Human Molecular Genetics, 2001, Vol. 10, No. 4 371–382

Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the α globin cluster

Jonathan Flint[1], Cristina Tufarelli[1], John Peden[1], Kevin Clark[1], Rachael J. Daniels[1], Ross Hardison[2], Webb Miller[2], Sjaak Philipsen[3], Kian Chen Tan-Un[4], Tara McMorrow[3], Jonathan Frampton[1], Blanche P. Alter[5,+], Anna-Marie Frischauf[6] and Douglas R. Higgs[1,§]

[1]MRC Molecular Haematology Unit, Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DS, UK, [2]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA, [3]Faculteit der Geneeskunde en Gezondheids-Wetenschappen, Erasmus Universiteit, dr. Molewaterplein 50, Rotterdam, The Netherlands, [4]School of Professional and Continuing Education, University of Hong Kong, Pokfulam Road, Hong Kong, [5]Division of Pediatric Hematology/Oncology, Children's Hospital, University of Texas Medical Branch, Galveston, TX, USA and [6]Institut fuer Genetik und Allgemeine Biologie, Universitaet Salzburg, Austria

Received 4 October 2000; Revised and Accepted 21 December 2000

We have cloned, sequenced and annotated segments of DNA spanning the mouse, chicken and pufferfish α globin gene clusters and compared them with the corresponding region in man. This has defined a small segment (~135–155 kb) of synteny and conserved gene order, which may contain all of the elements required to fully regulate α globin gene expression from its natural chromosomal environment. Comparing human and mouse sequences using previously described methods failed to identify the known regulatory elements. However, refining these methods by ranking identity scores of non-coding sequences, we found conserved sequences including the previously characterized α globin major regulatory element. In chicken and pufferfish, regions that may correspond to this element were found by analysing the distribution of transcription factor binding sites. Regions identified in this way act as strong enhancer elements in expression assays. In addition to delimiting the α globin chromosomal domain, this study has enabled us to develop a more sensitive and accurate routine for identifying regulatory elements in the human genome.

INTRODUCTION

Understanding the function of large stretches of genomic sequences is a major challenge of the post-genomic era: 4 years after its completion, two-thirds of the yeast genome has still to be assigned a function (1). The knowledge gap is even greater for mammalian genomes. A critical resource will be the availability of genomic regions that have been extensively characterized at a number of levels, allowing genomic sequence to be related to function. We have previously characterized the structure (2), epigenetic modifications (3–6) and function (summarized in refs 2 and 7) of a contiguous (376 kb) segment of DNA extending from the telomeric repeats of human chromosome 16p. This gene-rich region contains the α-like globin genes (tel-ζ-α2-α1-cen), which are expressed in a tissue- and developmental stage-specific manner, embedded within a variety of widely expressed genes (Fig. 1).

Observations of the normal human α globin cluster and its mutants *in vivo*, together with experimental analysis of constructs containing various segments (1.5–120 kb) of the α cluster in cell lines and transgenic mice, have identified some key *cis*-acting regulatory elements but, to date, it has not been possible to achieve fully regulated expression of the human α globin genes in transgenic mice (summarized in ref. 7). Nevertheless, the α globin genes can be expressed at high levels from a single copy of human chromosome 16 in an interspecific hybrid with a mouse erythroid cell background (7–9). Therefore, it appears that the transgenic constructs analysed to date may not span the entire chromosomal 'domain', defined here as a region containing all of the *cis*-acting elements required for fully regulated globin gene expression. Characterization of such domains will be of importance in understanding globin gene regulation and for understanding the general relationship between chromosome structure and function.

A comparison of the sequences spanning the human α cluster with the syntenic regions of other organisms might delimit such a chromosomal domain. Presumably critical *cis*-acting regulatory elements will have remained together throughout evolution or, alternatively, equivalent elements would have

[+]Present address: Clinical Genetics, NCI, Rockville, MD 20852, USA
[§]To whom correspondence should be addressed. Tel: +44 1865 222393; Fax: +44 1865 222500; Email: drhiggs@molbiol.ox.ac.uk

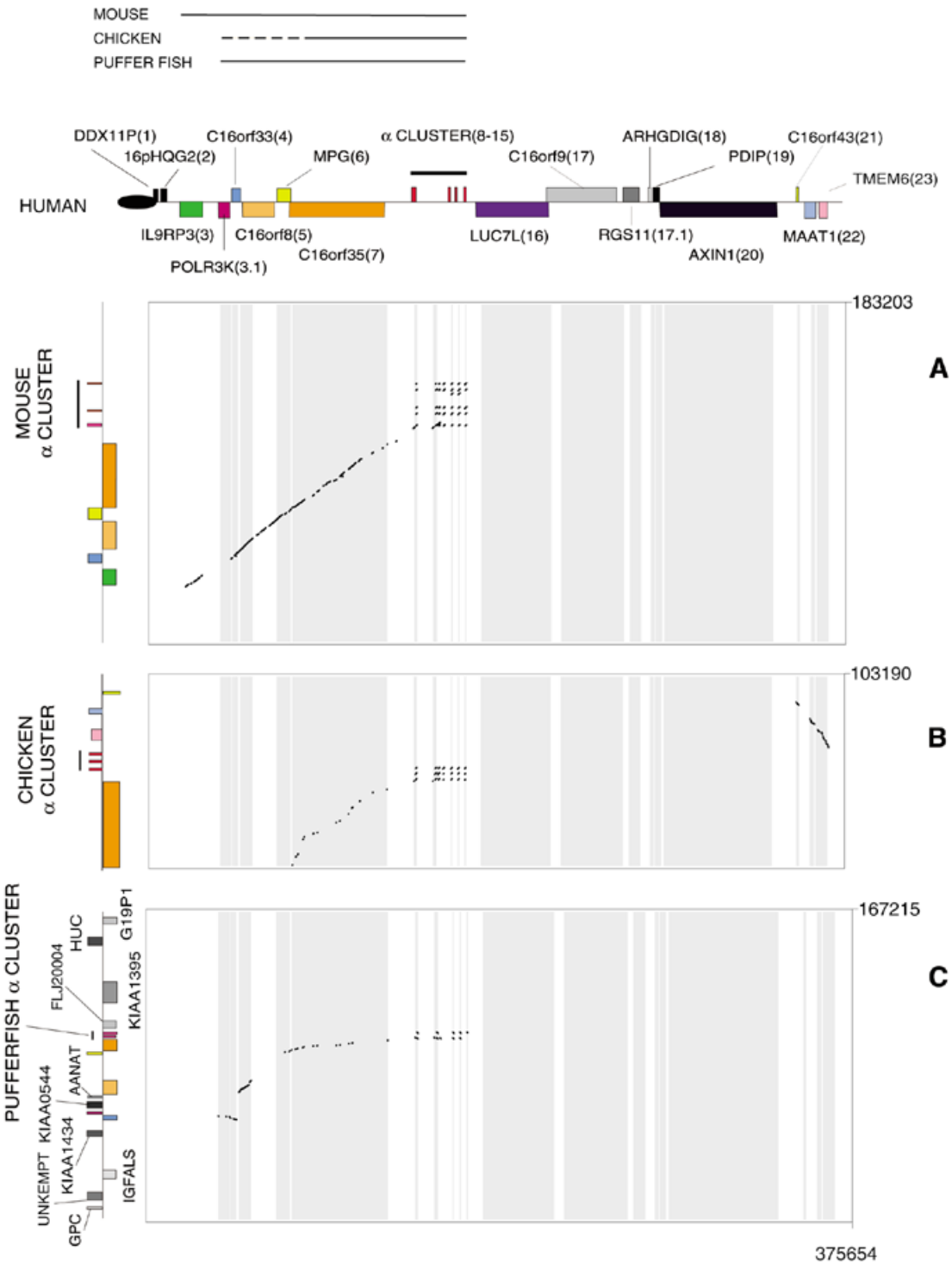

Figure 1. Comparison of the human (AE005175), mouse (AY016021 and AY016022), chicken (AY016020) and pufferfish (*S.nephelus* AY016023 and *F.rubripes* AY0160214) α globin clusters. (for further details see http://molbiol.ox.ac.uk/~haem/Syn/Fig1.jpg). The position and annotation of the human genes (Materials and Methods) are shown and summarized in Table 2. The direction of transcription towards the centromere (above the line) or telomere (below) is indicated. Orthologous genes in each species are represented in corresponding colours. Other non-syntenic genes are labelled. Comparative dotplots were created using the Pip analysis software (http://bio.cse.psu.edu/pipmaker) (27); (**A**) human/mouse, (**B**) human/chicken and (**C**) human/pufferfish. The extent of synteny of each cluster with respect to the human α globin cluster is represented by horizontal lines at the top of the figure. The scale in base pairs is indicated. The full extent of synteny in the chicken has not yet been determined (dashed line).

Table 1. Comparisons of the α-like globin gene clusters across the regions of synteny

	Length (bp)	GC % (average)[a]	Exons (%)[b]	Introns (%)[b]	All repeat types (%)	SINES (%)[b]	LINES (%)[b]
Human	152 040	54.5 (41.0)	8.88	49.65	36.1	22.9	5.7
Mouse	111 962	49.4 (44.4)	11.45	56.72	21.1	13	2.3
Chicken	53 599	47.6 (47.0)	5.37	67.31	2.1	1.45	0
Spheroides	32 054	49.1 (47.4)	18.19	24.12	1.7	0	0

[a]The average GC contents (in parentheses) were calculated from the analysis of current releases of human (1197 Mb), mouse (108 Mb), chicken (14 Mb) and pufferfish (22 Mb) sequences.
[b]The proportion of the structure compared with the total sequence analysed.

been recruited. Therefore, cross-species comparisons may also highlight the critical regulatory elements within such a region.

In this study, we have analysed contiguous segments of DNA spanning the α globin clusters of mouse, chicken and two species of pufferfish, enabling us to define a small segment of synteny and conserved gene order which has been maintained throughout evolution. In the human cluster, this conserved segment spans ~135–155 kb including the globin genes, their major regulatory element (αMRE or HS-40) and several widely expressed genes. When comparing the human and mouse sequences using previously defined routines (10–13), we did not clearly identify regulatory elements. However, by ranking sequence identities, the most conserved non-coding matches included the αMRE, the ζ globin promoter and a previously characterized, erythroid-specific DNAse1-hypersensitive site (HS-33): none of these elements could be identified by aligning human with chicken or pufferfish sequences. However, using prior knowledge of known erythroid-restricted transcription factor (TF) binding sites (14–16) we have shown that in chicken and pufferfish an α globin enhancer lies in a position corresponding to the αMRE of human and mouse. In addition to delimiting the α globin chromosomal domain, these findings highlight the potential value and limitations of comparative analysis for interpreting primary DNA sequence.

RESULTS

Overall comparison of the extended sequences

Clones containing orthologues of the human α-like globin genes were identified and contigs extending in either direction from these recombinants were assembled (Materials and Methods). For comparison we sequenced long segments of DNA spanning the mouse (*Mus musculus*; 183 kb), chicken (*Gallus gallus*; 103 kb) and pufferfish [*Spheroides nephelus*; 167 kb and *Fugu rubripes*; ~36 kb (data not shown)] α globin gene clusters and compared these with the previously analysed human [376 kb (2,17)] sequence (Fig. 1 and Table 1).

It has been shown previously that the α globin clusters of several species (e.g. man, rabbit and horse), lie in a telomeric position (18–20). In man, the α cluster commonly lies only 150 kb from the 16p telomere (allele A) (18). However, the mouse α cluster clearly lies at an interstitial position on chromosome 11 (21–23). The precise chromosomal location of chicken (linkage group E35; http://www.ri.bbsrc.ac.uk) is currently unknown. We found no telomeric or subtelomeric DNA repeats in the mouse, chicken or pufferfish α globin regions.

The GC content of the α clusters and their surrounding DNA is greater than the average for each respective species (Table 1). The α clusters lie in the most GC-rich isochores (24) of the human (H3, ~54% GC) and mouse (H2, ~50% GC) genomes. The chicken cluster also lies in a GC-rich isochore (H2, 50% GC) although isochores with higher GC content (H3 and H4) are found in this species (25). The pufferfish cluster has a similar GC content to those of the average segments of its genome. Comparisons of the proportions of each sequence representing exons, introns and repeat elements (Table 1 and Fig. 1) showed relative compaction of the α clusters in pufferfish and chicken due to fewer repeats (chicken and pufferfish) and smaller introns (pufferfish). In the pufferfish *S.nephelus*, the region of conserved synteny is compacted 5-fold, commensurate with the 7.5-fold smaller genome of the closely related fish *F.rubripes* (26).

Defining the segments of conserved synteny

We have previously reported the fully annotated sequence of the human α globin gene cluster (2,17) (summarized in Fig. 1 and Table 2). Using a similar approach (legend to Fig. 1) we initially analysed and annotated sequences from the other species. This showed that in addition to the α-like genes, several previously identified genes surrounding the human α cluster were also found flanking the α clusters of mouse, chicken and the pufferfish *S.nephelus* (Fig. 1 and Table 2).

Comparison of the human and mouse sequences using PipMaker (27) showed extensive matches (~155 kb in human and 112 kb in mouse) between the two clusters (Fig. 1A). Similarly, homology was seen comparing the chicken (Fig. 1B) and pufferfish (Fig. 1C) sequences with human, but at a lower percentage identity.

More detailed comparisons of the sequences (e.g. Fig. 2) showed that the clearest matches occurred in the syntenic, orthologous genes for which the number of coding exons, their sizes and sequences are well conserved (Fig. 2), as are their predicted proteins (data not shown). We also noted strong identity between the human and chicken sequences between coordinates 350859 and 352206 (in the human cluster). It seems most likely that this represents a new gene, although this was not predicted by any of the current programs and there are no expressed sequence tag (EST) hits associated with this putative gene. In summary, these analyses clearly demonstrated that sequence comparison between any two of the

Table 2. Summary of syntenic genes

Gene no.	Gene name	Known EST extent	Known or possible function	Mouse		Chicken		*Spheroides*	
1	*DDX11P*	1685–4086	Helicase pseudogene						
2	(16pHQG2)	4037–9529	Related to *CXYORF1*						
3	*IL9RP3*	17109–29489	Interleukin 9 receptor pseudogene	m3*	40171–30569				
3.1	*POLR3K*	36979–43634	RNA polymerase III subunit					s3.1	55865–56335
4	*C16orf33*	43871–47669	Unknown	m4	45575–48693			s4	55166–54356
5	*C16orf8*	48058–66371	Possible role in signal transduction	m5	56211–49759			s5	75541–69245
6	*MPG*	68247–75845	*N*-methylpurine-DNA glycosylase	m6	66660–71947			s6	90619–91649
7	*C16orf35*	74305–128837	Unknown	m7	107496–72907	c7	489–41783	s7	97493–92205
8	*HBZ*	142854–144504	Zeta globin (ζ)	m8	116561–117903	c8	47614–46183		
9	*HBZ'P*	152946–155254	Zeta globin pseudogene (ζψ)						
10	*HBAP2*	155997–156768	Alpha globin pseudogene (ψα2)						
11	*HBAP1*	158655–159453	Alpha globin pseudogene (ψα1)						
12	*HBA2*	162875–163709	Alpha globin 2 (α2)	m12 (5′α)	123603–124285	c12 (α^D)	49886–50692	s12 (A1)	98883–98247
13	*HBA1*	166674–167521	Alpha globin 1 (α1)	m13 (3′α)	136507–137189	c13(α^A)	53558–52881	s13 (A2)	102065–101295
14	*(ROP)*	168553–168654	HUMCRHY3 scRNA pseudogene						
15	*HBQ1*	170335–171177	Theta globin 1 (θ1)						
16	*LUC7L*	178971–219373	Possibly an RNA-binding protein						
17	*C16orf6*	222698–256658	Unknown						
17.1	*RGS11*	258849–266406	Regulator of G protein signalling						
18	*ARHGDIG*	271219–274238	Rho GDP-dissociation inhibitor (GDI)						
19	*PDIP*	273743–277753	Protein disulphide isomerase precursor						
20	*AXIN1*	277979–343000	Inhibits embryonic axis formation						
21	*C16orf43*	350849–352206				c21	87583–89164		
22	*MAAT1*	360509–358221	Melanoma antigen			c22	77373–80040		
23	*TMEM6*	361326–367701				c23	65741–72122		

species studied detected all of the genes including their alternatively spliced exons.

Defining the boundaries of conserved synteny

Closer inspection defined the 5′ breakpoints in homology between the α clusters of human and other species (Fig. 1). Comparing human and mouse, homology ends following the *IL-9* receptor (human gene 3), which is a functional gene in mouse (28) but a pseudogene in man (29). In addition, there is no match to *POLR3K* (human gene 3.1) in the mouse, suggesting that a further rearrangement within the region of synteny occurred after divergence of these species. In pufferfish the last point of homology extends to the end of *C16orf8* (human gene 5). Beyond this, orthologues for *POLR3K* and *C16orf33* (genes 3.1 and 4) are found, but rearranged with respect to the human cluster. At present there are insufficient sequence data to define the extent of homology upstream of the chicken cluster.

At the downstream end of the cluster (right hand boundary in Figs 1 and 2), the sequences of mouse, chicken and pufferfish diverge from the human sequence within a very narrow region between co-ordinates 171000 and 171186 of the human cluster (Fig. 1). Beyond this region, the sequences of each species appear quite different. In the mouse cluster, the only identifiable sequences lying adjacent to the α cluster consist of repetitive elements related to retroviral sequences. The orthologue of *LUC7L* (human gene 16), which lies immediately to the centromeric side of the human α cluster, is located on mouse chromosome 17 linked to a pseudoalpha gene but quite separate from the functional mouse α cluster on chromosome 11 (30). Downstream of the α genes in the pufferfish we identified a group of genes (FLJ20004, KIAA1395, HUC and G19P1) with homologues on human chromosome 19. In the

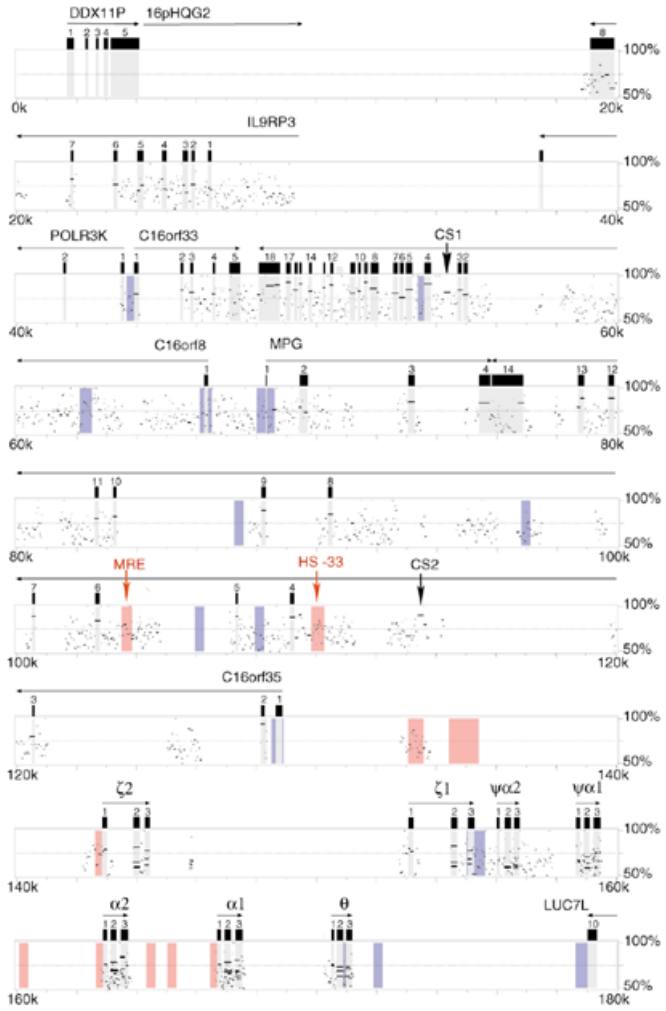

Figure 2. Detailed comparison of the human and mouse sequences generated using PipMaker (for further details see http://molbiol.ox.ac.uk/~haem/Syn/Fig2.jpg). Each gene is indicated with its extent and orientation shown by an arrow. Grey boxes represent exons, red boxes represent erythroid-specific HSs and blue boxes represent constitutive HSs. HS-33 is a previously described erythroid-specific HS and αMRE (HS-40) is the α globin MRE. The scale (kb) is shown. CS1 and CS2 indicate the positions of the two highly conserved sequences described in the text.

chicken there appears to have been an inversion, removing the orthologues of the human genes 16–19 and placing genes 23–20 in an inverted orientation next to the α genes (Fig. 1).

Despite the long period of evolution (~500 million years) throughout which these species have diverged from each other, the extant α globin clusters are arranged in a remarkably similar manner. In addition to the α genes themselves, several genes lying upstream of the clusters have been conserved. It seems likely that the boundaries of the α cluster, created independently in all four species, define the maximum limits of the chromosomal segment containing all of the *cis*-acting sequences required for fully regulated α gene expression.

Comparison of the four α globin clusters

To further our understanding of how the α globin gene cluster is normally regulated, we analysed the α-like genes in detail in each species (Fig. 3). Where known, the genes are arranged along each chromosome in the order in which they are expressed in development. Genes predominantly expressed in the embryonic period are found in human (ζ), mouse (ζ) and chicken (π). Similarly, genes expressed in the fetal/adult period are present in all species, interspersed with pseudogenes. As in all mammals studied to date (31–35 and R. Hardison, unpublished data), apparently functional θ genes are found in mouse and man. Although they could not carry oxygen (33) their role, if any, is currently unknown. At present, the pattern of expression of the pufferfish α-like globins is unknown. However, transcripts corresponding to A1 are represented multiple times in a *Fugu* cDNA database containing 3213 genes (http://fugu.hgmp.mrc.ac.uk), but no ESTs corresponding to A2 were found even though this appears to be a structurally normal gene. A2 is possibly expressed at a very low level or may be an embryonic gene. Another functional α-like gene (A3) is present in *F.rubripes* on another chromosome lying 2 kb from the *Fugu* β globin gene (T. McMorrow and S. Philipsen, unpublished data). It is interesting that the pufferfish α-like globin genes lie in the reverse orientation with respect to the linked non-globin genes compared with the human, mouse and chicken clusters (Fig. 3). It is also interesting that this is the first observation indicating a physical separation between the α and β clusters in fish or amphibians; in previous reports these genes are linked in these orders.

Evolution of CpG islands

We analysed the association of CpG islands with the α-like globin genes and the widely expressed genes flanking the cluster. In the human α cluster equally prominent CpG islands were associated with each functional α-like gene and with each of the widely expressed genes flanking the cluster (2) (Fig. 3). In the mouse, prominent CpG islands were seen associated with the widely expressed genes (e.g. the *C16orf35* gene in Figure 3 and data not shown) but only minor enrichment of CpG dinucleotides was found at the α-like genes themselves (Fig. 3). A very similar pattern was observed in the chicken cluster. However, no distinct CpG islands could be detected at either the globin genes or the widely expressed genes in the syntenic region of the pufferfish (Fig. 3 and data not shown) due to the high background of CpG dinucleotides found in fish (36).

A search for regulatory elements

It has been suggested that, in mammals, regulatory elements can be found by searching for ungapped alignments outside known exons (11–13,37). Most recently it has been suggested that regulatory elements should be found in non-coding regions with >70% identity over at least 100 bp (10). Comparing human and mouse sequences, two matches fulfilled these criteria (Fig. 2, CS1 and CS2). CS1 (co-ordinates 54332–54440) probably corresponds to an alternatively spliced exon of *C16orf8* (human gene 5) as we subsequently identified an EST sequence (HS804279) which matches CS1 and exons 1 and 2 of *DIST1*. The other match, CS2 (113455–113584), is a conserved sequence in intron 3 of *C16orf35* (human gene 7). Although this element does not contain an open reading frame it may contain part of a non-coding exon since, in both mouse and human, a portion of CS2 has been shown to match the 3′

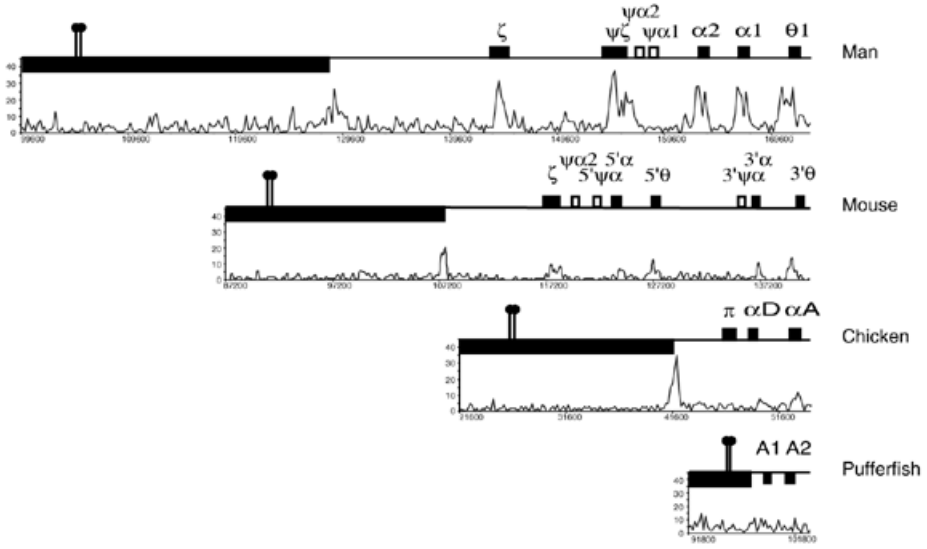

Figure 3. Structures of the α globin clusters of man, mouse, chicken and pufferfish (*S.nephelus*). Globin genes are shown as solid boxes and pseudogenes as open boxes. The long black boxes upstream of each cluster represent *C16orf35* and its orthologues. The direction of transcription is indicated by placing the gene above or below the line. The approximate positions of MARE sites are shown as stalks capped with filled circles. Below each cluster are plots of the CpG frequency in a 200 bp stepping window.

end of exon 1 of *C16orf35* (referred to as the *PROX* gene in ref. 38). It is also possible that this region plays a regulatory role, although it appears not to regulate expression of the α globin genes (7), does not coincide with any currently known HS and has no enhancer activity in transient assays. Therefore, in our initial search for regulatory elements we found only two potentially interesting sequences that probably correspond to alternatively spliced exons.

The criteria of Loots *et al.* (10) did not identify any of the previously characterized regulatory elements in the α cluster (e.g. the αMRE or the globin gene promoters). However, when gapped alignments between human and mouse were simply ranked by alignment scores (based on length and identity) the highest matches included CS1, CS2, αMRE, the promoters of the *ζ* and *ψζ* genes, HS-33 (a previously mapped erythroid-specific HS), a constitutive HS and only one unknown sequence (Table 3). To test this analysis further we analysed the human and mouse β globin clusters in the same way. Again, the highest matches identified most of the previously described *cis*-acting elements regulating the human β globin cluster (Table 3), validating this analytical routine.

At other loci, sequence comparisons of human with chicken (270 million years) and pufferfish (400 million years) DNA sequences have identified regulatory elements (39,40). For the α globin clusters, comparing chicken and pufferfish with human sequences showed homology in the exons of syntenic genes (data not shown). The only other match to the human sequence in either of these species coincided with CS1 (in pufferfish) and CS2 (chicken and pufferfish), which may represent exons of highly conserved genes around the α cluster. Therefore, preliminary comparisons of these sequences did not reveal candidates for regulatory elements.

Since none of the sequences associated with previously characterized regulatory elements or HSs appeared to be strongly conserved in chicken or pufferfish, we examined the distribution of binding sites for TFs known to play a role in globin gene regulation (41). Many of these sites (e.g. GATA binding sites) are very frequent in all sequences analysed here and were not enriched in known erythroid-specific HSs. In contrast, Maf recognition elements (MAREs; YGCTGASTCAY) (42) are relatively infrequent (~1 in 3000–8000 bp) and yet the HSs corresponding to the αMREs of man (HS-40) and mouse (HS-26), both located in intron 5 of *C16orf35*, each contain two MAREs positioned exactly 21 bp apart (Fig. 4). No other 'paired' sites, within 60 bp of each other, were found in human and only one other was found in mouse (coordinates 74892–74904).

We next searched the chicken sequence and the only paired MARE site was again found in intron 5 of the *C16orf35* gene. In this case, the sites were also 21 bp apart (Fig. 4). To analyse this further, we examined the position of HSs throughout the chicken cluster and found that the paired MAREs are associated with an HS in HD3 erythroid cells (Fig. 5). In pufferfish, one paired MARE site was found upstream of the *MPG* gene (gene *s6*). A second paired site was again found in intron 5 of the *C16orf35* gene (gene *s7*). In this paired site, the MARE motif (Fig. 5, right) lies in the opposite orientation to that observed in other species and is 43 rather than 21 bp distant from the second site. Closer inspection revealed a similar distribution of TF binding sites between all four species in this region (Fig. 4).

These findings strongly suggested that the paired MARE sites in intron 5 of the *C16orf35* gene in chicken and pufferfish (Fig. 3) may serve a similar role to that of the human and mouse αMREs. We therefore tested whether these regions, with minimal sequence identity but conserved TF binding sites, have maintained any functionally similar role.

Functional characterization of putative regulatory elements

It has previously been shown that the human and mouse αMREs behave as strong enhancers in erythroid cells (15,43–46).

Table 3. Highest scoring sequence pairs (HSPs) α globin cluster

Human		Mouse		Identity (%)	Length (bp)	Score	Annotation
Start	End	Start	End				
α globin cluster							
113388	113576	99056	99244	90	189	775	cs2
54268	54449	55466	55647	81	182	599	cs1
68517	68657	66694	66834	76	141	400	HS
103580	103664	91019	91103	80	85	272	HS-40
113588	113672	99252	99336	80	85	272	cs2
109705	109830	95941	96066	68	126	267	HS-33
109523	109611	95786	95874	76	89	253	HS-33
152943	153033	116373	116463	75	91	250	Promoter HBZ1
142729	142821	116371	116463	74	93	247	Promoter HBZ2
91131	91217	82271	82357	75	87	239	Unknown
β globin cluster							
41944	42071	37805	37932	83	128	444	HS2
52159	52327	46881	47049	61	169	251	Near HBE1
95280	95379	91577	91676	69	100	221	HBB promoter
44452	44604	40668	40820	60	153	214	Unknown
24861	24957	19190	19286	68	97	205	HS6
52704	52772	47443	47511	77	69	202	HBE1
48532	48643	45093	45204	62	112	176	Unknown
97933	97980	94288	94335	85	48	175	3′enh-b
31552	31649	21641	21738	64	98	172	3′HS5
37709	37793	31365	31449	66	85	164	HS3
55783	55837	55600	55654	76	55	156	Unknown
91799	91854	10494	10549	75	56	154	Unknown
97671	97773	94029	94131	61	103	153	Enhancer
91800	91853	95196	95249	76	54	153	Unknown
47763	47826	44684	44747	70	64	147	Unknown
47033	47105	41764	41836	66	73	141	HS1
37934	38007	31561	31634	65	74	136	3′HS3
31283	31334	21387	21438	69	52	114	HS5

Scores were calculated as described in Materials and Methods.

We used a transient transfection assay to investigate the effects of the putative regulatory elements from pufferfish and chicken on the chicken αA and pufferfish A2 promoter, respectively. Each promoter was inserted next to the renilla luciferase gene in the vector pRLnull and activity was measured in HD3 chicken cells. Results from the promoter construct were compared with those obtained when the putative MREs from either the pufferfish or chicken sequences were inserted immediately 5′ to their respective promoters (Fig. 6). We also investigated whether CS2 had any enhancer activity in these assays, again by inserting the chicken and *Spheroides* CS2 sequence next to their respective globin promoters. In each experiment, the putative regulatory elements were inserted in both orientations (Fig. 6).

The chicken sequence from intron 5 of the chicken orthologue of the *C16orf35* gene increased luciferase activity almost 20-fold over levels observed in the promoter alone. The equivalent pufferfish sequence also increased activity >20-fold. In both cases the magnitude of the effect was orientation dependent (Fig. 6). No effects were seen for CS2.

To determine whether the MAREs were critical for these enhancer effects we introduced mutations into each site (M1 and M2) and also constructed double mutants (M12). We found that, in the chicken, mutating the 5′ site reduces activity by a third and, in the pufferfish, to a sixth of wild-type levels. Mutating the second site has a slightly greater effect and the double mutant reduces expression still further. Although the double mutant still

Figure 4. Sequence alignment of the MRE. Consensus binding sites for MAREs were found in man (45 matches in 375 kb), mouse (30 matches in 183 kb), chicken (16 matches in 103 kb) and pufferfish (30 matches in 99 kb). However, two sites within 60 bp of each other (paired sites) were rare (human, 1; mouse, 2; chicken, 1; and pufferfish, 2) and in each species, one or the only paired site was located in intron 5 of the *C16orf35* orthologue. Alignment of these sequences between mouse and man showed significant matching but much less so between human and chicken and between human and pufferfish, where only the sites themselves appear to be conserved.

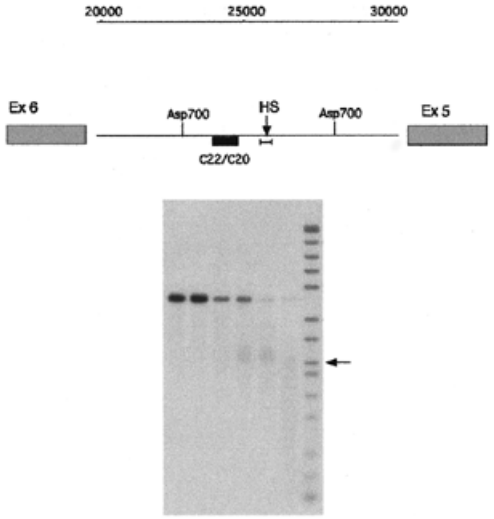

Figure 5. Analysis of DNAse1 HSs in the *CGTHBA* gene of the chicken cluster. The position of DNAse1 HS was mapped in HD3 (erythroid cells) and chicken embryonic fibroblasts. In the example shown, a prominent HS (horizontal arrow) was seen in HD3 cells following digestion with Asp700 and hybridization with the probe C22/20.

shows increased levels over the promoter construct, the MARE sites account for almost 90% of the enhancer activity.

Therefore, it appears that these poorly conserved elements, not readily identified by conventional sequence comparisons, may in fact underwrite some, or maybe all, of the role of the αMRE in these species.

DISCUSSION

Comparison of the α globin clusters of five distantly related species has shown that synteny and conserved gene order extend over a short distance (~135–155 kb). Inspection of the boundaries of this conserved chromosomal unit suggests that although the breakpoints lie in very similar locations, they have been defined by quite independent evolutionary events in each species. Whatever the mechanism(s), this is of considerable practical interest, as these breakpoints may delimit a chromosomal segment containing all of the critical *cis*-acting elements required for fully regulated expression of the α-like globin genes. It is interesting that even the 135 kb region representing the human/pufferfish synteny contains all the previously mapped erythroid-specific HSs around the human α globin cluster and broadly corresponds to a segment of chromatin which becomes hyperacetylated when the α genes are fully active in erythroid cells (E. Anguita *et al.*, in preparation). At present, none of the fragments containing the α cluster that have been tested in transgenic mice span this entire region. Since all of these constructs are expressed suboptimally (7), it remains to be seen whether a fragment containing this entire region would be fully regulated in transgenic mice. In general, these findings suggest that cross-species comparisons may address the important question of whether chromosomes are organized into discrete structural and functional domains.

Current software is capable of finding coding regions of the human genome. Programs for identifying regulatory elements adjacent to the genes have been less refined, nevertheless primary sequence analysis readily detects CpG-rich islands which associate with the promoters of virtually all housekeeping genes and approximately half of all tissue-specific genes (47,48). Since CpG islands co-localize with the promoters of genes, they may be involved in regulating their transcription (49) and specifically in influencing expression of the α-like globin genes (3). Although prominent CpG islands are associated with each of the widely expressed genes flanking the α clusters (human, mouse and chicken), the CpG islands associated with the α-like genes themselves have been substantially eroded during the evolution of mouse and chicken, demonstrating that they are not absolutely required for globin gene regulation in these species. Similar erosion of the CpG islands associated with tissue-specific genes has been noted before (49,50) and, therefore, this cross-species comparison confirms that CpG islands may be less useful for identifying tissue-specific genes in these species and adds to the evidence that CpG islands are not fundamentally important in regulating gene expression but reflect some other aspect(s) of genome structure, function or evolution as discussed by Antequera and Bird (49).

To annotate the human genome in full, it will be important to develop routines for identifying the key regulatory elements (e.g. promoters, enhancers and silencers) from the DNA sequence alone. Ultimately one might hope to identify other chromosomal elements; e.g. origins of replication, boundary elements and signals for nuclear sub-compartmentalization. At

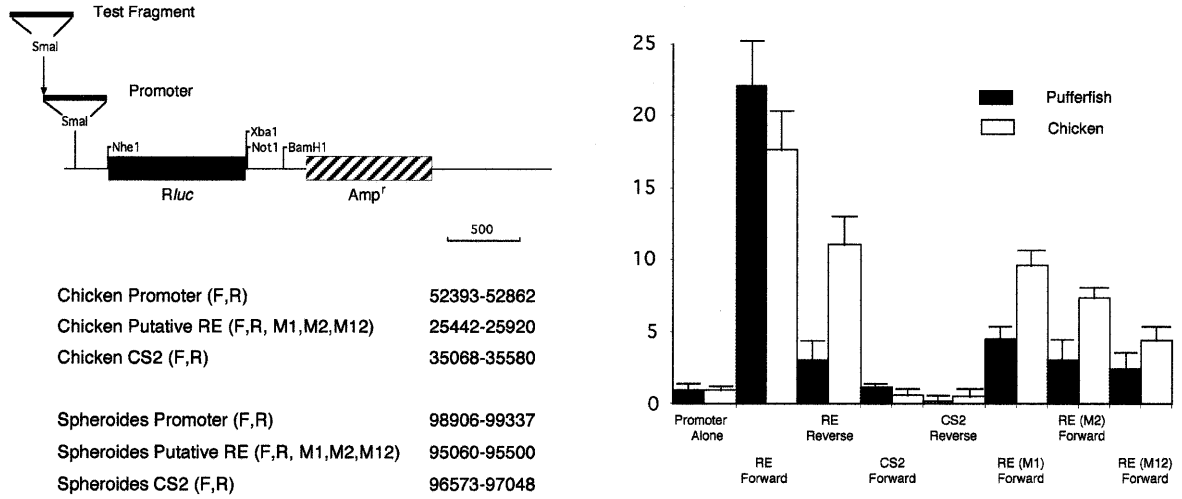

Figure 6. Construction and analysis of vectors containing putative regulatory elements (Materials and Methods). Normalized data, showing the relative levels of expression of each construct. Vertical lines represent 1 SD.

present it is not clear to what extent cross-species sequence comparisons can achieve this since, to date, very few studies have identified such elements by sequence analysis alone. Because many of the *bona fide* elements regulating α globin expression *in vivo* in man and mouse have been previously characterized (reviewed in refs 7,44 and 45), this study allowed us to test the ability of current routines to identify such elements. Using criteria set out by others (10) we were unable to distinguish even the most important regulatory elements (the globin promoters and MREs) of the human or mouse α globin clusters from many other non-coding regions. However, simply ranking the identity scores of conserved non-coding sequences identified many of the known regulatory elements in the α globin cluster. Similarly, using this approach to analyse the human and mouse β globin gene clusters we identified most of the known regulatory elements (Table 3), whereas previously defined criteria (10) identified only HS2 (51). Therefore, rather than setting fixed, arbitrary criteria, which may differ from one comparison to another, identifying the most highly scoring non-coding matches between two species, as set out here, may provide the most efficient guide to regulatory elements. This simple approach may be made even more informative by parallel analysis of several closely related species.

Although this approach was not informative in more distantly related species, using prior knowledge of the TF binding sites in the human and mouse αMREs it was possible to identify enhancer elements at a conserved position (intron 5 of *C16orf35*) in chicken and pufferfish. Although it is not yet clear whether these enhancers are functionally equivalent to the αMREs, our data strongly indicate that this type of regulatory element was present in this position in the ancestral α globin cluster. These findings suggest that future cross-species searches for regulatory elements might also include routines searching for significant conservation of motifs corresponding to chosen TF binding sites of interest separated by non-conserved gaps.

MATERIALS AND METHODS

Construction of contigs

The human contig has been published previously (2,17). In mouse we started with the contig containing 2CF2, P3 and P48 (52). Gaps in this contig were closed with F1 827 and the contig was extended with P2 and o05238 (P1 clones from the ICRF P703 mouse P1 library). In chicken, we started with λG2, λG5 and λG7 (53,54). This was extended with D18260 and D0242 (Library 125, Resourcenzentrum, Germany). A cosmid library derived from *F.rubripes* genomic DNA was constructed in the pTCF cosmid vector (55). Cosmids A and F representing the *F.rubripes* α globin locus were identified by screening this library with an A1 globin cDNA isolated from a cDNA library from peripheral blood. Subsequently, using probe 129 (coordinates 34062–34626 of the *Fugu* sequence) we isolated the P1s PACFF-74p9, -56M21 and -51D12 from the *S.nephelus* library (Genome Systems).

Isolation of probes

Human (6,56) and mouse (45,57) probes were as previously described or were amplified from primers (available on request) designed from the sequence. Probes to the chicken (C22/20), *Fugu* (129) and *nephelus* clusters were amplified using primers (available on request) designed from the sequence. Probes corresponding to A1 and β globin (*Fugu*) were isolated from cDNA made from peripheral blood RNA from a 3- to 4-month-old fish screened with salmon globin cDNA probes.

Sequence analysis

Recombinants were sequenced as previously described (2). Some data were previously available for mouse (21,22,38,44, 45,58,59) and chicken (53,54,60,61) but new sequence contigs were assembled without reference to these. Each newly assembled sequence was masked for repeats and initially annotated by sequence homology using the BLAST suite (62) to search

nucleotide [dbEST (63) and EMBL (64)] and protein [SwissProt and Trembl (65)] sequence databases and our unpublished cDNA sequences. In addition, sequences were analysed with the gene prediction programs GRAIL and GENSCAN (66,67). All sequences and analyses were processed using an automated annotation system and stored in ACEDB (http://www.acedb.org/). Sequences were compared by sequence identity using PipMaker (http://bio.cse.psu.edu/pipmaker). G+C% and the frequency of CpG dinucleotides plots were calculated using a 200 bp stepping window. Putative regulatory elements were identified by using BLASTZ (an integral part of PipMaker) to align genomic sequences that had been masked for known exons and repeats. Localized regions of sequence conservation, i.e. ungapped high scoring pairs (HSPs), were returned by PipMaker using the 'concise text' output option. A score for each ungapped HSP was calculated (+5 for each identical base and –4 for each mismatch) and used to rank the HSPs. Specific TF binding sites were identified using MacVector software and the Transfac database (http://transfac.gbf-braunschweig.de/TRANSFAC/index.html).

Identification of DNAse1-hypersensitive sites

Nuclei were prepared from 1×10^8 chicken HD3 (erythroid) cells and DNAse1-hypersensitive sites were identified as previously described by Higgs *et al.* (4).

Plasmid constructs

A 500 bp fragment containing the chicken αA-globin promoter (coordinates 52393–52862) was amplified using the primers 5′-CCTTCCCGGGGGTTGCACCTCTGTGTT-3′ and 5′-GTGCTCCTGAACCTACAG-3′. The PCR product was kinased and ligated into the *Sma*I site of the pRL-NULL plasmid (Promega); the orientation and sequence were confirmed. Similarly, a 500 bp fragment immediately 5′ of the pufferfish A2 α globin gene was amplified and cloned (primers 5′-CCTTCCCGGGGCCAAACTGGTGCTCAATTT-3′ and 5′-TTGTTGTTGGGGGTGTTTTT-3′; coordinates 988906–99337). Primers at the 5′ end of the insert (with respect to the luciferase gene) contain an *Sma*I site allowing the subsequent insertion of the following PCR products: the putative chicken RE (primers 5′-TTGTTGTTGGGGGTGTTTTT-3′ and 5′-TTGTTGTTGGGGGTGTTTTT-3′), chicken CS2 (primers 5′-GAACTGAAATGCCACCAACC-3′ and 5′-CCACACTCATTCTGGTTACCC-3′), putative pufferfish RE (primers 5′-CTGTAGAAAGTGCTTAGAAGTGAA-3′ and 5′-GCAAAGTAGTCTTCTTTACATTTTT-3′) and pufferfish CS2 (primers 5′-TTTGCAGCACGTTTATCCAA-3′ and 5′-GATTCACAGAATCCGCTGGT-3′). Mutations were introduced into the putative MRE by PCR-based, site-directed mutagenesis (QuikChange; Stratagene). All modifications were confirmed by sequence analysis. All PCRs were performed using Pfu DNA polymerase (Stratagene).

Transient transfection and functional assays

Chicken HD3 cells were maintained in Dulbecco's modified Eagle's medium with 10% fetal calf serum transfected by electroporation using a 950 μF capacitor array charged at 300 V. For each transfection, 1×10^6 cells were resuspended in 0.25 ml of RPMI 1640 with 10 μg of a test construct in pRL-NULL and 5 μg of a transfection control plasmid pCMVβGal. After discharge, cells were left on ice for 10 min before resuspending in growth medium and incubating at 37°C for 2 days. Subsequently, luciferase activities were determined according to the manufacturer's instructions (Promega). Relative β galactosidase activity in lysates was measured using *O*-nitrophenyl-D-galactopyranoside (0.67 mg/ml) as substrate in a 0.1 M phosphate buffer pH 7.0 containing 10 mM KCl and 1 mM $MgSO_4$ incubated at 37°C for 15–45 min. The A420 was determined after stopping the reaction by the addition of 0.3 M sodium carbonate. All constructs were tested in triplicate in at least three independent transfection experiments.

ACKNOWLEDGEMENTS

We are grateful to Professor Sir D.J. Weatherall for continued support and encouragement. We are grateful to Dr G. Elgar for helpful advice. We thank Drs W.G. Wood, R.J. Gibbons and V. Buckle for their comments on the manuscript. We are grateful to Julianne Ellu for providing HD3 cells. We also thank H. Ayyub for excellent technical assistance. Mouse cosmids were provided by Dr Q. Zhao and chicken recombinants were provided by Dr J. Dodgson. J. Flint and J. Frampton are Wellcome Trust Senior Fellows. B.A. was supported in part by the National Institutes of Health.

REFERENCES

1. Vukmirovic, O.G. and Tilghman, S.M. (2000) Exploring genome space. *Nature*, **405**, 820–822.
2. Flint, J., Thomas, K., Micklem, G., Raynham, H., Clark, K., Doggett, N.A., King, A. and Higgs, D.R. (1997) The relationship between chromosome structure and function at a human telomeric region. *Nature Genet.*, **15**, 252–257.
3. Bird, A.P., Taggart, M.H., Nicholls, R.D. and Higgs, D.R. (1987) Non-methylated CpG-rich islands at the human α-globin locus: implications for evolution of the α-globin pseudogene. *EMBO J.*, **6**, 999–1004.
4. Higgs, D.R., Wood, W.G., Jarman, A.P., Sharpe, J., Lida, J., Pretorius, I.-M. and Ayyub, H. (1990) A major positive regulatory region located far upstream of the human α-globin gene locus. *Genes Dev.*, **4**, 1588–1601.
5. Smith, Z.E. and Higgs, D.R. (1999) The pattern of replication at a human telomeric region (16p13.3): its relationship to chromosome structure and gene expression. *Hum. Mol. Genet.*, **8**, 1373–1386.
6. Vyas, P., Vickers, M.A., Simmons, D.L., Ayyub, H., Craddock, C.F. and Higgs, D.R. (1992) *Cis*-acting sequences regulating expression of the human α globin cluster lie within constitutively open chromatin. *Cell*, **69**, 781–793.
7. Higgs, D.R., Sharpe, J.A. and Wood, W.G. (1998) Understanding α globin gene expression: a step towards effective gene therapy. *Semin. Hematol.*, **35**, 93–104.
8. Deisseroth, A. and Hendrick, D. (1978) Human α-globin gene expression following chromosomal dependent gene transfer into mouse erythroleukemia cells. *Cell*, **15**, 55–63.
9. Zeitlin, H.C. and Weatherall, D.J. (1983) Selective expression within the α-globin gene complex following chromosome dependant transfer into diploid mouse erythroleukaemia cells. *Mol. Biol. Med.*, **1**, 489.
10. Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M. and Frazer, K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288**, 136–140.
11. Popperl, H., Bienz, M., Studer, M., Chan, S.K., Aparicio, S., Brenner, S., Mann, R.S. and Krumlauf, R. (1995) Segmental expression of Hoxb-1 is controlled by a highly conserved autoregulatory loop dependent upon exd/pbx. *Cell*, **81**, 1031–1042.
12. Hardison, R.C., Oeltjen, J. and Miller, W. (1997) Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.*, **7**, 959–966.

13. Lamerdin, J.E., Montgomery, M.A., Stilwagen, S.A., Scheidecker, L.K., Tebbs, R.S., Brookman, K.W., Thompson, L.H. and Carrano, A.V. (1995) Genomic sequence comparison of the human and mouse XRCC1 DNA repair gene regions. *Genomics*, **25**, 547–554.
14. Strauss, E.C., Andrews, N.C., Higgs, D.R. and Orkin, S.H. (1992) *In vivo* footprinting of the human α-globin locus upstream regulatory element by guanine/adenine ligation-mediated PCR. *Mol. Cell. Biol.*, **12**, 2135–2142.
15. Jarman, A.P., Wood, W.G., Sharpe, J.A., Gourdon, G., Ayyub, H. and Higgs, D.R. (1991) Characterization of the major regulatory element upstream of the human α-globin gene cluster. *Mol. Cell. Biol.*, **11**, 4679–4689.
16. Rombel, I., Hu, K.-Y., Zhang, Q., Papayannopoulou, T., Stamatoyannopoulos, G. and Shen, C.-K.J. (1995) Transcriptional activation of human adult α-globin genes by hypersensitive site-40 enhancer: function of nuclear factor-binding motifs occupied in erythroid cells. *Proc. Natl Acad. Sci. USA*, **92**, 6454–6458.
17. Horsley, S.W., Daniels, R.J., Anguita, E., Raynham, H.A., Peden, J.F., Villegas, A., Vickers, M.A., Green, S., Chui, D.H.K., Ayyub, H. *et al.* (2001) Monosomy for the most telomeric, gene-rich region of human chromosome 16p causes minimal phenotypic effects. *Eur. J. Hum. Genet.*, in press.
18. Wilkie, A.O.M., Higgs, D.R., Rack, K.A., Buckle, V.J., Spurr, N.K., Fischel-Ghodsian, N., Ceccherini, I., Brown, W.R.A. and Harris, P.C. (1991) Stable length polymorphism of up to 260 kb at the tip of the short arm of human chromosome 16. *Cell*, **64**, 595–606.
19. Xu, J. and Hardison, R.C. (1991) Localization of the α-like globin gene cluster to region q12 of rabbit chromosome 6 by *in situ* hybridization. *Genomics*, **9**, 362–365.
20. Oakenful, E.A., Buckle, V.J. and Clegg, J.B. (1993) Localization of the horse (*Equus caballus*) α-globin gene complex to chromosome 13 by fluorescence *in situ* hybridization. *Cytogenet. Cell Genet.*, **62**, 136–138.
21. Leder, A., Swan, D., Ruddle, F., D'Eustachio, P. and Leder, P. (1981) Dispersion of α-like globin genes of the mouse to three different chromosomes. *Nature*, **293**, 196–200.
22. Leder, A., Weir, L. and Leder, P. (1985) Characterization, expression and evolution of the mouse embryonic ζ-globin gene. *Mol. Cell. Biol.*, **5**, 1025–1033.
23. Tan, H. and Whitney, J.B.I. (1993) Genomic rearrangement of the α-globin gene complex during mammalian evolution. *Biochem. Genet.*, **31**, 473–484.
24. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953–957.
25. Bernardi, G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene*, **241**, 3–17.
26. Venkatesh, B., Gilligan, P. and Brenner, S. (2000) Fugu: a compact vertebrate reference genome. *FEBS Lett.*, **476**, 3–7.
27. Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. and Miller, W. (2000) PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.*, **10**, 577–586.
28. Vermeesch, J.R., Petit, P., Kermouni, A., Renauld, J.-C., Van Den Berghe, H. and Marynen, P. (1997) The IL-9 receptor gene, located in the Xq/Yq pseudoautosomal region, has an autosomal origin, escapes X inactivation and is expressed from the Y. *Hum. Mol. Genet.*, **6**, 1–8.
29. Kermouni, A., Van Roost, E., Arden, K.C., Vermeesch, J.R., Weiss, S., Godelaine, D., Flint, J., Lurquin, C., Szikora, J.-P., Higgs, D.R. *et al.* (1995) The IL-9 receptor gene (IL9R): genomic structure, chromosomal localization in the pseudoautosomal region of the long arm of the sex chromosomes, and identification of IL9R pseudogenes at 9qter, 10pter, 16pter and 18pter. *Genomics*, **29**, 371–382.
30. Tufarelli, C., Frischauf, A.-M., Hardison, R., Flint, J. and Higgs, D.R. (2001) Characterization of a widely expressed gene (*LUC7-LIKE*) defining the centromeric boundary of the human α globin domain. *Genomics*, in press.
31. Hsu, S.-L., Marks, J., Shaw, J.-P., Tam, M., Higgs, D.R., Shen, C.C. and Shen, C.-K.J. (1988) Structure and expression of the human θ_1 globin gene. *Nature*, **331**, 94–96.
32. Shaw, J.-P., Marks, J. and Shen, C.-K.J. (1987) Evidence that the recently discovered θ1-globin gene is functional in higher primates. *Nature*, **326**, 717–720.
33. Clegg, J.B. (1987) Can the product of the θ-gene be a real globin? *Nature*, **329**, 465.
34. Marks, J., Shaw, J.-P. and Shen, C.-K.J. (1986) Sequence organization and genomic complexity of primate θ1 globin gene, a novel α-globin-like gene. *Nature*, **321**, 785–788.
35. Hardison, R. (2001) Organization, evolution and regulation of the globin genes. In Steinberg, M.H., Forget, B.G., Higgs, D.R. and Nagel, R.L. (eds), *Disorders of Hemoglobin: Genetics, Pathophysiology and Clinical Management.* Cambridge University Press, Cambridge, MA.
36. Elgar, G., Clark, M.S., Meek, S., Smith, S., Warner, S., Edwards, Y.J.K., Bouchireb, N., Cottage, A., Yeo, G.S.H., Umrania, Y. *et al.* (1999) Generation and analysis of 25 Mb of genomic DNA from the pufferfish *Fugu rubripes* by sequence scanning. *Genome Res.*, **9**, 960–971.
37. Oeltjen, J.C., Malley, T.M., Muzny, D.M., Miller, W., Gibbs, R.A. and Belmont, J.W. (1997) Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.*, **7**, 315–329.
38. Kielman, M.F., Barradeau, S., Smits, R., Harteveld, C.L. and Bernini, L.F. (1996) Characterization and localization of the mProx1 gene directly upstream of the mouse alpha-globin gene cluster: identification of a polymorphic direct repeat in the 5′UTR. *Mamm. Genome*, **7**, 877–880.
39. Göttgens, B., Barton, L.M., Gilbert, J.G.R., Bench, A.J., Sanchez, M.-J., Bahn, S., Mistry, S., Grafham, D., McMurray, A., Vaudin, M. *et al.* (2000) Analysis of vertebrate *SCL* loci identifies conserved enhancers. *Nature Biotechnol.*, **18**, 181–186.
40. Aparicio, S., Morrison, A., Gould, A., Gilthorpe, J., Chaudhuri, C., Rigby, P., Krumlauf, R. and Brenner, S. (1995) Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc. Natl Acad. Sci. USA*, **92**, 1684–1688.
41. Orkin, S.H. (1995) Hematopoiesis: how does it happen? *Curr. Opin. Cell Biol.*, **7**, 870–877.
42. Andrews, N.C., Erdjument-Bromage, H., Davidson, M.B., Tempst, P. and Orkin, S.H. (1993) Erythroid transcription factor NF-E2 is a haematopoietic-specific basic-leucine zipper protein. *Nature*, **362**, 722–728.
43. Zhang, Q., Reddy, P.M.S., Yu, C.-Y., Bastiani, C., Higgs, D., Stamatoyannopoulos, G., Papayannopoulou, T. and Shen, C.-K.J. (1993) Transcriptional activation of human ζ2 globin promoter by the α globin regulatory element (HS-40): functional role of specific nuclear factor-DNA complexes. *Mol. Cell. Biol.*, **13**, 2298–2308.
44. Gourdon, G., Sharpe, J.A., Higgs, D.R. and Wood, W.G. (1995) The mouse α-globin locus regulatory element. *Blood*, **86**, 766–775.
45. Kielman, M.F., Smits, R. and Bernini, L.F. (1994) Localization and characterization of the mouse α-globin locus control region. *Genomics*, **21**, 431–433.
46. Pondel, M.D., George, M. and Proudfoot, N.J. (1992) The LCR-like α-globin positive regulatory element functions as an enhancer in transiently transfected cells during erythroid differentiation. *Nucleic Acids Res.*, **20**, 237–243.
47. Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.
48. Cross, S.H. and Bird, A.P. (1995) CpG islands and genes. *Curr. Opin. Genet. Dev.*, **5**, 309–314.
49. Antequera, F. and Bird, A. (1999) CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr. Biol.*, **9**, R661–R667.
50. Matsuo, K., Clay, O., Takahashi, T., Silke, J. and Schaffner, W. (1993) Evidence for erosion of mouse CpG islands during mammalian evolution. *Somat. Cell Mol. Genet.*, **19**, 543–555.
51. Hardison, R.C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.*, **16**, 369–372.
52. Zhao, Q.Z., Liang, X.L., Mitra, S., Gourdon, G. and Alter, B.P. (1996) Cloning and characterization of the mouse alpha globin cluster and a new hypervariable marker. *Mamm. Genome*, **7**, 749–753.
53. Engel, J.D. and Dodgson, J.B. (1980) Analysis of the closely linked adult chicken α-globin genes in recombinant DNAs. *Proc. Natl Acad. Sci. USA*, **77**, 2596–2600.
54. Dodgson, J.B. and Engel, J.D. (1983) The nucleotide sequence of the adult chicken alpha-globin genes. *J. Biol. Chem.*, **258**, 4623–4629.
55. Grosveld, F.G., Lund, T., Murray, E.J., Mellor, A.L., Dahl, H.H. and Flavell, R.A. (1982) The construction of cosmid libraries which can be used to transform eukaryotic cells. *Nucleic Acids Res.*, **10**, 6715–6732.
56. Nicholls, R.D., Fischel-Ghodsian, N. and Higgs, D.R. (1987) Recombination at the human α-globin gene cluster: sequence features and topological constraints. *Cell*, **49**, 369–378.
57. Holmquist, G.P. (1992) Chromosomal bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet.*, **51**, 17–37.
58. Kielman, M.F., Smits, R., Hof, I. and Bernini, L.F. (1996) Characterization and comparison of the human and mouse *Dist1/a-globin* complex reveals a tightly packed multiple gene cluster containing differentially expressed transcription units. *Genomics*, **32**, 341–351.

59. Nishioka, Y. and Leder, P. (1979) The complete sequence of a chromosomal mouse α-globin gene reveals elements conserved throughout vertebrate evolution. *Cell*, **18**, 875–882.
60. Engel, J.D., Rusling, D.J., McCune, K.C. and Dodgson, J.B. (1983) Unusual structure of the chicken embryonic α-globin gene, π'. *Proc. Natl Acad. Sci. USA*, **80**, 1392–1396.
61. Sjakste, N., Iarovaia, O.V., Razin, S.V., Linares-Cruz, G., Sjakste, T., Le Gac, V., Zhao, Z. and Scherrer, K. (2000) A novel gene is transcribed in the chicken α-globin gene domain in the direction opposite to the globin genes. *Mol. Gen. Genet.*, **262**, 1012–1021.
62. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
63. Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nature Genet.*, **4**, 332–333.
64. Baker, W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G. and Tuli, M.A. (2000) The EMBO nucleotide sequence database. *Nucleic Acids Res.*, **28**, 19–23.
65. Bairoch, A. and Apweiler, R. (2000) The Swiss-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
66. Xu, Y., Mural, R., Shah, M. and Uberbacher, E. (1994) Recognizing exons in genomic sequence using GRAIL II. *Genet. Eng.*, **16**, 241–253.
67. Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.