motif matches, thereby selecting for the putative functional instances of the motif. Hence, our procedure predicts only evolutionarily conserved targets. Of crucial importance for the efficiency of the procedure is the second filtering step, where diverged genomes are analyzed for the presence of the motif in an alignment-independent way. Recently, a comprehensive list of putative regulatory motifs was identified using annotated vertebrate genomes [23,32], but this work did not identify the direct target genes linked to the motifs.

Our benchmarking analysis demonstrates that our method significantly enriched ($p < 0.00001$) for true target genes of a transcription factor when compared to an experimental data set. Thus, our procedure provides a list of putative targets that have a high probability of being relevant. This list, as illustrated by our Ath5 target gene prediction, represents a valuable starting point for a downstream analysis of this transcriptional network. The use of distantly related fish species for the filtering procedure also implies that the list of predicted targets contains only genes from which the regulation through the transcription factor studied has been retained from mammals to fish. Considering the entire target gene set of a given transcription factor in one species, the nonconserved target genes will be missed using this procedure. This loss and the apparently low sensitivity (89% for the benchmarking using E2F) are intended, and are an unavoidable consequence of comparative studies aiming at high specificity using evolutionarily distant species. With the addition of more entirely sequenced genomes resulting from the ongoing sequencing efforts of many vertebrate species, the sensitivity issue will be improved while retaining similar specificity [33]. This will also allow clade-specific innovations to be addressed, rather than just conserved functions.

A prerequisite for using this procedure is an established binding site for the transcription factor studied. We experimentally identified an Ath5 binding site, relying on the direct Ath5 autoregulation, which is necessary for the upregulation of its expression in RGC precursors [34,35]. Based on this 7-bp Ath5 binding site, we identify 73 putative Ath5-regulated target genes. A recent microarray study on Ath5-regulated genes [26] compared wild-type and Ath5 mutant mouse retinae. The significant ($p = 5 \times 10^{-5}$) but limited (nine genes) overlap between our data set and the microarray study is not surprising, given the different approaches used. While our approach predicts direct targets of a given transcription factor, the microarray analysis does not distinguish between direct and more indirect responses and provides a more global view of the transcriptional differences. This is well supported by our benchmark analysis.

More recently, in a candidate gene approach, a number of transcriptional targets shared by the transcription factors Ath5 and NeuroD in *Xenopus* was reported [36]. Three out of the four *Xenopus* Ath5 target genes with clear orthologs in other vertebrate species were also identified by our procedure, further supporting the significance of our results.

Ath5 is one of the earliest transcription factors specifically expressed in terminally differentiating RGCs, suggesting its key position in the underlying regulatory network. The fact that within the target genes we find a strong enrichment of the GO term "transcription factor activity" is in good accordance with this and provides further evidence for the significance of our results. Within the predicted target genes, we find a strong enrichment of genes acting in cell cycle control, axonal guidance, and neuronal function. Considering that Ath5 is required for the differentiation of neurons that provide axonal connectivity, this finding is in good agreement with the developmental role of Ath5. For example, Ath5 is upregulated shortly before final mitosis [35], and cell cycle exit is a prerequisite for neuronal differentiation. The suggested role of Ath5 in this process is underscored by the enrichment of target genes acting in cell cycle control.

Our target gene validation by whole mount in situ hybridization revealed a coexpression with Ath5 in 63% of the cases analyzed. Furthermore, we show in vivo activation of targets by Ath5. This activation is specific for Ath5, whereas other related bHLH transcription factors fail to activate these targets. This strongly suggests that the identified Ath5 binding site is specifically recognized by Ath5 to activate transcription. Furthermore, in *ath5* mutant *lakritz* embryos the expression of all predicted target genes analyzed is absent from the retina, demonstrating their dependence on Ath5 function (Figure S4; unpublished data). Finally, target gene promoters are occupied in vivo by Ath5 at the time of retinal differentiation, as has been shown for the single established Ath5 target gene, *NachR* [18].

In summary, we present a novel in silico approach that predicts target genes of a given transcription factor. Our benchmarking and experimental application and validation on a novel binding site shows the high predictive power to identify in vivo relevant target genes.

## Materials and Methods

**Computational prediction: Sequence retrieval and orthology assignment.** The 5 kb upstream (1,000 bp upstream regions for the E2F benchmarking) of all annotated genes were retrieved in *Homo sapiens, Mus musculus, Rattus norvegicus, Takifugu rubripes,* and *Danio rerio* using Ensembl version 17. The sequences were repeat masked and exon masked (for possible annotated exon, upstream of the annotated gene start). The gene start was considered to be the annotated start of the longest transcript for each gene. Orthologous gene pairs were taken from the Compara database (version 17) and all the possible pairs were considered; best reciprocal hits as well as Reciprocal Hit based on Synteny around.

**Computational prediction: Alignment of humans and rodents using Promoterwise.** For each human upstream sequence retrieved, the 5-kb orthologous regions in rat and mouse were identified using the downstream gene orthology mapping described above. Pair-wise alignments between human and mouse and human and rat were done using Promoterwise [32]. A conserved region is defined as a region with significant alignments. A significant alignment is defined has having a *promoterwise* hit higher than 25 bitscore. See [32] for the justification of such a cutoff.

**Computational prediction: Filtering procedure for conserved motif.** A conserved site between human and mouse (or rat) is defined as a sequence that satisfies the motif description in both species in one position of the significant alignment. A conserved site between human and a fish is defined as a sequence that satisfies the motif description in both species' 5-kb (or 1-kb for the E2F benchmarking) orthologous region but is not necessarily located in a significant alignment between these two species. The motif description can either be a discrete motif or a position weight matrix (PWM). Both the forward and reverse strand were analyzed.

**Computational prediction: Benchmarking the computational procedure.** The ChIP data from [8] was used to benchmark the computational procedure. From the ChIP data, we used the 130 genes described in Table 3 in [8] as the positive set. The corresponding Ensembl identification numbers were retrieved from the gene annotation (113 genes). The total set corresponds to the entire array used in the experiment ( 1,449 genes from Table S1, of which 1,342 have an Ensembl identification number).

The E2F PWMs (M00516, M00050; Transfac [20],) were used to search E2F target genes as described in the filtering procedure section of the Materials and Methods. The sites were located using the perl module TFBS::pwm [37] with variable score cutoff (ranging from 75% to 100%).

The sensitivity and specificity for each PWM hit cutoff was calculated by comparing the result obtained from the filtering approach to the reference data from [8].

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}),$$

$$\text{Sensitivity} = \text{TN}/(\text{FP} + \text{TN}),$$

with TP (true positive) being the number of genes overlapping between the positive gene set in [8] (113 genes) and the gene set from the filtering procedure (*x* genes depending on the matrix cutoff). FN (false negative) is the number of genes in the positive gene set in [8] (113 genes) minus the TP. FP (false positive) is the number of genes overlapping among the gene sets from the filtering approach and the negative set of [8] (1,342 − 113) and TN (true negative) is the the negative set minus FP.

**Receiver operating characteristic curve plotted from the preceding data.** *Randomization:* A set of genes was randomly sampled from the genes analyzed by [8]. The number of genes in that random set corresponds to the real number of genes found by the computational procedure to overlap with the set of genes analyzed by [8]. The overlap between the random set and the positive set of [8] was assessed and compared with the real overlap obtained using the computational procedure. This randomization procedure was repeated 100,000 times. For example, the filtering dataset using the PWM M00516 with a cutoff of 85% gave 38 candidate genes, out of which 14 overlapped with the 1,342 genes studied by [8] and 12 overlapped with the positive set (113 genes, POS). We randomly picked 14 genes from the genes set studied by [8] (1,342 genes, ALL) and calculated the overlap of this random set with the positive set (113 genes). The procedure was repeated 100,000 times. The average overlap and maximum overlap was assessed.

**Computational prediction: GO category enrichment.** For each GO term identification number (from cellular component, molecular function, biological process), we calculated the number of genes annotated with the GO identification number in the positive set (166 predicted target genes of Ath5) and in the entire human gene set (Ensembl version 17). The enrichment of each GO term identification number was evaluated using hypergeometry distribution [38]. Only GO categories with more than three genes in the positive set were further analyzed.

**Computational prediction: Distribution of the Ath5 motif relative to the transcription start sites.** The positions of the Ath5 motif (CCACCTG) and its reverse complement motif are located on the upstream sequences of the human genes and the distance relative to the annotated start site is calculated (in bp from the longest transcript, Ensembl version 17). The distribution of these relative positions is then analyzed for all the annotated genes in the human genome and compared with the same distribution obtained using only the 166 predicted target genes of Ath5 (see Table S1).

**Medaka and zebrafish stocks.** The Cab strain of wild-type *Oryzias* *latipes* from a closed stock at EMBL-Heidelberg was kept as described [39]. Embryos were staged according to Iwamatsu [25]. Zebrafish *lak* mutants were obtained by crosses of heterozygous *lak*$^{th241}$ carriers.

**Molecular cloning and mutagenesis of the medaka Ath5 promoter.** A fragment of about 60 bp encoding medaka *ath5* homolog was amplified from a 3-d-old embryo cDNA library using degenerate PCR primers (forward ATGCARGGIYTNAAYACNGC, reverse TSICCCCAYTGIGGNACNAC). The PCR conditions were: 5 cycles at 95 °C for 1 min, 50 °C for 1 min and 72 °C for 1 min, followed by 30 annealing cycles at 55 °C. The PCR product was cloned into TOPO TA vector (Invitrogen) and sequenced. Based on this sequence, we designed specific primers for amplifying the full-length cDNA using standard PCR techniques. Full-length *ath5* sequence was cloned in the eukaryotic expression vector pCS2+ for overexpression, in vitro translation, and fluorescein-labeled probe synthesis (see below). The medaka *ath5* cDNA was used to screen a medaka genomic cosmid library. The 5 kb of *ath5* genomic sequence immediately 5′ of the coding region was then cloned into pGL3 (Promega) or into a promoterless GFP reporter (F. Loosli and J. W., unpublished results). The second vector contains recognition sequences for I-SceI meganuclease for efficient transgenesis [40]. Deletion constructs containing 4, 3, 2, or 1.5 kb of 5′ *ath5* genomic region were created by PCR (primer sequences are available upon request). Point mutations in the two Ath5 binding motifs were generated using the QuickChange XL kit (Stratagene). Primer sequences are as follows: WT (GGGGGCGGGCCTCCACCTGCTGCCACCTGTTTGTCTGCTGCG), M (GGGGGCGGGCCTCCAATTGCTGCCACCTGTTTGTCTGCTGCG), N (GGGGGCGGGCCTCCACCTGCTGCCATATGTTTGTCTGCTGCG), NM (GGGGCGGGCCTCCAATGCTGCCATATGTTTGTCTGCTGCG), H (GGGGGCGGGCCTCAAGCTTCTGCCACCTGTTTGTCTGCTGCG), P (GGGGGCGGGCCTCCACCTGCTGCCGATCGTTTGTCTGCTGCG), and HP (GGGGGCGGGCCTCAAGCTTCTGCCGATCGTTTGTCTGCTGCG). See also Table S3.

The Tbx2 and Dlx1 fugu 5′ genomic regions were identified in Ensembl. Two PCR products of 2.6 and 2.3 kb, containing the Ath5 binding motif, were amplified from fugu genomic DNA (Medical Research Council, Rosalind Franklin Centre for Genomics Research) using specific primers (Tbx2 forward GAA CCT CAC GGT GTT GCT CAA AGG CAC and reverse CCT GTT TAT TTG GAC CCG AAA CGA GCG; Dlx1 forward TTG AAT GTG GTG ACC TTT CTG CAG AAG and reverse GGA CGG CTC CCA ATT TAA GTC GAA CTG) and cloned into pGL3. All constructs were verified by sequencing.

**Transgenic procedure.** Transgenic fish embryos were generated as previously described [40]. As previously reported, due to the early integration of the reported construct, we observed a very low or null degree of mosaicism in the injected fish allowing the direct analysis of F0 embryos. Identical patterns of expression were maintained in the following generations (up to F2). Injection of reporter constructs differing in the e-box consensus led to different transgenesis efficiencies. WT: 44/110 embryos reproduce endogenous GFP expression pattern, 40% maximum reachable transgenesis efficiency. Variant H: 44/115, 38%; variant P: 26/107, 24%; and variant HP: 9/111, 8%. See Table S3 for primer sequences. To test the activity of *Xenopus* Ath5 promoter, the pG1X5 3.3-kb construct [34] was injected into embryos at the one-cell stage at a concentration of 20 ng/μl, and embryos were scored 4 d later for GFP expression.

**In situ hybridization.** Double whole mount in situ analysis on medaka embryos was performed using a fluorescein probe for *ath5*, revealed with fast red (Roche). Digoxygenin probes for the other target genes were revealed with the NBT/BCIP substrate (Roche) using standard protocols. The sequences of the medaka homologs of the genes were obtained by blasting the fugu coding region on the medaka genome sequence at http://medaka.utgenome.org/. Partial cDNA sequences were amplified by PCR from a cDNA library and cloned with TOPO TA vector kit (Invitrogen). All the clones where confirmed by sequencing and submitted to the European Molecular Biology Laboratory (EMBL) database. Primer sequences are available upon request. Embedding and sectioning was performed according to standard procedures as described previously [41]. Zebrafish in situs were performed using standard protocols. The sequences for Zebrafish *Brn3C, Gfi-1, CD166, and Adam11* orthologs were retrieved from Ensembl, sequences were amplified using standard PCR reactions from zebrafish 72 h-post-fertilization cDNA, and partial coding sequences were cloned into pCRII-TOPO vector (Invitrogen), following manufacturer's instructions. Primers and constructs sequences are available upon request.

**Transient DNA overexpression and mosaic analysis.** Injection of expression plasmids into one-cell stage fish embryos leads to mosaic distribution and expression with cosegregation of different constructs [29]. Medaka embryos at the one-cell stage were injected with a solution containing 50 ng/μl of the 2 kb *ath5* 5′ genomic region driving GFP expression and 50 ng/μl of either the *ath5, Xath5, Xath3/NeuroM,* or *Xash1* coding region in pCS2+ or else the pCS2+ empty vector [42].

A medaka *ath5* morpholino oligonucleotide (TCG ACG GGA CTT CAT GGT TTC TGT G) was coinjected at a concentration of 0.1 mM as indicated. We checked the specificity and efficacy of this morpholino oligo in standard control experiments [43,44]. At the tested concentrations, the morpholino injections faithfully phenocopied the zebrafish *lak/ath5* mutant phenotype. As judged by histolgical criteria and molecular marker analysis, no signs of ganglion cell differentiation were detected after up to 5 d of development (unpublished data). No additional morphological abnormalities were observed.

Injected embryos were allowed to develop until stage 22 (2 d, [25]) before fixation and in situ hybridization followed by fluorescent fast red detection (probes used are indicated in the figure legend and in the main text). GFP was detected using anti-GFP antibody (rabbit polyclonal, Molecular Probes) at 1:250 dilution detected with anti-rabbit secondary antibody Alexa-488 conjugated (Molecular Probes, 1:500). Nuclei were counterstained with DAPI and embryos analyzed using confocal microscopy (Leica TCS-SP).

**ChIP.** ChIP has been performed on chick dissected retina and optic tectum as previously described [18]. Primers and genomic sequences are available upon request.

**EMSA and luciferase assays.** EMSA and luciferase assays were performed using standard protocols. Briefly, each reaction contained 1μg of salmon sperm DNA, 1μg of poly(dC-dI), and ~1ng of DNA probe (see Table S3 for sequences) end-labeled using T4 polynucleotide kinase with [γ-$^{32}$P]dATP. Ath5 was in vitro translated (Promega TnT sp6 coupled reticulocyte lysate system) according to the manufacturer's specifications. of The Ath5 transcription translation reaction (5 μl) or mock reaction was added to each sample in 20 μl of total volume in water. Competition was performed with 10, 100, or 1,000-fold molar excess cold competitor DNA added to the reaction on ice 10 min before the radiolabeled DNA was added for an additional 20 min. The 20-μl reaction was run on a 5% non-denaturing polyacrylamide gel in 0.5× TBE buffer, at 250 V for 4z6 h at 4 °C. After electophoresis, the gel was dried and visualized by autoradiography. Luciferase transcription assay was performed using the Dual-Luciferase reporter system (Promega) according to the manufacturer's specifications. Cos7 cells were plated on 24-well plates and transfected at 50% confluence using Fugene6 (Roche). Each well received 20 ng of Ath5-pGL3 reporter vector DNA or mutant constructs and 5 ng of pRL DNA. In addition, 0 ng, 3 ng, 10 ng, 30 ng, 100 ng, or 300 ng of Ath5pCS2+-expressing vector was added. Total DNA transfected was kept constant by adding the appropriate amounts of pCS2+ empty vector. Cells were lysed after 24 h and lysates were then assayed for luciferase activity. Tbx2 and Dlx1 promoter assays were performed using 200 ng or 40 ng of reporter pGL3 vector and lysed after 24 h or 48 h, respectively. Each experiment was performed in quadruplicate and results were confirmed at least in two independent experiments. Results were independently reproduced in BHK21 cells (unpublished data).

## Supporting Information

**Figure S1.** Ath5 Promoter Characterization

(A) Binding ability of Ath5 on different mutant oligos comprising the promoter region where the two binding motifs are located (see Table S3 and Materials and Methods for oligo sequences), as revealed by EMSA. For each oligo, the first lane (0) is the control with no Ath5 protein added and the other two lanes include 0.5 μl and 5 μl of a TNT reaction (as indicated). (B) Competition of the binding of the wild-type radiolabeled probe to Ath5 (2.5 μl of TNT added to each lane, first lane control reaction without competitor) by different cold oligos as indicated at the top. Numbers at the top indicate molar excess of cold competitors. (C) Transgenic embryo expressing GFP under the control of the mutated form (MN) of the Ath5 promoter (presented pattern was similarly found for variants M and N).

Found at doi:10.1371/journal.pgen.0030159.sg001 (383 KB PDF).

**Figure S2.** Position Bias of the Ath5 Binding Motif and GO Analysis

(A) Distribution of the Ath5 binding motif within 5 kb of upstream genomic region in all annotated human genes and in the candidate target set. Distances are given as bp upstream of the gene start site as annotated in Ensembl. On the *y*-axis, the total number of occurrences of the Ath5 motif is indicated. Note the enrichment in the target set in close proximity to the annotated gene start. (B–D) GO annotation of the target gene set in reference to the entire human genome. The filtering strategy efficiently reduces the number of hits to the Ath5 motif upstream of annotated genes from 13,000 in the human genome to 166 candidates after the evolutionary double filtering. When analyzing the GO annotation of these targets with respect to cellular components (B), biological processes (C), and biological function (D), respectively, we find a marked increase in nuclear/transcription factor activity as well as extracellular/signal transducer activity. The terms enriched for the biological function are development (2% in the human genome/9% in the data set), neurogenesis (1%/3%), and cell adhesion (2%/4%)

Found at doi:10.1371/journal.pgen.0030159.sg002 (143 KB PDF).

**Figure S3.** Xath3/NeuroM Does Not Directly Activate the Predicted Ath5 Target Gene *ElavC*

(A–C) Upper left: *ElavC* fluorescent in situ, upper right: GFP reporter indicating Ath5 expression, lower left: DAPI nuclear staining, lower right: merge of GFP and in situ signals. (A) Medaka embryo control coinjected with the reporter construct containing the 2-kb Ath5 promoter driving GFP expression together with the empty pCS2+ expression vector. (B) Medaka embryo coinjected with the reporter construct containing the 2-kb Ath5 promoter driving GFP expression together with the *Xenopus* Xath5 pCS2+ expression vector. (C) Medaka embryo coinjected with the reporter construct containing the 2-kb Ath5 promoter driving GFP expression together with the *Xenopus* Xath3/NeuroM pCS2+ expression vector. Note that only in the injection with Xath5 is *ElavC* expression induced in the GFP/XAth5 positive cells. While Xath3 can, to some extent, activate the Ath5 reporter, it does not induce the Ath5 target gene *ElavC*, highlighting the specificity of the interaction.

Found at doi:10.1371/journal.pgen.0030159.sg003 (143 KB PDF).

**Figure S4.** Gene Expression in the Ath5 Mutant *lakritz*

Expression of target genes in the zebrafish Ath5 mutant *lakritz (lak),* which dos not express functional Ath5, at 48 h post fertilization (A–D, G–L) and at 72 h post fertilization (E, F). *Brn3C* (A, B), *CD166* (C, D), *Adam11* (E, F), *Gfi-1* (G, H), *HuC* (I, J), and *NN1* (K, L) expression was detected by whole mount in situ hybridization in wild-type (A, C, E, G, I, and K) and *lak* (B, D, F, H, J, and L) zebrafish larvae. Note the complete absence of target gene expression in the ganglion cell layer of mutant retinae.

Found at doi:10.1371/journal.pgen.0030159.sg004 (1.1 MB PDF).

**Table S1.** Ath5 Target Genes

Genes were classified based on their annotation. Genes speculated to function in cell cycle signaling, axon guidance, and neuron function as well as those enriched GO annotation terms (GO:0007275, development and GO:0003700, transcription factor activity) or with described expression in the retina are classified on the top of the list. In red are the genes we experimentally analyzed.

Found at doi:10.1371/journal.pgen.0030159.st001 (255 KB DOC).

**Table S2.** Ectopic Overexpression of the Related bHLH Transcription Factors

Found at doi:10.1371/journal.pgen.0030159.st002 (33 KB DOC).

**Table S3.** PCR and Mutagenesis Primer Sequences

Found at doi:10.1371/journal.pgen.0030159.st003 (32 KB DOC).

**Text S1.** Detailed Analysis and Benchmarking of the Computation Pipeline and Promoter Analysis

Found at doi:10.1371/journal.pgen.0030159.sd001 (180 KB DOC).

**Accession Numbers**

The Ensembl (http://www.ensembl.org/) accession number for medaka *ath5* is ENSORLG00000013722.

## Acknowledgments

We thank Felix Loosli for critical discussion, his great help with the manuscript, and for the GFP reporter plasmid. The data about the medaka genome has been freely provided by the National Institute of Genetics and the University of Tokyo for use in this publication only. We thank the Sanger systems support for computational support, K. Brown and W. Norton for critically reading the manuscript, M Vetter for the pG1X5 3.3-kb construct, W. Harris for the Xath5, Xath3, and Xash1 expression constructs, all the members of the Wittbrodt group for their critical input, and E. Grzebisz and A. Nowicka for fish husbandry. In the final part of the study, FDB received a Human Frontier Science Program long-term fellowship. Work in HB's lab was funded by a Sandler Opportunity Award. Work in JW's lab was supported by the European Union (FP6 STREP Hygeia) and the Deutsche Forschungsgemeinschaft, Collaborative Research Centre 488.

**Author contributions.** FDB, LE, and JW conceived and designed the experiments and wrote the paper. FDB, LE, and DSK performed the experiments. FDB, LE, JMM, EB, and JW analyzed the data. FDB, LE, DSK, HB, and JMM contributed reagents/materials/analysis tools.

**Funding.** The authors received no specific funding for this study.

**Competing interests.** The authors have declared that no competing interests exist.

**References**

1. Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5: 276–287.
2. Jenuwein T, Allis CD (2001) Translating the histone code. Science 293: 1074–1080.
3. Zinzen RP, Senger K, Levine M, Papatsenko D (2006) Computational models for neurogenic gene expression in the Drosophila embryo. Curr Biol 16: 1358–1365.
4. Wasserman WW, Fickett JW (1998) Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol 278: 167–181.
5. Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, et al. (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res 16: 656–668.
6. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, et al. (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 124: 47–59.
7. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, et al. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203: 439–455.
8. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, et al. (2002) E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes Dev 16: 245–256.
9. Brown NL, Kanekar S, Vetter ML, Tucker PK, Gemza DL, et al. (1998) Math5 encodes a murine basic helix-loop-helix transcription factor expressed during early stages of retinal neurogenesis. Development 125: 4821–4833.
10. Masai I, Stemple DL, Okamoto H, Wilson SW (2000) Midline signals regulate retinal neurogenesis in zebrafish. Neuron 27: 251–263.
11. Perron M, Kanekar S, Vetter ML, Harris WA (1998) The genetic sequence of retinal development in the ciliary margin of the Xenopus eye. Developmental Biology 199: 185–200.
12. Wang SW, Kim BS, Ding K, Wang H, Sun D, et al. (2001) Requirement for math5 in the development of retinal ganglion cells. Genes Dev 15: 24–29.
13. Brown NL, Patel S, Brzezinski J, Glaser T (2001) Math5 is required for retinal ganglion cell and optic nerve formation. Development 128: 2497–2508.
14. Kay JN, Finger-Baier KC, Roeser T, Staub W, Baier H (2001) Retinal ganglion cell genesis requires lakritz, a zebrafish atonal homolog. Neuron 30: 725–736.
15. Kanekar S, Perron M, Dorsky R, Harris WA, Jan LY, et al. (1997) Xath5 participates in a network of bHLH genes in the developing Xenopus retina. Neuron 19: 981–994.
16. Liu W, Mo Z, Xiang M (2001) The Ath5 proneural genes function upstream of Brn3 POU domain transcription factor genes to promote retinal ganglion cell development. Proc Natl Acad Sci U S A 98: 1649–1654.
17. Matter-Sadzinski L, Matter JM, Ong MT, Hernandez J, Ballivet M (2001) Specification of neurotransmitter receptor identity in developing retina: the chick ATH5 promoter integrates the positive and negative effects of several bHLH proteins. Development 128: 217–231.
18. Skowronska-Krawczyk D, Ballivet M, Dynlacht BD, Matter JM (2004) Highly specific interactions between bHLH transcription factors and chromatin during retina development. Development 131: 4447–4454.
19. Massari ME, Murre C (2000) Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. Mol Cell Biol 20: 429–440.
20. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: D108–110.
21. Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, et al. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. Nucleic Acids Res 34: D95–97.

22. Levy S, Hannenhalli S, Workman C (2001) Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics 17: 871–877.
23. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434: 338–345.
24. Consortium TGO (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25–29.
25. Iwamatsu T (1994) Stages of normal development in the medaka Oryzias latipes. Zoo Sci 11: 825–839.
26. Mu X, Klein WH (2004) A gene regulatory hierarchy for retinal ganglion cell specification and differentiation. Semin Cell Dev Biol 15: 115–123.
27. Wang SW, Mu X, Bowers WJ, Kim DS, Plas DJ, et al. (2002) Brn3b/Brn3c double knockout mice reveal an unsuspected role for Brn3c in retinal ganglion cell axon outgrowth. Development 129: 467–477.
28. de Melo J, Du G, Fonseca M, Gillespie LA, Turk WJ, et al. (2005) Dlx1 and Dlx2 function is necessary for terminal differentiation and survival of late-born retinal ganglion cells in the developing mouse retina. Development 132: 311–322.
29. Zukerberg LR, Patrick GN, Nikolic M, Humbert S, Wu CL, et al. (2000) Cables links Cdk5 and c-Abl and facilitates Cdk5 tyrosine phosphorylation, kinase upregulation, and neurite outgrowth. Neuron 26: 633–646.
30. Grabher C, Wittbrodt J (2004) Efficient activation of gene expression using a heat-shock inducible Gal4/Vp16-UAS system in medaka. BMC Biotechnol 4: 26.
31. Materials and methods are available as supporting material on Science Online.
32. Ettwiller L, Paten B, Souren M, Loosli F, Wittbrodt J, et al. (2005) The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. Genome Biol 6: R104.
33. Eddy SR (2005) A model of the statistical power of comparative genome sequence analysis. PLoS Biol 3: e10. doi:10.1371/journal.pbio.0030010
34. Hutcheson DA, Hanson MI, Moore KB, Le TT, Brown NL, et al. (2005) bHLH-dependent and -independent modes of Ath5 gene regulation during retinal development. Development 132: 829–839.
35. Matter-Sadzinski L, Puzianowska-Kuznicka M, Hernandez J, Ballivet M, Matter JM (2005) A bHLH transcriptional network regulating the specification of retinal ganglion cells. Development 132: 3907–3921.
36. Logan MA, Steele MR, Van Raay TJ, Vetter ML (2005) Identification of shared transcriptional targets for the proneural bHLH factors Xath5 and XNeuroD. Dev Biol 285: 570–583.
37. Lenhard B, Wasserman WW (2002) TFBS: Computational framework for transcription factor binding site analysis. Bioinformatics 18: 1135–1136.
38. Johnson N, Kotz S, Kemp A (1992) Univariate discrete distributions. New York: Wiley.
39. Köster R, Stick R, Loosli F, Wittbrodt J (1997) Medaka spalt acts as a target gene of hedgehog signaling. Development 124: 3147–3156.
40. Thermes V, Grabher C, Ristoratore F, Bourrat F, Choulika A, et al. (2002) I-SceI meganuclease mediates highly efficient transgenesis in fish. Mech Dev 118: 91–98.
41. Loosli F, Kmita-Cunisse M, Gehring WJ (1996) Isolation of a Pax-6 homolog from the ribbonworm Lineus sanguineus. Proc Natl Acad Sci U S A 93: 2658–2663.
42. Tessmar K, Loosli F, Wittbrodt J (2002) A screen for co-factors of Six3. Mech Dev 117: 103–113.
43. Carl M, Loosli F, Wittbrodt J (2002) Six3 inactivation reveals its essential role for the formation and patterning of the vertebrate eye. Development 129: 4057–4063.
44. Del Bene F, Tessmar-Raible K, Wittbrodt J (2004) Direct interaction of geminin and Six3 in eye development. Nature 427: 745–749.