Homeobox gene clusters and the human paralogy map

Cornel Popovici[a,b], Magalie Leveugle[a], Daniel Birnbaum[a,b,c,*], François Coulier[a,c]

[a] *Laboratoire d'Oncologie Moléculaire, U119 Inserm, IFR57, 27 bd. Leï Roure, 13009 Marseille, France*
[b] *Institut Paoli-Calmettes, Marseille, France*
[c] *Atelier de Bio-informatique, U119 Inserm, Marseille, France*

Received 7 December 2000; revised 24 January 2001; accepted 24 January 2001

First published online 9 February 2001

Edited by Matti Saraste

Abstract Homeobox genes encode important developmental control proteins. In vertebrates, those encoding the proteins of the HOX class and their most closely related families, including paraHOX and metaHOX classes, are clustered in paralogous regions (or paralogons). We show that the majority of the other homeobox genes (we called contra*HOX*) can also be clustered and belong to paralogons in humans. This suggests that they duplicated during vertebrate evolution along the same processes as the *HOX* genes. We tentatively assembled several paralogons in superparalogons. One of the superparalogons contains the contra*HOX* genes. These observations were extended to hundreds of genes, and allowed to describe a primary human genome paralogy map.

Key words: Duplication; Evolution; Genome; Homeobox gene; MetaHOX gene; Paralogy

1. Introduction

Proteins with homeodomains are major regulators of embryonic development in metazoans. They belong to several classes that can be recognized structurally and phylogenetically ([1]; http://www.sciencemag.org/feature/data/985556s2.gif). *Hox* and para*Hox* genes are constraint by rules of colinear expression that allow them to specify the anterior-polarity axis of ectoderm and endoderm derivatives in deuterostomians. *Hox* genes are clustered in most analyzed genomes: one *Hox* cluster exists in non-vertebrates whereas four *Hox* clusters exist in mammals. *Amphioxus* has one para*Hox* gene cluster [2] and mammals four remnants of para*Hox* clusters [3,4]. It was thought that the other homeobox genes were dispersed, so-called orphan genes. However, two recent articles have shown that this may not be the case. A meta*Hox* gene cluster has been described in *Drosophila* [5]. The ancestor of this cluster (also called *NKL*) has duplicated in the vertebrate lineage and four clusters can be found in mammals [4,5]. *HOX*, para*HOX* and meta*HOX* genes are closely related and can be grouped in a large superfamily sometimes called the 'ANTP' (for the *Drosophila Antennapedia* gene) homeobox genes superclass [4], which we will simply designate here as the *HOX* superclass of homeobox genes (http://www.sciencemag.org/feature/data/985556s2.gif).

The number of *HOX* genes in vertebrates is in agreement with the hypothetical model of genome duplications thought to have led to increase in gene number in this lineage [6–11]. It is believed that the vertebrate ancestor had a single gene corresponding to a gene family of two, three or four members in present-day tetrapods due to large-scale genome duplication events. This is sometimes known as the 'one-to-four model'. This mode of evolution is not however definitely acknowledged [12–15]. It has been hypothesized that two rounds of large-scale duplication occurred after the divergence of vertebrates from the cephalochordates [9] but this may be an oversimplification [16]. More duplications have occurred in the teleost fish lineage [17,18].

While the majority of homeogenes encoding proteins of the HOX superclass belong to clusters in mammals, many homeobox genes encoding proteins that belong to other classes than the HOX superclass exist in their genomes (http://www.sciencemag.org/feature/data/985556s2.gif). Because they do not belong to any gene cluster and no concerted expression has ever been demonstrated, these genes are considered dispersed genes. However, this was before extensive sequence and cytogenetical data provide information on their localization. We describe here other potential homeobox gene clusters. These clusters are included in large regions of paralogy, which allowed us to establish a primary human paralogy map.

2. Materials and methods

2.1. Database searches

We used the OMIM (http://www.ncbi.nlm.nih.gov/Omim), Genatlas (http://bisance.citi2.fr/GENATLAS/), and NCBI (http://www.ncbi.nlm.nih.gov/LocusLink/) databases to search for gene localization and gene sequences. Blast searches against the unfinished human genome sequence (http://www.ncbi.nlm.nih.gov/blast/) were done. The Mouse Genome Informatics (http://www.informatics.jax.org/), the Fly base (http://flybase.bio.indiana.edu:82/), the genome annotation database of *Drosophila* (http://hedgehog.lbl.gov:8000/cgi-bin/annot/query) and the Sanger Center (http://www.sanger.ac.uk/Projects/C_elegans/wabace_front_end.shtml) databases were consulted for information on orthologs. In the tables, genes are designated by an OMIM, Genatlas (GEN) or NCBI (ID) number.

Paralogs included in the tables were selected using several criteria: (i) a search for human genes with consequent structural similarity (roughly over 50%) that are localized in potential paralogous regions (as defined in tables); (ii) each amino acid sequence derived from these genes was blasted (http://www.ncbi.nlm.nih.gov/blast/blast.cgi) against protein sequences in *Drosophila melanogaster*; the selected paralogs should be related to a common sequence (or common sequences in case of duplication in the fly lineage) in the *Drosophila* genome. However, this was not always the case. (iii) In the absence of obvious ortholog, a phylogenetic tree of the family was derived using mammalian (human or mouse) and fly sequences. None of these criteria was absolute.

*Corresponding author. Fax: (33)-4-91 26 03 64.
E-mail: birnbaum@marseille.inserm.fr

Abbreviations: PG, paralogon

PII: S0014-5793(01)02187-1

Table 1
Clusters of homeobox genes in human genome PGs

2q/12q/17q PG	2q11-35	12q12-22	17q12-25	7	*Drosophila*		
Homeoprotein class, superclass	D	C	B	A			
				7p15-21			
HOX*			MEOX1 (11–58)	MEOX2 (12–20)	btn (3R–94B)		
HOX clusters, HOX*	HOXD(2–45)	HOXC(15–57)	HOXB (11–56)	HOXA (6–26)	HOM-C (3R–84/89)		
HOX*	EVX2 (2–45)			EVX1 (6–26)	eve (2R–46D)		
				7q21-35			
HOX*	DLX1, 2 (2–44)		DLX3, 4 (?)	DLX5, 6 (6–2)	dll (2R–60E)		
HOX*	EN1 (1–64)			EN2 (5–15)	en (2R–48A)		
HOX*	GBX2 (1–65)			GBX1 (5–15)	unpg (2R–45B)		
HOX*				HLXB9 (5)	CG8254 (3L–66A)		
1/9/19p PG	1	19p12-13	9q13-34	5p15-q23	6p21-23	*Drosophila*	
Homeoprotein class, superclass	A	B	C	C	D		
	1p21-36						
HOX*	BARHL2 (?)		BARHL1 (2–17)			B-H2 (X–15F), B-H1 (X–16A)	
LIM class, ContraHOX	LHX8 (3)		LHX6 (?)			Awh (3L–63F)	
	1q12-25						
LIM class, ContraHOX	LHX9 (1–82)		LHX2 (?)			ap (2R–41E)	
LIM class, ContraHOX	LHX4 (1–82)		LHX3 (2–16)			Lim3 (2L–37C)	
LIM class, ContraHOX	LMX1A (1–88.2)		LMX1B (2–21)			CG10432, CG4328 (3L–66A)	
PRD class, ContraHOX	PMX1/PHOX1 (1–85.4)		PRX2 (2–19)			CG9876 (2R–59C)	
TALE class, ContraHOX	PBX1 (1–88.1)		PBX3 (2–22)		PBX2 (?)	exd (X–13F)	
POU class, ContraHOX	POU2F1 (1–87.2)	POU2F2 (7–6)			POU5F1 (17–19)	nub, pdm2 (2L–33E)	
TALE class, ContraHOX				IRX1, 2, 4 (13–43)			
11q/14q/19q PG	19q13	1q23-44	2p22-25	20p11-13	14q12-32	15q11-26	*Drosophila*
Homeoprotein class, superclass	B	C	C	C	D	D	
PRD class, ContraHOX					GSC (12–52)		Gsc (2L–21C)
PRD class, ContraHOX				VSX (?)	CHX10 (12–38)		CG4136, CG15782 (X–5A)
NKX2 class, ContraHOX				(Nkx2-4; 2–76)	(Nkx2-9; 12–26)		vnd (X–1B)
NKX2 class, ContraHOX				NKX2B (2–76)	NKX2A/TITF1 (12–26)		vnd (X–1B)
PRD class, ContraHOX	CRX (7–8.5)		OTX1 (11–12)		OTX2 (14–19)		oc (X–7F)
SIX class, ContraHOX	SIX5 (7–4)		SIX2, 3 (17–45.5)		SIX1, 4, 6 (12–31)		so (2R–43C), Optix (2R–44A)
CUT class, ContraHOX	ONECUT3 (?)					ONECUT1 (9–42)	onecut (4–102C)
TALE class, ContraHOX	(Meis 3; 7–6)		MEIS1 (11–11)			MEIS2 (2–64)	hth (3R–86B)
Prospero class, ContraHOX		PROX1 (1–106)					pros (3R–86E)
HOX*		HLX1 (1–99.5)					H2.0 (2L–26B)
1/11/12 PG	1p11-21	11p11-15	19q13	12q13-24	*Drosophila*		
Homeoprotein class, superclass	A	B	C	D			
LIM class, ContraHOX		LHX1 (11–48)		LHX5 (5–64)	Lim1 (X–8A)		
PRD class, ContraHOX	ALX3 (3–48.8)	ALX4 (2–52)		CART1 (10–55.5)	al, Pph13 (2L–21C), CG10614 (2R–57B)		
6p/11q/12p/19q PG	11q13	6p	12p	19q13	*Drosophila*		
Homeoprotein class, superclass	A	B	C	D			

Table 1 (*continued*)

PRD class, ContraHOX		ARIX/PHO2A (7–50)			PHDP (2R–60A)
4q/5q/Xq/13q PG	4q12-35	5q31-35	Xq12-8	13q12-34	*Drosophila*
Homeoprotein class, superclass	A	B	D	C	
ParaHOX class, HOX*	GSH2 (5–41)			GSH1 (5–82)	ind (3L–71B)
ParaHOX class, HOX*		CDX1 (18–30)	CDX4 (X–42.5)	CDX2 (5–82)	cad (2L–38E)
ParaHOX class, HOX*				PDX1/IPF1 (5–82	?
POU class, ContraHOX	POU4F2 (X)	POU4F3 (18–24)		POU4F1 (14)	acj6 (X–13C)
4/5q/10q PG	4	5q14-35	10q21-26	8p/2p	*Drosophila*
Homeoprotein class, superclass	A	D	C	B	
	4p13-16			**2p11-23**	
MetaHOX class, HOX*			VAX1 (19–53.5)	VAX2 (6–35.5)	CG9930 (3R–88A)
MetaHOX class, HOX*			EMX2 (19–53.5)	EMX1 (6–35.5)	ems (3R–88A)
MetaHOX class, HOX*			HPX42B (?)		?
MetaHOX class, HOX*	HMX1 (5–18)		HMX2, 3 (7–65)		NK7.1 (3R–88C), CG5832 (3R–90B)
MetaHOX class, HOX*			HHEX (19–47.5)		CG15696 (3R–93A), CG7056 (3R–93C)
MetaHOX class, HOX*		CSX/NKX2-5 (17–13)	NKX2-3 (19–42)		tin/NK4 (3R–93D)
MetaHOX class, HOX*		TLX3 (?)	HOX11/TLX1 (19–43)	TLX2 (6–35.5)	C15 (3R–93E)
				8p12-22	
MetaHOX class, HOX*	BAPX1 (5–23)			NKX3A (?)	bap/NK3 (3R–93E)
MetaHOX class, HOX*			LBX1 (19–47.5)		lbl, lbe (3R–93E)
MetaHOX class, HOX*	(Sax2/Nkx-1; 5–18)		SAX1 (7–?)		slou/NK1 (3R–93E)
MetaHOX class, HOX*	MSX1 (5–21)	MSX2 (13–32)	(Msx3; 7–68)		msh (3R–99B)
	4q21-25				
PRD class, ContraHOX	PITX2 (3–57.7)	PITX1 (13–34)	PITX3 (19–46.5)		Ptx (3R–100B)
MetaHOX class, HOX*	NKX6A (?)		(Nkx6-2; 7–68)		CG4745 (3L–70E)

Most homeobox genes that belong to PG are listed. The meta*HOX* genes, which are found on a 4/5q/10q PG (described in [5]), have been added. A total of 24 homeobox clusters, some currently reduced to only one gene, can thus be described in the human genome. The complete sets of genes in the PGs are given in tables 2–9 (http://www.elsevier.nl/febs/show/).

For each gene, the localization of the mouse ortholog is indicated in parentheses (chromosome–cM). In the right column is indicated the localization of the *Drosophila* ortholog (map position: chromosome–segment). Paralogous clusters are designated by letters (see tables 2–9). Homeobox proteins from HOX, paraHOX and metaHOX classes collectively belong to the HOX* superclass. The other classes are as previously defined (http://www.sciencemag.org/feature/data/985556s2.gif) and are collectively designated as contraHOX.

The 2q/12q/17q PG contains a series of genes encoding homeobox proteins: mesenchyme homeobox proteins (OMIM 600147, 600535), homeobox proteins of the four HOX clusters, distal-less homeobox proteins (OMIM 126255, 600028, 600029, 600030, 600525, 601911), even-skipped homeobox proteins (OMIM 142991, 142996), engrailed homeobox protein (OMIM 131290, 131310), gastrulation brain homeobox proteins (OMIM 601135, 603354) and HLXB9 (OMIM 142994). The 1/9/19p PG includes homeobox genes encoding: BarH-like homeobox (OMIM 605211, 605212), LIM homeobox (OMIM 600298, 600577, 602575, 603759, ID 26468, 56956), paired mesoderm homeobox (OMIM 167420, ID 51450), PBX pre-B cell leukemia homeobox (OMIM 176310, 176311, 176312) and iroquois-related homeobox proteins (ID 50805), and POU domain transcription factors (OMIM 164175, 164176, 164177). Some homeobox genes have been added in this PG after blasting the emerging human sequence with the mouse corresponding sequences.

The 11q/14q/19q PG contains genes encoding homeoproteins of various classes (OMIM 138890, 142993, 142995, 600036, 600037, 601546, 601739, 601740, 602225, 600635/603245, 604164, 604612, 605020, ID 29782), including sine oculis homeoproteins (OMIM 600963, 601205, 603714, ID 4990, 10736, 51804). The mouse *Nkx2-4*, *Nkx2-9* and *Meis3* genes, not found in humans, are also indicated in parentheses. The 1/11/12 PG currently contains five homeobox genes (OMIM 601527, 601999, OMIM 605420, ID 257, 64211), and the 6p/11Q/12p/19q PG the *ARIX* gene (OMIM 602753).

The 4q/5q/Xq/13q PG (a part of a super PG described in table 9) contains genes encoding the paraHOX GSX/GSH, PDX1/IPF1/XLOX, CDX (OMIM 300025, 600297, 600733, 600746) proteins, and POU domain transcription factors (OMIM 113725, 601632, 602460).

The 4/5q/10q PG (described in [4,5]) contains homeobox genes encoding metaHOX proteins VAX (OMIM 604294, 604295), EMX (OMIM 600034, 600035), HPX42B (ID 27287), HMX/H6 (OMIM 142992, 600647, ID 3168), HHEX (OMIM 604420), CSX/NKX2.5 and NKX2-3 (OMIM 600584, ID 53631), TLX/HOX11 (OMIM 186770, 604240, 604640), NKX3A and NKX3B/BAPX1 (OMIM 602041, 602183), LBX1 (OMIM 604255), SAX1 (identified by blast search, AW075638) and MSX1/HOX7 and MSX2/HOX8 (OMIM 123101, 142983), NKX6A (OMIM 602563), as well as PITX paired-class homeoproteins (OMIM 601542, 602149, 602669). The mouse genes *Nkx6-2*, *Sax2/Nkx-1*, *Msx3*, which have not been characterized in humans, belong to the cluster; they are indicated in parentheses.

2.2. Phylogenetic trees

Phylogenetic trees were constructed from sequence alignments obtained with Clustal W (Blosum 30 matrix), using the neighbor-joining algorithms implemented in Clustal W. A number of phylogenetic trees were also retrieved from the literature [28–34].

3. Results and discussion

3.1. Definitions

Two terms are commonly used to define relationships between genes [19]: two genes are 'orthologs' if they diverged due to a speciation event; they are 'paralogs' if they diverged due to duplication within a lineage. Chromosomal regions that contain paralogs are paralogous regions. We proposed the name 'paralogon' (PG) [5] to designate a series of paralogous regions that could be recognized as deriving from a common ancestor region along the hypothetical model described above. The name 'proto PG' applies to the ancestor of these regions, before duplication. According to the 'one-to-four rule' that is thought to have governed tetrapod genome evolution, a typical PG would be constituted of four clusters or regions of paralogy in mammals. We will call 'paralogs' only the genes from the same subfamily that belong to a PG and have a phylogenetic relationship in agreement with the 'one-to-four' rule (e.g. *ERBB1-4* tyrosine kinase receptor genes, or *EVX1 and EVX2* genes); other genes of the same family will be called 'metalogs' (e.g. *EVX1* and *OTX1* genes). Thus paralogy, orthology and metalogy define relatedness within a species and within a subfamily, across species and within a subfamily, within a species and across subfamilies, respectively. According to this model, the number of paralogs should be four in humans, but it may be less if the subfamily has been subjected to gene loss during evolution and more if it has been amplified. Additional terms will be defined later (see also our web site: http://u119.marseille.inserm.fr/Db/paralogy.html).

We will consider here the model of large-scale duplications in early vertebrate ancestry as a working hypothesis, and gene families that potentially derive from these duplications, independently of the number of rounds of duplication which may have occurred [16].

3.2. Identification of PGs

We searched the human genome for chromosomal regions that contain paralogs. Several paralogous regions, i.e. PGs, have been recognized in earlier works [8,11,20–25]. We recently described two such PGs [3,5]. We describe here other potential PGs. Genes belonging to these PGs, i.e. paralogs, are listed in the tables: the homeobox genes found in the PGs are listed in Table 1; the complete sets of genes are shown in tables 2–9 as supplementary material on the web version (http://www.elsevier.nl/febs/show/ or http://www.elsevier.nl/PII/S0014579301021871).

The '2q/12q/17q PG' is described here as one of the best known and recognized example of PG (see table 2 on the web version). It includes regions containing the four *HOX* gene clusters (on 2q, 7p, 12q and 17q) and can be extended to two other chromosomal regions (3p and 7q). In Table 1 and table 2 on the web version, the paralogous clusters are labeled A–D according to the *HOX* clusters. Other homeobox genes were included in this PG and were grouped with the HOX clusters (Table 1). The location of the mouse orthologs of the homeobox genes are listed in parentheses. The right column of the table shows potential orthologs in the *Drosophila* genome with their localization (chromosome–segment). Some degree of putative endo-*cis*-duplication may be observed for this PG (see table 2 on the web version, #, ##, @, @@), thus defining a cluster of nuclear hormone receptor genes.

In table 3 (on the web version), we show a list of genes mapping to other paralogous regions, thus defining another PG. Part of this PG has been recognized earlier [21,22,26]. Like for the 2q/12q/17q PG, we have extended the previous observations on this PG. The participation of chromosome arms 1p and 1q, 9p and 9q, and 5p and 5q shows that a region of paralogy can span the centromere (this was obvious also for chromosome 7 in table 2).

Table 4 (on the web version) shows series of gene families that are again distributed over seven chromosomal regions. This would not fit the one-to-four rule. However, on closer look, regions on chromosome arms 1q, 2p and 20p on the one hand, and 14q and 15q on the other hand, can be grouped together in single clusters of genes (labeled respectively C and D in the table). Split clusters were already evident in the two previously described PGs (3p/7 and 6p/15q). This last example shows best that a cluster of paralogy may be split over at least three different portions of genome. The three portions can be put together in a single cluster (C) that can now fit more conventionally into a PG (hereafter designated 11q/14q/19q PG). A split cluster might correspond to a single ancestral segment. Comparison of genetic maps in other species could be helpful in this context. Thus, identification of split clusters can be used as a general method to link various regions of chromosomes and establish contiguity independently of the physical islands that represent the chromosomes in order to help reconstitute portions of an ancestral genome [27].

In tables 5 and 6 (on the web version), we describe PGs that contain regions from chromosome arms 1p/1q, 11p, 12p/12q and 19q and from chromosome arms 6p, 11p/11q, 12p and 19q, respectively. Some of these locations (i.e. on chromosomes 1, 11 and 19) are shared by several PGs described above (tables 2–6 on the web version).

3.3. PGs and homeobox gene clusters

Each of the above described PGs contains a series of clustered homeobox genes (except the 6p/11/12/19q which contains only one such gene). They are listed in Table 1. The 2q/12q/17q PG contains genes encoding the *HOX* genes and related homeobox genes (http://www.sciencemag.org/feature/data/985556s2.gif). This clustering is also apparent in the mouse. These genes and the para*HOX* and meta*HOX* genes – the latter two are on two other PGs recently defined [3–5] (see also tables 8 and 9 on the web version, and Fig. 1) – are homeobox genes of the *HOX* superclass (HOX* in Table 1).

Tables 3–6 (on the web version) contain homeobox genes (boxed) that do not belong to the *HOX* superclass but to other classes: CUT, LIM, POU, PRD, Prospero, SIX and TALE classes. These classes are more related between themselves than to the HOX superclass (http://www.sciencemag.org/feature/data/985556s2.gif). Some of these genes do not have paralogs: the *PROX1* gene (see table 9 on the web version) could be placed in the 11q/14q/19q PG or in the 1/11/12 PG. We favor the location in the former due to the close linkage of *hth* and *pros* genes in *Drosophila*. We placed *HLX1* next to *PROX1* as in the mouse. A blast search of

the emerging human sequence with the mouse sequences identified the human *LHX4* and *LHX8* genes on chromosome 1 (AC073499, AC019260) and *IRX1* (Table 1) on chromosome 5 (AC016595). There remains a number of apparently truly dispersed homeobox genes, such as *BARX2* or *PKNOX1* (see table 9).

3.4. Various degrees of paralogy: paralogs and coparalogs

The PGs contain various subfamilies of genes which share different degrees of relatedness. The 2q/12q/17q PG for example, contains a series of homeobox genes, and a series of genes encoding steroid hormone receptors and related molecules (see table 2 on the web version, @@). We propose to designate different genes from subfamilies of paralogs (e.g. *RARA*, *RARB* and *RARG* are paralogs, and so are *EVX1* and *EVX2*) that belong to the same PG under the name '*coparalogs*'. Coparalogs may (e.g. *EVX* and *EN* genes) or may not (e.g. the *RAR* genes and the *EVX* genes) be related. Clustered genes described in the tables are coparalogs.

The basis for coparalogy is worth discussing. In the case of homeobox genes of the *HOX* class, it has been found to be an important basis for colinearity in expression. It is likely to have occurred through *cis*-duplications and to have been maintained through selective pressure. The presence of a cluster of hormone nuclear receptors in the 2q/12q/17q PG (table 2, @@), or of the amyloid precursor proteins and a class of enzymes involved in their processing in the same PG (see table 9) is worth mentioning as other examples of coparalogy of related genes or of genes involved in the same developmental or cellular processes.

3.5. Supersets of PGs and potential linkage of PGs

A certain degree of linkage and/or overlap of the 1/9/19p (table 3), 11q/14q/19q (table 4), 1/11/12 (table 5) and 6p/11/12p/19q (table 6) PGs is visible: they have clusters located on the same chromosomes (1, 6, 11, 12, 15 and 19).

We can thus establish a 'superset of PGs': four PGs can be merged in one 'super PG'. We will collectively designate the homeobox genes that do not belong to the *HOX* superclass as 'contra*HOX* genes' (meaning 'aside from *HOX*'). The majority of the contra*HOX* genes are contained in the super PG made of 1/9/19p, 11q/14q/19q, 1/11/12 and 6p/11/12p/19q PGs, which can thus be described as the 'contra*HOX* super PG'. These observations are summarized in Fig. 1. Thus, the homeobox genes belong schematically to two superclasses, the *HOX** (*HOX*, para*HOX* and meta*HOX*) and contra*HOX* (CUT, LIM, POU, PRD, Prospero, SIX and TALE) superclasses, which have co-evolved and for the majority, have been maintained in clusters.

Comparison with the genome organization of the mouse may help building supersets of PGs. Thus, in the mouse the homeobox clusters that have homologs in the 1/9/19p and 11q/14q/19q PGs may be close: *Pou2f2*, *Crx, Meis3* and *Six5* map to the same location, and *Pbx1, Lmx1a, Pou2f1* are close to *Prox1* (see table 7 on the web version). Similarly, the PGs containing the *Hox* and meta*Hox* gene clusters can be assembled (see table 8 on the web version).

3.6. Extension of the PG supersets

The possibility to assemble PGs in supersets is further exemplified in table 9 (on the web version), which shows a superset of three small PGs. One of these PGs contains the para*HOX* gene clusters [3,4]. This new super PG, which has clusters on chromosomes 11 and 19 also, was tentatively placed adjacent to the previously described super PG (Fig. 1).

We further placed the meta*HOX* cluster adjacent to this second super PG. The PG that contains the *HOX* genes (table 2) could be placed in contiguity with the meta*HOX* PG (Fig. 1); ancestral linkage of the two groups of *HOX* families are suggested by the mouse data (table 8) and has indeed already been proposed [4,5].

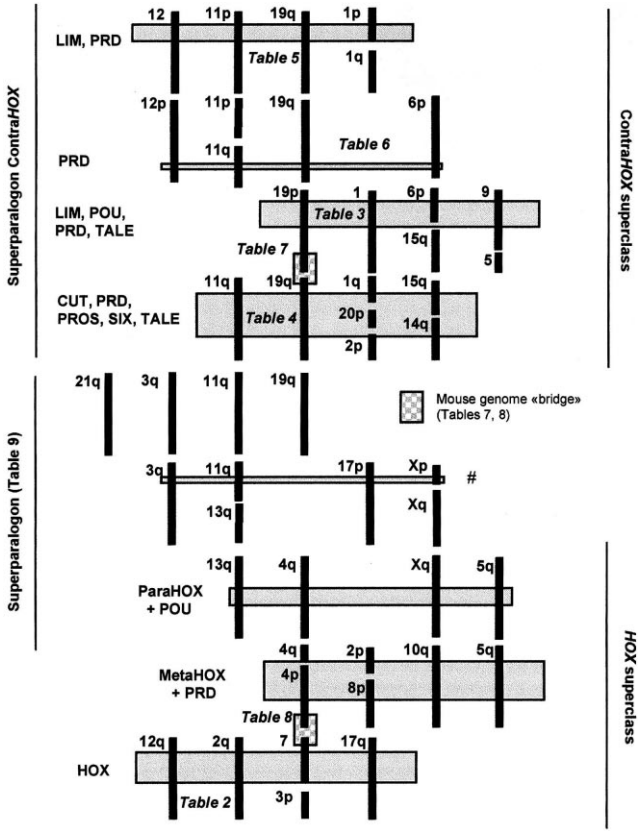

Fig. 1. Putative supersets of PGs containing homeobox gene clusters and a first tentative human genome paralogy map. PGs are schematized by series of four vertical bars representing clusters of genes, either uninterrupted or interrupted (split clusters). Chromosomal localizations of the clusters are indicated at the left of each bar. Gray boxes represent homeobox gene clusters; they are also described in Table 1. The correspondence with the tables found on the web version is indicated; see supplementary material, http://www.elsevier.nl/febs/show/. Starting from the top, four different PGs are placed in an adjacent manner according to overlapping chromosomal localizations in the human genome (their respective description in the tables is indicated). They contain homeobox genes (gray boxes) we collectively named contra*HOX* genes (described in tables 3–6, http://www.elsevier.nl/febs/show/). Together they constitute the contra*HOX* super PG. Then, a series of additional PGs is tentatively placed contiguously below: first, a super PG (described in table 9) made of three potentially adjacent PGs, one of which (4q/5q/Xq/13q) contains the para*HOX* gene clusters [3], and another one the *SHOX* genes (# box), second, the PG that contains the meta*HOX* gene clusters [5]. Collectively, these PGs contain most of the homeobox genes of the *HOX* superclass (see Table 1). Generally, from one PG to another, the bars representing individual clusters are contiguous when these clusters are on the same chromosome. Gray squares, i.e. 'mouse genome bridges' refer to data presented in tables 7 and 8 which help position the supersets of PGs in humans.

3.7. Concluding remarks: towards a human genome paralogy map

Even if some families, regions or PGs are misplaced, and if the relationship between PGs may be as much intrication as contiguity, our work provides an evolving frame within which more families and clusters could take place. More work is required to firmly establish the supersets and to position additional PGs. We have also found a 8q/16q/18q/20q PG and a 16p/17/22q PG (not shown) (see also [25]). With these, it appears that all chromosome arms are involved in paralogous relationships. Thus, large-scale duplications do not concern only limited regions but the whole human genome.

This led us to propose a first tentative physical 'human genome paralogy map' (Fig. 1). Assembling the supersets is no proof that an initial linkage existed in a ancestor. However, the comparison of this map with a similar map in other species, could distinguish what is most likely to have been linked in a common ancestor; additional evidence could be looked for in the non-vertebrate cephalochordate *Amphioxus*, which is supposed to derive from an ancestor prior to the large-scale duplications which have taken place in the vertebrate lineage.

In conclusion, we have provided in this report evidence for two main lines of thought. First, most of the homeobox genes are clustered (Table 1). This could have functional implications in terms of spatial and temporal patterns of gene expression, although clustering does not necessarily imply colinearity in expression. Second, the described homeobox clusters were used as a starting point to develop a more general view of the human genome and set up a large canvas of related regions (Fig. 1). It is apparent from the number of paralogs and chromosomal regions involved that the paralogy map covers the entire genome. This is in favor of a mechanism that involves massive large-scale duplications or tetraploidizations at the origin of the increase in gene number in vertebrates. Finding organization and patterns in genomes is a primary goal of paleogenomics [27]. This approach could usefully complement phylogenetical analyses to help deduce the evolution of the vertebrate genomes.

Acknowledgements: This work was supported by Inserm and Institut Paoli-Calmettes. We thank L. Abi Rached, C. Mawas and P. Pontarotti for help and advice.

References

[1] Ruvkun, G. and Hobert, O. (1998) Science 282, 2033–2041.
[2] Brooke, N.M., Garcia-Fernandez, J. and Holland, P.W.H. (1998) Nature 392, 920–922.
[3] Coulier, F., Burtey, S., Chaffanet, M., Birg, F. and Birnbaum, D. (2000) Int. J. Oncol. 17, 439–444.
[4] Pollard, S. and Holland, P.W.H. (2000) Curr. Biol. 10, 1059–1062.
[5] Coulier, F., Popovici, C., Villet, R. and Birnbaum, D. (2000) J. Exp. Zool. (Mol. Dev. Evol.) 288, 345–351.
[6] Schughart, K., Kappen, C. and Ruddle, F.H. (1989) Proc. Natl. Acad. Sci. USA 86, 7067–7071.
[7] Ohno, S. (1970) Evolution by gene duplication, Springer Verlag, Berlin.
[8] Lundin, L.G. (1993) Genomics 16, 1–19.
[9] Holland, P.W.H., Garcia-Fernandez, J., Williams, N.A. and Sidow, A. (1994) Development (Suppl.) 125–133.
[10] Sidow, A. (1996) Curr. Opin. Genet. Dev. 6, 715–722.
[11] Spring, J. (1997) FEBS Lett. 400, 2–8.
[12] Hughes, A.L. (1998) Mol. Biol. Evol. 15, 854–870.
[13] Skrabanek, L. and Wolfe, K.H. (1998) Curr. Opin. Genet. Dev. 8, 694–700.
[14] Martin, A.P. (1999) Am. Nat. 154, 111–128.
[15] Smith, N.G.C., Knight, R. and Hurst, L. (1999) Bioessays 21, 697–703.
[16] Wang, Y. and Gu, X. (2000) J. Mol. Evol. 51, 88–96.
[17] Amores, A., Force, A., Yan, Y.L., Joly, L., Amemiya, C., Fritz, A., Ho, R., Langeland, J., Prince, V., Wang, Y.L., Westerfield, M., Ekker, M. and Postlethwait, J.M. (1998) Science 282, 1711–1714.
[18] Aparicio, S. (2000) Trends Genet. 16, 54–56.
[19] Fitch, W.M. (1970) Syst. Zool. 19, 99–113.
[20] Rosnet, O., Stephenson, D., Mattei, M.-G., Marchetto, S., Shibuya, M., Chapman, V. and Birnbaum, D. (1993) Oncogene 8, 173–179.
[21] Kasahara, M., Hayashi, M., Tanaka, K., Inoko, H., Sugaya, K., Ikemura, T. and Ishibashi, T. (1996) Proc. Natl. Acad. Sci. USA 93, 9096–9101.
[22] Katsanis, N., Fitzgibbon, J. and Fisher, E.M. (1996) Genomics 35, 101–108.
[23] Ollendorff, V., Mattei, M.-G., Fournier, E., Adélaïde, J., Lopez, M., Rosnet, O. and Birnbaum, D. (1998) Int. J. Oncol. 13, 1159–1161.
[24] Pébusque, M.-J., Coulier, F., Birnbaum, D. and Pontarotti, P. (1998) Mol. Biol. Evol. 15, 1145–1159.
[25] Giles, R.H., Dauwerse, H.G., Van Ommen, G.J. and Breuning, M.H. (1999) Am. J. Hum. Genet. 63, 1240–1242.
[26] Abi-Rached, L., McDermott, M.F. and Pontarotti, P. (1999) Immunol. Rev. 167, 33–45.
[27] Birnbaum, D., Coulier, F., Pébusque, M.-J. and Pontarotti, P. (2000) J. Exp. Zool. (Mol. Dev. Evol.) 288, 21–22.
[28] Costache, M., Apoil, P.-A., Cailleau, A., Elmgren, A., Larson, G., Henry, S., Blancher, A., Iordachescu, D., Oriol, R. and Mollicone, R. (1997) J. Biol. Chem. 272, 29721–29728.
[29] Jékely, G. and Friedrich, P. (1999) J. Mol. Evol. 49, 272–281.
[30] Kaestner, K.H., Knöchel, W. and Martinez, D.E. (2000) Genes Dev. 14, 142–146.
[31] Laporte, J., Blondeau, F., Buj-Bello, A., Tentler, D., Kretz, C., Dahl, N. and Mandel, J.-L. (1998) Hum. Mol. Genet. 7, 1703–1712.
[32] Laudet, V. (1997) J. Mol. Endocrinol. 19, 207–226.
[33] Laudet, V., Hänni, C., Stéhelin, D. and Duterque-Coquillaud, M. (1999) Oncogene 18, 1351–1359.
[34] Popovici, C., Roubin, R., Coulier, F., Pontarotti, P. and Birnbaum, D. (1999) Genome Res. 9, 1026–1039.