Journal of Cellular Biochemistry 52:23–36 (1993)

Homeodomain Protein Binding Sites, Inverted Repeats, and Nuclear Matrix Attachment Regions Along the Human β-Globin Gene Complex

Teni Boulikas

Institute of Molecular Medical Sciences, Palo Alto, California 94306

Abstract β-Globin genes in primates arose during evolution by duplication of an ancestral gene, and their order of arrangement along the DNA is related to their timing of expression during development. We believe that nuclear matrix anchorage sites (MARs) along the β-globin gene complex considered to be mass binding sites for transcription protein factors, some of which are developmental stage specific and others ubiquitous, play a decisive role in cell memory by determining the developmental stage-specific expression of the genes. The AT-rich class of MARs appears to possess a significant number of ATTA and ATTTA motifs known to be mass binding sites for homeodomain proteins that determine body formation in development. MARs also appear to harbor origins of replication, to be enriched in inverted repeats (dyad symmetry motifs) and were proposed to include the DNase I hypersensitive sites of a particular gene determined at the chromatin level. This study is an attempt to finely identify MARs at the nucleotide level along the β-globin gene complex. Searches of a contiguous stretch of about 73.3 kb of human sequences comprising and surrounding the ε-, γ^G-, γ^A-, δ-, and β-globin genes of the human β-globin gene complex for homeotic protein binding sites as well as for inverted repeats has shown that these elements are clustered nonrandomly at particular sites within the β-globin gene complex. These sites are presumed to be the AT-rich class of MARs of the β-globin gene complex. The inverted repeats which are characteristic of origins of replication and some promoter/enhancer regions and the homeotic protein sites seem to include the DNase I hypersensitive sites of the gene complex. Indeed, dyad symmetry sequences are present close to the four DNase I HS sites in the locus control region (LCR) of the gene complex as well as in the 5′ flanking regions and the large introns of the δ- and β-globin genes. A search of the putative MAR regions of the gene complex suggests that, in addition to their enrichment in ATTA motifs, palindromes, and DNase I hypersensitive sites, these regions may comprise TG-rich motifs and potential Z-DNA as well as polypurine and polypyrimidine blocks.

From the positions of palindromes and clusters of homeodomain protein sites along the complex we propose that an extended origin of replication able to initiate at several sites is present in the LCR and two others surrounding the δ- and β-globin genes. Furthermore, we propose that DNase I HS sites, potential Z-DNA, polypurine and polypyrimide stretches, TG-boxes, homeodomain protein sites, and dyad symmetry motifs may be features diagnostic of MARs. This analysis supports a model which predicts that facultative matrix anchorage sites containing homeotic and other transcription protein factor binding sites might anchor the ε-, γ-, and β-globin genes in embryonic, fetal, and adult tissue, respectively, and might thus regulate the ordered developmental expression of the genes in the β-globin gene complex.

Key words: nuclear matrix, β-globin gene, homeotic protein site, palindromes, origins of replication

Scaffold attachment regions (SARs) along genomic DNA were first observed by Paulson and Laemmli [1977] on electron micrographs of mitotic chromosomes depleted of their histones. Interphase chromatin, like mitotic chromosomes, is also organized into loops or domains [see Hancock and Boulikas, 1982; Berezney, 1991]. This organization is brought about by the anchorage of specific DNA sequence landmarks termed MARs (matrix-attached regions) to a network of protein cross-ties of the nuclear ma-

Abbreviations used: bp, base pairs; DNase I HS site, DNase I hypersensitive site; HOMEO, homeotic protein binding site; HSV, Herpes Simplex Virus; kb, kilobase pairs; LCR, locus control region; MAR, matrix attachment region; nt, nucleotides; ORI, origin of replication; 3′ UTR, 3′ untranslated region.

Received December 21, 1992; accepted December 24, 1992.

Address reprint requests to Teni Boulikas, Ph.D., Institute of Molecular Medical Sciences, 460 Page Mill Road, Palo Alto, CA 94306.

trix [Mirkovitch et al., 1984, 1988; Cockerill and Garrard, 1986]. MARs are believed to facilitate the proper expression and replication of DNA during the cell cycle, to govern the cell type-specific expression of genes and the activation of different origins of replication during development, to define the borders between vicinal chromatin domains, and to play a major role in DNA recombination [Dworetzky et al., 1990; see Bodnar, 1988; Fey et al., 1991; Villarreal, 1991; Boulikas, 1992a,b].

MARs may be located in the vicinity of transcriptional enhancers [Cockerill and Garrard, 1986] and may display enhancer activity in transfection and transgenic experiments [Xu et al., 1989; Blasquez et al., 1989a; Stief et al., 1989; Phi-Van et al., 1990; Klehr et al., 1991]. Ends of transcription units are flanked, in some cases, by a MAR [Mirkovitch et al., 1984; Dijkwel and Hamlin, 1988; Phi-Van and Strätling, 1988]. This MAR, as a permanent component of the nuclear matrix, is supposed to bring the promoter region of its gene in contact with transcription factors that are linked with the MAR element on the nuclear matrix, causing looping out of the DNA between the MAR and the 5′ end of the gene [Boulikas, 1992a]. This makes it likely that MARs have a functional role. In addition, MARs harbor origins of replication (ORIs) (see Discussion). The differential activation of ORIs during development might be responsible for establishing cell memory (i.e., the cell types of the body and the differential gene expression) [Spradling and Orr-Weaver, 1987]. In agreement with this idea, active genes depend on a 5′ flanking ORI for their replication, whereas inactive genes are believed to be transcribed from an ORI to the 3′ end of the gene [Leffak and James, 1989]. Our view of the nuclear matrix DNA (MARs) is that of a mass binding site of transcription and replication factors; indeed, nuclear matrix is the principal site of transcription, replication, repair, and recombination [Berezney, 1991; Boulikas, 1987, 1992a–d].

β-Globin genes seem to have arisen by duplication of a single ancestral β-globin gene about 200 million years ago. The 5′ member of these duplicated genes is the ancestor of the ε and γ genes and the 3′ member is the ancestor of the δ and β genes. The ε and γ genes, which separated about 120 million years ago, and the δ and β genes, which separated at about the same time, gave rise to an ancestral four-gene system that appears to be the β-globin complex precursor in all mammalian species [Hardies et al., 1984; see Collins and Weissman, 1984]. Short, direct repeats flanking the coding regions might have played an important role in the generation of deletions in the coding and noncoding regions of genes in the β-globin gene complex [Efstratiadis et al., 1980]. Genes in the human β-globin gene complex have been sequenced [see Collins and Weissman, 1984]. Sequence data include 16 kb of DNA 5′ to the human ε-globin gene [Li et al., 1985], the embryonic ε-globin gene [Baralle et al., 1980], the fetal γ^G and γ^A genes [Shen et al., 1981], the pseudogene ψ^β [Shen and Smithies, 1982], and the δβ-globin genes [Poncz et al., 1983]. β-Globin genes have been extensively studied and present a unique model system to study the ordered expression of developmentally controlled genes.

In the present study, the contiguous 73.3 kb of sequence information spanning the locus of the β-globin gene complex is analyzed for the presence of inverted repeats (also called palindromes or dyad symmetry motifs) and homeodomain protein binding sites which were suggested to be abundant in the AT-rich class of MARs [Boulikas, 1992b]. This search suggests the positions of putative MARs along the gene complex, information useful for a better understanding of the mechanisms that control the coordinate expression of the genes in the complex during development.

RESULTS

ATTA and ATTTA Motifs Are Characteristic of Origins of Replication (ORIs), Homeodomain Protein Binding Sites, and the AT-Rich Class of MARs

A striking similarity exists between AT-rich sequence motifs which are recognition sites of homeotic proteins, matrix attachment regions, and elements in the core sequence of replication origins [see Tables I–III in Boulikas, 1992b]. For example, the ATTA and ATTTA elements are present in the recognition sequences of the human homeotic HOX4D protein [Zappavigna et al., 1991], the mouse Hox-1.3 homeotic protein [Odenwald et al., 1989], the Oct-1 homeobox-containing transcription factor [Herr et al., 1988], several Drosophila homeotic proteins [Beachy et al., 1988; Desplan et al., 1988; Florence et al., 1991; Qian et al., 1991]. In addition, the ATTA and ATTTA core elements occur in the matrix attachment region of the Chinese hamster DHFR gene [Käs and Chasin,

1987], the human HPRT gene [Sykes et al., 1988], and within the topoisomerase II cleavage sites of the mouse κ gene [Blasquez et al., 1989b] and in the chicken α-globin gene [Farache et al., 1990]. What is more striking, the 3′ MAR of the human apolipoprotein B gene is a mosaic of TAAT and TAAAT motifs [Fig. 1 in Boulikas, 1993]. The ATTA and ATTTA motifs are also found in the origin of replication of adenoviruses 2 and 4 [Hay, 1985], ORI_S of HSV1 [Stow and McMonagle, 1983], ORI_L of HSV1 [Weller et al., 1985], the ORI_{L1} and ORI_{L2} of HSV2 [Lockshon and Galloway, 1986], the origins of replication of SV40 and polyoma viruses [Prives et al., 1987], and in several other ORIs. Thus, the ATTA and ATTTA elements are shared between ORI, MAR, and homeotic protein recognition sequences, suggesting that nuclear matrix is involved in the differential activation of origins of replication during cell type determination in development [Boulikas, 1992b].

Search for Matrix Anchorage Binding Sites Along the β-Globin Gene Complex

We have screened a contiguous stretch of 73.3 kb of human sequences including and surrounding the five globin genes for the presence of potential homeotic protein binding sites (HOMEOs), MARs, and origins of replication (ORIs). This article points out that HOMEOs, MARs, and ORIs are related and may, in some cases, be represented by a common sequence. To do so, we have applied a set of several rules [Boulikas, 1993], most of which ought to be satisfied by a stretch of 100–1,000 bp of DNA in order for this sequence to be classified as a MAR. These rules were set from sequence analysis of the MARs identified in the various genes in experimental work by others (see introduction) as well as from work in our own laboratory using random sequencing of cloned human, mouse, rat, and cat matrix attachment sites [Boulikas and Kong, in press].

Potential Homeodomain Protein Binding Sites Within the β-Globin Gene Complex

Figure 1 shows homeotic protein binding sites along the 73.3 kb β-globin gene locus. The core recognition sites of homeodomain proteins, origins of replication, and MARs listed in Tables I–III in Boulikas [1992b] were utilized to screen the β-globin gene complex using a Sun workstation computer. Sequences were 6–8 nucleotides long. These include homeodomain protein binding sites on promoter sequences from mammalian, *Xenopus* and *Drosophila* species. Sequences from homeotic protein footprints usually having a length of 8–24 nt were broken down to 8 nt motifs. The core sequences of the actual DNase I footprints of the homeotic proteins were used here. Such an analysis is limited by the fact that only a few homeotic protein binding sites have been determined today, most of which are from *Drosophila*. However, the fact that the homeobox sequence, which is the actual DNA binding site of homeodomain proteins, is highly conserved among species as distant as flies and humans [see Levine and Hoey, 1988] and the fact that many homeodomain proteins recognize the core ATTA sequence [Scott et al., 1989] suggests that the results of our search are significant.

Figure 1 shows that homeodomain protein binding sites (HOMEOs) are nonrandomly distributed but are clustered at characteristic sites within the gene complex. Several clusters of such motifs are present in the LCR region (1.0–16.0 kb region in Fig. 1). Three other clusters of HOMEOs are in the ϵ-γ^A intergenic region (around the 22.6, 32.0, and 37.4 kb regions), two 3′ to the ψ^β gene (around positions 48.0 and 51.0 kb of the gene map), and one each in the large intron of δ- and β-globin genes. Two eminent clusters of HOMEOs are in the δ-β intergenic region (around the 58.0 and 60.5 kb regions). Finally, one HOMEO cluster is present in the 69.5 kb site within an L1 repeat and another in the 72.8 region to the 3′ side of L1.

Palindromic Sequences Within the β-Globin Gene Complex

Figure 2 shows potential cruciform structures along the 77.3 kb domain of the β-like globin genes. Using a program developed by Dr. Alexander Milosavljevic, 22 potential palindromic structures were detected which appear to be nonrandomly distributed along the β-globin gene complex. Of these, the ten first palindromes fall within the LCR region, most of them very close to the four DNase I hypersensitive sites. Of the others, #11 lies ~ 2 kb to the 5′ end of the ε-globin gene, #13 and #14 lie ~ 2.3 kb to the 5′ end of the ψ^β gene, #16 lies ~ 1 kb to the 5′ end of δ-globin gene, #17 and #18 within the large intron of the δ-globin gene, #19, #20, #21 at, respectively, ~ 1.8, ~ 0.57, and ~ 0.55 kb 5′ to the β-globin gene, and finally, #22 within the large intron of the β-globin gene (Fig. 2). Recent

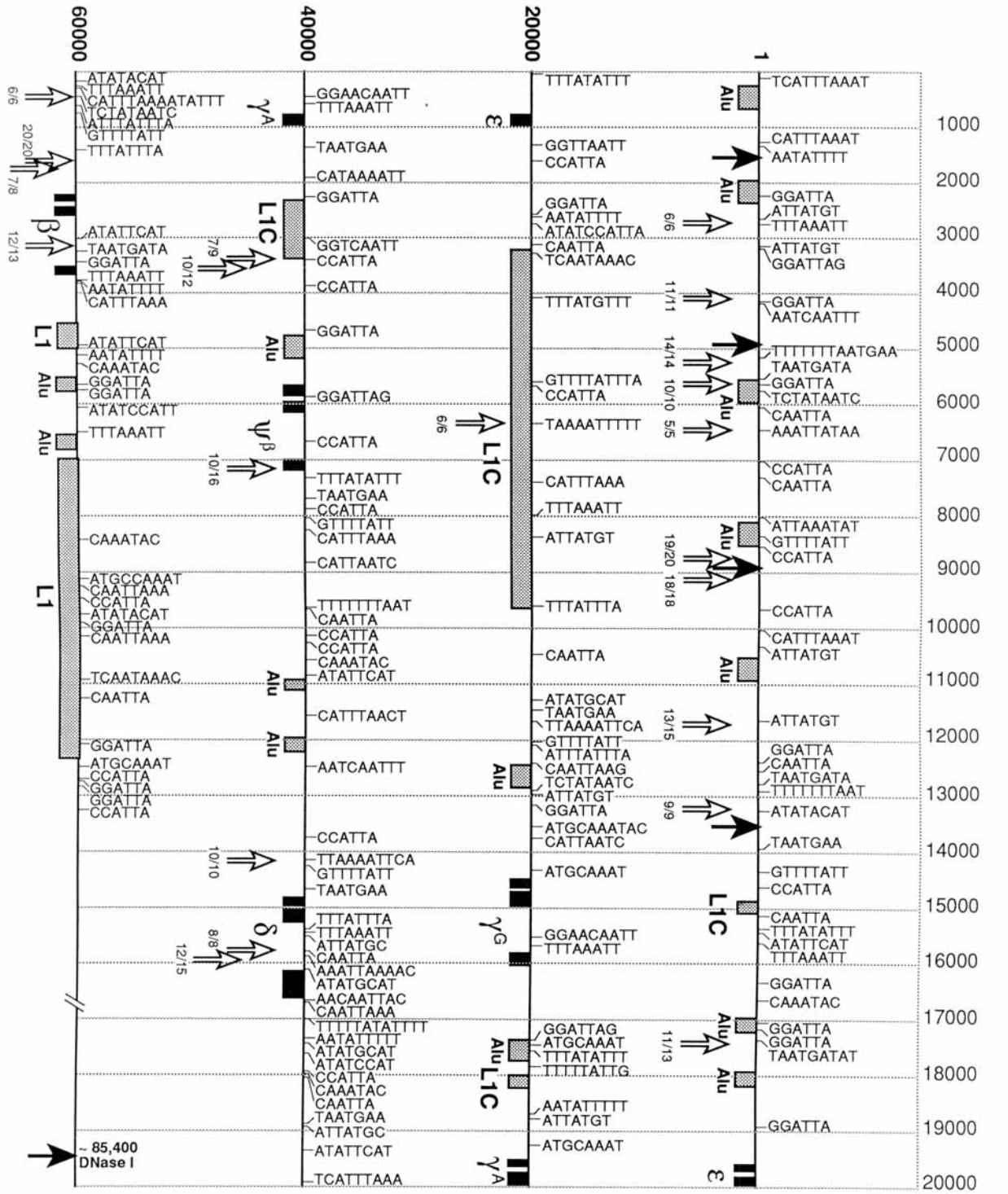

Fig. 1. Homeotic protein binding sites, MARs, ORIs, and palindromes in the human β-globin gene locus This figure is in scale Each of the four horizontal lines represents 20 kb of DNA The ε, γ^G, γ^A, ψ^β, δ, and β gene exons, three in each gene, are shown as solid boxes The *Alu* and L1 repeats are shown in gray *L1C* (*L1* complementary) represents an L1 repeat in the opposite orientation Solid arrows represent the DNase I hypersensitive sites, open arrows indicate the position of palindromes (see also Fig 2) The first number before the slash at the base of open arrows represents the number of matches, and the second the total nucleotides in the stem of each palindrome (see also Fig 2) The short sequence motifs derived from homeotic protein binding sites, MARs, and ORIs of other known genes are shown, with the thin lines representing their exact position on the map

data by Dayn and collaborators [1992] have shown that the sequence $d(A\text{-}T)_{16}$, similar to palindromes 6 and 16 (Fig. 2), when inserted into the promoter region of a gene, can provoke its conversion into a cruciform structure in bacteria by transcription-driven negative supercoiling.

On the basis of the specific sites where strong palindromic sequences are present, it is suggested that at least three potential origins of replication are present: a broad one in the LCR; a second in the δ-globin; and finally a third in the β-globin gene (large intron and 5′ flanking region of the β- and δ-globin genes). The LCR origin might be active in cells of the embryonic yolk sac which express the ε-gene whereas the putative β- and δ-globin gene origins might be actively used in adult bone marrow tissue cells expressing the β- and δ-globin genes. A fourth putative origin in the distal 3′ flanking region of the β-globin gene might be used in nonerythroid tissues for the replication of the entire β-globin gene complex.

Potential MAR/Homeodomain Protein Binding Sites in the LCR Region

The locus control region (LCR) located about 50 kb upstream from the β-globin gene and ~ 7 kb upstream from the ε-globin gene is a region of several kb that contains four or five DNase I hypersensitive sites and dictates the ordered expression of the five globin genes (ε, γ^G, γ^A, δ, and β) [Tuan et al., 1985; Forrester et al., 1986, 1987; Grosveld et al., 1987; see Townes and Behringer, 1990].

About 1.0 kb of 5′ and 3′ flanking sequences of human γ- and β-globin genes will drive a low-level correct tissue-specific and developmental stage-specific expression of these genes in transgenic mice [Townes et al., 1985; Chada et al., 1986; Kollias et al., 1986]. However, the LCR region is essential for the high-level expression of these genes. Thus, LCR seems both to organize the chromatin of the β-globin gene complex into an open conformation as well as to act as a powerful enhancer for the transcription of ε-, γ-, and β-globin genes [see Townes and Behringer, 1990; Felsenfeld, 1992].

Figures 3 and 4 show part of DNA sequence motifs from the LCR which include the DNase I hypersensitive sites 4 and 2. These putative MARs contain palindromes (overlined in Figs. 3, 4), GA-rich and TC-rich stretches with the potential to form triplex DNA, and TG-boxes. In addition, the TG, TA, and CA dinucleotides known to kink DNA when spaced at 0.5 or 1.0 helical turns and to be overrepresented in protein transcription factor recognition and binding sites [see Boulikas, 1993] are outlined. These regions also possess a certain number of ATTA and ATTTA motifs (boxed) (see also Fig. 1).

Figures 3 and 4 do not exhaust all potential MARs of the β-globin gene complex. Figure 1, indeed, shows that additional clusters of homeodomain protein binding sites which might be potential MARs are present in the region surrounding the DNase I-HS site 1, the 5′ flanking region of γ^G, a flanking region to the 3′ side of the β-globin gene which falls within an L1 repeat, and others.

It will be interesting to theoretically search, in addition to the homeodomain protein sites, for the binding sites of all transcription factors known today along the β-globin gene complex.

DISCUSSION

Comparison of Our Finding With Others

Our search along the sequenced 73.3 kb long stretch of DNA comprising and surrounding the β-globin gene complex suggested that globin genes may contain MAR-like sequences in their upstream large intron and downstream regions. Long MAR-like sequences are present in the upstream locus control region (LCR), including the DNase I hypersensitive sites that are suggested to delineate the 5′ border of the β-globin gene domain. Strong internal MARs (facultative or functional) are suggested to be present in the upstream γ^G region, as well as in the upstream large intron and 3′ flanking regions of the δ and β genes.

Jarman and Higgs [1988] have studied the distribution of segments from the human β-globin gene complex among matrix-bound (pellet) and nonbound (soluble) fractions after fractionation of nuclei. Their method of nuclear matrix isolation includes removal of histones from DNA with 25 mM LIS (lithium diiodosalicylate) followed by incubation of the resulted halos of nuclei with restriction enzymes (EcoRI, BglII, or HindIII) for 3–5 h, a method resulting in solubilization of ~ 75% of the DNA, leaving 25% of DNA crosscomplexed with nuclear matrix proteins. In spite of the possibility that the long time of incubation of nuclei halos with restriction enzymes (3–5 h) can cause nuclear matrix disassembly into protein and DNA components leading to potential artifacts, their data have

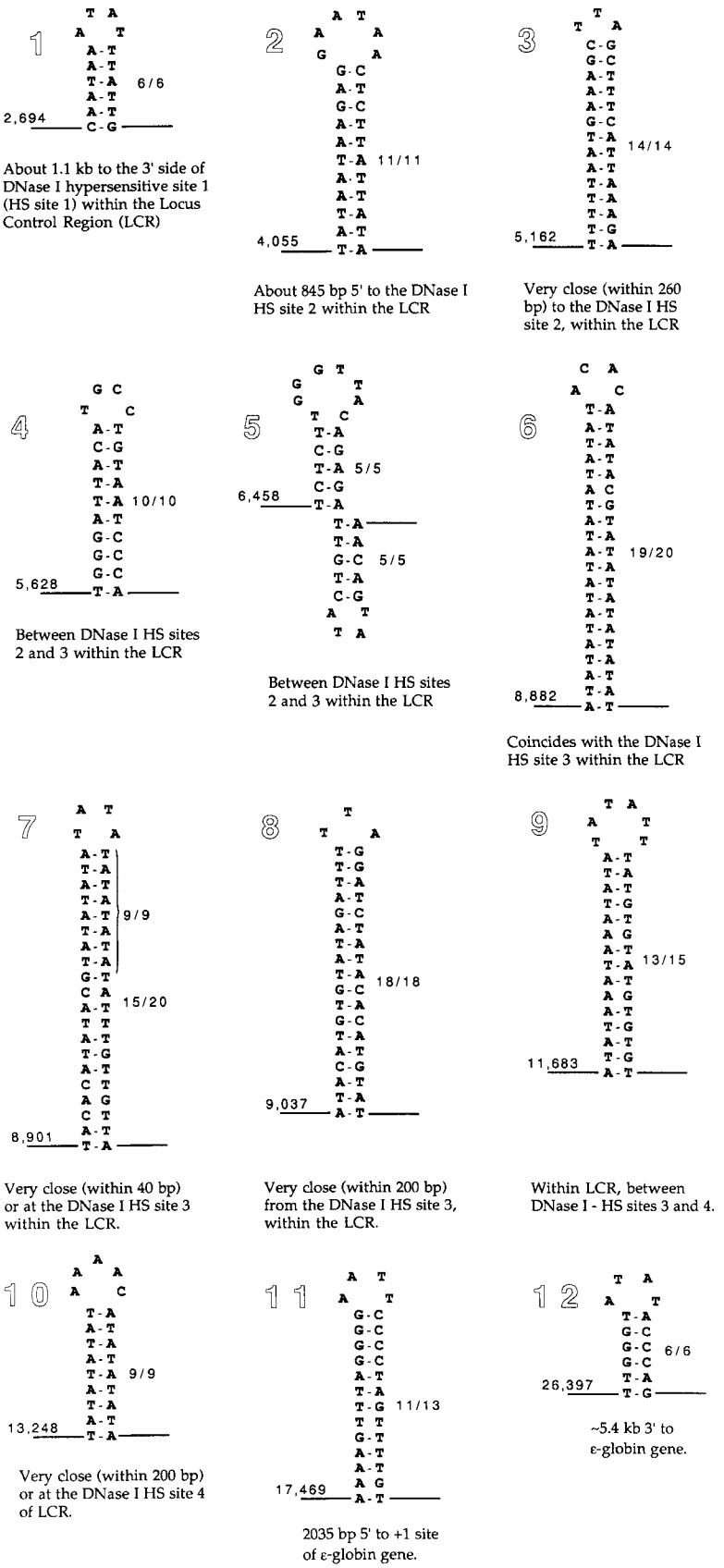

Fig. 2. Potential cruciform structures within the β-globin gene domain, formed under torsional strain on the DNA. These structures are numbered 1–22 according to their position in the locus which is indicated at the left horizontal line. The figure 6/6 in the palindrome 1 represents 6 matches over 6 nucleotides in the stem. The legend to each palindrome describes its position with respect to the DNase I HS site and the genes.

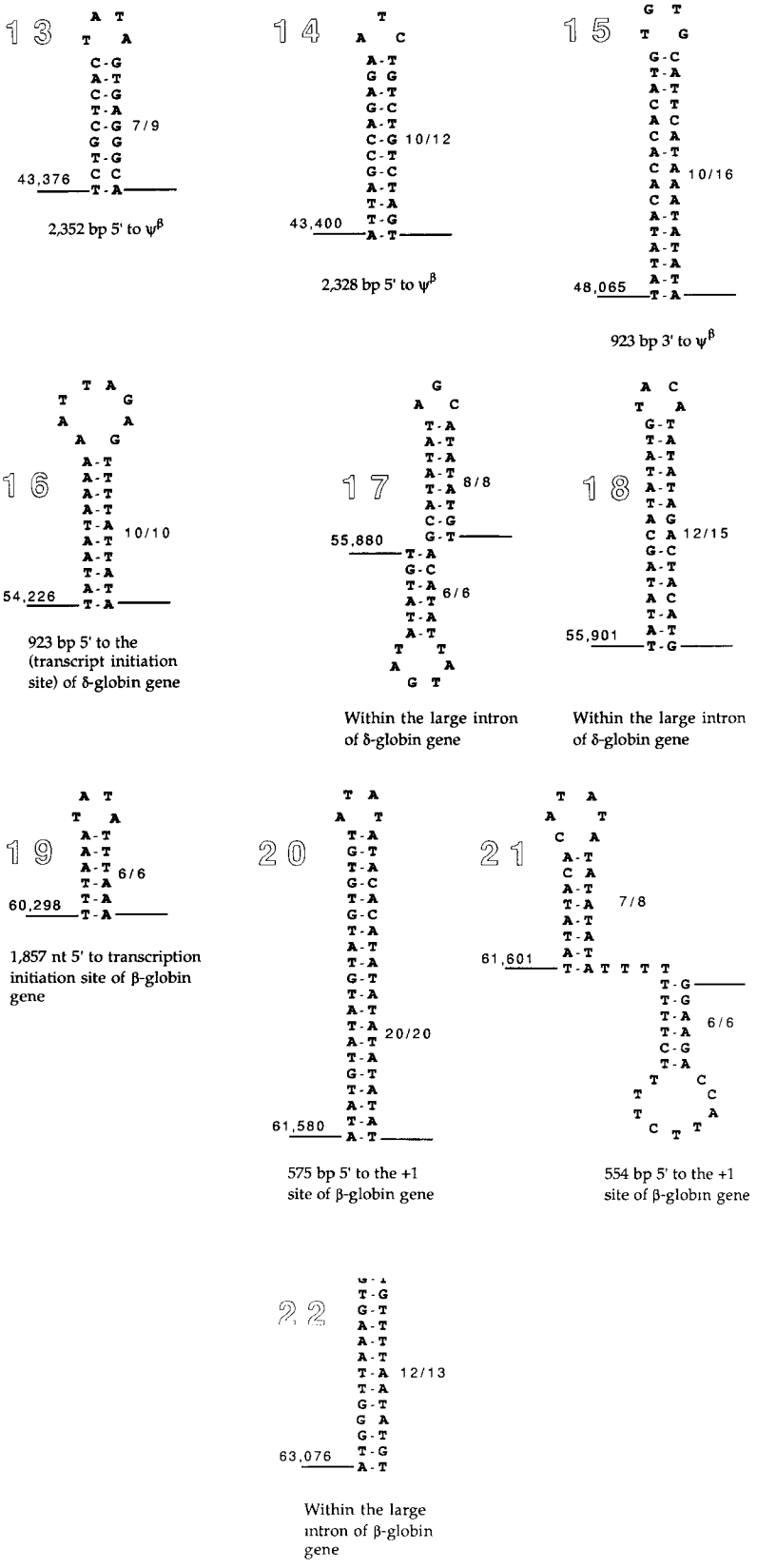
13
7/9
43,376
2,352 bp 5' to ψβ
14
10/12
43,400
2,328 bp 5' to ψβ
15
10/16
48,065
923 bp 3' to ψβ
16
10/10
54,226
923 bp 5' to the (transcript initiation site) of δ-globin gene
17
8/8
6/6
55,880
Within the large intron of δ-globin gene
18
12/15
55,901
Within the large intron of δ-globin gene
19
6/6
60,298
1,857 nt 5' to transcription initiation site of β-globin gene
20
20/20
61,580
575 bp 5' to the +1 site of β-globin gene
21
7/8
6/6
61,601
554 bp 5' to the +1 site of β-globin gene
22
12/13
63,076
Within the large intron of β-globin gene

Figure 2.

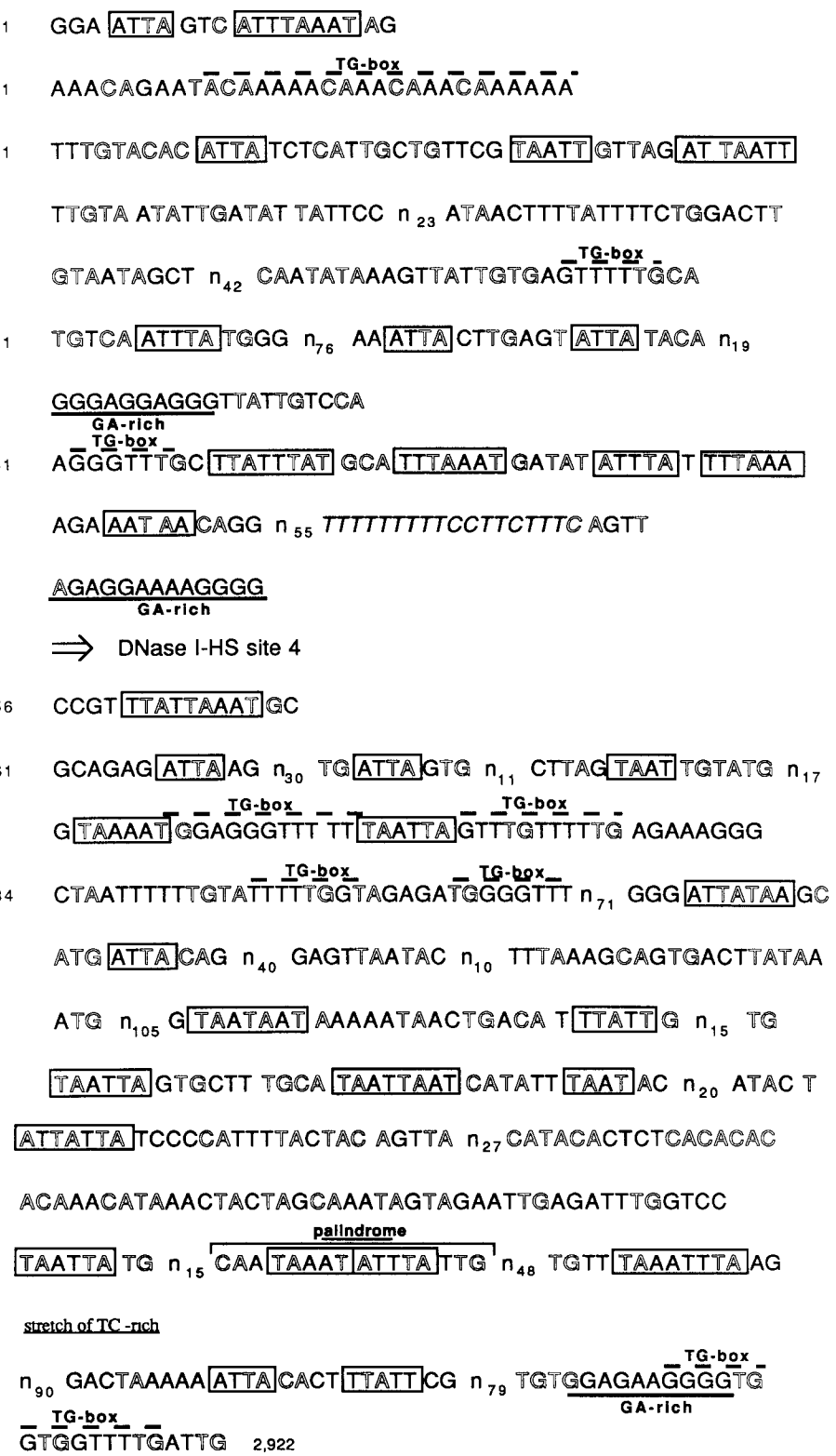

Fig. 3. Potential MAR/homeotic protein binding sites (HOMEOs) in the LCR region surrounding the DNase I hypersensitive site 4 (far upstream) of the human β-globin gene complex HOMEOs in region from nucleotide 91 to nucleotide 2,922 such as ATTA and ATTTA are boxed, whereas TA, TG, and CA dinucleotides which cause kinks and are overrepresented in protein recognition sites [see Boulikas, 1993] are outlined An n followed by a number indicates the number of nucleotides between two sequence motifs The DNase I HS site 4 is indicated with an open arrow Potential left-handed DNA is underlined and palindromes are double-overlined

Fig. 4. Potential MAR/homeotic protein binding sites (HOMEOs) in the LCR region surrounding the DNase I hypersensitive site 2 (far upstream) of the human β-globin gene complex. (region 6,842–10,939). GA-rich stretches are in thick characters and underlined. TG boxes are overlined. An *Alu* sequence is shown. See legend to Figure 3 for other symbols.

identified large segments (several kb in size) within the β-globin gene complex associated with the nuclear matrix. However, in a recent study, no matrix attachment sites were found within a region starting 7.5 kb 5′ to the ε-globin gene and ending 4 kb 3′ to the β-globin gene [Bartjeliotou and Dimitriadis, 1992].

Our computer-assisted searches for HOMEOs (Fig. 1) and palindromes (Fig. 2) in the human β-globin gene complex give clusters of such motifs at positions, most of which coincide to those regions identified by Jarman and Higgs [1988]. These authors have detected strong attachment sites within two large EcoRI restriction fragments of 5.5 and 3.8 kb surrounding the adult-specific β-globin gene. In addition, matrix attachment sites were detected at a 8.2 kb BglII fragment surrounding and including the δ-globin gene as well as at a 4.1 kb BglII fragment composed of the entire ψ^{β} gene and some downstream sequences. Weak MARs were found at the γ^{G}-gene, and no sites at the γ^{A}-gene. Finally, a strong MAR was detected in the far upstream of the ε-gene region but not within the immediate upstream and immediate downstream region of the ε-globin gene; in this study, due to use of restriction enzyme sites along the gene complex, rather large DNA fragments (i.e., 3.8–8.2 or larger) were identified as possessing MAR-like activity [Jarman and Higgs, 1988]. The theoretical predictions of this study, directed at identifying potential MAR sequences at the nucleotide level, complement the searches of Jarman and Higgs [1988] for MARs on this gene complex.

MARs as ORIs

That newly replicated DNA is specifically located on the nuclear matrix has been demonstrated in several laboratories [Berezney and Coffey, 1975; Pardoll et al., 1980; Buongiorno-Nardelli et al., 1982; Aelen et al., 1983; van der Velden et al., 1984; Razin et al., 1986; Carrì et al., 1986; Dijkwel and Hamlin, 1988; Vaughn et al., 1990]. Amati and Gasser [1988, 1990] have shown that MARs from yeast coincide with putative origins of replication and that *Drosophila* MARs can drive the autonomous replication of plasmids in yeast [see also Amati et al., 1990]. In addition, isolation and cloning of putative origins of replication from monkey cells in culture [Zannis-Hadjopoulos et al., 1985; Frappier and Zannis-Hadjopoulos, 1987] have shown them to possess sequence homology with MARs [Rao et al., 1990]. These data rather conclusively demonstrate that origins of replication are nested in the nuclear matrix DNA. Furthermore, the work of Boulikas [1992b] pointing out that origins of replication, nuclear matrix anchorage sites, and homeotic protein recognition and binding sites share the ATTA, ATTTA, and ATTTTA motifs suggests that the differential activation of origins of replication which is thought to establish the differential gene expression, cell memory, and cell type formation in development [Callan, 1973; Spradling and Orr-Weaver, 1987] might be regulated on the nuclear matrix. The LCR region predicted from this study to be an extensive MAR (see Figs. 3, 4) appears to harbor a broad origin of replication. At least two other ORIs are proposed to reside around the δ- and β-globin genes, including palindromes 16–22 (Figs. 1, 2).

MARs May Contain Inverted Repeats

Palindromic (or inverted repeats or dyad symmetry) sequences are able to convert into cruciform structures upon introduction of torsional strain on the DNA [Panayotatos and Wells, 1981; Dayn et al., 1992]. We have recently cloned and sequenced a non-AT-rich type of MARs and have shown it to be enriched in inverted repeats [Boulikas and Kong, in press]. Additionally, many origins of replication contain palindromic sequences. Inverted repeats are found in replication origins of *E. coli* [Meijer et al., 1979], bacteriophages λ and φ 80 [Hobom et al., 1979], and mammalian viruses including SV40 [Subramanian et al., 1977], polyoma [Soeda et al., 1978; Hendrickson et al., 1987], the human papovavirus BKV [Dhar et al., 1978], ORI_S and ORI_L of HSV1 [Stow and McMonagle, 1983; Weller et al., 1985], HSV2 [Lockshon and Galloway, 1986], Epstein-Barr virus [Reisman et al., 1985], JC virus [Frisque, 1983], and mammalian ORI sequences [Frappier and Zannis-Hadjopoulos, 1987; Iguchi-Ariga et al., 1988]. These studies, together with numerous findings vigorously showing that origins of replication are attached to the nuclear matrix (see above) support the idea that palindromic sequences may indeed be characteristic structures of MARs. Our search for palindromes in the human β-globin gene complex (Fig. 2) has indeed shown that palindromes are preferentially located near clusters of homeodomain protein binding sites (Fig. 1), presumably coinciding with matrix attachment sites.

Permanent and Functional MARs

Two types of MARs have been described. 1) The permanent (constitutive) MARs are present in all cell types at all times of embryonic development and stages of cell cycle irrespective of the transcriptional state of the gene [Razin et al., 1986; Levy-Wilson and Fortier, 1989]. Permanent MARs include the ORIs according to Razin and coworkers [1986]. Permanent MARs in our opinion also delineate the border of an entire domain. 2) The functional (or facultative or transient) MARs appear temporarily and are related to transcription, repair, and a particular stage of development. Facultative MARs appear only in cell types which express a given gene but are absent from cell types harboring the same gene into inactive chromatin structures [Levy-Wilson and Fortier, 1989]. Our study is aimed at predicting the location of both types of MARs in the human β-globin gene complex (except those related to repair of damaged sites). Whether or not a specific MAR is constitutive or facultative depends absolutely on the expression of the specific protein transcription factors that interact with one another and with the DNA of the MAR leading to the formation of a supramolecular complex. We consider MARs as weak and strong mass binding sites for protein transcription factors, including homeodomain proteins.

The experimental identification of the functional MARs surrounding the ε-globin gene would be fully feasible when nuclear matrices from embryonic yolk sac cells are used. Thus, according to our model, nuclear matrix preparations from cells expressing the ε-gene ought to compete more effectively for binding to the functional MAR surrounding the ε-gene. However, nuclear matrix proteins from cells at other developmental stages are expected to bind to the MAR of the ε-gene with very low affinity. This low affinity is due to similarities in structure between homeotic proteins assessed from the degree of homology of their genes [e.g., Scott et al., 1989], and from the fact that multiple binding sites of transcription factors (most of which, according to our model, are nuclear matrix proteins) are present within a MAR; several transcription factors are constitutively expressed at all stages of development though at varying levels [e.g., Saffer et al., 1991]. In addition, nuclear matrices prepared from fetal liver cells are expected to identify the facultative matrix anchorage sites of the γ^G- and γ^A-globin genes. Jarman and Higgs [1988] have used the K562 human myelogenous erythroleukemia cell line which is known to express the fetal stage-specific genes [Purucker et al., 1990]. Nuclear matrices from adult erythroid cells are expected strongly to identify the facultative MARs surrounding the adult δ- and β-globin genes. Greenstein [1988] using the mouse β-globin gene complex and Farache and coworkers [1990] using adult chicken erythrocytes have most probably identified the functional MARs around the adult state-specific β- and α-globin genes as well as the permanent MARs in these genes.

We hope that future studies will be directed at screening gene sequences from databanks for inverted repeats, homeodomain proteins, and other transcription factor binding sites, to predict putative positions of MARs, and by consequence, positions of transcriptional enhancers and origins of replication. Such theoretical data could then be correlated with experimental work aimed at defining origins of replication, DNase I HS sites at the gene's chromatin, and enhancer elements.

ACKNOWLEDGMENTS

The valuable input of Drs. Aleksandar Milosavljevic, Jerzy Jurka, Emile Zuckerkandl, and Ed Trifonov is greatly appreciated. Special thanks to Sherman Yang and Dr. Chuang Fong Kong for help with the computer searches, to J. Walichiewicz for computer graphics, and to D. Read and D. Brooks for preparation of the manuscript.

REFERENCES

Aelen JMA, Opstelten RJG, Wanka F (1983): Organization of DNA replication in *Physarum polycephalum*. Attachment of origins of replicons and replication forks to the nuclear matrix. Nucleic Acids Res 11:1181–1195.

Amati B, Gasser SM (1990): *Drosophila* scaffold-attached regions bind nuclear scaffolds and can function as ARS elements in both budding and fission yeasts. Mol Cell Biol 10:5442–5454.

Amati B, Pick L, Laroche T, Gasser SM (1990): Nuclear scaffold attachment stimulates, but is not essential for ARS activity in *Saccharomyces cerevisiae:* Analysis of the *Drosophila ftz* SAR. EMBO J 9:4007–4016.

Amati BB, Gasser SM (1988): Chromosomal ARS and CEN elements bind specifically to the yeast nuclear scaffold. Cell 54:967–978.

Baralle FE, Shoulders CC, Proudfoot NJ (1980): The primary structure of the human ε-globin gene. Cell 21:621–626.

Bartjeliotou AJ, Dimitriadis GJ (1992): The association of the human ε-globin gene with the nuclear matrix: A reconsideration Mol Cell Biochem 115:105–115.

Beachy PA, Krasnow MA, Gavis ER, Hogness DS (1988) An Ultrabithorax protein binds sequences near its own and the Antennapedia P1 promoter Cell 55 1069–1081

Berezney R (1991) The nuclear matrix A heuristic model for investigating genomic organization and function in the cell nucleus J Cell Biochem 47 109–123

Berezney R, Coffey DS (1975) Nuclear protein matrix Association with newly synthesized DNA Science 189 291–293

Blasquez VC, Sperry AO, Cockerill PN, Garrard WT (1989b) Protein DNA interactions at chromosomal loop attachment sites Genome 31 503–509

Blasquez VC, Xu M, Moses SC, Garrard WT (1989a) Immunoglobulin κ gene expression after stable integration I Role of the intronic *MAR* and enhancer in plasmacytoma cells J Biol Chem 264 21183–21189

Bodnar JW (1988) A domain model for eukaryotic DNA organization A molecular basis for cell differentiation and chromosome evolution J Theor Biol 132 479–507

Boulikas T (1987) Nuclear envelope and chromatin structure Int Rev Cytol Suppl 17 599–684

Boulikas T (1992a) Poly(ADP-ribosyl)ation, repair, chromatin and cancer Curr Persp Mol Cell Oncol 1 1–109

Boulikas T (1992b) Homeotic protein binding sites, origins of replication, and nuclear matrix anchorage sites share the ATTA and ATTTA motifs J Cell Biochem 50 111–123

Boulikas T (1992c) Chromatin and nuclear matrix in development and in carcinogenesis A theory Int J Oncol 1 357–372

Boulikas T (1992d) Evolutionary consequences of preferential damage and repair of chromatin domains J Mol Evol 35 156–180

Boulikas T (1993) Nature of DNA sequences at the attachment regions of genes to the nuclear matrix J Cell Biochem 52 14–22

Boulikas T, Kong CF (in press) Multitude of inverted repeats characterize a class of anchorage sites of chromatin loops to the nuclear matrix

Buongiorno-Nardelli M, Micheli G, Carri MT, Marilley M (1982) A relationship between replicon size and supercoiled loop domains in the eukaryotic genome Nature 298 100–102

Callan HG (1973) DNA replication in the chromosomes of eukaryotes Cold Spring Harb Symp Quant Biol 38 195–203

Carri MT, Micheli G, Graziano E, Pace T, Buongiorno-Nardelli M (1986) The relationship between chromosomal origins of replication and the nuclear matrix during the cell cycle Exp Cell Res 164 426–436

Chada K, Magram J, Costantini F (1986) An embryonic pattern of expression of a human fetal globin gene in transgenic mice Nature 319 685–689

Cockerill PN, Garrard WT (1986) Chromosomal loop anchorage of the kappa immunoglobulin gene occurs next to the enhancer in a region containing topoisomerase II sites Cell 44 273–282

Collins FS, Weissman SM (1984) The molecular genetics of human hemoglobin Prog Nucleic Acid Res Mol Biol 31 315–462

Dayn A, Malkhosyan S, Mirkin SM (1992) Transcriptionally driven cruciform formation in vivo Nucleic Acids Res 20 5991–5997

Desplan C, Theis J, O'Farrell PH (1988) The sequence specificity of homeodomain-DNA interaction Cell 54 1081–1090

Dhar R, Lai C-J, Khoury G (1978) Nucleotide sequence of the DNA replication origin for human papovirus BKV Sequence and structural homology with SV40 Cell 13 345–358

Dijkwel PA, Hamlin JL (1988) Matrix attachment regions are positioned near replication initiation sites, genes, and an interamplicon junction in the amplified dihydrofolate reductase domain of Chinese hamster ovary cells Mol Cell Biol 8 5398–5409

Dworetzky SI, Fey EG, Penman S, Lian JB, Stein JL, Stein GS (1990) Progressive changes in the protein composition of the nuclear matrix during rat osteoblast differentiation Proc Natl Acad Sci USA 87 4605–4609

Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel JK, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot N (1980) The structure and evolution of the human β-globin gene family Cell 21 653–668

Farache G, Razin SV, Targa FR, Scherrer K (1990) Organization of the 3′-boundary of the chicken α globin gene domain and characterization of a CR 1-specific protein binding site Nucleic Acids Res 18 401–409

Felsenfeld G (1992) Chromatin as an essential part of the transcriptional mechanism Nature 355 219–224

Fey EG, Bangs P, Sparks C, Odgren P (1991) The nuclear matrix Defining structural and functional roles CRC Crit Rev Euk Gene Express 1 127–143

Forrester WC, Takegawa S, Papayannopoulou T, Stamatoyannopoulos G, Groudine M (1987) Evidence for a locus activator region Nucleic Acids Res 15 10159–10177

Forrester WC, Thompson C, Elder JT, Groudine M (1986) A developmentally stable chromatin structure in the human β-globin gene cluster Proc Natl Acad Sci USA 83 1359–1363

Frappier L, Zannis-Hadjopoulos M (1987) Autonomous replication of plasmids bearing monkey DNA origin-enriched sequences Proc Natl Acad Sci USA 84 6668–6672

Frisque RJ (1983) Nucleotide sequence of the region encompassing the JC virus origin of DNA replication Virology 46 170–176

Greenstein RJ (1988) Constitutive attachment of murine erythroleukemia cell histone-depleted DNA loops to nuclear scaffolding is found in the β-major but not the α1-globin gene DNA 7 601–607

Grosveld F, Blom van Assendelft G, Greaves DR, Kollias G (1987) Position-independent high level expression of the human β-globin gene in transgenic mice Cell 51 975–985

Hancock R, Boulikas T (1982) Functional organization in the nucleus Int Rev Cytol 79 165–214

Hardies SC, Edgell MH, Hutchinson CA III (1984) Evolution of the mammalian β-globin gene cluster J Biol Chem 259 3748–3756

Hay RT (1985) Origin of adenovirus DNA replication Role of the nuclear factor I binding site in vivo J Mol Biol 186 129–136

Hendrickson EA, Fritze CE, Folk WR, DePamphilis ML (1987) The origin of bidirectional DNA replication in polyoma virus EMBO J 6 2011–2018

Herr W, Sturm RA, Clerc RG, Corcoran LM, Baltimore D, Sharp PA, Ingraham HA, Rosenfeld MG, Finney M, Ruvkun G, Horvitz HR (1988) The POU domain A large conserved region in the mammalian *pit*-1, *oct*-1, *oct*-2, and *Caenorhabditis elegans unc-86* gene products Genes Dev 2 1513–1516

Hobom G, Grosschedl R, Lusky M, Scherer G, Schwarz E, Kossel H (1979) Functional analysis of the replicator structure of lambdoid bacteriophage DNAs Cold Spring Harb Symp Quant Biol 43 165–178

Iguchi-Ariga SMM, Okazaki T, Itani T, Ogata M, Sato Y, Ariga H (1988) An initiation site of DNA replication with transcriptional enhancer activity present upstream of the c-*myc* gene EMBO J 7 3135–3142

Jarman AP, Higgs DR (1988) Nuclear scaffold attachment sites in the human globin gene complexes EMBO J 7 3337–3344

Kas E, Chasin LA (1987) Anchorage of the Chinese hamster dihydrofolate reductase gene to the nuclear scaffold occurs in an intragenic region J Mol Biol 198 677–692

Klehr D, Maass K, Bode J (1991) Scaffold-attached regions from the human interferon β domain can be used to enhance the stable expression of genes under the control of various promoters Biochemistry 30 1264–1270

Kollias G, Wrighton N, Hurst J, Grosveld F (1986) Regulated expression of human $^{A}\gamma$-, β-, and hybrid γβ-globin genes in transgenic mice Manipulation of the developmental expression patterns Cell 46 89–94

Leffak M, James CD (1989) Opposite replication polarity of the germ line c-*myc* gene in HeLa cells compared with that of two Burkitt lymphoma cell lines Mol Cell Biol 9 586–593

Levine M, Hoey T (1988) Homeobox proteins as sequence-specific transcription factors Cell 55 537–540

Levy-Wilson B, Fortier C (1989) The limits of the DNase I-sensitive domain of the human apolipoprotein B gene coincide with the locations of chromosomal anchorage loops and define the 5′ and 3′ boundaries of the gene J Biol Chem 264 21196–21204

Li Q, Powers PA, Smithies O (1985) Nucleotide sequence of 16-kilobase pairs of DNA 5′ to the human ε-globin gene J Biol Chem 260 14901–14910

Lockshon D, Galloway DA (1986) Cloning and characterization of ori_{L2}, a large palindromic DNA replication origin of Herpes Simplex Virus Type 2 Virology 58 513–521

Meijer M, Beck E, Hansen FG, Bergmans HEN, Messer W, von Meyenburg K, Schaller H (1979) Nucleotide sequence of the origin of replication of the *Escherichia coli* K-12 chromosome Proc Natl Acad Sci USA 76 580–584

Mirkovitch J, Gasser SM, Laemmli UK (1988) Scaffold attachment of DNA loops in metaphase chromosomes J Mol Biol 200 101–109

Mirkovitch J, Mirault M-E, Laemmli UK (1984) Organization of the higher-order chromatin loop Specific DNA attachment sites on nuclear scaffold Cell 39 223–232

Odenwald WF, Garbern J, Arnheiter H, Tournier-Lasserve E, Lazzarini RA (1989) The *Hox*-1 3 homeo box protein is a sequence-specific DNA-binding phosphoprotein Genes Dev 3 158–172

Panayotatos Wells RD (1981) Cruciform structures in supercoiled DNA Nature 289 466–470

Pardoll DM, Vogelstein B, Coffey DS (1980) A fixed site of DNA replication in eucaryotic cells Cell 19 527–536

Paulson JR, Laemmli UK (1977) The structure of histone-depleted metaphase chromosomes Cell 12 817–828

Phi-Van L, Stratling WH (1988) The matrix attachment regions of the chicken lysozyme gene co-map with the boundaries of the chromatin domain EMBO J 7 655–664

Phi-Van L, von Kries JP, Ostertag W, Stratling WH (1990) The chicken lysozyme 5′ matrix attachment region increases transcription from a heterologous promoter in heterologous cells and dampens position effects on the expression of transfected genes Mol Cell Biol 10 2302–2307

Poncz M, Schwartz E, Ballantine M, Surrey S (1983) Nucleotide sequence analysis of the δβ-globin gene region in humans J Biol Chem 258 11599–11609

Prives C, Murakami Y, Kern FG, Folk W, Basilico C, Hurwitz J (1987) DNA sequence requirements for replication of polyomavirus DNA in vivo and in vitro Mol Cell Biol 7 3694–3704

Purucker M, Bodine D, Lin H, McDonagh K, Nienhuis AW (1990) Structure and function of the enhancer 3′ to the human $^{A}\gamma$ globin gene Nucleic Acids Res 18 7407–7415

Qian S, Capovilla M, Pirrotta V (1991) The *bx* region enhancer, a distant *cis*-control element of the *Drosophila Ubx* gene and its regulation by *hunchback* and other segmentation genes EMBO J 10 1415–1425

Rao BS, Zannis-Hadjopoulos M, Price GB, Reitman M, Martin RG (1990) Sequence similarities among monkey *ori*-enriched (*ors*) fragments Gene 87 233–242

Razin SV, Kekelidze MG, Lukanidin EM, Scherrer K, Georgiev GP (1986) Replication origins are attached to the nuclear skeleton Nucleic Acids Res 14 8189–8207

Reisman D, Yates J, Sugden B (1985) A putative origin of replication of plasmids derived from Epstein-Barr Virus is composed of two *cis*-acting components Mol Cell Biol 5 1822–1832

Saffer JD, Jackson SP, Annarella MB (1991) Developmental expression of Sp1 in the mouse Mol Cell Biol 11 2189–2199

Scott MP, Tamkun JW, Hartzell GW (1989) The structure and function of the homeodomain Biochim Biophys Acta 989 25–48

Shen S-H, Slightom JL, Smithies O (1981) A history of the human fetal globin gene duplication Cell 26 191–203

Shen S-H, Smithies O (1982) Human globin $\psi\beta_2$ is not a globin-related sequence Nucleic Acids Res 10 7809–7818

Soeda (1978) Similarity of nucleotide sequences around the origin of DNA replication in mouse polyoma virus and simian virus 40 Proc Natl Acad Sci USA 75 162–166

Spradling A, Orr-Weaver T (1987) Regulation of DNA replication during *Drosophila* development Annu Rev Genet 21 373–403

Stief A, Winter DM, Stratling WH, Sippel AE (1989) A nuclear DNA attachment element mediates elevated and position-independent gene activity Nature 341 343–345

Stow ND, McMonagle EC (1983) Characterization of the TR_S/IR_S origin of DNA replication of Herpes Simplex Virus Type 1 Virology 130 427–438

Subramanian KN, Dhar R, Weissman SM (1977) Nucleotide sequence of a fragment of SV40 DNA that contains the origin of DNA replication and specifies the 5′ end of "early" and "late" viral RNA III Construction of the total sequence of *Eco*RII-G fragment of SV40 DNA J Biol Chem 252 355–367

Sykes RC, Lin D, Hwang SJ, Framson PE, Chinault AC (1988) Yeast ARS function and nuclear matrix association coincide in a short sequence from the human HPRT locus Mol Gen Genet 212 301–309

Townes TM, Behringer RR (1990) Human globin locus activation region (LAR) Role in temporal control Trends Genet 6 219–222

Townes TM, Lingrel JB, Chen HY, Brinster RL, Palmiter RD (1985) Erythroid-specific expression of human β-globin genes in transgenic mice EMBO J 4 1715–1723

Tuan D, Solomon W, Li Q, London IM (1985) The "β-like-globin" gene domain in human erythroid cells Proc Natl Acad Sci USA 82 6384–6388

van der Velden HMW, Van Willigen G, Wetzels RHW, Wanka F (1984) Attachment of origins of replication to the nuclear matrix and the chromosomal scaffold FEBS Lett 171 13–16

Vaughn JP, Dijkwel PA, Mullenders LHF, Hamlin JL (1990) Replication forks are associated with the nuclear matrix Nucleic Acids Res 18 1965–1969

Villarreal LP (1991) Relationship of eukaryotic DNA replication to committed gene expression General theory for gene control Microbiol Rev 55 512–542

Weller SK, Spadaro A, Schaffer JE, Murray AW, Maxam AM, Schaffer PA (1985) Cloning, sequencing, and functional analysis of ori_L, a Herpes Simplex Virus Type 1 origin of DNA synthesis Mol Cell Biol 5 930–942

Xu M, Hammer RE, Blasquez VC, Jones SL, Garrard WT (1989) Immunoglobulin κ gene expression after stable integration II Role of the intronic *MAR* and enhancer in transgenic mice J Biol Chem 264 21190–21195

Zannis-Hadjopoulos M, Kaufmann G, Wang S-S, Lechner RL, Karawya E, Hesse J, Martin RG (1985) Properties of some monkey DNA sequences obtained by a procedure that enriches for DNA replication origins Mol Cell Biol 5 1621–1629

Zappavigna V, Renucci A, Izpisua-Belmonte J-C, Urier G, Peschle C, Duboule D (1991) *HOX4* genes encode transcription factors with potential auto- and cross-regulatory capacities EMBO J 10 4177–4187