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

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. Here we investigate the role of amino acid tandem repeat expansions in the functional diversification of TF families originated at the 2R-WGD. We find that the number of low-complexity regions (LCRs) in duplicated gene copies is significantly higher than the number in single-copy TFs evolved during the same period of time. Overall, nearly half of the TF gene families (107 out of 237) have gained novel LCRs in one or more gene copies since the 2R-WGD, compared to only 15 of the 115 non-duplicated genes used as a control. In addition, duplicated TF preferentially accumulate certain repeat types, such as those enriched in alanine or glycine. We experimentally test the role of alanine repeats in two different TF gene families, LHX2/LHX9 and PHOX2A/PHOX2B. In both cases, the gain of the alanine repeat in one of the copies significantly increases the capacity of the protein to activate transcription. Taken together, the results support a key role of LCRs in the functional diversification of duplicated TFs.

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 (Edger and Pires 2009; Freeling and Thomas 2006; Birchler et al. 2005). At later evolutionary stages, these factors may diverge and acquire new functions. 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 (Huminiecki and Heldin 2010).

Recent large-scale studies on the evolution of gene duplicates have focused on the analysis of amino acid substitutions (Han, Demuth, McGrath, Casola, & Hahn, 2009; Pegueroles, Laurie, & Albà, 2013; Pich I Roselló & Kondrashov, 2014; Scannell & Wolfe, 2008; Zhang, Gu, & Li, 2003) or changes in expression patterns (Farré & Albà, 2010; Makova & Li, 2003). These studies have reported an increase in the accumulation 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.

However, none of the above studies have considered the possibility of amino acid repeats contributing to an accelerated evolution of gene duplicates. Amino acid repeats 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 (Simon and Hancock 2009; Albà et al. 2002). Both perfect and cryptic repeats, collectively called low-complexity regions (LCRs), can be detected by specific programs, such as SEG, on the basis of their compositional bias (Wootton and Federhen 1996). LCRs are especially abundant in transcription factors and other regulatory proteins (Karlin et al. 2002; Albà and Guigó 2004; Faux et al. 2005; Huntley and Clark 2007).

Several lines of evidence point to a functional role for the LCRs. First, tracts of repeated proline, glutamine, or alanine can affect the capacity of a protein to activate transcription (Gerber et al. 1994; Janody et al. 2001; Galant and Carroll 2002; Brown et al. 2005; Sauer et al. 1995). Second, histidine repeats are responsible for the targeting of the host protein to nuclear speckles, which are subnuclear structures that store regulatory proteins (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). Finally, although repeats can expand rapidly by triplet slippage (Ellegren 2000), they are often preserved by purifying selection once they reach a certain length (Mularoni et al. 2010).

Here, we investigate if LCRs have contributed to the functional diversification of human duplicated TFs formed during the 2R-WGD. Focusing on 2R-WGD gene families has the advantage that the set is large and highly homogeneous (550 duplicated proteins evolved for the same period of time), that the time past since the duplication is sufficiently long as to be able to detect many newly formed LCRs and that the functional effect of the LCR on transcriptional activity is

amenable to experimental testing. We show that the accumulation of LCRs in duplicated TFs is significantly higher than in single copy genes evolved during the same period of time. Moreover, the results of reporter-based experiments performed in two different gene families, LHX2/LHX9 and PHOX2A/PHOX2B, illustrate how the gain of a novel LCR in one of the gene copies increases the capacity of the protein to activate transcription, connecting the presence of the LCR to a functional activity.

Results

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

In order to investigate the role of repeats in the functional diversification of gene duplicates, we obtained a set of 550 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 (see Material and Methods). The TFs belonged to 237 different gene families with 2, 3, or, in two cases, 4 human gene members (Figure 1, duplicated transcription factors; Supplemental Table 1 and 2). 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 non-duplicated genes (Figure 1, single-copy transcription factors; Supplemental Table 3).

We quantified the number of LCRs in the proteins 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. We identified LCRs in 107 (45%) of the 237 gene families (Figure 1). This set comprised 222 LCRs in 155 different duplicated proteins out of the 550 analyzed (Supplemental Table 5 and 6). In contrast, we only found 20 LCRs in 15 single copy proteins out of the 115 analyzed (Figure 1, Supplemental Table 7). 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 $p=0.009$).

Also nNoteworthy, only 8 out of the 222 LCRs in the duplicated TFs displayed conservation patterns consistent with the existence of a repeat before the 2R-WGD. In addition, we could only detect three putative cases of chordate ancestral repeats using *Ciona intestinalis* orthologues. Therefore, we concluded that the vast majority of the LCRs (>95%) had been formed after the 2R-WGD.

Overrepresentation of alanine repeats in duplicated proteins

We calculated the frequency of different amino acids in the LCRs of the duplicated and non-duplicated gene sets. For this analysis, we labeled each LCR according to the most frequent amino acid or, in several cases in which two different amino acids were similarly abundant, the two most frequent amino acids (17% of LCRs) (see Supplemental Tables 5 and 6 for a complete list of LCR sequences). 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%)(Figure 2). In contrast, we only detected one alanine-rich LCR in the non-duplicated gene set (1 out of 20 LCRs; 5%).

We identified several disease-associated genes in our list of 42 duplicated proteins with alanine repeats (Supplemental Table 8). They included two genes in which expansions of alanine tracts result in an incorrect functioning of the protein: 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 (CCHS; OMIM#209880). This disorder involves several alterations in the development of the autonomic nervous system (Amiel et al. 2003). 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). Expansion of other alanine repeats in this set of genes may be causing 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 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. In PHOX2B the LCR contained a perfect alanine tract of size 20, whereas in LHX2, the perfect repeat was formed by 10 alanines. We experimentally tested the transcriptional activity of each of the gene constructs by using 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.

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). *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). *PHOX2A* is critical for the development of noradrenergic neurons and specific motor neuron nuclei (Pattyn et al. 1997).

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). To gain further insight into the evolution of the family, we estimated the sequence of the ancestral PHOX2A and PHOX2B proteins using a maximum likelihood approach (Yang 2007). We combined sequence information from *Xenopus*, chicken, and several mammalian species (Supplemental Table 9) to deduce the most likely ancestral sequences just before the separation of the amphibians, at a time when the LCR was not yet present. As shown in Figure 3, the ancestral PHOX2B sequence is very similar to the present day human sequence except for the expansion of the alanine-rich LCR.

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 accumulated within the nuclei of the transfected cells, suggesting that the repeat was not affecting

subcellular localization (Figure 5C). 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 4D, Supplemental Table 10). 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.

LHX2/LHX9

The LHX2 and LHX9 genes belong to the LIM homeobox (LHX) family, a family 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).

The domain structure of LHX2 and LHX9 is highly conserved; both are composed of 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 6). Consistent with the formation of LHX2 and LHX9 by gene duplication in an ancestral vertebrate, both proteins are found in different vertebrate species, 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 paralogous gene LHX2 in humans contains an alanine-rich LCR that is highly conserved in other mammalian orthologous sequences and that is not present in LHX9 (Figure 5, Supplemental Table 11).

To experimentally test the transcriptional activity of the two gene copies and the specific effect of the LHX2 alanine-rich LCR, we applied an equivalent experimental approach as with the PHOX2A/PHOX2B pair (Figure 6A). The two paralogues and the LCR deletion mutant of LHX2 were expressed at similar levels (Figure 6B) and the subcellular distribution was indistinguishable (Figure 6C). LHX2, containing the alanine-rich LCR, behaved as an activator in the one hybrid assay (Figure 6D, Supplemental Table 12). As with PHOX2B, this activity was severely impaired when the LCR was deleted.

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 (Cañestro et al. 2013; Lundin 1993; Dehal and Boore 2005). Many developmental genes were retained after the 2R-WGD and subsequently gained new functions, a process known as neofunctionalization. One 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 duplicated 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 had gained novel LCRs. The gain was significantly higher than expected if the genes had not been retained in multiple copies, judging by the results obtained in a control non-duplicated gene set. 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 one of the copies only (Green and Wang 1994; Radó-Trilla and Albà 2012; Ellegren 2000). One illustrative example is the gene family comprised of POU4F1, POU4F2 and, POU4F3. The three genes have two exons and encode a highly conserved POU domain at the protein C-terminus. We identified LCRs in two gene family members: POU4F1 and POU4F2. The first protein contained two LCRs, one enriched in alanines and another one in glycines. The second protein had a histidine-rich tract, previously shown to be required for nuclear speckle targeting (Salichs et al. 2009), and another LCR containing several serine and glycine runs. These different repeat expansions mapped to different exons and had clearly occurred independently in the two proteins.

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. This is consistent with earlier observations that both alanine and glycine-rich LCRs are easily gained in vertebrate proteins (Radó-Trilla and Albà 2012). These amino acids are encoded by GC-rich codons and their abundance has been previously related to nucleotide compositional constraints in vertebrate genes (Nakachi et al. 1997; Albà and Guigó 2004).

Experiments performed in the gene families PHOX2A/PHOX2B and LHX2/LHX9 showed that the gain of a novel LCR strongly impacts the ability of the protein to activate transcription (Figures 3 to 6). In our assay, the alanine repeats in both PHOX2B and LHX2 clearly behaved as transcriptional activation domains. This effect may be 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). We noted that, despite the fact that repeats can evolve very rapidly during evolution (Fondon and Garner 2004;

Mularoni et al. 2008), the size of these alanine-rich tracts was well conserved across mammals, suggesting that there are important size constraints. 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 (Di Lascio et al. 2013) and impairs the ability of the PHOX2B to regulate the transcription of its target genes (Bachetti et al. 2005; Trochet et al. 2005). 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 transactivation activity when assayed *in vitro* (Adachi et al. 2000), suggesting that when the tract becomes abnormally long it has a negative impact on transcriptional activity. 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).

Despite the positive link between transcriptional activation and the presence of a poly-alanine tract in the TFs tested, the evolutionary conservation of the LCR might be linked to other functional activities. In this line, 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). In fact, the analysis of Gene Ontology terms in proteins containing alanine-rich LCRs yielded a similar number

of genes annotated as "positive regulation of transcription" or annotated as "negative regulation of transcription" (Supplemental Table 6), suggesting that they can influence transcription in very different contexts.

In summary, the results of the present study indicate that LCR gain is greatly facilitated by the existence of additional gene copies and that, at least in the case of alanine repeats, this can result in changes in transcriptional activity. The LCR-mediated neofunctionalization events are likely to have been subject to positive selection, thereby contributing to the increased 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), and obtained 1,177 different genes. Subsequently, we used the paralogous gene dating information in Ensembl Compara (Vilella et al. 2009) to obtain 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 separation of jawed vertebrates (Euteleostomi paralogy type).

For each gene family member, orthologous sequences from *Mus musculus* (mouse), *Gallus gallus* (chicken), and *Dario 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.

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). We used alignments that included orthologous sequences 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* (Supplemental Table 9 and 10). Ancestral sequence reconstruction was based on alignments obtained with Prank +F to increase the accuracy of the position of indels (Löytynoja and Goldman 2008; Villanueva-Cañas et al. 2013; Laurie et al. 2012). The amino acid alignments were converted to coding sequence alignments using PAL2NAL program (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 (Supplemental Table 13) 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 (Supplemental Table S13) 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.

Supplemental material

All supplemental tables mentioned in the text are available from http://www.evolutionarygenomics.imim.es (Publications, Datasets).

Figure legends

Figure 1. Diagram showing the gene family datasets studied and the number of proteins with repeats in each of them. The repeats correspond to low-complexity regions (LCR) detected by the SEG algorithm. 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 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 23 had repeats (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 (Supplemental Table 10).

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 (Supplemental Table 12).

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.

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**: 539–54.

Albà MM, Guigó R. 2004. Comparative analysis of amino acid repeats in rodents and humans. *Genome Res* **14**: 549–54.

Albà MM, Laskowski RA, Hancock JM. 2002. Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics* **18**: 672–8.

Amiel J, Laudier B, Attié-Bitach T, Trang H, de Pontual L, Gener B, Trochet D, Etchevers H, Ray P, Simonneau M, et al. 2003. Polyalanine expansion and frameshift mutations of the paired-like homeobox gene PHOX2B in congenital central hypoventilation syndrome. *Nat Genet* **33**: 459–61.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25–9.

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**: 1815–24.

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**: 13290–302.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. *Trends Genet* **21**: 219–26.

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**: 411–20.

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**: 4392–403.

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**: 6635–42.

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**: 400–2.

Escriva H, Bertrand S, Germain P, Robinson-Rechavi M, Umbhauer M, Cartry J, Duffraisse M, Holland L, Gronemeyer H, Laudet V. 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**: 325–35.

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**: 537–51.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, 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**: 7037–45.

Fondon JW, Garner HR. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A* **101**: 18058–63.

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res* **16**: 805–14.

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**: 154–61.

Galant R, Carroll SB. 2002. Evolution of a transcriptional repression domain in an insect Hox protein. *Nature* **415**: 910–3.

Gatchel JR, Zoghbi HY. 2005. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet* **6**: 743–55.

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**: 808–11.

Green H, Wang N. 1994. Codon reiteration and the evolution of proteins. *Proc Natl Acad Sci U S A* **91**: 4298–302.

Guindon S, Lethiec F, Duroux P, Gascuel O. 2005. PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* **33**: W557–9.

Han, M. V, Demuth, J. P., McGrath, C. L., Casola, C., & Hahn, M. W. (2009). Adaptive evolution of young gene duplicates in mammals. Genome Research, 19(5), 859–67

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 MA, Clark AG. 2007. Evolutionary analysis of amino acid repeats across the genomes of 12 Drosophila species. *Mol Biol Evol* **24**: 2598–609.

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**: 2281–90.

Kadrmas JL, Beckerle MC. 2004. The LIM domain: from the cytoskeleton to the nucleus. *Nat Rev Mol Cell Biol* **5**: 920–31.

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**: 333–8.

De la Luna S, Allen KE, Mason SL, La Thangue NB. 1999. Integration of a growth-suppressing BTB/POZ domain protein with the DP component of the E2F transcription factor. *EMBO J* **18**: 212–28.

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**: 478–85.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**: 1632–5.

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, K. D., & Li, W.-H. (2003). Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Research, 13(7), 1638–45.

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**: 3681–96.

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**: 3557–64.

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**: 745–54.

Mularoni L, Toll-Riera M, Albà M. 2008. Trinucleotide repeats in human and ape genomes. In *Encyclopedia of Life Sciences*, John Wiley & Sons, Ltd, Chichester, UK.

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**: 1042–9.

Ohno S. 1970. *Evolution by gene duplication*. Springer New York.

Ohno S, Wolf U, Atkin NB. 1968. Evolution from fish to mammals by gene duplication. *Hereditas* **59**: 169–87.

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**: 4065–75.

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**: 366–70.

Pich I Roselló, O., & Kondrashov, F. A. (2014). Long-term asymmetrical acceleration of protein evolution after gene duplication. Genome Biology and Evolution, 6(8), 1949–55.

Pegueroles C, Laurie S, Albà MM. 2013. Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. *Mol Biol Evol*.

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**: 469–76.

Radó-Trilla N, Albà MM. 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**: 1825–8.

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**: 137–47.

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.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**: W609–12.

Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang J-M, Taly J-F, Notredame C. 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res* **39**: W13–7.

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**: 3697–708.

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**: 327–35.

Villanueva-Cañas JL, Laurie S, Albà MM. 2013. Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biol Evol* **5**: 457–67.

Wootton JC, Federhen S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* **266**: 554–71.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–91.

Zhang, P., Gu, Z., & Li, W.-H. (2003). Different evolutionary patterns between young duplicate genes in the human genome. Genome Biology, 4(9), R56.

Figure 1

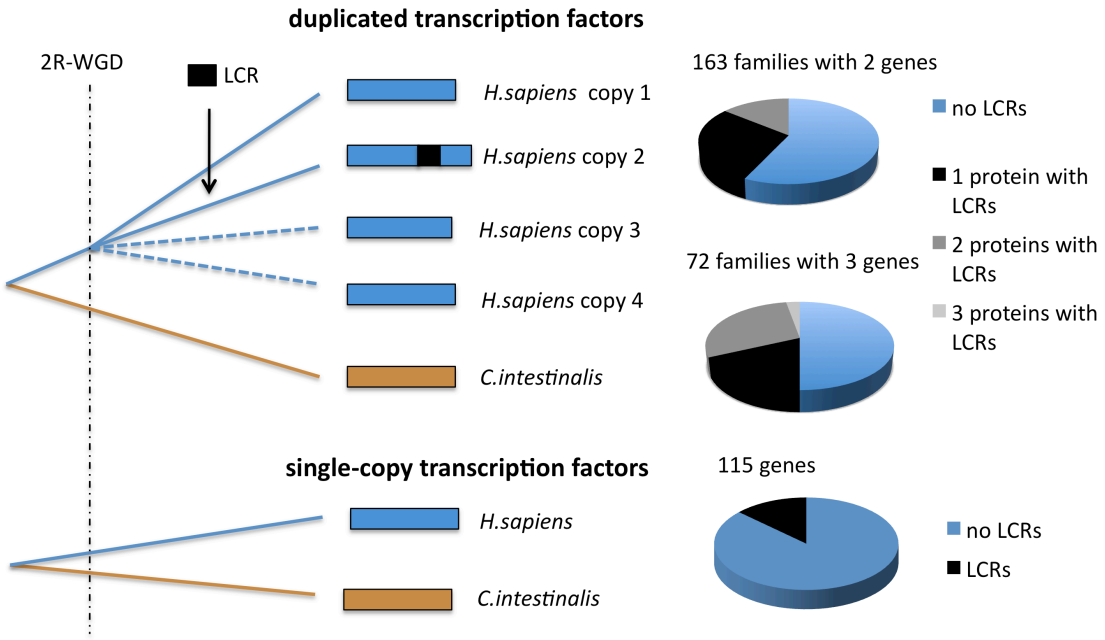

Figure 2

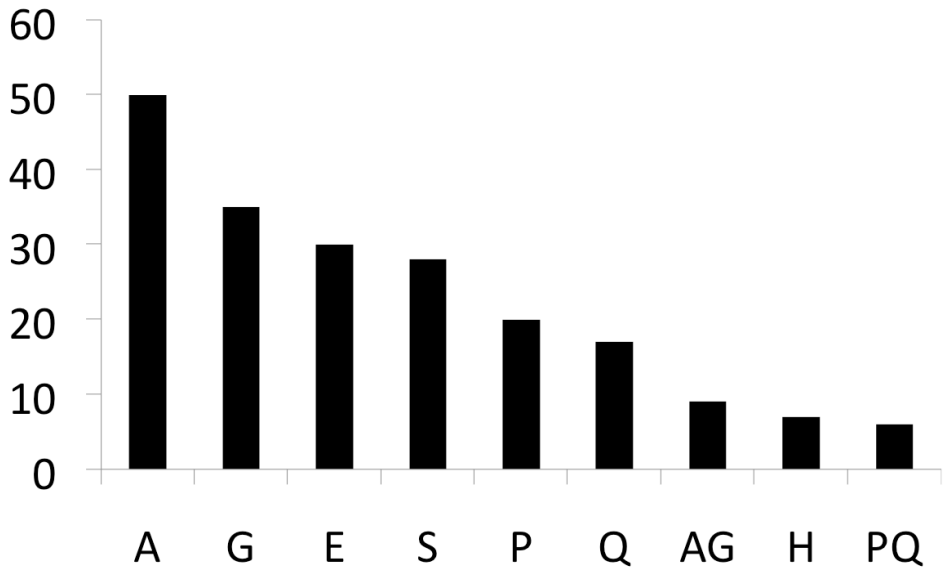

Figure 3

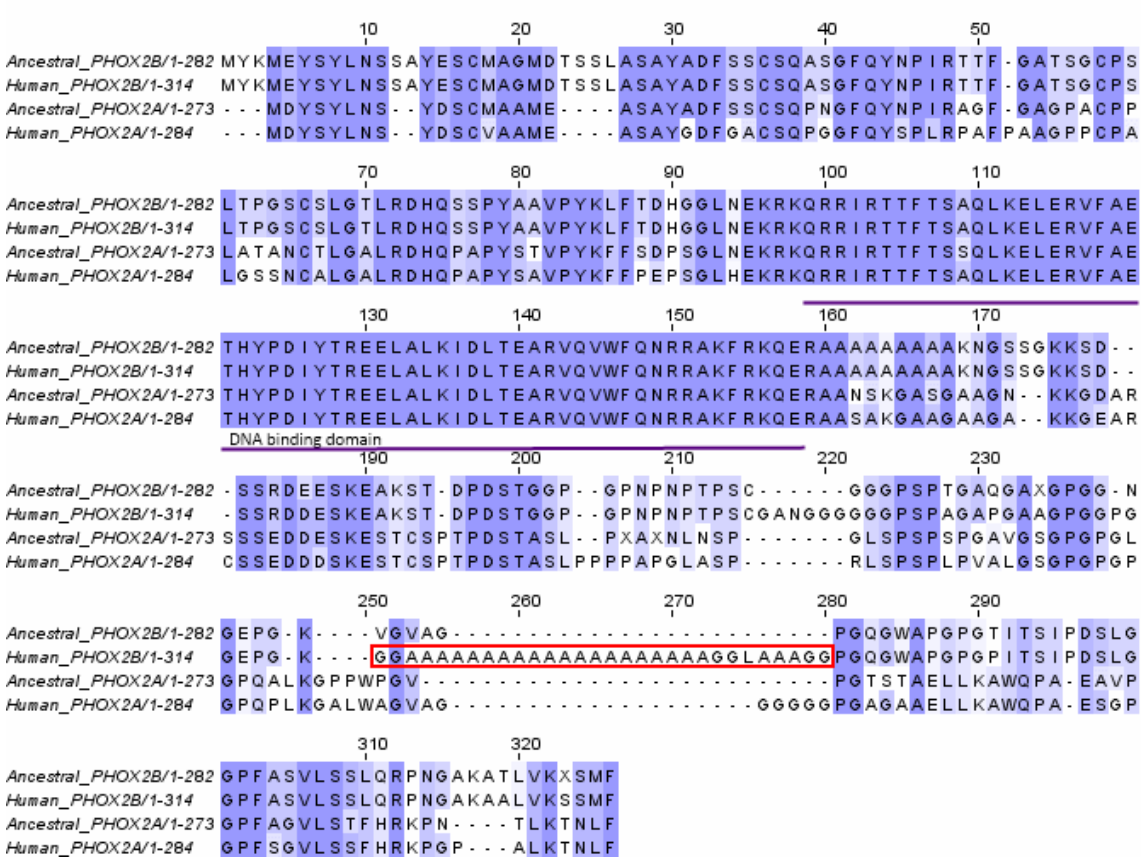

Figure 4

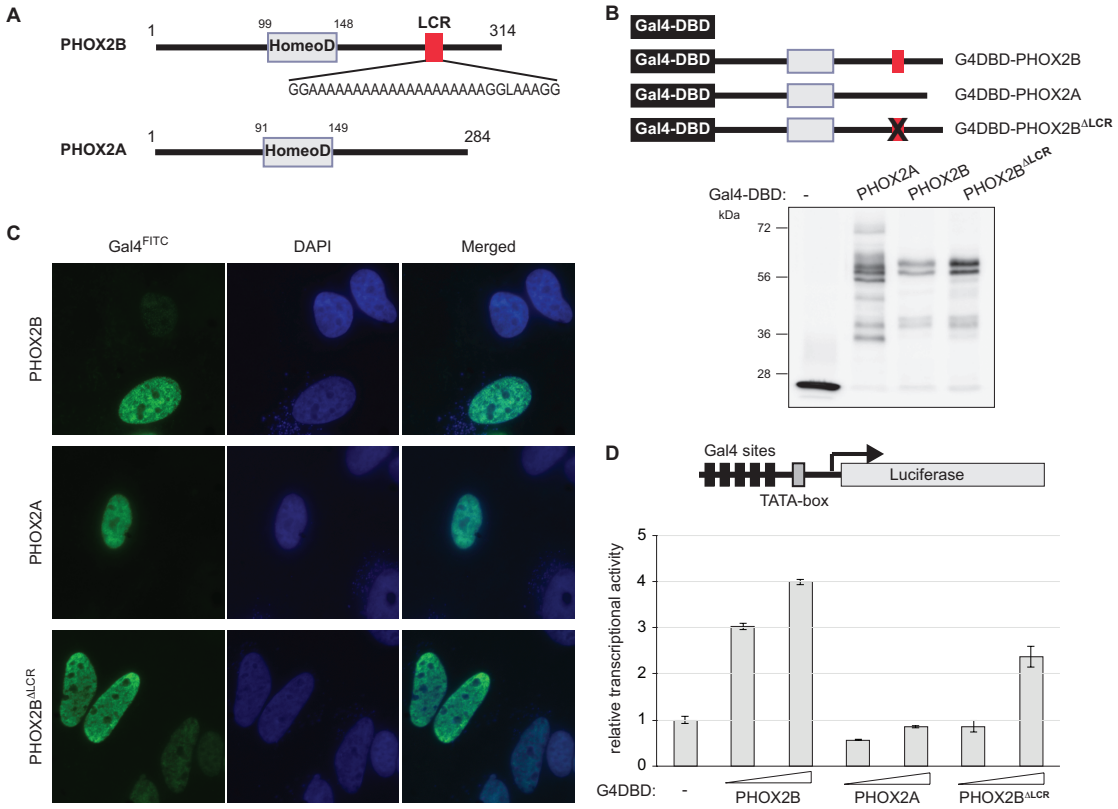

Figure 5

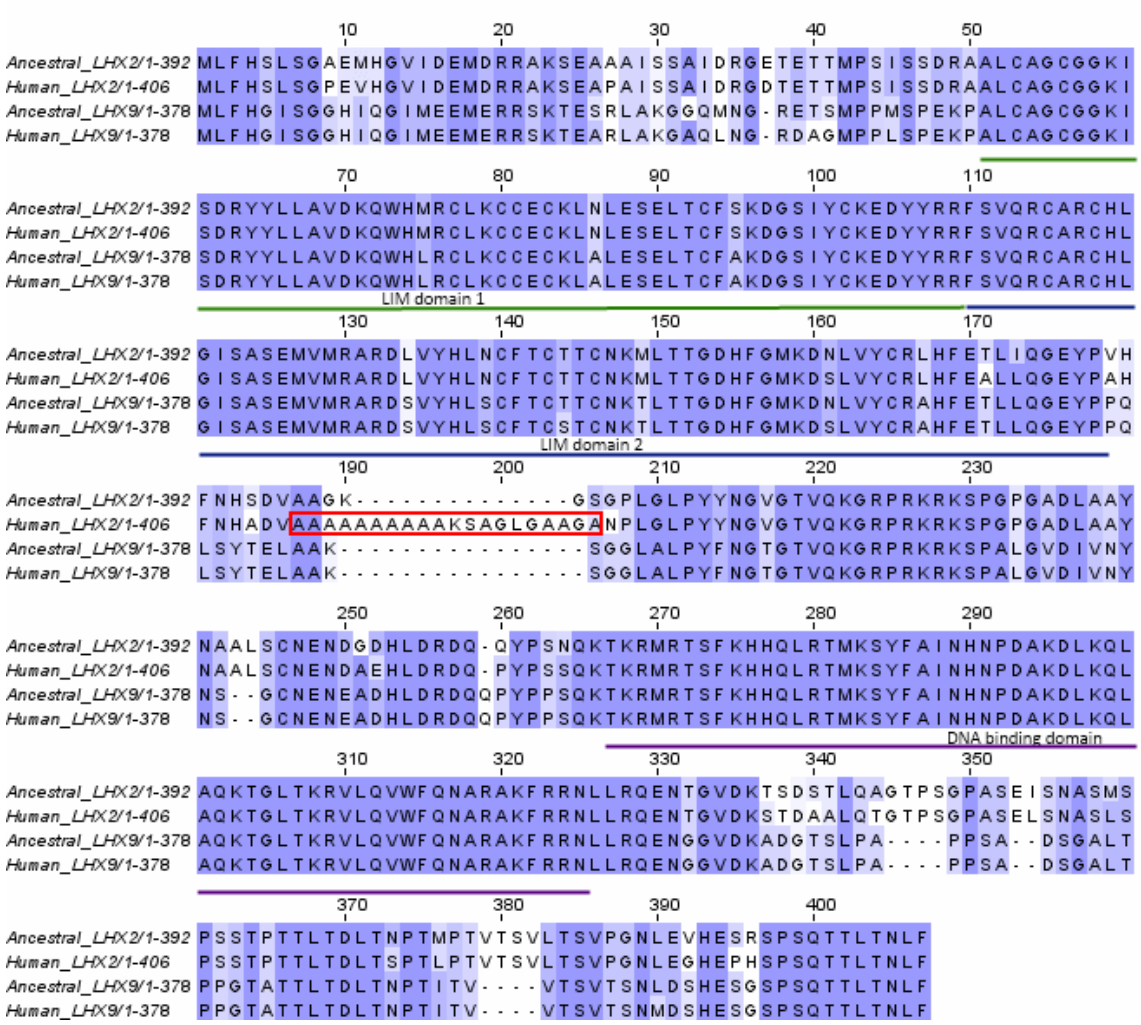

Figure 6