Key role of amino acid repeat expansions in the functional diversification of duplicated transcription factors

Núria Radó-Trilla[1], Krisztina Arató[2,3,4], Cinta Pegueroles[1,3], Alicia Raya[2,3,4], Susana de la Luna[2,3,4,5,*] and M.Mar Albà[1,2,5,*]

[1]Evolutionary Genomics Group, Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute (IMIM), Dr. Aiguader 88 08003 Barcelona, Spain.
[2]Universitat Pompeu Fabra (UPF), Dr. Aiguader 88 08003 Barcelona, Spain.
[3]Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain.
[4]Centro de Investigación Biomèdica en Red en Enfermedades Raras (CIBERER), Spain.
[5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain.

* To whom correspondence should be addressed

Keywords: gene duplication, low complexity region (LCR), transcription factor, PHOX2B, polyalanine

Corresponding authors contact information:

M. Mar Albà
ICREA Research Professor, Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute (IMIM) - Dr. Aiguader 88, 08003 Barcelona, Spain.
e-mail: malba@imim.es

Susana de la Luna
ICREA Research Professor, Gene Regulation, Stem Cells and Cancer Programme, Centre for Genomic Regulation (CRG) - Dr. Aiguader 88, 08003 Barcelona, Spain.
email: susana.luna@crg.es

Abstract

The high regulatory complexity of vertebrates has been related to two closely spaced whole genome duplications (2R-WGD) that occurred before the divergence of the major vertebrate groups. Following these events, many developmental transcription factors (TFs) were retained in multiple copies and subsequently specialized in diverse functions, whereas others reverted to their singleton state. TFs are known to be generally rich in amino acid repeats or low-complexity regions (LCRs), such as polyalanine or polyglutamine runs, which can evolve rapidly and potentially influence the transcriptional activity of the protein. Here we test the hypothesis that LCRs have played a major role in the diversification of TF gene duplicates. We find that nearly half of the TF gene families (107 out of 237) originated during the 2R-WGD contain LCRs, compared to only a small percentage of the non-duplicated TF genes used as a control (15 out of 115). At the individual gene level, we observe that twice as many duplicated TFs have gained LCRs as non-duplicated TFs. In addition, duplicated TFs preferentially accumulate certain LCR types, the most prominent of which are alanine repeats. We experimentally test the role of alanine-rich LCRs in two different TF gene families, PHOX2A/PHOX2B and LHX2/LHX9. In both cases, the presence of the alanine-rich LCR in one of the copies (PHOX2B and LHX2) significantly increases the capacity of the TF to activate transcription. Taken together, the results provide strong evidence that LCRs are important driving forces of evolutionary change in duplicated genes.

Introduction

Gene duplication is a major mechanism for the emergence of novel gene functions (Ohno 1970). Duplicated genes may arise from local genomic duplication events affecting just one or a few genes, or from whole genome duplications (WGDs). It has been observed that after a WGD there is preferential retention of duplicated genes that are in dosage balance, including many transcription factors and developmental regulatory proteins (Birchler et al. 2005; Freeling and Thomas 2006; Edger and Pires 2009). At later evolutionary stages, these factors diverge and acquire new functions (Escriva et al. 2006; Lynch et al. 2008). Many human transcription factors (TFs) encoded in the human genome date from the two rounds of whole genome duplication (2R-WGD) that occurred in early vertebrate evolution (Ohno et al. 1968; Lundin 1993; Dehal and Boore 2005; Blomme et al. 2006). The 2R-WGD increased the opportunities for new regulatory pathways to arise, contributing to the formation of complex structures such as the brain and the circulatory system (Wagner 2008; Huminiecki and Heldin 2010).

Many studies have reported an increase in the number of amino acid substitutions after gene duplication, which is consistent with relaxed purifying selection and/or gain of new functions in one or both gene copies (see for example Zhang et al. 2003; Scannell and Wolfe 2008; Han et al. 2009; Pegueroles et al. 2013; Pich I Roselló and Kondrashov 2014). Other works have examined the divergence of expression patterns of gene duplicates, which appears to be relatively independent of amino acid sequence change (Wagner 2000; Makova and Li 2003; Farré and Albà 2010).

However, there is a lack of studies on amino acid repeat expansions after gene duplication. These events often correspond to insertions or deletions in the

alignments and they become invisible when evolutionary rates are computed. Over time, perfect amino acid tandem repeats formed by triplet slippage tend to accumulate amino acid substitutions and become cryptic repeats or simple sequences (Green and Wang 1994; Huntley and Golding 2000; Albà et al. 2002; Simon and Hancock 2009). Both perfect and cryptic repeats, collectively called low-complexity regions (LCRs), can be detected on the basis of sequence compositional biases by specific programs, such as SEG (Wootton and Federhen 1996). It has been noted that LCRs are especially abundant in TFs compared to other functional classes (Karlin et al. 2002; Albà and Guigó 2004; Faux et al. 2005; Huntley and Clark 2007).

Several lines of evidence point to LCRs as a source of functional innovation. First, tracts of repeated proline, glutamine, or alanine fused to a DNA-binding domain can change the capacity of a protein to activate transcription (Gerber et al. 1994; Sauer et al. 1995; Janody et al. 2001; Galant and Carroll 2002; Brown et al. 2005). Second, histidine repeats can mediate the targeting of the protein to nuclear speckles, which are subnuclear structures for the storage of splicing factors and TFs (Salichs et al. 2009). Third, the uncontrolled expansion of glutamine repeats has been associated with a number of neurodegenerative diseases in humans (Gatchel and Zoghbi 2005), whereas mutations resulting in abnormally long alanine repeats cause several developmental disorders (Messaed and Rouleau 2009). Fourth, changes in the length of amino acid repeats have been associated with morphological changes in dog breeds (Fondon and Garner 2004). Finally, although repeats can expand rapidly by triplet slippage, there is also evidence that they are often preserved by purifying selection once they reach a certain length (Mularoni et al. 2010).

Here, we investigate the role of LCRs in the functional diversification of human duplicated TFs formed during the 2R-WGD. Using this set of families ensures that

the number of genes is sufficiently large and homogeneous for statistical analysis. In addition, we can experimentally test the functional consequences of LCRs in TFs using well-established transcriptional activity assays. We find that there is a very significant enrichment of LCR in duplicated TFs when compared to non-duplicated TFs. The results of the experiments performed in two different gene families, PHOX2A/PHOX2B and LHX2/LHX9, illustrate how the gain of a novel LCR in one of the gene copies alters the capacity of the protein to activate transcription, connecting the presence of the LCR to a gain of functional activity.

Results

Increased gain of low complexity regions (LCRs) in duplicated transcription factors (TFs)

In order to investigate the role of LCRs in the functional diversification of gene duplicates, we obtained a set of 550 human TFs that had been retained in duplicated form after the vertebrate 2R-WGD. For this, we used paralogy information from Ensembl (Vilella et al. 2009; Flicek et al. 2013) as well as phylogenetic tree reconstructions based on homologous sequences from diverse vertebrate species (see Material and Methods). The TFs belonged to 237 different gene families with 2, 3, or, in two cases, 4 gene members in the human genome (Figure 1, duplicated transcription factors; supplementary tables S1 and S2, Supplementary Material online). The scarcity of gene families of size 4 concurs with the observation that many duplicated genomic segments were lost following the 2R-WGD (Friedman and Hughes 2003; Dehal and Boore 2005). For comparison, we also obtained a set of 115 human non-duplicated TF genes (Figure 1, single-copy transcription factors; supplementary table S3, Supplementary Material online). These genes presumably reverted to their single copy gene state soon after the 2R-WGD.

We identified LCRs in TFs with the SEG algorithm (Wootton and Federhen 1996), using parameter settings that favor the detection of highly repetitive sequences (see Material and Methods). LCRs ranged between 9 and 59 amino acids in size, with an average of 22 residues. 107 of the 237 gene families (45%) contained LCRs in one or more gene copies (Figure 1). This set comprised 222 LCRs in 155 different duplicated proteins out of the 550 analyzed (supplementary tables S4 and S5, Supplementary Material online). In contrast, we only found 20 LCRs in 15 single copy proteins out of the 115 analyzed (Figure 1, supplementary table S6, Supplementary Material online). Therefore, the fraction of proteins with LCRs in the duplicated gene set was about twice as large as the fraction in the non-duplicated gene set (28% versus 13%, respectively, Chi-square test p=0.009; supplementary table S7, Supplementary Material online).

After careful examination of all multiple sequence alignments of the gene families with LCRs, we could only identify 8 cases (out of 222) in which the LCR conservation pattern was consistent with the existence of a repeat before the 2R-WGD. Similarly, using *Ciona intestinalis* orthologues we could only detect three putative cases of chordate ancestral repeats in our data set. Therefore, we concluded that the vast majority of the LCRs identified (>95%) had been formed after the 2R-WGD.

The most common scenario was gain of LCRs in one member of the gene family, but we also found several cases in which LCRs had been gained in an independent manner in several family members. One illustrative example is the gene family comprising *POU4F1*, *POU4F2* and, *POU4F3*. Each of these genes contains two exons each and encodes a highly conserved POU domain at the protein C-terminus. We identified LCRs in the POU4F1 and POU4F2 proteins, but not in POU4F3. POU4F1 contained one LCR enriched in alanines and another one in

glycines. POU4F2 had a histidine-rich tract, previously shown to be required for nuclear speckle targeting (Salichs et al. 2009), and another long LCR containing several serine and glycine runs. The LCRs from the two proteins mapped to different exons and were clearly unrelated to each other.

We identified 62 LCRs that were highly conserved in different vertebrate lineages and that had probably originated in the period between the 2R-WGD and the separation of the bony fishes, approximately 450 to 420 million years ago (Dehal and Boore 2005). In other cases the incomplete conservation of the LCR suggested a more recent time of origination, such as the alanine-rich LCRs in PHOX2B and LHX2, which apparently formed after the amphibian split and which are discussed in more detail below.

Overrepresentation of alanine-rich LCRs in duplicated TFs

Next, we investigated the amino acid composition of the LCRs in the duplicated and non-duplicated gene sets. Most LCRs were highly enriched in one or at most two amino acids (supplementary tables S4 to S6, Supplementary Material online). For this reason we decided to label each LCR according to the most frequent amino acid or, in those cases in which two different amino acids were similarly abundant, the two most frequent amino acids (17% of LCRs). The most commonly found amino acids in the non-duplicated gene set were glutamic acid and proline (6 and 4 cases, respectively). In contrast, in the duplicated gene set the most abundant amino acid was alanine, accounting for 50 of the 222 LCRs (22.5%, in 42 proteins)(Figure 2). Remarkably, we only detected one alanine-rich LCR in the non-duplicated gene set (1 out of 20 LCRs; 5%).

The list of duplicated proteins with alanine-rich LCRs included several disease-associated genes (supplementary table S8, Supplementary Material online).

Noteworthy, we identified two genes in which expansions of the alanine tract beyond the normal range result in an incorrect functioning of the protein in humans: Paired-like homeobox 2B (*PHOX2B*) and Aristaless related homeobox (*ARX*). The LCR in the PHOX2B protein contains a perfect run of 20 alanines that, when expanded to 25-29 residues, causes central hypoventilation syndrome, an alteration in the development of the autonomic nervous system (Amiel et al. 2003; Trochet et al. 2005). The ARX sequence includes four alanine repeats; short expansions beyond the wild type range in the first and second repeats are found in individuals with X-linked mental retardation and in several other X-linked developmental disorders (Shoubridge et al. 2010). Given these examples, it seems possible that mutations increasing the size of the alanine repeats in other genes of the set may be associated with developmental disorders of yet unknown origin.

The effect of alanine-rich LCRs on transcriptional activation

To evaluate the functional impact of the LCRs gained after gene duplication, we examined in detail two transcription factor families: PHOX2A/PHOX2B and LHX2/LHX9 (LIM homeobox 2/9). In both families one of the gene copies had gained an alanine-rich LCR (PHOX2B and LHX2). In PHOX2B the LCR included a perfect alanine tract of size 20, whereas in LHX2 there was a perfect repeat composed of 10 alanines. Whereas the polyalanine tract in PHO2B has been associated with central hypoventilation syndrome, no previous data exists on the possible functionality of the alanine repeat in LHX2.

We experimentally tested the transcriptional activity of different gene constructs, including the two paralogous proteins and a deletion mutant lacking the LCR. We employed one-hybrid assays on a reporter gene in which the DNA binding sites were placed on a minimal promoter structure (containing only a TATA-box) to

reduce promoter-specific effects. Moreover, we used a heterologous DNA binding domain (DBD) to tether the effector protein to the reporter DNA to avoid any confounding effect due to possible differential DNA binding affinities of the different family members.

PHOX2A/PHOX2B

The paralogous genes *PHOX2A* and *PHOX2B* encode transcription factors that play key roles in different developmental processes of vertebrates. Both genes are expressed in peripheral and central noradrenergic neurons and neural crest derivatives, with differences in the onset and extent of expression during development (Pattyn et al. 1997; Amiel et al. 2003). The association of each paralogue to a specific functional role is further complicated by the fact that *PHOX2B* activates *PHOX2A* as observed in gain-of-function experiments (Flora et al. 2001). However, both paralogues are not functionally redundant because *PHOX2A* cannot always replace *PHOX2B in vivo* (Coppola et al. 2005). *PHOX2A* is critical for the development of noradrenergic neurons and specific motor neuron nuclei (Pattyn et al. 1997). *PHOX2B* is essential for the development of the central and peripheral autonomic nervous systems and it is a master regulator for the differentiation of the visceral nervous system (Pattyn et al. 1999; Dauger et al. 2003).

The alignment of the two human paralogous proteins shows that the homeodomain, which functions as the DNA-binding domain (Benfante et al. 2007), is extremely well conserved between the two copies and that most differences accumulate in a region at the C-terminal end that has several amino acid repetitions; the most conspicuous of which is an alanine-rich LCR only present in *PHOX2B* (Figure 3).

Analysis of orthologous sequences from several vertebrates indicated that the alanine-rich LCR in PHOX2B had probably originated in an ancestral *Amniota*. We combined sequence information from *Xenopus*, chicken, human, macaque, mouse, rat and cow (supplementary table S9, Supplementary Material online) to deduce the most likely ancestral sequence at a time when the LCR was not yet present using a maximum likelihood approach (Yang 2007). We followed the same procedure for the PHOX2A protein. The sequence alignment of ancestral and modern proteins shows that the reconstructed ancestral PHOX2B is about 10% shorter than the human PHOX2B and lacks the expanded alanine repeat (Figure 3).

We generated expression plasmids to produce the complete PHOX2A and PHOX2B open reading frames which were N-terminally fused to Gal4-DBD (Figure 4A). We also constructed a plasmid to express a mutated version of PHOX2B that lacked the alanine-rich LCR (PHOX2BΔLCR). The three fusion proteins accumulated to a similar extent when over-expressed in HeLa cells, as shown by the results of Western blot analysis with an antibody against the Gal4-DBD (Figure 4B). Immunofluorescence analysis revealed that the proteins were located in the nuclei of the transfected cells, suggesting that the repeat was not affecting subcellular localization (Figure 4C). In reporter assays, PHOX2B showed a dose-dependent transcriptional activity over the basal Ga4-DBD levels, which was not detected in transfections with PHOX2A (Figure 4 D; supplementary table S10, Supplementary Material online for the results of three experimental replicates). In PHOX2BΔLCR the capacity to activate transcription was strongly impaired with relation to the wild type PHOX2B, providing strong evidence that the alanine-rich LCR is an activator domain. These results support the idea that the formation of the LCR had a strong impact in the functional divergence of the two gene copies.

LHX2/LHX9

The *LHX2* and *LHX9* genes belong to the LIM homeobox (LHX) family, a group of proteins which perform important roles in tissue-specific differentiation and body patterning during development in both vertebrates and invertebrates (Kadrmas and Beckerle 2004). In particular, both *LHX2* and *LHX9* are involved in the development of brain, limb, and eyes, as well as in hematopoiesis (Srivastava et al. 2010). Consistent with the formation of *LHX2* and *LHX9* by gene duplication during the 2R-WGD, the two paralogoues are present in different vertebrate lineages but there is only one orthologous gene, named *Apterous*, in the *Drosophila* genus. *Apterous* shares key functional properties with *LHX2*/*LHX9* genes during development (Hobert and Westphal 2000).

The LHX2 and LHX9 proteins contain two LIM domains, a type of zinc-finger known to mediate protein-protein interactions (Matthews et al. 2009), and a helix-turn helix forming a homeodomain that binds to specific DNA sequences in the promoters of target genes (Figure 5). The paralogous protein LHX2 additionally contains an alanine-rich LCR that is highly conserved in mammals. We reconstructed ancestral sequences for the two paralogues using a similar approach than with the PHOX2A/PHOX2B pair (supplementary table S11, Supplementary Material online). The alignment indicated that the modern and ancestral LHX2 proteins are vey similar except for the alanine-rich LCR (Figure 5).

We experimentally tested the transcriptional activity of the two gene copies and the specific effect of the LHX2 alanine-rich LCR using the one hybrid assay (Figure 6). The two paralogues and the LCR deletion mutant of LHX2 were expressed at similar levels (Figure 7B) and the subcellular distribution was indistinguishable (Figure 6C). LHX2 behaved as an activator in this experimental setting, but LHX9 did not show any significant difference with respect to the control containing only the DNA-binding domain (Figure 6D; supplementary table S12, Supplementary

Material online for the results of three experimental replicates). The LHX2 activity was severely impaired when the LCR was deleted. The results clearly link the formation of the LCR with the gain of a novel functionality in LHX2.

Discussion

The evolution of vertebrates is intimately related to the two rounds of whole genome duplication (2R-WGD) that occurred before the diversification of the group (Lundin 1993; Dehal and Boore 2005; Cañestro et al. 2013). Many developmental genes were retained after the 2R-WGD and subsequently gained new functions, a process known as neofunctionalization. One classic example is the retinoic acid receptor (RAR) family, which includes RAR alpha, RAR beta, and RAR gamma. Whereas RAR beta maintained its original function, the other two paralogues accumulated amino acid substitutions that modified their capacity to bind different retinoid compounds (Escriva et al. 2006).

Here, we turned our attention to modifications due to gain of LCRs. Except for a few examples, including a number of proteins containing histidine repeats (Salichs et al. 2009), the incidence of LCRs in the evolution of gene duplicates has not been examined to our knowledge. We found that nearly half the TF gene families (107 out of 237) originated at the 2R-WGD contained LCRs, and that the number of duplicated genes that had gained LCRs was significantly larger than the equivalent number of non-duplicated genes. In some cases we even found that several members of the same family had gained different LCRs.

In general, the LCRs corresponded to long gaps when aligned to the other gene copies, consistent with the action of replication slippage in expanding the repeat. In the absence of functionality, repeats will tend to accumulate amino acid substitutions and degenerate. However, the sudden disappearance of an LCR,

albeit possible, appears to be a rare event (Radó-Trilla and Albà 2012). Similarly, LCRs are more frequently found associated with insertions than with deletions in multiple sequence alignments of mammalian orthologous proteins (Laurie et al. 2012). Taken together, these observations point to an important role of repeat expansions in increasing protein length and complexity.

Our bioinformatics pipeline identified 50 different alanine-rich LCRs, ranging in size from 9 to 31 amino acids, in the duplicated TFs. Glycine-rich LCRs followed in order of abundance (30 gains) and there were additional 9 LCRs that were enriched in both alanine and glycine. Repeats of these two amino acids appear to expand with relatively ease in vertebrate proteins in general (Nakachi et al. 1997; Radó-Trilla and Albà 2012). The reasons are yet unclear but several factors, such as high slippage propensity of GC-rich triplets (Ellegren 2000), low amino acid synthesis cost, reduced toxicity, or high adaptive value of the runs formed by these amino acids, may all contribute to this bias.

Experiments performed in the gene families *PHOX2A*/*PHOX2B* and *LHX2*/*LHX9* linked the presence of alanine-rich LCRs to neofunctionalization. In both cases we found that the LCR significantly increased the ability of the protein to activate transcription. Although the precise molecular mechanisms of this activation remain to be determined, one possibility is that it is driven by the interaction with coactivator subunits of the basal transcription complex TFIID, as shown for the alanine-rich region of the *Drosophila* TF Bicoid (Sauer et al. 1995).

The impact of alanine repeats in the regulatory activity of PHOX2B, and more specifically the consequences of abnormal expansions of these tracts for human health, have been the focus of several studies. In humans, the heterozygotic expansion of the 20-alanine stretch in PHOX2B to 25-29 residues leads to congenital central hypoventilation syndrome (CCHS; OMIM 603851), a rare

disorder defined by a failure of the autonomic control of breathing (Amiel et al. 2003). It has been reported that several PHOX2B mutants with expansion of the alanine repeat suffer an alteration in the conformation of the PHOX2B protein that disrupts normal protein function and impairs the ability of the PHOX2B to regulate the transcription of its target genes (Bachetti et al. 2005; Trochet et al. 2005; Di Lascio et al. 2013). In addition, PHOX2B nonsense mutations leading to truncated proteins that lack the C-terminus, which contains the alanine repeat, have been found in neuroblastoma patients (Raabe et al. 2008). PHOX2B polyalanine expanded mutants show a significant reduction in their capacity to activate transcription when assayed *in vitro* (Adachi et al. 2000), suggesting that when the tract becomes abnormally long it has a negative impact on transcriptional activity. Interestingly, expansion of the alanine tract from 15 to 25 residues in the ZIC2 protein, which would mimic the mutation found in holoprosencephaly patients, also results in near complete loss of transcriptional activation (Brown et al. 2005).

In addition to the link between transcriptional activation and the presence of a poly-alanine tract, the evolutionary conservation of the LCR might be associated with other protein activities. For example, deletion of the alanine repeat in FOXL2 results in intranuclear protein aggregation (Moumné et al. 2005). In addition, the combination of alanine repeats with other LCR types may lead to complex functional outcomes. It has been reported that alanine repeats can exert an inhibitory effect over glutamine repeats, which are strong activators on their own (Janody et al. 2001; Galant and Carroll 2002). Analysis of GO terms in our set of duplicated TFs with yielded a similar number of genes annotated as "positive regulation of transcription" than of genes annotated as "negative regulation of transcription" among those containing alanine-rich LCRs (supplementary table S5, Supplementary Material online), suggesting that these repeats may influence transcription in very different contexts.

In summary, the results of the present study indicate that the rate of gain of LCRs in duplicated genes increases with respect to single copy genes. As shown for alanine repeats, this can result in protein neofunctionalization. This process is likely to have been shaped by selection, contributing to increase the complexity of regulatory networks in vertebrates.

Material and Methods

Sequence datasets and identification of gene families

We selected human genes whose molecular function in the Gene Ontology database matched ”transcription factor” in Ensembl v.64/66, using BioMart (Flicek et al. 2013). We annotated TFs with an activator or repressor role using the Gene Ontology (GO) terms “positive regulation of transcription” and “negative regulation of transcription” (Ashburner et al. 2000). We used the paralogous gene dating information in Ensembl Compara (Vilella et al. 2009) to extract human transcription factor families whose duplication time was consistent with the two rounds of whole genome duplication (2R-WGD) at the base of the vertebrates before the formation of jawed vertebrates (Euteleostomi paralogy type).

For each gene family member, orthologous sequences from *Mus musculus* (mouse), *Gallus gallus* (chicken), and *Danio rerio* (zebrafish) were retrieved from Ensembl. We obtained 539 proteins in mouse, 441 proteins in chicken, and 635 proteins in zebrafish, denoting some secondary loses and duplications (in fishes due to an additional WGD) with respect to the human dataset. We also obtained orthologous proteins in *C. intestinalis* whenever available (141 *C. intestinalis* proteins). We used these sequences to generate multiple protein alignments with T-Coffee (Di Tommaso et al. 2011). We subsequently built maximum likelihood

based trees with PhyML v.2.4.4 (Guindon et al. 2005) and only accepted tree topologies congruent with the hypothesis of 2R-WGD, with no more than two duplication events in the ancestral vertebrate branch and no subsequent duplications in the lineage leading to humans. The final set of duplicated proteins consisted of 237 transcription factor protein families: 163 families with two human gene copies, 72 families with three human gene copies, and 2 families with four human gene copies. The total number of human genes within these families was 550 (average 2.32 genes/family).

A set of 115 non-duplicated human transcription factors was also obtained from Ensembl for comparative purposes. These proteins corresponded to transcription factors for which no paralogues had been retained after the 2R-WGD.

Identification and characterization of low complexity regions (LCRs)

We identified sequences corresponding to low complexity regions (LCRs) using the program SEG (Wootton and Federhen 1996). This program divides sequences into contrasting segments of low sequence complexity, amino acid repeats, and segments of high sequence complexity. The low sequence complexity segments are identified because they depart strongly from a random residue composition. We used the parameters window=15, K_1=1.5, and K_2=1.8, which identify highly repetitive regions.

The amino acid repeats were annotated according to the most frequently occurring amino acid in the sequence identified by SEG. When the frequency of the second amino acid was more than half the frequency of the most abundant one (17% of the repeats), the two most frequently occurring amino acids were considered in the annotation of the repeats. In general these cases corresponded

to the presence of several single amino acid repeats of different composition next to each other.

We used an in-house Perl script to identify the position of all LCRs on the previously generated multiple sequence alignments. Conserved LCRs were defined as those that overlapped in the alignment and were enriched in the same amino acid.

Reconstruction of ancestral sequences

Vertebrate ancestral protein sequences, including ancestral LHX2, LHX9, PHOX2A, and PHOX2B, were reconstructed using Codeml (parameter RateAncestor = 1) in PAML v4.4 (Yang 2007). All multiple sequence alignments included orthologous proteins from *Homo sapiens*, *Macaca mulata*, *M. musculus*, *Rattus norvegicus*, *Bos taurus, Xenopus tropicalis*, *D. rerio,* and in the case of LHX2 and LHX9, also *G. gallus* (supplementary tables S9 and S11, Supplementary Material online). Ancestral sequence reconstruction was based on alignments obtained with Prank +F to increase the accuracy of the position of indels (Löytynoja and Goldman 2008; Laurie et al. 2012; Villanueva-Cañas et al. 2013). The amino acid alignments were converted to coding sequence alignments using PAL2NAL (Suyama et al. 2006).

Construction of plasmids

Full length cDNA clones for human *LHX2* and *LHX9* (IMAGE clones: IRATp970G12116D and IRCMp5012C0737Q, respectively) and *PHOX2A* and *PHOX2B* (IRAUp969B08103D and IRAUp969C0347D, respectively) were purchased from ImaGenes and checked by sequencing. The open reading frames were amplified by polymerase chain reaction with specific primers

(supplementary table S13, Supplementary Material online) and they were inserted in-frame into the EcoRI and XbaI sites of the expression plasmid pGal4-DBD (de la Luna et al. 1999) to generate N-terminally fused proteins to Gal4 DBD. The deletion of the LHX2 LCR -AAAAAAAAAAKSAGLGAAGA- was performed by site-directed mutagenesis (Stratagene) on pG4DBD-LHX2 with primer LHX2ΔLCR to generate pG4DBD-LHX2ΔLCR. A synthetic gene containing the human PHOX2B open reading frame without the LCR -GGAAAAAAAAAAAAAAAAAAAGGLAAAGG- was purchased from GeneArt (Life Technologies) and cloned into pGal4-DBD. All the plasmids were checked by DNA sequencing.

Cell culture and transfection

HeLa cells were obtained from the American Type Cell Culture Collection. Cells were maintained at 37ºC in DMEM supplemented with 10% fetal calf serum (FCS) and antibiotics. Transient transfections were performed using the calcium phosphate method and cells were processed 48 h after transfection.

Immunofluorescence

Cells grown in coverslips were fixed in 4% paraformaldehyde in phosphate-buffered saline (PBS) for 15 min, permeabilized in 0.1% Triton X-100 in PBS for 10 min and blocked with 10% FCS for at least 30 min. Cells were incubated with a mouse monoclonal antibody to Gal4-DBD (1:100 in PBS-1% FCS; Santa Cruz Biotechnology) as the primary antibody for 1 h. After washing extensively with PBS-1% FCS, the coverslips were incubated with an Alexa Fluor 488 anti-mouse antibody (1:2000 in PBS-1% FCS; Invitrogen) for 1 h, washed repeatedly with PBS-1% FCS and mounted onto slides using Mowiol-DAPI (4', 6'-diamino-2-phenylindole, 1 g/ml; Vector Laboratories, Inc). All procedures were done at

room temperature. Images were captured with an Inverted ZEISS Microscope with fluorescence.

Western blot

Whole-cell extracts were prepared in 25 mM Tris-HCl, pH 7.5, 1 mM EDTA, and 1% sodium dodecyl sulfate (SDS). Samples were resolved by SDS-PAGE, transferred onto nitrocellulose membrane (Hybond C; GE Healthcare), and blocked with 10% skimmed milk in Tris-buffered saline (TBS) (10 mM Tris-HCl, pH 7.5, and 100 mM NaCl) containing 0.1% Tween 20 (TBS-T). Membranes were incubated with anti-Gal4-DBD antibody (in 5% skimmed milk in TBS-T) overnight at 4º C. After washing with TBS-T, membranes were incubated for 1 h at room temperature with horseradish peroxidase-conjugated polyclonal rabbit anti-mouse antibody (in 5% skimmed milk in TBS-T) and then washed again with TBS-T. Proteins were detected by chemiluminiscence with Western Lightning Plus-ECL (Perkin Elmer) in a LAS-3000 image analyzer (Fuji PhotoFilm).

One-hybrid gene reporter assays

HeLa cells were transfected with the pG5E1B-luc reporter (de la Luna et al. 1999) in which *Firefly* luciferase expression is driven by five repeats of yeast Gal4-binding sites introduced upstream of the minimal adenovirus E1B promoter, together with the plasmids to express unfused Gal4-DBD, or different Gal4-DBD fusion proteins as indicated in the Figure Legends. A *Renilla* luciferase plasmid (pCMV-RNL, Promega) was used as an internal control for transfection efficiency. Cells were lysed 48 h post-transfection and the activity of both luciferase enzymes was measured with the Dual-Luciferase Reporter Assay kit (Promega). Transfections were done in triplicates, and the experiments repeated independently three times.

Supplementary Material Online

Supplementary tables are in Rado-Trilla_etal_2015_sup.xlsx, available from evolutionarygenomics.imim.es, Publications/Datasets links.

Figure legends

Figure 1. Diagram showing the gene family datasets studied and the number of proteins with LCRs in each of them. The number of duplicated gene families was 237: 163 had 2 members, 72 had 3 members, and only 2 had 4 members (one protein in this latter subset had LCRs, not shown in the Figure). The total number of proteins with repeats in the duplicated gene set was 155. The number of single copy genes in the non-duplicated datset was 115, of which 15 had LCRs.

Figure 2. Number of LCRs annotated by the most abundant amino acid(s) in the duplicated gene set. The list has been shortened to LCR types occurring more than 5 times.

Figure 3. Alignment of human PHOX2B and PHOX2A proteins, together with the corresponding ancestral, non-LCR containing, proteins. The low-complexity region (LCR) identified in the human PHOX2B by SEG is framed. Sequence reconstruction was performed by maximum likelihood (see Material and Methods for more details). We employed sequences from Xenopus, chicken, cow, rat, mouse, macaque, and human. Zebrafish sequences were not included because the examination of sequences from the different species clearly indicated that the LCR had been gained after the separation from the amphibians.

Figure 4. A) Schematic representation of PHOX2A and PHOX2B. HomeoD: homeodomain; LCR: low-complexity region. **B)** Total cell extracts from cells

transfected with equal amounts of the plasmids to express the indicated Gal4DBD-fusion proteins (as indicated in the scheme) were analyzed by Western blot with anti-Gal4-DBD to show the expression levels of the fusion proteins. **C)** The subcellular localization of the indicated PHOX2-fusion proteins (as in B) was analyzed by indirect immunofluorescence with anti-Gal4-DBD (left panels) in HeLa transfected cells. Nuclei were counterstained with DAPI (middle panels). **D)** One-hybrid reporter assays using a 5xGal4 sites/luciferase reporter and increasing amounts (150 ng, 450 ng) of expression plasmids to the effector proteins indicated. The graph represents the transcriptional activation of Gal4-DBD fusions to PHOX2A, PHOX2B wild-type or to the PHOX2B LCR deletion mutant, as indicated, relative to that of unfused Gal4-DBD arbitrarily set as 1. Error bars represent standard deviations of triplicated transfected plates. The experiment shown is representative of 3 performed.

Figure 5. Alignment of human LHX2 and LHX9 proteins, together with the corresponding ancestral non-LCR containing proteins. The low-complexity region (LCR) identified in the human LHX2 protein by SEG is framed. Sequence reconstruction was performed by maximum likelihood (see Methods for more details). We employed sequences from Xenopus, chicken, cow, rat, mouse, macaque, and human. Zebrafish sequences were not included because the examination of sequences from the various species clearly indicated that the LCR had been gained after the separation from the amphibians.

Figure 6. A) Schematic representation of LHX2 and LHX9. HomeoD: homeodomain; LCR: low-complexity region; LIM: LIM domain. **B)** Total cell extracts from cells expressing transfected with equal amounts of the plasmids to express the indicated Gal4DBD-fusion proteins (as indicated in the scheme) were analyzed by Western blot with anti-Gal4-DBD to show the expression levels of the fusion proteins. **C)** The subcellular localization of the indicated fusion proteins (as in B) was analyzed by indirect immunofluorescence with anti-Gal4-DBD in

transfected HeLa cells (left panels). Nuclei were counterstained with DAPI (middle panels). **D)** One-hybrid reporter assays using a 5xGal4 sites/luciferase reporter and increasing amounts (50 ng, 150 ng) of expression plasmids to the effector proteins indicated. The graph represents the transcriptional activation of Gal4-DBD fusions to LHX9, LHX2 wild-type or to the LHX9-LCR deletion mutant, as indicated, relative to that of unfused Gal4-DBD arbitrarily set as 1. Error bars represent standard deviations of triplicated transfected plates. The experiment shown is representative of 3 performed.

Acknowledgements

We received funding from Ministerio de Innovación y Tecnología (BIO2009-08160 to M.M.A.), Ministerio de Economía y Competitividad (BFU2012-36820 to M.A. and BFU2010-15347 and 'Centro de Excelencia Severo Ochoa 2013-2017'-SEV-2012-0208 to S.L.), the Secretariat of Universities and Research-Government of Catalonia (2014SGR1121 TO M.M.A. and 2014SGR674 to S.L.), Fundació Javier Lamas Miralles (fellowship to N.R-T.), and Institució Catalana de Recerca i Estudis Avançats (ICREA contract to M.M.A. and S.L.). We acknowledge Will Blevins for his useful comments on the manuscript.

References

Adachi M, Browne D, Lewis EJ. 2000. Paired-like homeodomain proteins Phox2a/Arix and Phox2b/NBPhox have similar genetic organization and independently regulate dopamine beta-hydroxylase gene transcription. DNA Cell Biol. 19:539–554.

Albà MM, Guigó R. 2004. Comparative analysis of amino acid repeats in rodents and humans. Genome Res. 14:549–554.

Albà MM, Laskowski RA, Hancock JM. 2002. Detecting cryptically simple protein sequences using the SIMPLE algorithm. Bioinformatics 18:672–678.

Amiel J, Laudier B, Attié-Bitach T, et al. 2003. Polyalanine expansion and frameshift mutations of the paired-like homeobox gene PHOX2B in congenital central hypoventilation syndrome. Nat. Genet. 33:459–461.

Bachetti T, Matera I, Borghini S, Di Duca M, Ravazzolo R, Ceccherini I. 2005. Distinct pathogenetic mechanisms for PHOX2B associated polyalanine expansions and frameshift mutations in congenital central hypoventilation syndrome. Hum. Mol. Genet. 14:1815–1824.

Benfante R, Flora A, Di Lascio S, Cargnin F, Longhi R, Colombo S, Clementi F, Fornasari D. 2007. Transcription factor PHOX2A regulates the human alpha3 nicotinic receptor subunit gene promoter. J. Biol. Chem. 282:13290–13302.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. Trends Genet. 21:219–226.

Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 7:R43.

Brown L, Paraso M, Arkell R, Brown S. 2005. In vitro analysis of partial loss-of-function ZIC2 mutations in holoprosencephaly: alanine tract expansion modulates DNA binding and transactivation. Hum. Mol. Genet. 14:411–420.

Cañestro C, Albalat R, Irimia M, Garcia-Fernàndez J. 2013. Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. Semin. Cell Dev. Biol. 24:83–94.

Coppola E, Pattyn A, Guthrie SC, Goridis C, Studer M. 2005. Reciprocal gene replacements reveal unique functions for Phox2 genes during neural differentiation. EMBO J. 24:4392–4403.

Dauger S, Pattyn A, Lofaso F, Gaultier C, Goridis C, Gallego J, Brunet J-F. 2003. Phox2b controls the development of peripheral chemoreceptors and afferent visceral pathways. Development 130:6635–6642.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3:e314.

Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res. 17:699–717.

Ellegren H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. Nat. Genet. 24:400–402.

Escriva H, Bertrand S, Germain P, et al. 2006. Neofunctionalization in vertebrates: the example of retinoic acid receptors. PLoS Genet. 2:e102.

Farré D, Albà MM. 2010. Heterogeneous patterns of gene-expression diversification in mammalian gene duplicates. Mol. Biol. Evol. 27:325–335.

Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, de la Banda MG, Whisstock JC. 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res. 15:537–551.

Flicek P, Ahmed I, Amode MR, et al. 2013. Ensembl 2013. Nucleic Acids Res. 41:D48–55.

Flora A, Lucchetti H, Benfante R, Goridis C, Clementi F, Fornasari D. 2001. Sp proteins and Phox2b regulate the expression of the human Phox2a gene. J. Neurosci. 21:7037–7045.

Fondon JW, Garner HR. 2004. Molecular origins of rapid and continuous morphological evolution. Proc. Natl. Acad. Sci. U. S. A. 101:18058–18063.

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. 16:805–814.

Friedman R, Hughes AL. 2003. The temporal distribution of gene duplication events in a set of highly conserved human gene families. Mol. Biol. Evol. 20:154–161.

Galant R, Carroll SB. 2002. Evolution of a transcriptional repression domain in an insect Hox protein. Nature 415:910–913.

Gatchel JR, Zoghbi HY. 2005. Diseases of unstable repeat expansion: mechanisms and common principles. Nat. Rev. Genet. 6:743–755.
Gerber HP, Seipel K, Georgiev O, Höfferer M, Hug M, Rusconi S, Schaffner W. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 263:808–811.
Green H, Wang N. 1994. Codon reiteration and the evolution of proteins. Proc. Natl. Acad. Sci. U. S. A. 91:4298–4302.
Han M V, Demuth JP, McGrath CL, Casola C, Hahn MW. 2009. Adaptive evolution of young gene duplicates in mammals. Genome Res. 19:859–867.
Hobert O, Westphal H. 2000. Functions of LIM-homeobox genes. Trends Genet. 16:75–83.
Huminiecki L, Heldin CH. 2010. 2R and remodeling of vertebrate signal transduction engine. BMC Biol. 8:146.
Huntley M, Golding GB. 2000. Evolution of simple sequence in proteins. J. Mol. Evol. 51:131–140.
Huntley MA, Clark AG. 2007. Evolutionary analysis of amino acid repeats across the genomes of 12 Drosophila species. Mol. Biol. Evol. 24:2598–2609.
Janody F, Sturny R, Schaeffer V, Azou Y, Dostatni N. 2001. Two distinct domains of Bicoid mediate its transcriptional downregulation by the Torso pathway. Development 128:2281–2290.
Kadrmas JL, Beckerle MC. 2004. The LIM domain: from the cytoskeleton to the nucleus. Nat. Rev. Mol. Cell Biol. 5:920–931.
Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. 2002. Amino acid runs in eukaryotic proteomes and disease associations. Proc. Natl. Acad. Sci. U. S. A. 99:333–338.
Di Lascio S, Bachetti T, Saba E, Ceccherini I, Benfante R, Fornasari D. 2013. Transcriptional dysregulation and impairment of PHOX2B auto-regulatory mechanism induced by polyalanine expansion mutations associated with congenital central hypoventilation syndrome. Neurobiol. Dis. 50:187–200.
Laurie S, Toll-Riera M, Radó-Trilla N, Albà MM. 2012. Sequence shortening in the rodent ancestor. Genome Res. 22:478–485.
Lynch VJ, Tanzer A, Wang Y, Leung FC, Gellersen B, Emera D, Wagner GP. 2008. Adaptive changes in the transcription factor HoxA-11 and evolution of pregnancy in mammals. Proc. Natl. Acad. Sci. USA 105:14928-14933.
Lundin LG. 1993. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics 16:1–19.
Makova KD, Li W-H. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. 13:1638–1645.
Matthews JM, Bhati M, Lehtomaki E, Mansfield RE, Cubeddu L, Mackay JP. 2009. It takes two to tango: the structure and function of LIM, RING, PHD and MYND domains. Curr. Pharm. Des. 15:3681–3696.
Messaed C, Rouleau GA. 2009. Molecular mechanisms underlying polyalanine diseases. Neurobiol. Dis. 34:397–405.
Moumné L, Fellous M, Veitia RA. 2005. Deletions in the polyAlanine-containing transcription factor FOXL2 lead to intranuclear aggregation. Hum. Mol. Genet. 14:3557–3564.
Mularoni L, Ledda A, Toll-Riera M, Albà MM. 2010. Natural selection drives the accumulation of amino acid tandem repeats in human proteins. Genome Res. 20:745–754.
Nakachi Y, Hayakawa T, Oota H, Sumiyama K, Wang L, Ueda S. 1997. Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors. Mol. Biol. Evol. 14:1042–1049.
Ohno S, Wolf U, Atkin NB. 1968. Evolution from fish to mammals by gene duplication. Hereditas 59:169–187.
Ohno S. 1970. Evolution by gene duplication. Springer New York
Pattyn A, Morin X, Cremer H, Goridis C, Brunet JF. 1997. Expression and interactions of the two closely related homeobox genes Phox2a and Phox2b during neurogenesis. Development 124:4065–4075.
Pattyn A, Morin X, Cremer H, Goridis C, Brunet JF. 1999. The homeobox gene Phox2b is essential for the development of autonomic neural crest derivatives. Nature 399:366–370.
Pegueroles C, Laurie S, Albà MM. 2013. Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. Mol. Biol. Evol.
Pich I Roselló O, Kondrashov FA. 2014. Long-term asymmetrical acceleration of protein evolution after gene duplication. Genome Biol. Evol. 6:1949–1955.
Raabe EH, Laudenslager M, Winter C, Wasserman N, Cole K, LaQuaglia M, Maris DJ, Mosse YP, Maris JM. 2008. Prevalence and functional consequence of PHOX2B mutations in neuroblastoma. Oncogene 27:469–476.

Radó-Trilla N, Albà M. 2012. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. BMC Evol. Biol. 12:155.
Salichs E, Ledda A, Mularoni L, Albà MM, de la Luna S. 2009. Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. PLoS Genet. 5:e1000397.
Sauer F, Hansen SK, Tjian R. 1995. DNA template and activator-coactivator requirements for transcriptional synergism by Drosophila bicoid. Science 270:1825–1828.
Scannell DR, Wolfe KH. 2008. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome Res. 18:137–147.
Shoubridge C, Fullston T, Gécz J. 2010. ARX spectrum disorders: making inroads into the molecular pathology. Hum. Mutat. 31:889–900.
Simon M, Hancock JM. 2009. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. Genome Biol. 10:R59.
Srivastava M, Larroux C, Lu DR, Mohanty K, Chapman J, Degnan BM, Rokhsar DS. 2010. Early evolution of the LIM homeobox gene family. BMC Biol. 8:4.
Trochet D, Hong SJ, Lim JK, Brunet J-F, Munnich A, Kim K-S, Lyonnet S, Goridis C, Amiel J. 2005. Molecular consequences of PHOX2B missense, frameshift and alanine expansion mutations leading to autonomic dysfunction. Hum. Mol. Genet. 14:3697–3708.
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19:327–335.
Wagner A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proc. Natl. Acad. Sci. U. S. A. 97:6579–6584.
Wagner A. 2008. Gene duplications, robustness and evolutionary innovations. Bioessays 30:367–373.
Wootton JC, Federhen S. 1996. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 266:554–571.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.
Zhang P, Gu Z, Li W-H. 2003. Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol. 4:R56.

Figure 1

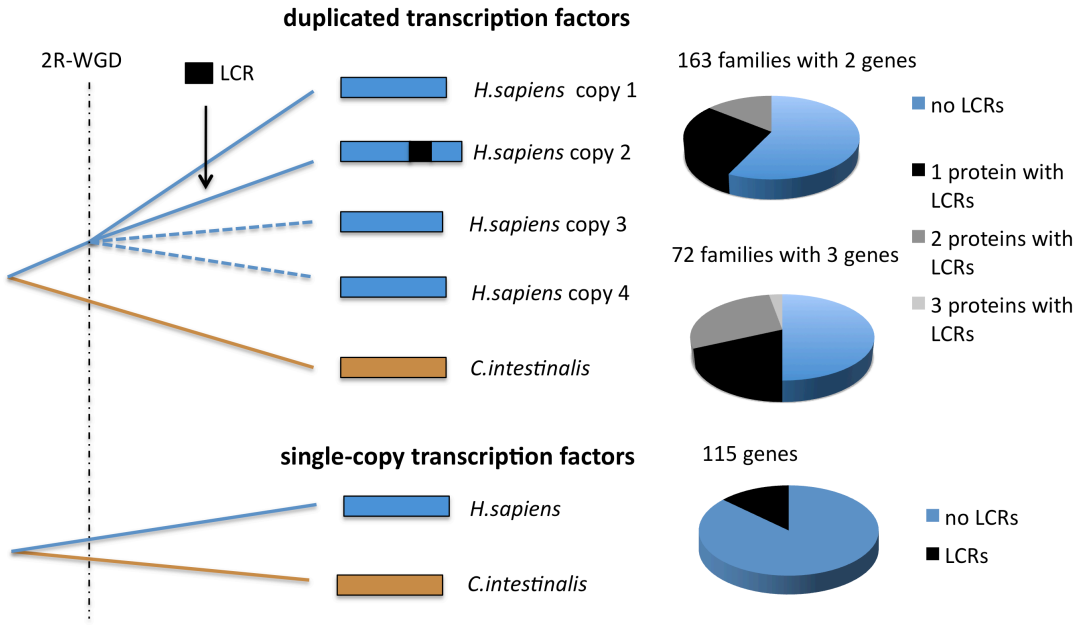

duplicated transcription factors
2R-WGD
LCR
H.sapiens copy 1
H.sapiens copy 2
H.sapiens copy 3
H.sapiens copy 4
C.intestinalis
163 families with 2 genes
no LCRs
1 protein with LCRs
2 proteins with LCRs
3 proteins with LCRs
72 families with 3 genes
single-copy transcription factors
H.sapiens
C.intestinalis
115 genes
no LCRs
LCRs

Figure 2

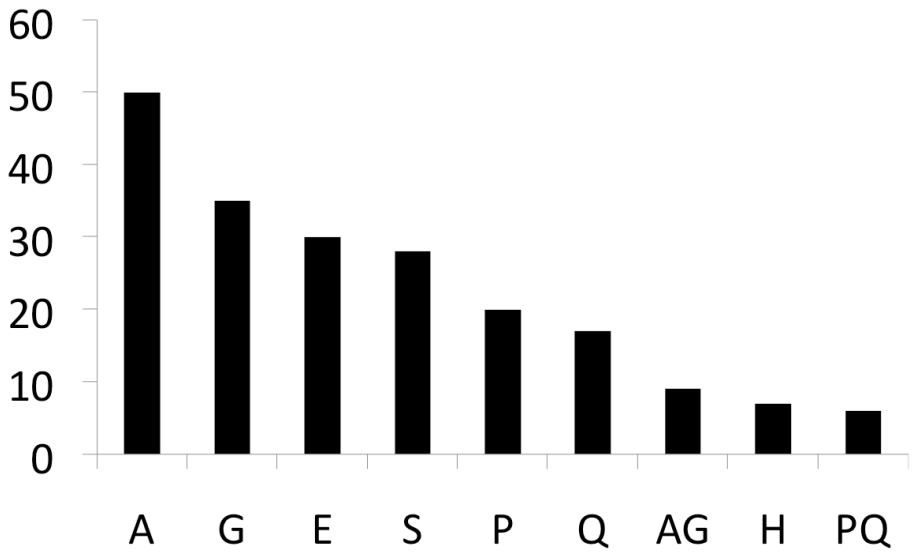


Figure 3

```
                                10        20        30        40        50
Ancestral_PHOX2B/1-282 MYKMEYSYLNSSAYESCMAGMDTSSLASAYADFSSCSQASGFQYNPIRTTF-GATSGCPS
Human_PHOX2B/1-314     MYKMEYSYLNSSAYESCMAGMDTSSLASAYADFSSCSQASGFQYNPIRTTF-GATSGCPS
Ancestral_PHOX2A/1-273 ---MDYSYLNS--YDSCMAAME----ASAYADFSSCSQPNGFQYNPIRAGF-GAGPACPP
Human_PHOX2A/1-284     ---MDYSYLNS--YDSCVAAME----ASAYGDFGACSQPGGFQYSPLRPAFPAAGPPCPA

                                70        80        90       100       110
Ancestral_PHOX2B/1-282 LTPGSCSLGTLRDHQSSPYAAVPYKLFTDHGGLNEKRKQRRIRTTFTSAQLKELERVFAE
Human_PHOX2B/1-314     LTPGSCSLGTLRDHQSSPYAAVPYKLFTDHGGLNEKRKQRRIRTTFTSAQLKELERVFAE
Ancestral_PHOX2A/1-273 LATANCTLGALRDHQPAPYSTVPYKFFSDPSGLNEKRKQRRIRTTFTSSQLKELERVFAE
Human_PHOX2A/1-284     LGSSNCALGALRDHQPAPYSAVPYKFFPEPSGLHEKRKQRRIRTTFTSAQLKELERVFAE

                               130       140       150       160       170
Ancestral_PHOX2B/1-282 THYPDIYTREELALKIDLTEARVQVWFQNRRAKFRKQERAAAAAAAAAKNGSSGKKSD--
Human_PHOX2B/1-314     THYPDIYTREELALKIDLTEARVQVWFQNRRAKFRKQERAAAAAAAAAKNGSSGKKSD--
Ancestral_PHOX2A/1-273 THYPDIYTREELALKIDLTEARVQVWFQNRRAKFRKQERAANSKGASGAAGN--KKGDAR
Human_PHOX2A/1-284     THYPDIYTREELALKIDLTEARVQVWFQNRRAKFRKQERAASAKGAAGAAGA--KKGEAR
                       DNA binding domain
                               190       200       210       220       230
Ancestral_PHOX2B/1-282 -SSRDEESKEAKST-DPDSTGGP--GPNPNPTPSC------GGGPSPTGAQGAXGPGG-N
Human_PHOX2B/1-314     -SSRDDESKEAKST-DPDSTGGP--GPNPNPTPSCGANGGGGGGPSPAGAPGAAGPGGPG
Ancestral_PHOX2A/1-273 SSSEDDESKESTCSPTPDSTASL--PXAXNLNSP-------GLSPSPSPGAVGSGPGPGL
Human_PHOX2A/1-284     CSSEDDDSKESTCSPTPDSTASLPPPPAPGLASP-------RLSPSPLPVALGSGPGPGP

                               250       260       270       280       290
Ancestral_PHOX2B/1-282 GEPG-K----VGVAG-------------------------PGQGWAPGPGTITSIPDSLG
Human_PHOX2B/1-314     GEPG-K----GGAAAAAAAAAAAAAAAAAAAAGGLAAAGGPGQGWAPGPGPITSIPDSLG
Ancestral_PHOX2A/1-273 GPQALKGPPWPGV-------------------------PGTSTAELLKAWQPA-EAVP
Human_PHOX2A/1-284     GPQPLKGALWAGVAG--------------------GGGGGPGAGAAELLKAWQPA-ESGP

                               310       320
Ancestral_PHOX2B/1-282 GPFASVLSSLQRPNGAKATLVKXSMF
Human_PHOX2B/1-314     GPFASVLSSLQRPNGAKAALVKSSMF
Ancestral_PHOX2A/1-273 GPFAGVLSTFHRKPN----TLKTNLF
Human_PHOX2A/1-284     GPFSGVLSSFHRKPGP---ALKTNLF
```

Figure 4

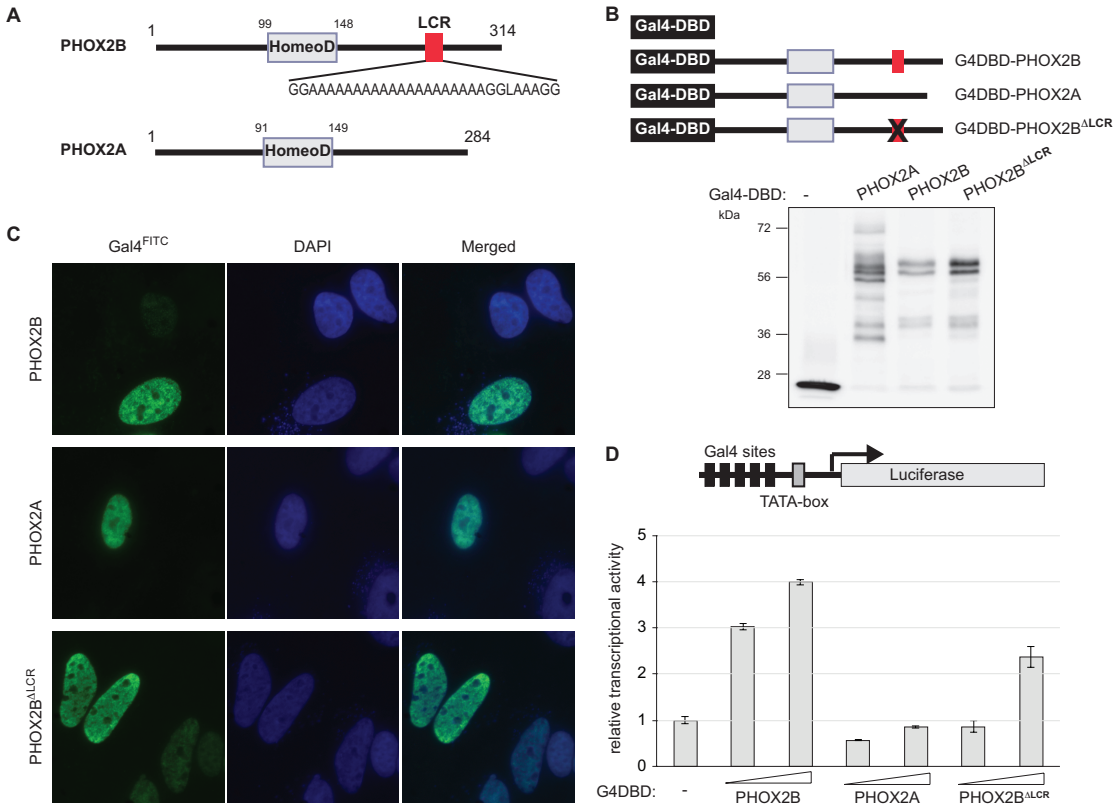

A
PHOX2B
1
99
148
LCR
314
HomeoD
GGAAAAAAAAAAAAAAAAAAAAGGLAAAGG
PHOX2A
1
91
149
284
HomeoD
B
Gal4-DBD
G4DBD-PHOX2B
G4DBD-PHOX2A
G4DBD-PHOX2B^ΔLCR
Gal4-DBD: -
PHOX2A
PHOX2B
PHOX2B^ΔLCR
kDa
72
56
36
28
C
Gal4^FITC
DAPI
Merged
PHOX2B
PHOX2A
PHOX2B^ΔLCR
D
Gal4 sites
Luciferase
TATA-box
relative transcriptional activity
5
4
3
2
1
0
G4DBD: -
PHOX2B
PHOX2A
PHOX2B^ΔLCR

Figure 5

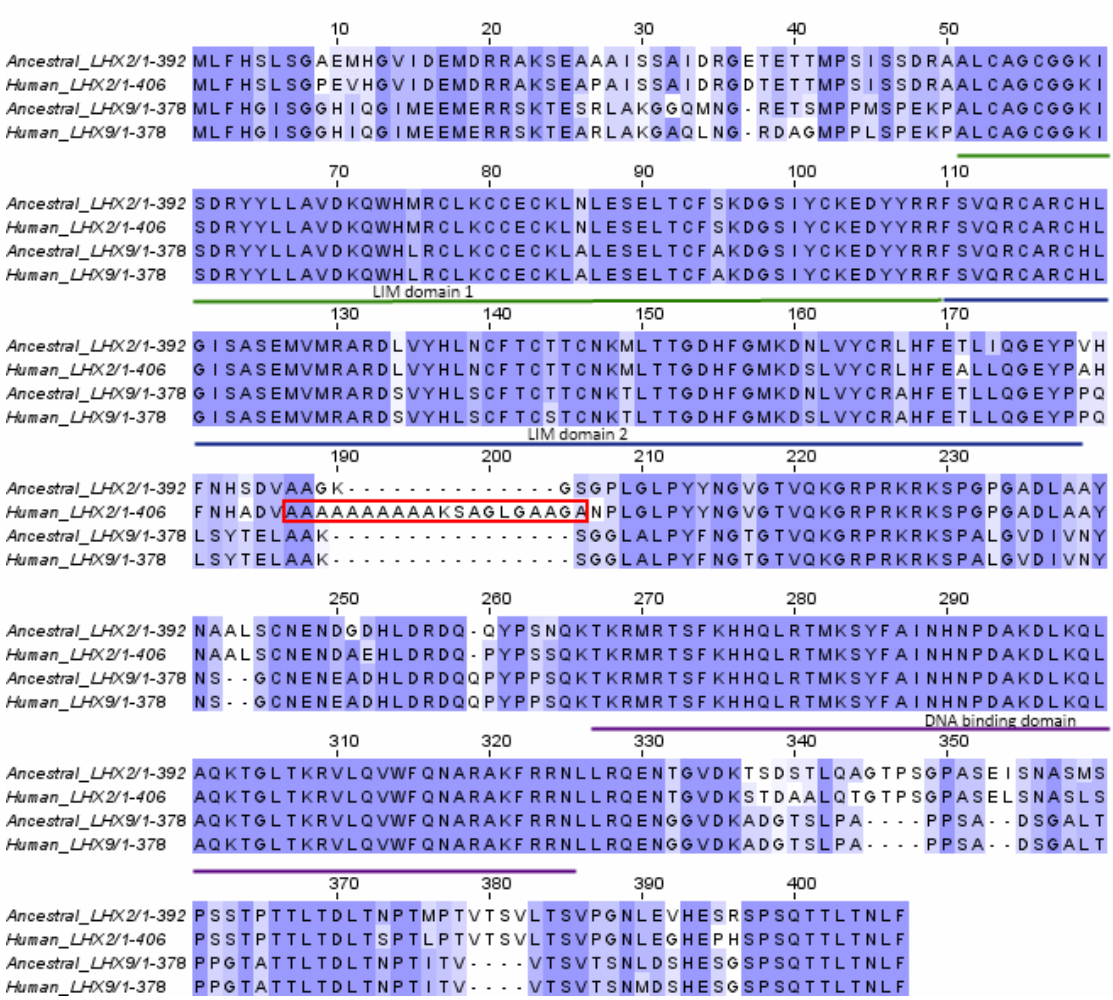
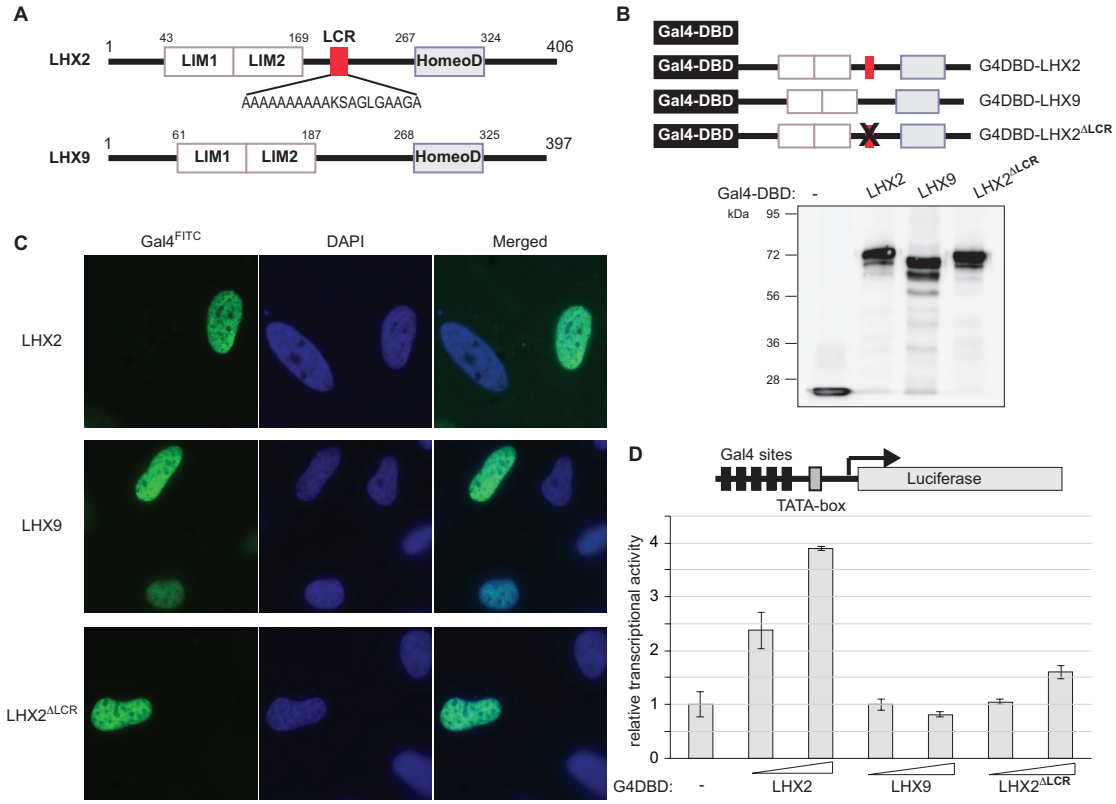

Ancestral_LHX2/1-392
Human_LHX2/1-406
Ancestral_LHX9/1-378
Human_LHX9/1-378
LIM domain 1
LIM domain 2
DNA binding domain

Figure 6


A
LHX2
1
43
169
LCR
267
324
406
LIM1
LIM2
HomeoD
AAAAAAAAAAKSAGLGAAGA
LHX9
1
61
187
268
325
397
LIM1
LIM2
HomeoD
B
Gal4-DBD
G4DBD-LHX2
G4DBD-LHX9
G4DBD-LHX2ΔLCR
Gal4-DBD: -
LHX2
LHX9
LHX2ΔLCR
kDa
95
72
56
36
28
C
Gal4FITC
DAPI
Merged
LHX2
LHX9
LHX2ΔLCR
D
Gal4 sites
Luciferase
TATA-box
relative transcriptional activity
0
1
2
3
4
G4DBD: -
LHX2
LHX9
LHX2ΔLCR