CLONING AND CHARACTERISATION OF THE HUMAN CARBONIC ANHYDRASE I GENE AND MAPPING OF THE CARBONIC ANHYDRASE GENE CLUSTER ON CHROMOSOME 8

A Thesis Submitted in Partial Fulfilment of the Requirements for Admission to the Degree of Doctor of Philosophy of the University of London

by

Nick Lowe
Department of Biochemistry
University College London

ABSTRACT

The carbonic anhydrase I gene (*CA1*) is part of a multigene family, the protein products of which are enzymes characterised by their ability to catalyse the reversible hydration of CO_2. It exhibits a tissue specific pattern of expression, notably being expressed at high levels in erythroid cells in humans. It also shows regulation at a developmental level, with the CAI protein at very low concentration in the blood of the foetus till a few weeks before birth.

cDNA and genomic clones encoding CAI were isolated and charachterised. Analysis of cDNA clones showed the presence of an occasional exon in the 5' leader of the transcript, while at the 3'-end two polyadenylation sites could be used. while analysis of genomic clones showed that *CA1* is atypical amongst the carbonic anhydrases in having a large intron of 36 kb separating the erythroid specific promoter from the coding region, making the entire gene some 50 kb in length.

Pulsed-field gel electrophoresis was used in the analysis of the physical linkage relationship between *CA1* and the *CA2* and *CA3* genes. Both of these genes lie 5' to *CA1* and are transcribed away from it. *CA3* lies centrally in this cluster separated from the 5' end of *CA1* by 80 kb, while the 5'-end of *CA2* lies some 20 kb downstream of the 3' end of *CA3*.

The DNA methylation state of the gene in several erythroid and non-erythroid cell lines was examined. This showed that in the majority of these cell lines, which do not express the *CA1* gene, extensive regions around *CA1* were largely demethylated. In contrast, DNA from the only *CA1* expressing cell line, HEL, appeared highly methylated at all *Hpa*II sites tested apart for one site at the erythroid promoter and another at the 3'-end of the gene. High levels of methylation of the *CA1* gene were also found in DNA from untransformed cells. The possible implications of this are discussed.

ACKNOWLEDGEMENTS

This thesis would not have been written without the help and advice from a number of people over the years. I would firstly like to thank my supervisor Peter Butterworth for his support and advice throughout the time I have worked with him at U.C.L.

Many thanks also go to the other members of the group including Jon Barlow and Hugh Brady with whom I worked closely in the early stages of this project and Mina Edwards who made all tissue culture work possible. I would also like to thank Jane Sowden for her friendship and helpful discussions during this work.

Many other people also helped with the numerous technical and scientific problems encountered, but I would particularly like to thank Louise Sefton, Peter Rowe and Mike Larkum for their help with pulsed-field electrophoresis. I would especially like to thank Yvonne Edwards not only for her collaborative help, but also for her critical assessment of this manuscript.

ABBREVIATIONS

A	Adenine
amp	ampicillin
APRT	adeninephosphoribosyl transferase
ATP	adenine triphosphate
bp	DNA base pairs
BSA	bovine serum albumin
C	cytosine
CAI	carbonic anhydrase I
CA1	gene encoding carbonic anhydrase I
cDNA	complementary DNA
CIP	calf intestinal phosphatase
ddNTP	dideoxynucleotide triphosphate
DHFR	dihydrofolate reductase
dNTP	deoxynucleotide triphosphate
DDW	double distilled water
DEPC	diethylpyrocarbonate
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
EBV	Epstein Barr virus
EDTA	ethylenediamine tetra-acetic acid
FCS	foetal calf serum
FIGE	field-inversion gel electrophoresis
G	Guanine
hr	hours
HSV	Herpes Simplex Virus
IPTG	isopropyl ß-D-thiogalactopyranoside
kb	kilobases / kilobase pairs
kd	kilodaltons
Klenow	Klenow fragment of DNA polymerase I
µCi	microcuries
min	minutes
mRNA	messenger RNA
M.W.	molecular weight
nt	nucleotide
OFAGE	orthagonal field agarose gel electrophoresis

OD_x	optical density at x nm
PAGE	polyacrylamide gel electrophoresis
PBGD	porphobilinogen deaminase
PBSa	phosphate buffered saline
PEG	polyethylene glycol
PFGE	pulsed-field gel electrophoresis
pfu	plaque forming units
phage	bacteriophage
PMSF	phenyl methyl sulphonylfluoride
PNK	polynucleotide kinase
poly(A+)	polyadenylated
Pu	purine
Py	pyrimidine
RNA	ribonucleic acid
RNase	ribonuclease
rpm	revolutions per minute
RT	room temperature
SDS	sodium dodecyl sulphate
SSC	saline sodium citrate
T	Thymine
TCA	trichloroacetic acid
T_d	dissociation temperature
TEMED	N,N,N',N'-tetramethyl ethylenediamine
tk	thymidine kinase
tRNA	transfer RNA
u	unit(s)
uv	ultraviolet
wk	week(s)
Wu	Weiss units
X-gal	5-bromo-4-chloro-3-indoyl ß-D-galactopyranoside

CONTENTS

FIGURES

CHAPTER ONE: INTRODUCTION

The results described in this thesis have been divided into four sections; 1) Isolation and characterisation of cDNA clones encoding human carbonic anhydrase I; 2) Isolation and characterisation of the gene for carbonic anhydrase I (*CA1*); 3) Analysis of the physical relationship between *CA1* and the closely linked genes *CA2* and *CA3*; 4) Analysis of the methylation state of *CA1, CA2* and *CA3* in various cell types. This work took place within the context of several distinct fields of study and this introductory chapter has therefore been divided into three main sections.

The first of these deals with the carbonic anhydrase isozymes, a family of proteins distinguished by their ability to catalyse the reversible hydration of carbon dioxide ($H_2O + CO_2 -- H_2CO_3$). This relatively simple reaction plays an important role in a number of physiological processes such as respiratory CO_2 transport, pH balance and ion exchange in a wide variety of tissues. Each of the isozymes in this family has its own set of kinetic and physical properties and in addition shows a distinct pattern of tissue specific distribution. This variation presents evolutionary biologists with a model system to study the changes which have taken place within the family of carbonic anhydrase proteins (CAI, CAII, CAIII, etc) and the genes (*CA1, CA2, CA3,* etc) which encode them.

The interest in *CA1* as a subject for study by the techniques of molecular biology stemmed from two features of its expression in man, namely the tissue specific nature of its expression and the temporal control of its expression. CAI is found in a number of tissues of the body, but is found at its highest levels in the red blood cell in man and is one of the earliest detectable markers which appears during the process of formation of the erythrocyte from pluripotent haematopoietic stem cells. *CA1* expression also exhibits developmental changes, *vis* a dramatic increase in expression late in foetal life. Both of the above characteristics - high levels of expression in erythroid

tissues and fetal to adult developmental activation of gene expression - are shared with the globins, possibly the best characterised gene family to date. A brief description of the process of erythropoiesis and the place of CAI within it is therefore given in the second part of the Introduction.

The last section deals with a general description of the control of gene expression at a transcriptional level, with an emphasis on erythroid tissue specificity.

1.1 The Carbonic Anhydrases

1.1.1 The Historical Background

The discovery of carbonic anhydrase was prompted by essentially theoretical considerations. It had been pointed out as early as 1928 that the exchange of carbon dioxide in the lungs could not be accounted for by the known rates of hydration and dehydration of CO_2 (Henriques, 1928). Subsequent experiments involving the exposure of blood to a vacuum showed that the release of CO_2 from the blood was in fact extremely rapid. At this time there was dispute over whether CO_2 was transported through the blood as bicarbonate, or whether (like oxygen) there was some substance with which it could be complexed. One of the suggested candidates for this role was haemoglobin, which was one of the first characterised red cell proteins and was known to be responsible for oxygen transport. This confusion over the role of haemoglobin in CO_2 transport in the blood was compounded by the finding that "purified" haemoglobin could catalyse the hydration reaction (van Slyke and Hawkins, 1930). This error was of course due to contaminating carbonic anhydrase in the preparation. It may also have been an attractive unifying concept to think of globin as having a role in transporting both the oxygen needed for respiration, and the waste product of that respiration. This dispute about the role of globin in CO_2 transport and catalysis in the blood was finally settled with the

discovery of carbonic anhydrase in 1933 when it was shown that the catalytic activity could be separated from the haemoglobin content of blood during extraction with chloroform (Meldrum and Roughton, 1933).

The identification of carbonic anhydrase together with assays for its activity was followed by the isolation of the enzyme from tissues other than blood, firstly in gastric mucosa (Davenport and Fisher, 1938) and shortly afterwards in the kidney (Davenport and Wilhelmi, 1941). It was at about this time that a discovery was made which has been of enormous importance for those working on the physiological role of CA, namely the specific inhibition of CA activity by the sulfonamide antibiotics (Mann and Keilin, 1940). This finding, followed by the synthesis of a number of related compounds with more effective inhibitory action, has facilitated a wide range of studies to determine the physiological role of the enzyme in a variety of tissues.

From the 1940's onwards, there has been a steady increase in the number of tissues identified as CA-containing. The application of CA inhibitors to these tissues and the study of the resulting physiological changes also brought an appreciation that carbonic anhydrase was not only functioning to facilitate the exchange of respiratory CO_2 but was involved in a number of other processes centred on the production of H^+ or bicarbonate ions and often used in the generation of trans-membrane concentration gradients (Maren, 1988).

Over the same period it was also found that what had been termed simply as carbonic anhydrase could be resolved into a number of distinct isozymes. This began when Nyman (1961), at about the same time as a number of other workers, began applying more sophisticated methods of chromatography to preparations of carbonic anhydrase and demonstrated the presence of two distinct forms in human blood. One of these had a relatively low activity and was termed CA B (later to become known as CAI), while the other had a very high activity and was called CA C (now known as CAII). A third

isozyme - CAIII - was discovered in the late 70's in muscle (Holmes, 1976; Koestler *et al.* 1977; Register *et al.* 1978).

The three isozymes mentioned above are all cytoplasmic forms of the enzyme. Of great interest has been the more recent discovery of non-cytoplasmic CAs. These include a membrane-bound form which has been found in both kidney and lung (CAIV), a mitochondrial form (CAV) and a secreted salivary form (CAVI) (see Fernley, (1988) for a review of the non-cytoplasmic CA isozymes). A seventh isozyme (CAVII) has been proposed to exist based on the analysis of genomic clones containing a carbonic anhydrase-like gene (Montgomery *et al.*, 1986).

1.1.2 General characteristics of the carbonic anhydrases

The cytoplasmic isozymes CAI, II and III are the best characterised of the carbonic anhydrases. They are all monomeric zinc metalloenzymes with a molecular weight of 28-30 kd and possess 259 or 260 amino acids. The essential zinc atom lies at the bottom of the active site cleft and is coordinated by three histidine side chains (numbers 94, 96, and 119 which are invariant in all known CA isozymes) and one H_2O or OH^- ion (Nostrand *et al.*, 1974; Kannan, 1980; Eriksson, 1988). It is thought that the zinc ion binds a water molecule and acts to stabilise the -OH group formed, at a pH where OH^- is not usually present in quantity. It is this hydroxyl ion which then adds to the CO_2 substrate. The zinc therefore acts to facilitate the generation of an attacking base in the reaction (Woolley, 1975). Structural considerations together with comparison of protein sequence of CAs from a number of species have identified those residues needed for either catalytic activity or structural integrity of the protein. These are shown in Fig. 1.1. The active site residues are found on four of the seven exons, but apart from this there is no strong correlation between exon structure and functional domains of the protein (Venta *et al.*, 1985). (see Fig 2 which shows the exon to structure relationship of CAII).

Although their general structure is similar, the cytoplasmic isozymes vary considerably in their catalytic properties. CAI has a CO_2 hydration turnover number of approximately 2×10^5 s^{-1} and will also catalyse the hydration of acetaldehyde though at a thousand-fold slower rate (Khalifa, 1971, Sanyal and Maren 1981). CAII has a turnover number of 10^6 - the highest known turnover number of

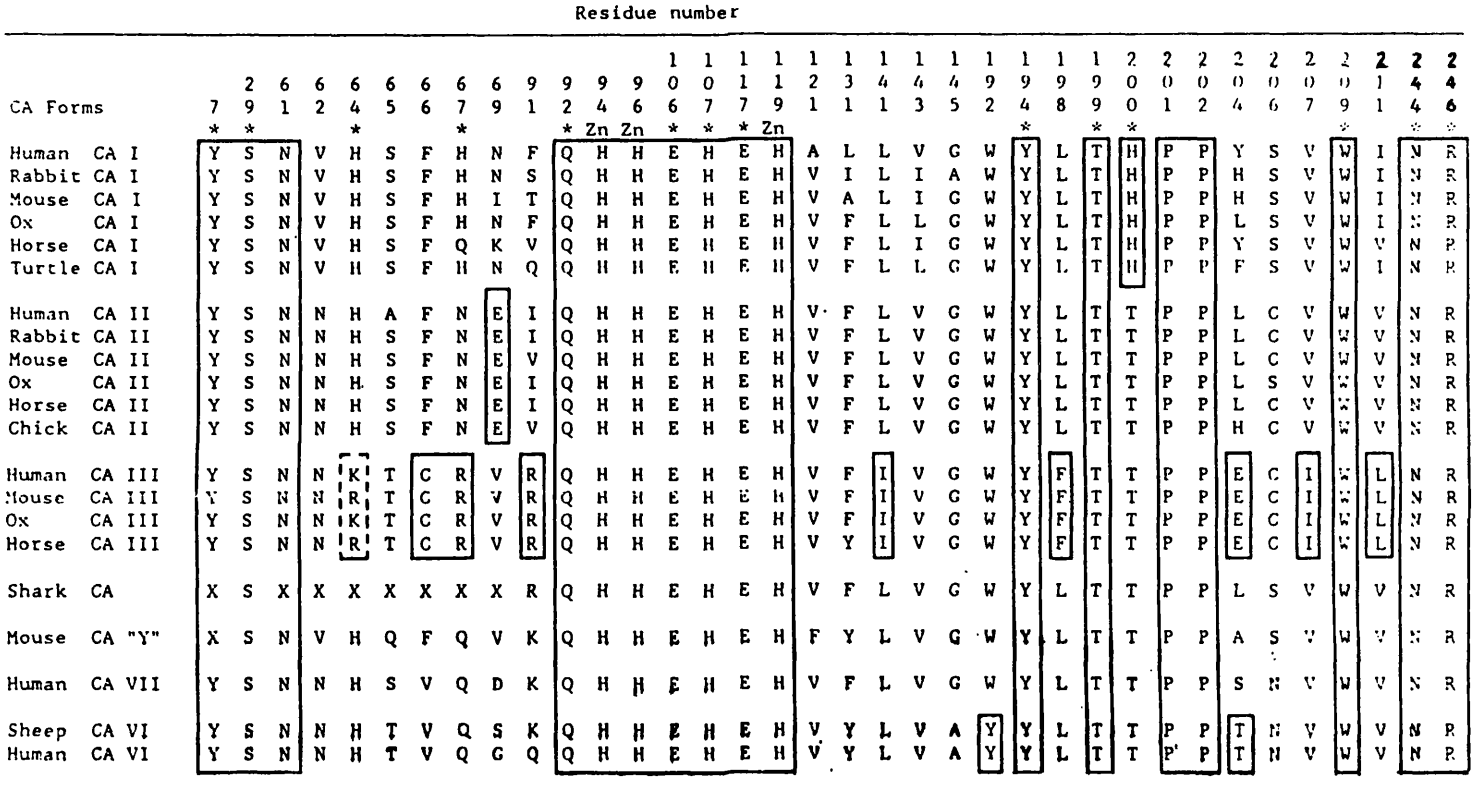


Fig. 1.1 Conserved and active site residues of carbonic anhydrase. Residues common to all amniote CA sequences have been boxed. Other boxes indicate invariant and unique residues for particular isozymes. Residues thought to be hydrogen-bonded to Zn-bound solvent molecule, or to the three Zn-liganded His residues (Zn) are indicated by an asterisk (*). (From Tashian, 1989).

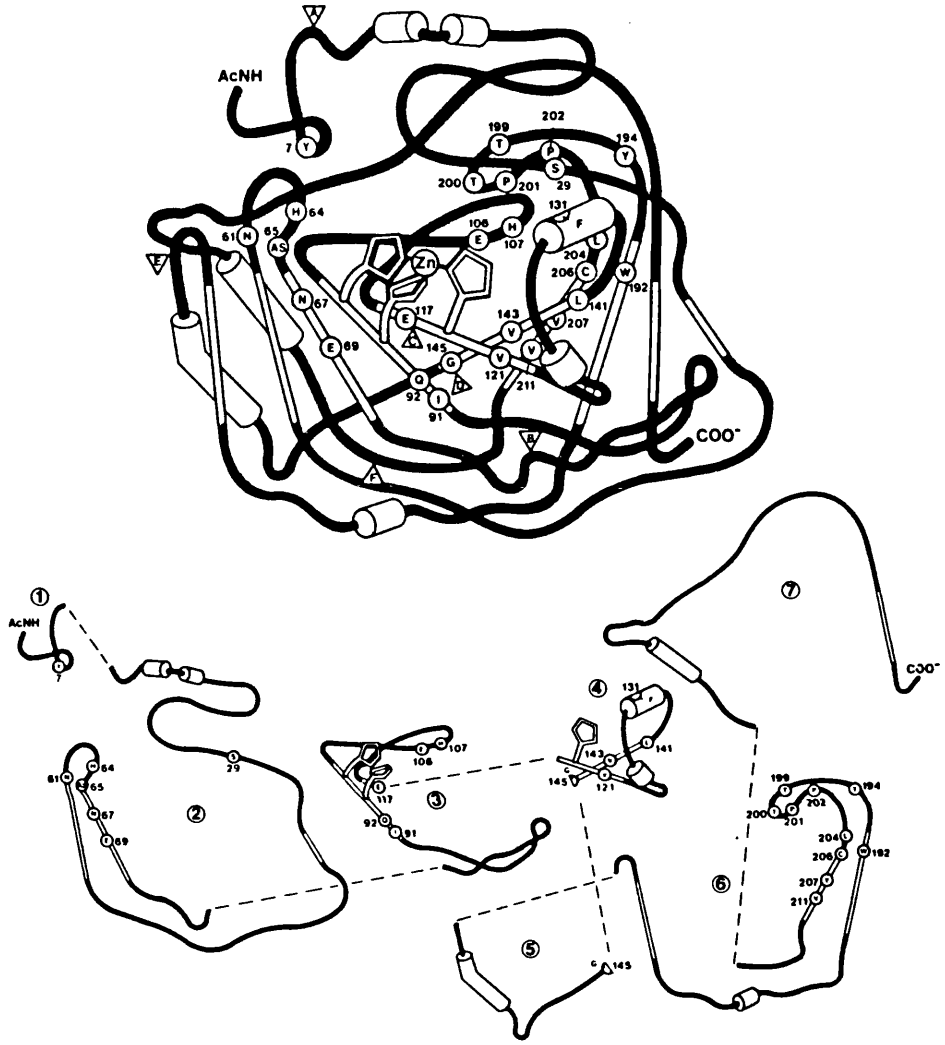


Fig. 1.2 Exon to protein domain relationship of carbonic anhydrase. The figure depicts human carbonic anhydrase II which has an almost identical crystal structure to CAI. Helical structures are depicted by cylinders and ß-segments by unshaded straight sections. Residues whose side chains project into the active site are designated by the one letter code of Dayhoff. The positions of the introns are shown by triangles in the intact molecule and the seven exons have been drawn below, separated in an exploded form. The numbering of the residues is based on the human CAI sequence (From Venta *et al.*, 1985).

any enzyme (Khakifa 1971, Sanyal and Maren, 1981), while the catalytic efficiency of CAIII is much lower than this with a turnover number of approximately 4 X 10^3 (Tu *et al.*, 1983). This isozyme lacks the His 64 residue found in all other known CAs. This may be responsible for the rather low activity of this isozyme and may also be responsible for its resistance to inhibition by sulfonamides. The non-cytoplasmic CAs differ from the cytoplasmic CAs in being of different sizes and/or post-translationally modified. These features are described in the next section which deals with each isozymes distribution and possible metabolic role.

1.1.3 The distribution of the isozymes

CAI is the second most abundant protein in the red blood cell at about 10ug/mg haemoglobin, (Nyman, 1961; Funakoshi and Deutsch, 1971) and there are indications that in this location it may be complexed with a 24,000 molecular weight protein similar to the band 3 spectrines or glycophorin A (Shäfer and Dietsch, 1984). CAII is also found in this tissue but at only 1/5 of this concentration (Nyman, 1961). CAI is not however a universal feature of the mammalian reticulocyte, for example, ox blood contains only one type of carbonic anhydrase - the high activity CAII (Lindskog, 1960).

CAI has also been identified in various parts of the alimentary canal. In the rodent, CAI has been shown to exist at reasonably high levels in the caecum and proximal colon and at lower levels in the small intestine (Carter and Parsons, 1971). More recently, more sensitive methods of immunohistochemistry has shown the isozyme to exist in particular cell types in a number of other tissues (Spicer *et al.*, 1984; Tashian *et al.*, 1984).

CAII has a more widespread tissue distribution than CAI; examination of a number of mammalian species shows that it is invariably present in red blood cells, but almost every tissue examined contains the enzyme (though it may only be

found in particular cells) (Tashian *et al.*, 1984).

Interestingly several of the sites containing CAI are those cell types from which CAII is absent. These include (in humans) vascular endothelium (Kumpulainen and Korhonen, 1978; Ryan *et al.*, 1982), myoepithelial cells of the secretory coil of the sweat gland (Briggman *et al.*, 1983), luminal and basal cells of the coiled and straight segment of the sweat gland duct (Spicer *et al.*, 1982; Kumpulainen, 1981), non goblet columnar cells in the colon (Lonnerholm, 1984; Spicer *et al.*, 1982), corneal endothelium (Wistrand, P. J. 1984b) and submandibular gland acini in the rat (Spicer *et al.*, 1984). This pattern is not universal and the cellular distribution of CAI and II often overlap so it is unclear as to what specific role each isozyme might play. It is fairly clear however that the rapid production of H^+ and HCO_3^- is required for a number of processes involving secretion and pH balance. In all of these processes, bicarbonate ions or protons may be required as counter ions in transport across membranes. CA is thus acting in a secondary role providing the ions which other cellular transport systems use to generate pH or ionic gradients across cell membranes. See Fig. 1.3.

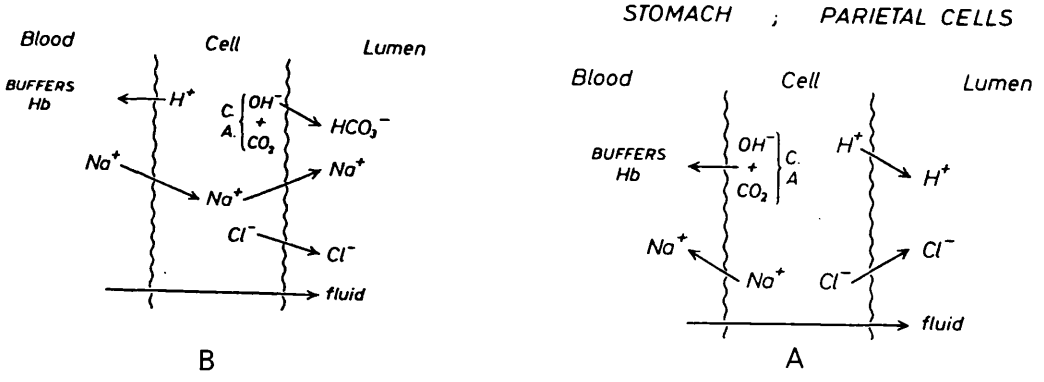


Fig. 1.3 The role of CA in secretion. A) The role of proton separation and carbonic anhydrase in gastric secretion. Primary process is linkage of active Cl^- movement and H^+ production. B) A general scheme for secretion of pancreas, salivary glands, cerebro-spinal fluid, aqueous humour and sweat. Sodium secretion linked to ATP-ase activity operates in concert with ^-HCO formation. (From Maren 1984)

CAIII is found in muscle, notably in type 1 fibres (Register *et al.*, 1978) and the male rat liver where it appears to be under hormonal control (Carter *et al.* 1981). Recent work also indicates that it is found in adipocytes and the notochord (Lyons *et al.*, 1990)

Less well characterised than CAI, II and III (though attracting increasing interest) are the non-cytoplasmic carbonic anhydrases. The membrane bound CAIV is found in both lung and kidney (Whitney and Briggle, 1982; Wistrand and Knuuttila, 1989) and the isozyme appears to be located on the luminal surface of the brush border of the proximal kidney tubule and of the pulmonary endothelial cells of the lung (Ryan *et al.*, 1982). Recent work shows the enzyme to be bound to the membrane via a phosphatidylinositol-glycan linkage (Sato *et al.*, 1990).

Mammalian hepatocyte and kidney mitochondria exhibit low levels of a carbonic anhydrase termed CAV contained entirely within the inner mitochondrial membrane (Dodgson *et al.*, 1983). The protein has been partially sequenced showing that it lacks the first 21 N-terminal amino acids found in the other CAs (Hewett-Emmett *et al.*, 1986). In this location, it is thought that the enzyme is playing a physiological role distinct from those already mentioned (Dodgson *et al*, 1984; Dodgson and Contino, 1988).

The salivary isozyme (CAVI) was first isolated in 1979 and has an apparent molecular mass of about 45,000. This is significantly larger than the cytoplasmic CAs. However much of this is due to the fact that the enzyme is glycosylated (Murakami and Sly, 1987; Fernley *et al.*, 1988a). Purification and sequencing of the protein has shown it to have 307 amino acid residues (compared to 260 for the cytoplasmic isozymes). This additional length is due to a 45 residue carboxy-terminal extension which may be responsible for its secretory properties (Fernley *et al*,. 1988b). It is unclear what the role of CAVI in saliva may be, although it is a high activity form like CAII (Fernley 1988a) and may play a role in the formation of bicarbonate and buffering in the

oral cavity.

1.1.4 Clinical aspects of the carbonic anhydrases

1.1.4.1 Deficiencies in CA levels

Familial deficiency states of CAI in which the enzyme is absent from erythrocytes has been reported in both humans (Kendall and Tashian, 1977) and the pigtailed macaque (Ferrell *et al.*, 1981). In both these instances there were no clinical symptoms. This asymptomacy is probably due to the presence of CAII which is probably carrying out the same role, or capable of substituting for CAI, in the red blood cell. It should be pointed out however that the other tissues in which CAI is present were not examined for presence of the enzyme. By contrast CAII deficiency has been shown to be responsible for symptoms of osteopetrosis with renal tubular acidosis in both humans (Sly *et al.*, 1983) and mice (Lewis *et al.*, 1988a). Changes in the level of CAI found in the red blood cell has been found in a number of medical states. Patients suffering from thyrotoxicosis have been shown to have a CAI concentration only about one third that of normal controls (Funakoshi and Deutsch, 1971), while a similar drop in CAI levels has been found in premature babies suffering from respiratory distress syndrome (Sell and Petering, 1974). A low activity form of CAI has also been reported to exist in erythrocytes of some patients suffering from renal tubular acidosis (Kondo *et al.*, 1978) and primary aldosteronism (Kondo *et al.*, 1984).

The apparent asymptomacy of CAI deficiency coupled with the fact that many species lack the enzyme in erythrocytes raises the question of the necessity of CAI (at least in the red blood cell) for normal physiological function. It should also be borne in mind that the various inhibitors used clinically (either as CA inhibitors, or as antibiotics) are often applied at concentrations which should be sufficient to inactivate CA in most of the sites discussed above.

Although side effects of these drugs are seen, often requiring administration of supplementary amounts of sodium and potassium salts to balance losses through poor kidney reabsorption, it seems surprising (given the numbers of processes in which CA has been implicated) that they are not more toxic. This is probably a reflection of the fact that spontaneous hydration, dehydration and diffusion of CO_2 is sufficient for the unstressed metabolism.

1.1.4.2 CA inhibition in clinical treatment

Inhibition of CA activity is the basis of a number of clinical treatments including the treatment of gastric and duodenal ulcers, hydrocephaly and oedema. By far the most common use of CA inhibitors however is in the treatment of glaucoma, accounting for over 95% of prescribed usage (Wistrand, 1984b). The enzyme is required in the process of aqueous humour formation in the eye, and inactivation causes a reduction in intraocular pressure by reducing the rate of entry of sodium and bicarbonate into the aqueous humour (Maren, 1988).

1.1.5 Evolutionary relationships of the CAs

The carbonic anhydrases are an ancient class of enzymes and have been found in organisms of all phyla examined. It is clear that the genes for these isozymes (*CA1, CA2, CA3*...etc) are descended via duplication events from a single ancestor and that over time the duplicated genes have evolved distinct set of functions, patterns of tissue distribution and temporal regulation. This makes the CAs a rich vein of data for evolutionary biologists and comparative studies of the isozymes and their genes have been carried out for some time. The identification of homologous isozymes in different species and comparison with tiger shark CA indicate that the duplication events which gave rise to these genes occurred at some time between the

divergence of the elasmobranchs (450 million years ago (mya)) and the divergence of the amniotes (300 mya)(Hewett-Emmett *et al.*, 1984). Fraser and Curtis (1986) on the basis of nucleic acid sequence comparison of mouse *CA1* and *CA2* cDNAs suggest a divergence time of 320 mya. Comparison of the protein sequences of CAI, CAII and CAIII suggests that *CAII* and *III* are more closely related to each other than either is to *CAI* (Lloyd et al. 1986., Hewett-Emmett and Tashian, 1990).

Until recently much less was known about either the protein sequence or gene localisation of the non-cytoplasmic CAs. In last three years however these isozymes have been extensively purified and in the case of CAVI completely sequenced (Fernley *et al.*, 1988b) while CAIV and CAV has been partially sequenced (Zhu *et al.*, 1988; Hewett-Emmett *et al.*, 1986). Not surprisingly considering their different subcellular location, immunogenic and physical properties, such comparisons have shown these isozymes to be more distantly related to CAII than either CAI or III (Hewett-Emmett and Tashian 1990). Apart from the genes known to be associated with identified proteins, molecular genetic studies have identified recombinant clones encoding two other potential CAs these have been designated CAVII (Montgomerey *et al.*, 1987) and CAY (cited in Tashian, 1989).

The chromosomal location of several of the genes for these isozymes has been determined. *CA1, CA2* and *CA3* are known to be tightly linked within a few hundred kb of each other in both mouse and man (Kearney *et al.*, 1987; Venta *et al.*, 1987). In man they lie on the long arm of chromosome 8 (8q22) (Edwards *et al.*, 1986a,b; Davis *et al.*,1987; Nakai *et al.*, 1987) while the mouse genes lie on chromosome three (band 3A) (Eicher *et al.*, 1976; Beechy *et al.*, 1990). Other members of the family have been assigned to other chromosomal locations, *CAY* being found on chromosome 1 (Tashian 1989) and CAVII on chromosome 16 in humans (Montgomery *et al.*, 1987).

1.2 Erythropoiesis

1.2.2 Erythropoietic Differentiation

The mature mammalian erythrocyte is an enucleate cell lacking RNA and incapable of regeneration. Such cells are just one of a number of cell types - macrophage, granulocyte and lymphoid produced from a single pluripotent haematopoietic stem cell (Weatherall and Clegg, 1981) (Fig. 1.4).

During the process of differentiation an undifferentiated stem cell receives a stimulus - such as the hormone erythropoietin - which directs it towards the erythrocytic series. The individual cells moving down this pathway in this early stage of differentiation are not distinguished by any visible morphological changes, but can be distinguished by their appearance following *in vitro* culture e.g. burst forming units (BFU-E), colony-forming units (CFU-E) (Dexter, 1979; Lajtha, 1989). This process in which cells become irreversibly committed is generally termed determination and is associated with a general loss of regenerative capacity. Once determined, the stem cell becomes a proerythroblast the first morphologically distinct cell type of the erythrocytic series, following which a process of maturation takes place (Fig. 1.4). During maturation a fixed number of cell divisions occurs and visible phenotypic changes take place as erythroid-specific proteins (such as spectrin, glycophorin A and the globins) accumulate. The nucleus is lost at about the same time as the cell is released into the circulation. The reticulocytes, which in a normal human make up only 2 % of the circulating red cells, eventually lose their RNA to become mature erythrocytes.

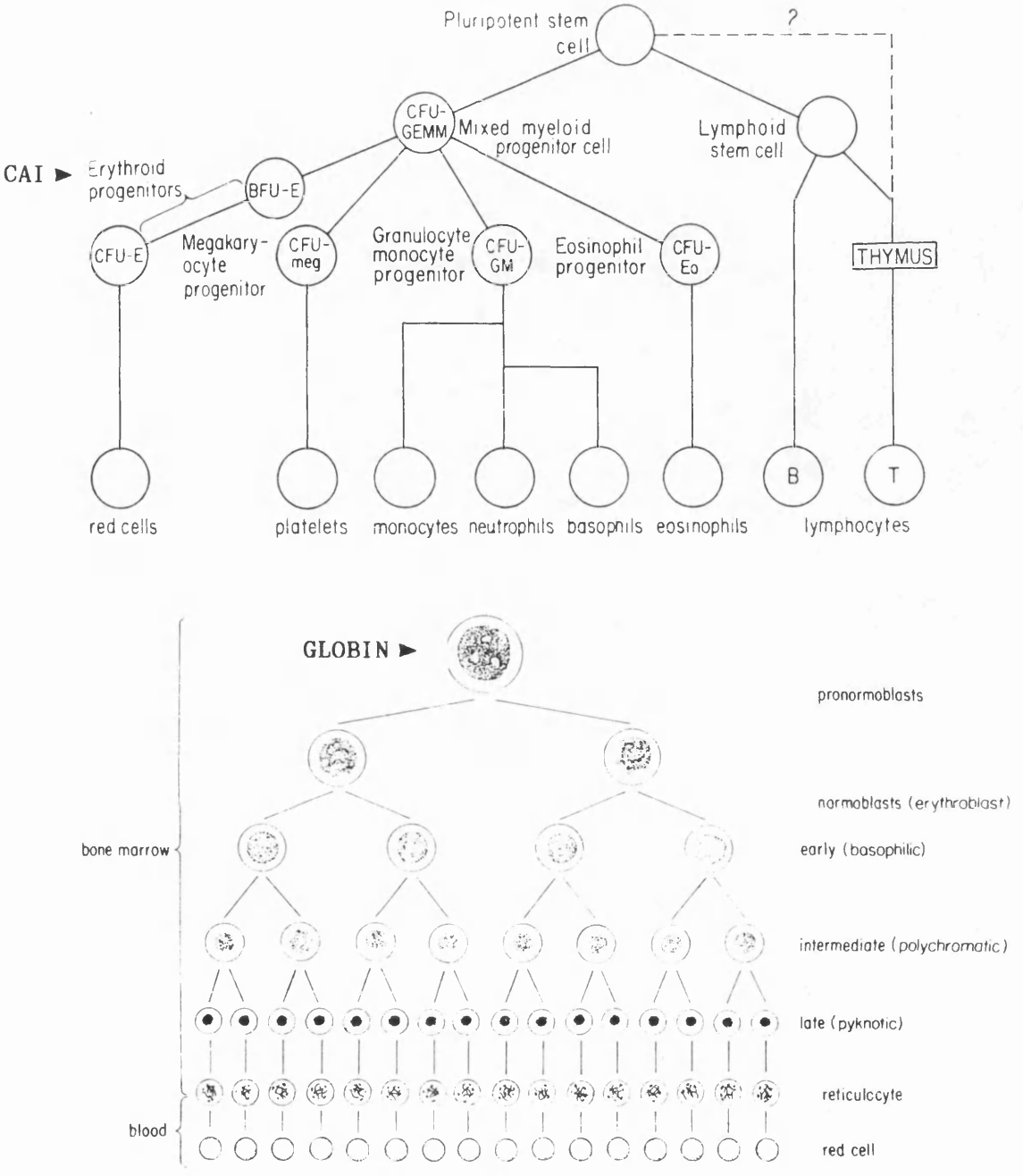


Fig. 1.4 The process of erythropoeisis. The erythroid lineage is one of a number of cell types formed from a regenerating pluripotent haematopoietic stem cell. Once committed to the erythroid lineage cells go through a number of stages such as BFU-E and CFU-E till the erythroid pronormoblast is reached, following which a visible process of maturation takes place culminating in the loss of the nucleous and extrusion into the circulation.

Although differentiating haematopoietic stem cells can be isolated from bone marrow (Dexter, 1979; Lajtha, 1979), the extent to which various lineages are "committed" or can be re-dedicated to another lineage is difficult to assess. This problem of determination *vs* pluripotencey is of course not specific to the haematopoietic system, but is a more general problem encountered in developmental biology. Answers to such problems often require (in the absence of obvious visual distinctions between differentiation pathways) the identification of gene products which can act as specific "markers" for that lineage. In this respect the globins are particularly unhelpful as indicators of lineage commitment, being produced relatively late in the pathway of erythropoiesis at the pronormoblast stage. CAI on the other hand has been identified in cells of the BFU-E stage or earlier, prior to the appearance of other red cell markers such as spectrin or glycophorin A (Villeval *et al.*, 1985). Because of this early activation, it is hoped that isolation of the *CA1* gene and characterisation of its regulatory mechanisms may provide more information regarding the early processes of erythroid commitment.

1.2.3 Developmental Changes

The site of erythropoiesis in man (as in most other vertebrates) changes throughout ontogeny. Erythropoietic progenitor cells are first found in the foetal yolk sack. After 4 or 5 weeks however erythropoiesis shifts to the developing liver and by 12 weeks this is the exclusive site of erythropoiesis. By the twentieth week of fetal life onwards, the liver is gradually replaced by the adult site of erythropoiesis - the bone marrow. The spleen also shares this role during later foetal life (Weatherall and Clegg 1981, Karlsson and Nienhuis, 1985).

In parallel to this change in the site of erythropoiesis the type of globin found in circulating erythrocytes also changes. Each haemoglobin tetramer consists of two chains of α-like globin (ζ- or α-globin) and two of β-like globin (ε-,

γ- or ß-/ε-globin). In the early embryo the $\zeta_2\varepsilon_2$ form predominates together with some of the $\alpha_2\varepsilon_2$. By the twelfth week of foetal life however this has been almost exclusively replaced by the foetal haemoglobin $\alpha_2\gamma_2$. Late foetal life sees a third change or "switch" in which the foetal γ-globin is replaced by adult ß-globin (Karlsson and Nienhuis, 1985).

Despite the similarity in timing between changes in the site of erythropoiesis and changes in the type of globin produced, there does not seem to be a causal link between the two. Thus erythroid cells from different haematopoietic tissues can produce both foetal and adult globins (Stamatoyannopoulos *et al.*, 1987). The timing of the switch may be dictated by some internal mechanism of the erythropoietic cells themselves rather than response to external stimuli (Wood *et al.*, 1985; Melis *et al.*, 1987)

CAI levels also change during development, being virtually absent in early foetal life but rising markedly from about 34 weeks of gestation (Boyer *et al.*, 1983). The fact that CAI and ß-globin levels both increase in late foetal life prompts speculation that common genetic control elements may be involved and indeed CAI deficiency in adult life is often associsated with persistence of foetal haemoglobin (Eng and Tarail, 1966; Alter *et al.*, 1976; Sheridan *et al.*, 1976). It should be noted however that the developmental changes in CAI levels differ somewhat from that of globin in being relatively gradual and prolonged, continuing for several years after birth (Weatherall and McIntyre, 1967; Sell and Petering, 1974; Boyer *et al.*, 1983), while the switch from γ- to ß-globin is more rapid (Wood *et al.*, 1977).

1.3 The Control Of Gene Expression

The machinery which directs eukaryotic gene transcription is complex and must allow expression in the correct temporal and cell-type specific manner. Every protein-coding gene has a unique sequence, but beneath this diversity it is apparent that many genes, together with the elements associated with modulating their transcription share common features. The gene itself can, at its most basic level, be represented as a simple stretch of DNA awaiting transcription - the first and probably most critical stage for controlling expression of the gene as a protein. For simplicity, the various elements responsible for control of transcription can be divided into two parts: those which are an integral part of the gene - the *cis*-acting DNA sequences - and those which are part of the nuclear environment in which the gene is expressed - the *trans*-acting (usually protein) factors.

Cis-acting DNA sequences are conventionally divided into two types, promoters and enhancers - although it is clear that the term enhancer can embrace sequences which operate in different ways. The essential difference between these two classes of sequence is that promoters have a fixed positional relationship to the gene, lying at its 5'-end, while enhancers have a looser relationship and can exert their effects from considerable distances (for review see Serfling *et al* 1985; Dynan 1989;).

The promoter region is responsible for accurate initiation of transcription, allowing association of RNA polymerase II with a specific region of the DNA in a multi-protein initiation complex. This complex contains at least five proteins. One of these, termed TFIID, binds to the highly conserved TATA box sequence motif found 25-30 bases upstream from the point where transcription is initiated (Breathnach and Chambon, 1981; LaThange and Rigby, 1988). Other common elements found upstream of the site of initiation include the CAAT box and often a GC rich element GGGCGG (Lee *et al.*, 1987). These sequences are less highly

conserved and are usually found at a distance of up to 100 bases from the transcription start point. Other less common elements can be found in the upstream region which have been implicated in mediating the responses of the gene to particular signals, for example hormone responsiveness (Evans, 1988) or response to heat shock (Davidson *et al.*, 1983).

The term enhancer was first used to describe a region of DNA found in the genome of the eukaryote virus SV40. This region could not in itself act as the site of initiation of RNA transcription but, when linked in *cis*, could increase (enhance) the level of transcription from not only the SV40 promoters with which it is normally associated, but the promoters of non-SV40 cellular genes such as ß-globin. It was also found that it could exert its effect across several thousand bp of intervening sequence and in either orientation relative to the gene (Banerji *et al.*, 1981).

1.3.1 Transcription Factors

The last few years has seen the adoption of techniques designed to examine interactions between nucleic acids and proteins. For example the bandshift assay (in which proteins bound to DNA retard its migration through polyacrylamide gels, Fried and Crothers, 1981; Strauss and Varshavsky, 1984) and DNA footprinting (where bound proteins protect DNA from nuclease digestion, Galas and Schimitz, 1978). These techniques have shown that many of the *cis*-acting sequences identified from mutagenesis studies as being important in transcription exert their effect in combination with sequence-specific, DNA-binding proteins. These proteins can be ubiquitous, found in all cells, or restricted to particular cell types.

Purification and cloning of transcription factors has revealed that they are diverse in their physical and chemical properties. They have been found to range in size from 36 kD to 100 kD. They can act as dimers (Landshulz *et*

al.,1988; Kumar *et al.*, 1988) or monomers (Kadonga *et al.*, 1987) and have been found to use a variety of cofactors (for reviews see: Jones *et al.*, 1988; Ptashne 1988; Mitchell and Tijan 1989; Latchman 1990). They can however be divided into a number of classes principally defined by the mechanism by which they interact with DNA (Struhl 1989). This is usually accomplished via a domain of basic residues which can interact with DNA in a sequence specific manner. Another region of the protein often consists of acidic residues which are thought to carry out the "activating" function of the transcription factor by interacting with the transcription initiation complex or other proteins (Ptashne 1988; Sigler, 1988). These two domains are often functionally independent such that mutants in the activating region which are no longer capable of modulating transcription are still able to bind to their DNA sequence. This independance allows elegant "domain swap" experiments to be carried out in which proteins containing the DNA-binding region from one factor and the activating region from another can be linked together (e.g. Green and Chambon, 1987; Kumar et al., 1987; Brent and Ptashne, 1985).

Promoter and enhancer regions of genes can contain binding sites for many different transcription factors. Some of these factors can show synergy or be mutually exclusive in their binding sites. Activity of a gene can thus be modulated in quite complex ways by the presence of such factors in different combinations.

Erythroid-Specific Transcription Factors

The last two or three years has seen the identification of an erythroid-specific transcription factor which is highly conserved from species to species both at a protein sequence level and in terms of the DNA sequence recognised (Trainor *et al.*, 1990). This sequence - centering on a GATA core sequence - is found, sometimes in multiple copies, at both the promoter and enhancer region of many, if not all,

globins and non-globin erythroid-specific genes (Evans *et al.*, 1988; Plumb *et al.*, 1989; Wall *et al.*, 1988).

This factor has been investigated by several research groups and has been given a variety of names such as GF1 (mouse) (Martin *et al.*, 1989), Eryf1 (chicken) (Evans *et al.* 1988), NF-E1 (human) (Wall *et al.*, 1988), EF-1 (mouse) (Barnhart *et al.*, 1989), and EF a (human) (Gumucio *et al.*, 1988) and has recently been designated GATA-1 (Pevny *et al.*, 1991). Both the mouse and human GATA-1 consist of a 413 amino acid polypeptide containing two Cys-Cys domains reminiscent of "zinc finger" structures (Tsai *et al.* 1989; Zon *et al.*, 1990), while the chicken factor is slightly smaller but shares extensive homology. (Evans and Felsenfeld, 1989; Trainor *et al.* 1990). Such zinc fingers have been found in a number of transcription factors and seem to be a common structural motif used in DNA binding (Evans and Hollenberg, 1988).

GATA-1 is found in the nuclei of mammalian erythroid tissues at all stages of development and its binding sites are found in erythroid genes expressed in embryonic, foetal and adult life (Plumb *et al.*, 1989; Wall *et al.*, 1988). It has more recently been found in a number of other cell types, including megakaryocyte and mast cell lineages (Martin *et al.*, 1990; Romeo *et al.*, 1990), HeLa, embryonic carcinoma cells and lymphoid cells (Perkins *et al.*, 1990). In all of these cases the factor is found at much lower levels than in erythroid cells.

At least one other mammalian erythroid transcription factor has been identified and termed NF-E2 (Mignotte *et al.*, 1989). Although this has not been cloned and analysed in the same detail as GF1, this factor appears to bind to the same site as the factor AP1 though with a slightly different specificity, and appears to increase in concentration following treatment of the mouse erythroleukaemic cell line MEL with DMSO to induce high-level expression of ß-globin (Mignotte *et al.*, 1989).

The promoters of many globins and other erythroid

specific genes contain a so-called CACCC motif, deletion of which causes 5 to 10 fold reduction of ß-globin transcription *in vivo* (Orkin *et al.*, 1984; Orkin *et al.*,1982). It is unclear whether the protein which binds this motif is erythroid-specific. Despite the assertion by Mantovani *et al.* (1989) that the factor which binds this sequence is erythroid-specific, the evidence is inconclusive and is contradicted by the finding that deletion of this motif in transient expression constructs causes a reduction in transcription in the non-erythroid mouse 3T6 cells, a fibroblast cell line (Dierks *et al.*, 1983).

As yet little progress has been made towards identification of those factors responsible for the developmental switching program seen in mammalian globins. More successful has been the search for developmental stage-specific proteins which control the similar switch seen in the chicken. Three related GATA motif binding transcription factors have been identified in this organism and one of these (NF-E1c) is found only in adult erythrocytes (Yamamoto *et al.*, 1990). In addition to this, non-GATA-binding proteins showing temporal variation have also been described. One erythroid-specific protein "BGP1", binding to a poly(dG) motif upstream of the adult ß-globin gene is only detected late in development (from 5-9 days of incubation) and is possibly a factor associated with the switch from expression of embryonic to adult globin (Lewis *et al.*, 1988a). No obvious counterpart of either the poly(dG) motif or BGP1 has been found in mammals. Another stage specific erythroid DNA-binding protein has also been reported in the chicken termed NF-E4 (Gallarda, 1989), while it has been proposed that the adult ß-globin expression levels are also regulated by two proteins PAL and CON which vary reciprocally during development. CON has also been described as the CACCC binding protein in this organism (Emerson *et al.*, 1989).

1.3.2 Chromatin structure and gene expression

The majority of DNA in the chromosome exists in a highly

condensed form, associated with the histones H1-H4 and highly resistant to attack by nucleases. Actively transcribed chromatin by contrast exhibits an increased sensitivity to digestion by nucleases (Weintraub and Groudine, 1976; Garel and Axel 1976) and an altered appearance when examined electron microscopically (Foe *et al.*, 1976). This sensitivity is usually interpreted as a reflection of the DNA existing in a more "relaxed" or open conformation, in which the gene is accessible to the machinery of transcription. It has been suggested that this change in conformation may be the result of acetylation or other modification of nucleosomal histones (Vidali *et al.*, 1978; Hutcheon *et al.*, 1980), absence of histone H1 (Huang and Cole 1984; Schlissel and Brown, 1984) or the presence of nonhistone proteins (in particular high mobility group proteins; Weisbrod and Weintraub 1979; Sandeen *et al.*, 1980; Gazit *et al.*, 1980) or methylation (see section 1.3.3.)

These more open regions or chromatin domains are often large, covering several tens of kb of DNA and may be demarcated by chromosomal matrix attachment regions (Cockerill and Garrard, 1986; Loc and Stratling, 1988).

More detailed analysis of these regions of active chromatin often reveals the presence of sites which are at least one order of magnitude more sensitive to nuclease digestion than the surrounding DNA (Elgin, 1981). These DNase I hypersensitive sites (HSS) frequently lie in the promoter and enhancer regions of genes, regions identified as being the sites of interaction with transcription factors (see Gross and Garrard, 1988 for review).

In the human ß-globin gene, for example hypersensitive sites exist in the body of the gene within an intragenic enhancer as well as at the 5' end of the gene in the -200 region and at the 3' end of the gene (Groudine *et al.*, 1983). The 5' site lies within the promoter region of the gene, while the intragenic site and the 3' site lie within regions defined as having enhancer activity (Antoniou *et al.* 1988; Behringer *et al.*, 1987; Kollias *et al.*, 1987). The other genes

in this cluster also possess HSS in their promoter regions. Both these and the sites near the ß-globin gene are more susceptible to digestion in cells of the developmental stage in which they are expressed (Groudine *et al.*, 1983; Tuan *et al*,. 1985).

In addition to these sites located in or close to the body of the gene, there is another set of sites which have to be considered in any survey of the globin genes. These sites, lying 5-20kb upstream of the ß-globin gene, are present in erythroid tissues at all stages of development and appear to contain sequences which are capable of defining a region of DNA as competent for expression in erythroid cells (Grosveld *et al.* 1987). Stable expression constructs of the ß-globin gene which contain these superhypersensitive sites are able to escape the so-called "position effect" of integration on expression (Gordon and Ruddle, 1985; Palmiter and Brinster 1985), producing high level copy number-dependant expression levels (Grosveld *et al.*, 1987). This region has been termed the Dominant Control Region (DCR) (Grosveld *et al.* 1987) (Talbot *et al.*, 1989) or the Locus Activating Region (Forrester *et al.*, 1987). Sequences which have similar properties to the DCR/LAR have been found in other genes, for example the human CD2 gene (Greaves *et al.*, 1989) the chicken lysozyme gene (Steif *et al.*, 1989). More recently the α-globin gene cluster has been shown to have a DCR-like element lying 30-50 kb upstream of it (Higgs *et al.*, 1990).

The sequences flanking the lysozyme gene are the same as the matrix attatchment sites which demarcate the boundaries of the genes chromatin domain (see above) and it would appear that these regions are also sites of interaction with topoisomerase II. This observation leads to the idea that these sequences serve to allow the introduction of torsional stress into regions of DNA, encouraging the changes noted in chromatin conformation and serving to open up the DNA duplex (Cockerill and Garrard, 1986; Loc and Stratling, 1988).

1.3.3 Methylation and Gene Expression

Within the AT-rich mammalian genome (60% A+T), the dinucleotide CG is found at only one fifth the frequency predicted from base composition (Russell *et al.*, 1976). Of those CG dinucleotides which are found, the majority are methylated at the 5- position of the cytosine pyrimidine ring. These two observations are probably related since the (spontaneous) deamination of methyl cytosine produces the base thiamine, producing G:T mismatches and subsequent depletion of cytosine (Green *et al.*, 1990; Wang *et al.*, 1982). Several lines of evidence indicate that this methylation of cytosine in CpG dinucleotides plays a role in gene regulation.

The first of these concerns the role of CpG rich HTF (HpaII tiny fragment) "islands" which make up about 1% of the nuclear DNA in vertebrates (Bird *et al.*, 1985). These HTF islands are G,C-rich regions which have been shown to be associated with the 5'region of most housekeeping genes and a number of tissue-specific genes. CpG dinucleotides within such islands are extensively demethylated and thus susceptible to digestion by the methylation sensitive restriction enzyme HpaII (Bird, 1987).

Many genes which are expressed in a cell-type specific manner do not contain CpG islands and exhibit methylation levels which vary from tissue to tissue, typically showing hypomethylation only in the tissue in which the gene is expressed (Cedar, 1988). CpG sites in the globin genes are usually undermethylated only in erythroid cells and generally show demethylation of specific sites close to genes only in the developmental stage in which they are expressed (Van der Ploeg and Flavell, 1980; Mavilio *et al.*, 1983; Enver *et al.*, 1988b).

The notion that methylation of CpGs may affect gene expression is supported by several other lines of evidence.

1.3.3.1 Azacytidine treatment of cells

The application of the nucleoside analog 5-Azacytidine (5-Aza-CR) to cells in culture is capable of producing enormous changes in expression levels of genes. This compound does not appear to exert its effect via mutagenesis, but rather to act as an inhibitor of DNA methylation (Jones, 1984). The changes in expression are often permanent and can lead to a complete reprogramming of the differentiated state of the cell. So for example the mouse embryonic fibroblast cell line C3H10T1/2 can be converted to myoblasts with a high frequency (Taylor and Jones, 1979), while teratocarcinoma-derived mesenchymal cells can be converted into epithelial cells (Darmon *et al.* 1984). This reproducible reprogramming of cell phenotype at a high frequency suggests that particular key genes which may be important determinants of a cells differentiation state are being activated by demethylation (Jones, 1984).

1.3.3.2 Introduction of methylated genes into cell lines.

One way in which it was hoped to determine whether methylation can directly affect expression was to compare exression levels from methylated and unmethylated genes following reintroduction into cell lines. This has been done with a number of tissue-specific and non-tissue-specific genes such as APRT and DHFR (Stein *et al.*, 1982) γ-globin (Busslinger *et al.*, 1983) thymidine kinase (Wigler *et al.*, 1981) genes. In all of these cases, it was found that methylation acted to block transcription in transient transfection assays. This block to transcription was also found to be localised such that particular sites, often (but not always) close to the 5'-end of the gene were more effective in reducing transcription levels when methylated. (Busslinger *et al.*, 1983; Keshet *et al.*, 1985; Murray and Grosveld, 1987).

1.3.4 Methylation and transcription factors

Because inhibition of transcription by methylation is often localised to areas known to bind transcription factors, the simplest hypothesis to explain this effect is that CpG methylation affects factor binding. This has been explored experimentally by looking at factor binding *in vitro* to methylated and non-methylated templates. In this way several factors have been shown to be prevented from binding by methylation of CpGs in their recognition site (Iguchi-Ariga and Schaffner, 1989; Kovesdi *et al.*, 1987; Becker *et al.*, 1987; Comb and Goodman, 1990). It is clear however that this is not always the case, the factor Sp1 is unaffected in either its binding or activation function by methylation of the CpG in its binding site (Iguchi-Ariga and Schaffner, 1989). Because Sp1 sites are often associated with housekeeping promoters, Höller *et al.* (1988) have suggested that constitutive Sp1 binding may prevent *de novo* methylation which would be important for maintaining gene activity. Apart from this immunity to methylation, many factors do not have methylatable CpGs in their recognition sequences.

The inhibition of gene expression associated with methylation may be a secondary effect since methylated DNA has been shown to affect the formation of active chromatin (Keshet *et al.*, 1986) while introduced methylated DNA can be transcribed for several hours until association with nucleosomes occurs, when transcription ceases (Buschhausen *et al.*, 1987). It has also been shown in cell fusion studies in which the δ- to ß-globin switch takes place *in vitro* that methylation of the δ-globin gene may occur *after* transcription has ceased (Enver *et al.*, 1988a).

1.4 Aims of this work

A number of questions of a fairly general or long-term nature arise from what has been described in this introduction.

How might expression of the *CA1* gene be controlled in a tissue-specific manner - what features might it have in common with other genes expressed in erythroid tissues and why is it activated early in erythropoiesis?

Is the mechanism of the developmental switch observed for CAI the same as that found in globins?

In the context of the Carbonic Anhydrase family, how does *CA1* compare with other CA genes already isolated and can this throw light on their evolutionary relationships?

With these questions in mind the aims of this project were:

1. To isolate and characterise cDNA clones encoding human CAI.

2. Use the cDNA to isolate (and then characterise) genomic recombinants containing the human *CA-1* gene.

4. Determine the linkage relationship between the three human cytoplasmic genes *CA1*, *CA2* and *CA3*, clustered on the long arm of chromosome 8 using pulsed field gel electrophoresis.

3. Assess the methylation state of *CA1* in expressing and non-expressing cell lines.

CHAPTER TWO: MATERIALS AND METHODS

2.1 Materials

Standard reagents: AR grade supplied by BDH, Fisons and Sigma.

Enzymes: Restriction enzymes from New England Biolabs, Betheseda Res. Labs. (BRL) and Anglian Biotechnology. Other enzymes used were from the same sources except RNase-free DNAase, and calf intestinal phosphatase from Boehringer Mannheim.

Electrophoresis reagents: Acrylamide from BDH.
Agarose and TEMED from Sigma.
Ultra-pure Urea from BRL.

Tissue Culture: media and cultureware - sources described in Section 2.22.

Miscellaneous: dNTPs, NTPs and NAP-5 columns supplied by Pharmacia.
Yeast tRNA and spermidine from Sigma.
NA45 paper from Schleicher and Schuell.
RNAsin from Promega.
rNTPs from Boehringer Mannheim.
Radiochemicals from New England Nuclear (NEN), or Amersham,
Nuclease-free Bovine Serum Albumin from BRL.
Guanidinium hydrochloride and thiocyanate from Fluka.

Common Buffers:

1 X TBE: 89mM Tris. HCl, 89mM Boric acid, 2mM EDTA.

1 X TE: 10mM Tris. HCl and 1mM EDTA, pH 7.4.

20 X SSC: 175.3g NaCl, 88.2g of sodium citrate / litre.

2.2 Bacterial strains and microbiology media

a) bacterial strains

HB101 (general recA host for plasmids): F- *supE*44 *hsdS*20(r_B-, m_B-) *recA*13 *ara*-14 *proA*2 *lacY*1 *galK*2 *rpsL*20 (Sm^r) *xyl*-5 *mtl*-1 λ-

LE392 (host for λcharon 4A): *hsdR- hsdM+ supE44 supF thi met lacY.*

JM101 (M13, pUC and Bluescript host): *supE thi* δ(*lac-proAB*), F'[*traD*36 *proAB*$^+$ *lacI*^{q} *lacZ*δM15], r+m+.

Q358 (host for λ2001): *hsdR supE.*

Y1090 (host for λgt11): *supF hsdR araD*139 *lon lac*U169 *rpsL trpC*22::Tn*10(tet)* pMC98

Y1088 (host for λgt11): *supE supF hsdR metB trpR tonA*21 *proC*::Tn5

b) microbiology media

Bacto tryptone, Yeast extract and agar from Difco

LB (Luria-Bertini medium)

Per liter:
10g Tryptone
5g Yeast Extract
10g NaCl

2YT media

Per liter:
16g Tryptone
10g Yeast extract
5g NaCl

H media

Per liter:
10g Tryptone
8g NaCl
12g agar

Glucose/minimal medium plates
(These plates select for retention of the F' episome of JM101 required for M13 propagation)

add to 1 litre of sterile M9 salts:
1ml 1M $MgSO_4$
1ml 0.1M $CaCl_2$
1ml 1M Thiamine HCl
10ml 20% glucose

M9 minimal salts
Per liter:
6g Na_2HPO_4
3g KH_2PO_4
1g NH_4Cl
0.5g NaCl

Agar plates made by adding 12g/lit. of bacto agar to the above broths.
Top agar overlays made by adding 7g/lit. agar
Top agarose overlays for phage lifts made by adding 7g/lit. agarose

SM (bacteriophage lambda storage medium)
Per liter:
5.8g NaCl
2g $MgSO_4 \cdot 7H_2O$
50ml 1M Tris HCl (pH 7.5)
5ml 2% gelatin solution

2.3 General methods

All solutions and (siliconised) glassware for nucleic acid work were sterilised by autoclaving at 15 p.s.i. for 20 min, unless heat labile. For RNA work the extra precaution of overnight pre-incubation at 37°C with 0.1%

diethylpyrocarbonate (DEPC) was taken. Disposable gloves were worn for all experimentation. The basis for most of the methods used in this thesis can be found in Maniatis *et al.* (1982), Sambrook *et al.* (1989) or Perbal (1988). All autoradiography involving ^{32}P was carried out by exposure to X-ray film (Fuji) with an intensifying screen at -70°C.

2.4 Genomic DNA preparation

Genomic DNA was isolated from tissue or cultured cells using protease K treatment followed by organic extractions, essentially by the method of Maniatis *et al.* 1982. Blocks of tissue weighing 1-5g were frozen with dry ice or liquid nitrogen. Single blocks were ground to a fine powder in the prescence of solid CO_2 using a coffee grinder. Powdered tissue (1g in 50 ml) or cultured cells (1 x 10^7 cells in 10 ml) were suspended in 100 mM EDTA, 0.5% SDS and 100 µg/ml Protease K. Protease K treatment was carried out at 55°C overnight. Following protease digestion, the solution was mixed gently with an equal volume of phenol/chloroform for 1 hour at room temperature. The phases were separated by centrifugation (5,000g 15min) and the aqueous phase removed, taking care to take as little interface material as possible. The extraction was repeated once more with phenol/chloroform and once with chloroform. DNA was precipitated by addition of 1/10 vol. 3M sodium acetate and 2 vols. ethanol. Large quantities of DNA (>250µg) were spooled onto a glass rod or hooked out as a floating precipitate, while smaller quantities were centrifuged. DNA was rinsed once with 70% ethanol and resuspended over several days in T.E. usually at a concentration of 0.2-0.5µg/µl.

2.5 Preparation of DNA in agarose for pulsed-field gel electrophoresis

For pulsed-field gel electrophoresis, the cells used to prepare the DNA were spun down and resuspended in phosphate-buffered saline at a concentration of 5x10^6-1x10^7/ml An equal

volume of molten 1% agarose (Ultrapure, Bio-Rad) was added, and the suspension pipetted into 80 µl block formers (Pharmacia LKB). Solidified blocks were digested in >10 volumes of 0.5% SDS, 100 mM EDTA, 100 µg/ml proteinase K at 55°C for at least 6 hours. Following proteinase digestion cells were washed (20 min 55°C) twice in >10 volumes TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA), twice in TE plus 2.5 mM phenylmethyl sulphonyl fluoride and twice more in TE.

2.6 Restriction enzyme digests

a) DNA in solution

Restriction enzyme buffers were those suggested by New England Biolabs and for most enzymes were: 10mM Tris.HCl (pH7.5), 10mM $MgCl_2$, 100µg/ml BSA together with 0, 50 or 100mM NaCl depending on the enzyme. For *Sma*I the buffer was 10mM Tris.HCl (pH 8.0), 20mM KCl, 10mM $MgCl_2$, 100µg/ml BSA. Some digests were carried out using Boehringer Mannheim incubation buffers for restriction enzymes, as recommended by the manufacturer. For most digests sufficient enzyme was added to give a 10-times overdigestion according to the unit activity given in the data sheet for the enzyme.

b) DNA in agarose for PFGE

For pulsed-field gel electrophoresis, half an 80 µl agarose block containing DNA (Section 2.10) was used per restriction enzyme digestion. After pre-equilibration in 400 µl of the appropriate reaction buffer, 40 units of enzyme was added to the block in a reaction volume of 150 µl and incubated for at least 4 hr. If a second digestion was required, using a different reaction buffer, the first buffer was removed and the block was equilibrated with the second buffer prior to digestion. Following digestion blocks were equilibrated with stop buffer/dye-mix and inserted into wells.

Stop buffer/dye mixes were: a)0.5 x TBE, 50mM EDTA 0.1% bromophenol blue, 0.1% xylene cyanol or b) 7M urea, 50% sucrose, 50mM EDTA, 0.25% bromophenol blue and 0.25% xylene

cyanol.

Size markers for PFGE were lambda concatamers purchased from Pharmacia LKB

2.7 DNA modification reactions

Blunt-ending of overhangs generated by restriction enzymes was needed for many cloning reactions. 5'-overhangs were filled in with the Klenow fragment of *E.Coli* DNA Polymerase I. Reactions were performed at 37°C for 30 min in a solution containing 10mM Tris.HCl, pH 7.5, 100mM NaCl, 10mM $MgCl_2$, 200µM of all four dNTPs and 0.25u/µl of Klenow. Reactions were terminated by heating to 65°C for 15 min. 3'-overhangs were blunted using either klenow polymerase or T4 polymerase using the same reaction conditions.

Dephosphorylation of plasmid vectors to prevent self-ligation was acheived using calf intestinal phosphatase (CIP). Typically 2-5ug of vector would be digested in 20ul. Following digestion, 1u of CIP was added to the reaction and the reaction allowed to proceed for 1 hr. The reaction was stopped by addition of EDTA to 20mM and incubation at 65°C for 20 min. The DNA was then precipitated from 0.3M NaOAc with 2 volumes of ethanol, washed with 70% Ethanol and resuspended in TE at 10-50ng/ul. Dephosphorylated vector was used for all blunt-ended cloning and most sticky-end ligations especially where insert was at low concentration.

10-50ng of vector were used for most subcloning operations. Ligations were carried out in 10ul of a solution containing 50mM Tris.HCl (pH7.4), 10mM $MgCl_2$, 10mM DTT, 1mM spermidine, 1mM ATP and 100µg/ml BSA. Reactions were performed at 4-8°C overnight and generally contained about a 3-fold molar excess of insert DNA to vector DNA together with 15 Weiss units of T4 DNA ligase.

2.8 Agarose gel electrophoresis and recovery of DNA fragments

DNA was resolved on 0.7-2.0% neutral agarose gels prepared in 18 X 18 cm or 9 X 9 cm flat-bed moulds. The gels

were generally run at 2V/cm overnight or at up to 10V/cm for shorter periods. Gels were made up and run in TBE buffer (90mM Tris.HCl, 90mM boric acid & 2mM EDTA). DNA fragments were visualised by ethidium bromide staining and uv transillumination.

DNA fragments required for subcloning procedures and for use as probes were isolated from agarose gels as described in Young et al (1985). In this method, DNA is run out in a normal agarose gel. A cut is then made in the gel below the band of interest and into is placed a piece of Schleicher & Schull NA45 ion-exchange paper. Electrophoresis is then resumed so that the DNA band runs into the paper to which it binds. The paper is then transferred to an Eppendorf centrifuge tube containing 0.45ml of a 1M NaCl/50mM arginine (free base) solution which is then heated to 70°C for 30-60 min to elute the DNA. The paper is then removed and, after a single phenol/chloroform extraction, the DNA is recovered from solution by ethanol precipitation at -70°C for 15 min.

2.9 Preparation and transformation of competent *E.Coli*

Competent *E.Coli* HB101 and JM101 cells were prepared using the $CaCl_2$ procedure as described by Maniatis *et al* (1982). 30ml of L-broth were inoculated with 0.5 ml of an overnight culture of the appropriate strain. This was incubated at 37°C with shaking until the culture reached an OD_{550} of 0.4-0.5. The culture was then chilled on ice for 10 min. The cell suspension was centrifuged at 500 rpm at 4°C in a Sorvall SS34 rotor. The bacterial suspension was resuspended in 15ml of an ice-cold, sterile, 50mM $CaCl_2$ solution and kept on ice for 15 min. The cells were then pelleted as before and resuspended in 3.0ml ice-cold, sterile, 50mM $CaCl_2$.

The resuspended cells were stored on ice for at least 2 hours before proceeding. For transformation 200-300ul of competent cells were aliquoted into pre-cooled tubes and up to 50ng of DNA added. This mix was left on ice for 30 min and heat shocked at 37°C for 2 min before plating out. The selection media onto which the transformed cells were plated

depended on the vector used, that for M13 is described below in Section 2.14. When using plasmids based on pBR322, the cells were pre-expressed in 0.5ml L-broth at 37°C for 30 min and then plated onto L-agar plates containing 125µg/ml ampicillin (amp) and incubated overnight at 37°C. When using pUC or Bluescripts plasmid utilising blue/white screening 50ul of 2% X-Gal and 20 ul of 100mM IPTG were spread onto amp plates. This was allowed to dry before spreading the transformed bacteria

2.10 Large scale and rapid plasmid preparations

Bacterial plasmids were isolated from large-scale, liquid cultures by the alkaline lysis method of Birnboim and Doly (1979) as modified by Ish-Horowitz and Burke (1981) and carried out exactly as described in Maniatis *et al.* (1982). Supercoiled plasmid DNA was purified on a CsCl-ethidium bromide gradients. Following centrifugation, the supercoiled DNA was recovered by puncturing the side of the centrifuge tube and drawing out the supercoil band (visualised under uv light) with a hypodermic syringe. Ethidium bromide was removed by repeated extractions with isopropanol equilibrated with saturated CsCl solution. Supercoiled DNA was removed from solution by repeated ethanol precipitations then quantified and analaysed for purity by spectrophotometry at 220-300nm and by gel electrophoresis. Rapid plasmid isolations were performed by the small-scale version of the above technique, also exactly as described in Maniatis *et al* (1982).

2.11 Pulsed-field gel electrophoresis

The two pulsed-field electrophoresis systems used in this work were the orthagonal field electrophoresis system (OFAGE)(Carle and Olson, 1984) using a double inhomogeneous field and the field inversion system (FIGE) (Carle *et al.*, 1986).

OFAGE was carried out on a LKB Pulsaphor system

(Pharmacia LKB). Gel concentration was 1.4% agarose in 0.5x TBE, run at 330v for 30-40 hours with a switching interval of 90 seconds. Electrodes were set as follows. Anodes: 90mm (North West), 90mm (North East). Cathodes: 5, 95, 180mm (South West) and 5, 95, 180mm (South East).

Field inversion was carried out using a commercial (Flowgen) or workshop built apparatus, both of which gave similar results. Gels (18cm in length) were 1.1-1.4% agarose in 0.5x TBE run at 7.5 volts/cm. Switching intervals were linearly ramped from 3s forward, 1s reversed at the start of the run to 120s forward 40s reverse after 48 hours (although runs were often terminated after 24 hours). If separation of smaller fragments was required, the switching pattern was restarted after 16 hr and allowed to run for a further 4 hr.

Transfer of DNA to hybridisation membranes was as for un-pulsed gels

2.12 Southern analysis

Southern analysis (Southern, 1975) was carried out essentially as suggested for Hybond N membranes (Amersham):

Following electrophoresis, agarose gels were rinsed briefly in distilled water. The gel was then immersed in a solution of 0.25M HCl and rocked gently for 20-30 min. The gel was rinsed again in distilled water and immersed in denaturing solution and rocked gently for at least 30 min. For Hybond N^+ charge modified nylon membranes (Amersham) denaturing solution was 0.5M NaOH and for Hybond N or Genescreen membranes (Dupont) was 1.5M NaCl/0.5M NaOH. Following denaturation DNA was transferred by standard capillary blotting (Maniatis *et al.*, 1982) to the hybridisation memebrane using denaturing solution as the transfer buffer. Following transfer the membrane was rinsed in 2 X SSC following which Hybond N^+ membranes were used for hybridisation. DNA on Hybond N and Genescreen membranes was UV cross-linked following removal of excess liquid by placing face down on a short wavelength (245nm) UV transilluminator (UV Products) for 2 min.

Prehybridisaton and hybridisation was in 4 x SSC, 10 x Denhardts solution, 0.1% SDS, 10 mM sodium phosphate pH 6.8, 6% polyethylene glycol (mol wt. 6000) and 100 µg/ml denatured salmon sperm DNA at 65°C. Prehybridisation was carried out for at least 2 hours. Fifty to two hundred nanograms of DNA probe was used per hybridisation, labelled by the random oligo primer technique (Feinberg and Vogelstein, 1983b) to a specific activity of $>5x10^8$ cpm/µg using a Boehringer Mannheim kit. After overnight hybridisation, filters were washed for 20 min in 2x SSC, 0.1% SDS once at room temp and once at 65°C and once in 0.1x SSC, 0.1% SDS for 20 min, 65°C. If probes contained repetitive sequence, sonicated, denatured human DNA was added to the hybridisation at a concentration of 20µg/ml.

2.13 Isolation of total and mRNA

Total RNA was isolated from human peripheral blood (provided by the Haematology Department, UCMHMS) of patients with an elevated reticulocyte count and also from cord blood. The blood cells were washed three times in phosphate-buffered saline (PBSa; 8g/l NaCl, 2g/l KCl, 1.5g/l Na_2HPO_4, 2g/l KH_2PO_4), the buffy coat removed and the remaining cells pelleted. The cell pellet was rapidly homogenised in 4 volumes of 6M urea/3M LiCl (Auffray & Rougeon, 1980) and left at 4°C overnight to precipitate the RNA. The RNA was collected by centrifugation at 16000g for 30 min at 4°C. The pellet was resuspended in 9ml of 8M guanidium.HCl, 0.4% sodium acetate and 0.1M ß-mercaptoethanol (Chirgwin *et al.*,1979). This solution was layered on top of 2.5ml 5.7M CsCl / 0.1M EDTA, pH8.0, in a 14ml polypropylene centrifuge tube. Following centrifugation for 18 hours at 33000 rpm and 20°C in a swing-out rotor the contents of the tube were removed to leave an RNA pellet at the bottom which was resuspended in 0.1% DEPC-treated DDW and recovered by two ethanol precipitations from 0.3M NaOAc (pH5.2).

The RNA was finally resuspended in DDW and its concentration and purity measured by spectrophotometry at

220-300nm.

mRNA was purified from total RNA using oligo-dT cellulose chromatography based on the method of Aviv and Leder, 1972. Briefly, total RNA was dissolved in DDW, heated at 65°C for 5 min adjusted to binding buffer concentration (0.5M NaCl, 10mM Tris.HCl, pH7.5, 1mM EDTA, 0.1% SDS) and loaded onto an oligo-(dT) column (oligo-$(dT)_{12-18}$ cellulose from BRL) equilibrated with the same buffer concentration. The RNA was recycled twice through the column and the column was then washed with binding buffer until the A_{260} of the eluant was <0.05. The column was then given a further wash with 0.1M NaCl, 10mM Tris.HCl, pH7.5, 1mM EDTA, 0.1% SDS and the bound RNA eluted with elution buffer (10mM Tris.HCl, pH 7.5, 1mM EDTA & 0.05% SDS). The eluted RNA was then mixed with an equal volume of 2 X binding buffer and the binding, washing and elution procedure repeated once more. Finally, the poly (A+) mRNA was adjusted to 0.3M NaOAc, pH5.2, and precipitated with 2.5 volumes of ethanol.

2.14 Northern analysis

Formaldehyde gels were used in northern blot analysis, as described by Fourney *et al.* (1988) with some modifications.

20-30µg of total RNA was resuspended in 5µl DEPC-treated DDW to which was added 25µl of an electrophoresis sample buffer made up of 0.75ml deionized formamide, 0.24ml formaldehyde, 0.1ml DEPC-treated DDW, 0.1ml glycerol, 0.08ml of a 10% bromophenol blue solution and 0.15ml 10 X MOPS buffer (0.2M 3-N-morpholinopropanesulphonic acid, 50mM NaOAc, 10mM EDTA, pH7.0). The mixture was heated at 65 C for 15-20 min and chilled on ice for 5 min. 2µl of a 1mg/ml ethidium bromide solution was added before the sample was loaded on a denaturing agarose gel. The gel consisted of 2.2g agarose, 18ml 10 X MOPS, 152.7ml DDW and 9.3ml 37% formaldehyde. Electrophoresis was for 16-18 hours in 1 X MOPS buffer at 30V.

Following electrophoresis the RNA was visualised by uv transillumination and photographed. The gel was then

soaked in 10 X SSC for 2 X 20 min before transfer of the RNA to a membrane filter, either GeneScreen Plus (DuPont-NEN) or Hybond N-Plus (Amersham), by the standard capillary blot technique (Maniatis *et al.*,1982) in 10 X SSC. After transfer GeneScreen Plus filters were baked at 80°C for 2 hours whereas Hybond N-Plus filters were treated with 0.05M NaOH for 5 min to fix the RNA. Both types of filter were pre-hybridised and hybridised in a solution containing 6 X SSC, 5 X Denhardt's, 0.5% SDS and 100µg/ml single stranded herring sperm DNA at a temperature determined by the probe. Pre-hybridisation was for 4-5 hours and hybridisation for 14-16 hours. Filters were then washed in 6 X SSC / 0.1% SDS twice for 10 min at RT and twice in 0.6 X SSC / 0.1% SDS for 15 min at 5 to 10 degrees below the hybridisation temperature depending on the experiment before autoradiography at -70°C.

2.15 Oligonucleotide synthesis and purification

Oligonucleotides were synthesised using the phosphoramidite method by either Dr C. J. Taylorson (London Biotechnology Ltd., U.C.L.) on a Cyclone DNA synthesizer (Biosearch, Inc) or Dr Y.-Z. Xu. (Biochemistry Dept., U.C.L.) on a Cruachem PS200 DNA synthesizer.

Oligonucleotides were cleaved from the support matrix on which they were synthesised using 1ml 0.88M ammonia. Ammonia solution was either injected into the synthesis column and left for 1h or the support material was decanted into an eppendorf tube and ammonia added. The ammonia/oligonucleotide solution was heated at 65°C for 4 h. or 50°C overnight in a sealed container. If the oligonucleotide synthesis was finished in the trityl-off state, the ammonia was removed by lyophilization and the oligonucleotide resuspended in 1ml TE buffer. The small protecting groups were removed from the oligonucleotide by passing the 1ml solution of oligonucleotide in TE down a NAP-5 (Pharmacia) desalting column. The column was pre-

equilibrated in 20% ethanol. The 1ml of TE was passed down the column and this was followed by 1.5ml of 20% ethanol. The eluant was lyophilized and the resulting purified oligonucleotide resuspended in 0.5ml of sterile double distilled water or TE. If synthesis had been finished in the trityl-on state, the oligonucleotide was purified on a NENSORB PREP cartridge (NEN-DuPont) which selectively binds fully elongated trityl-on oligos. The ammonia treated oligo was partially lyophilised to remove most of the ammonia (to give a volume of <0.5ml) and the volume made up to 4ml with 0.1M tri-ethylammonium acetate (TEAA). The TEAA solution was then loaded onto the cartridge and washed through with 10ml acetonitrile/0.1M TEAA (10:90 v/v) to remove failure sequences and salts. 25ml of 0.5% tri-flouroacetic acid was then passed through the column to hydrolise the trityl group. After another wash with TEAA the oligo was eluted in 35% methanol, 65% water, lyophilysed and resuspended in water or TE.

2.16 ^{32}P labelling of DNA

a) 5'-end labelling with T4 polynucleotide kinase.

T4 polynucleotide kinase was used with [γ-^{32}P]ATP to end-label oligonucleotides, linkers and 5' overhangs left by restriction enzymes. Dephosphorylation of 5' overhangs to increase labelling efficiency was carried out using calf intestinal phosphatase as in section 2.7. Oligos and linkers did not require dephosphorylation since they were synthesised without 5'-phosphates. Labelling was carried out in a 20 ul reaction containing 50mM Tris.HCl (pH7.6, 10mM $MgCl_2$, 1mM EDTA, 1mM spermidine, 100µCi of [γ-^{32}P]ATP and 1u/ul T4 PNK. Up to 50pmol could be efficiently labelled (equivalent to approximately 300ng of a 20-mer oilgo or 80ug linear pBR322). The reaction was terminated by heating to 65°C for 15 min. For oligonucleotides labelled in this manner to be used in hybridisations, the unincorporated nucleotides were removed by ion-exchange chromatography on DE-52 cellulose (Wallace *et al.*,1981). The labelled oligonucleotide

was diluted 10-fold with TE buffer and loaded onto a 0.3ml DE-52 column equilibrated with TE. The column was then washed with 2ml TE buffer, the unincorporated ATP eluted with 3ml TE / 0.2M NaCl and finally the labelled oligonucleotide eluted with three 0.5ml aliquots of TE / 0.5M NaCl.

b) Continuously labelled probes.

Continuously labelled DNA probes using plasmid DNA or isolated restriction fragments were also labelled by either nick-translation (Rigby *et al.*,1977) or by random priming (Feinberg & Vogelstein, 1983). For most library screening, insert fragments were purified away from vector to reduce background hybridisation, this was especially necessary if the vector contained cloned bacterial sequences.

In both nick-translation and random-priming, unincorporated nucleotides were removed from the labelled probe using 1ml sephadex G-50 (Pharmacia) columns. the reaction volume was made up to 150µl and loaded onto a column pre equilibrated with the same volume of TE. These were spun at 2000 rpm for 3 min on a bench centrifuge.

The specific activity of the labelled DNA was assessed by removing 1/100 of the reaction into 200µl of 10% trichloroacetic acid (TCA) / 5% sodium pyrophosphate and 50µg of BSA carrier. This mix was kept on ice for 15 min and the solution passed through a Whatman GF/C filter. The filter was then washed three times with a large volume of 5% TCA / 5% sodium pyrophosphate to remove unincorporated nucleotides. Following drying, radioactivity was measured by scintillation counting. All probes were denatured by boiling prior to addition to hybridisation solution.

2.17 Lambda and cosmid library plating and transfer to hybridisation membranes.

2.17.1 Lambda library lifts.

Human genomic clones were isolated from a human genomic library kindly provided by Dr T.H. Rabbitts, Laboratory of

Molecular Biology, Cambridge. This library (LeFranc *et al.*, 1986) was constructed using DNA from an EBV-transformed B-lymphoblastoid cell line, SH. The DNA was partially digested with *Sau*3AI and ligated into *Bam*HI-digested λ2001 bacteriophage vector (Karn *et al.*, 1984) and propagated on the *E.Coli* Q358 strain (Karn *et al.*, 1980).

cDNA recombinants were isolated from two separate λgt11 libraries constructed using human reticulocyte RNA and propagated in *E.Coli* Y1088 or Y1090.

An amount equivalent to 25-40000 pfu was added to 1ml of an overnight culture of host bacteria grown in 10mM $MgCl_2$ and 0.2% maltose. The phage were allowed to adsorb to the bacteria for 15 min at 37°C. 12ml of soft agarose (0.6% agarose, kept at 50°C to prevent setting) was then added, mixed and plated onto a pre-dried L-agar/10mM $MgCl_2$ plate in a 140mm Petri-dish (Sterilin). This was done for 12 seperate plates giving a total of about 480,000 pfu to be screened. After setting of the agarose, the plates were inverted and incubated at 37°C for 7-9 hours depending on growth and ensuring that the plaques remained sub-confluent.

Lifts were then taken from each plate by laying a 132mm Biodyne (Pall) or Hybond (Amersham) transfer membrane on the soft agarose and marking its position by stabbing through both filter and agar with a hot wire. A duplicate filter lift was made for each plate. After 90 sec, each filter was removed and laid, plaque face up, on Whatman 3MM soaked in denaturing solution (0.5M NaOH, 1.5M NaCl). After 5 min, the filters were similarly laid on 3MM paper soaked in 3M NaOAc, pH5.5 (Biodyne) or 0.5M Tris.HCl pH7.5, 1.5M NaCl (Hybond). After a further 5 min the filters were removed from the 3MM and allowed to air dry for 15 min. The filters were baked overnight at 80°C. (Biodyne) or UV cross-linked (Hybond) to fix DNA to filters.

2.17.2 Cosmid library lifts.

Three cosmid libraries were screened for genomic CAI clones. The first library was constructed in the vector

LoristB (Cross and Little, 1986) in the laboratory of Dr Peter Little (Imperial College, London) by partial *Hind*III digestion of human DNA inserted into the *Hind*III site of the vector. This library had already been plated and transferred onto hybridisation membranes. The second library was constructed in the vector cos202 using DNA from the lymphocytic leukaemia cell line HPB-ALL (Kiouissis *et al.*, 1987) and was provided by Dr Paul Brickell (U.C.L. Biochemistry Department). The third library screened was constructed in the vector Lorist 6 (Gibson *et al.*, 1987) and was a gift of Dr Terry Rabbits. This library was constructed by Dr Laki Buluwela (MRC Laboratory of Molecular Biology, Cambridge), by insertion of partially digested DNA from the cell line COLO 320 HSR into the *Hind*III site of the vector. Each of these libraries were propagated and screened in the same way apart from using different antibiotic selection. Lorist libraries were grown on kanamycin (30µg/ml) and the cos202 library on ampicillin (100µg/ml) containing media

Growth of colonies and transfer to filters.

The cosmid screening method used was essentially that of Little 1987. The library was spread directly onto hybridisation membranes (Hybond N, Amersham) which had been laid on the appropriate selection media (L-agar plus antibiotic) and grown till colonies were 0.5mm across. These "master" membranes were then removed and placed onto antibiotic plates containing 25% glycerol and left 2-3 h. This filter was removed and placed face up onto three or four filter papers, the top one of which had been pre-wetted with L.broth plus 15% glycerol. A second hybridisation membrane was pre-wetted by laying onto a selective plate containing 25% glycerol and laid carefully on the first (master) membrane covered with colonies. Three or four filter papers, the first one of which was pre-wetted with L-broth plus glycerol, were then laid on top of the second membrane. This "sandwich" was then compressed to transfer the bacteria. The filter papers were then peeled away and

registration holes were made through both membranes using a sterile needle. After this the duplicate membrane was then laid on the appropriate selective agar plate to regrow till colonies were easily visible.

Membranes were then placed (colony side up) on pre-soaked filter papers as follows:

3 mins 10% SDS.
5 mins 1.5M NaCl, 0.5M NaOH.
5 mins 1.5M NaCl, 0.5M Tris HCl (pH 8.0).
5 mins 2x SSC.

Hybridisation membranes were then air dried for 1 hr, UV treated to cross link DNA to filters (laid DNA face down on shortwave UV transilluminator for 3mins) and baked at 80°C for at least two hours.

The above process was repeated to give a duplicate hybridisation filter. After this a third filter was placed onto the master and the sandwich compressed and stored frozen at -40°C. Following screening this third filter could be removed and re-grown as a source of recombinants, always being replaced with a fresh sterile membrane wetted in glycerol containing selective media.

Following screening to detect recombinants the appropriate region of the filter was identified and colonies scraped from an area of 1-2cm diameter using a sterile loop. Bacteria were suspended in L-broth and re-spread on large petri dishes to give 50-200 colonies per plate. These were then treated as above for the secondary screening stage.

2.18 Hybridisation of plaque and colony lifts.

Following fixing of DNA to filters, hybridisation membranes were washed in 1 X SSC, 0.1% SDS to remove bacterial debris.

If nick-translated or random oligo labelled probes were used, filters were pre-hybridised at 65°C for at least 2 hours in a solution containing 4 X SSC, 10 X Denhardts

solution (50 X Denhardt's = 1% BSA (Pentex), 1% Ficoll and 1% Polyvinylpyrrolidone), 50mM sodium phosphate and 100µg/ml denatured sonicated salmon sperm DNA (6% polyethylene glycol was also used if increased signal strength was needed). Probe was added after denaturation to a fresh aliquot of the same solution. The probe was labelled to a specific activity of 1-5 X 10^8 c.p.m /ug and 100µg to 500µg was used in a hybridisation volume of 20-50ml. Hybridisation was carried out overnight at 65°C.

Following hybridisation, filters were washed once at room temperature and once at 65°C in 2 X SSC, 0.1% SDS and once at 65°C in 0.1 X SSC, 0.1% SDS for 20 min each wash. Filters were then autoradiographed at -70°C

For oligonucleotide probes, prehybridisation and hybridisation was carried out in 6 X SSC, 50mM sodium phosphate (pH6.5), 100µg/ml denatured herring sperm DNA and 5 X Denhardt's solution at a temperature determined by the base content of the probe to be hybridised to the filters. The optimum hybridisation temperature was approximated from the formula T_d = 2 C X no. of A:T base pairs (bp) + 4 C X no. of G:C bp, where T_d is the temperature at which ½ of the duplexes are thermally denatured in the presence of 0.9M sodium ions. The oligonucleotide probes used to screen the library were labelled and separated from unincorporated ^{32}P-γ-ATP as described above.

The labelled oligonucleotide in TE / 0.5M NaCl was then added directly to the hybridisation solution, made up exactly as the pre-hybridisation solution. Both pre-hybridisation and hybridisation were at the same temperature, the latter being left for 12-16 hours. The filters were first rinsed 3 times with 6 X SSC at 4°C and then washed twice in 6 X SSC for 30 min at RT. The filters were now rinsed in a tetramethylammonium chloride solution (3M tetramethylammonium chloride, 50mM Tris.HCl, pH8.0, 2mM EDTA and 1mg/ml SDS) at 37°C (Wood *et al.*, 1985b), and then washed twice for 20 min at 2-3°C above the hybridisation temperature. The filters were finally rinsed in 6 X SSC at

RT before autoradiography.

Each plaque on the plates producing a positive signal on both primary and duplicate filters as judged by autoradiography was picked into 1ml of phage storage medium (SM) which is 100mM NaCl, 8mM $MgSO_2$, 50mM Tris.HCl, pH7.5 and 0.01% gelatin. Each isolate was then titred, replated on single 90mm Petri dishes (Sterilin) at 200 pfu / plate and rescreened as above. Positive plaques were again picked and screened for a third time at 50 pfu / plate so that single positive clones could be isolated.

2.19 High-Titre phage stocks from plate-lysis

Several ml of high-titre (1 x 10^{10}) phage stocks could be prepared by plating purified phage isolates at sufficient concentration to produce confluent lysis on 90ml petri dishes. Host bacteria were infected with sufficient phage to just produce confluent lysis in soft agar overlays. Following overnight growth the overlays were mashed up with 2 ml SM using a sterile glass spreader. The agar was then transferred to 1.5 ml eppendorf centrifuge tubes and 15ul of chloroform added. The mixture was vortexed briefly and allowed to stand for 1 h at room temperature. The tubes were spun at 12000rpm for 10 min and the supernatant removed and stored at 4°C in the prescence of chloroform (0.3%)

2.20 Large scale preparation of λ bacteriophage DNA

Large scale preparation of DNA from the phage isolates was necessary to facilitate further analysis. The stocks were titred and 5 X 10^8 pfu of each isolate were allowed to adsorb to 1 X 10^{10} *E.Coli* Q358. The same ratio of pfu to host was also used when preparing DNA of an isolate from a library made in the Charon 4A phage vector using *E.Coli* LE392 as host.

The phage were adsorbed for 20 min at 37°C. The mix was then added to 200ml of L-broth and 10mM $MgCl_2$ and incubated with shaking at 37°C for 8-9 hours. 2ml chloroform was added and shaking continued for a further 15 min to ensure

complete lysis of bacteria before being centrifuged in a Sorvall SS34 rotor at 7000 rpm and 4°C for 10 min to pellet bacterial debris. 8g NaCl, 200µg DNAase and RNAase were then added and the mix kept at RT for one hour. Polyethylene glycol (Mol Wt 6000) was added to a concentration of 10% (W/V) dissolved very gently and kept 2 hr to overnight at 4°C. The solution was then centrifuged as previously, the pellet formed resuspended in 10ml of SM and extracted with 10ml of chloroform. 0.75g/ml CsCl was added to the resulting aqueous phase which was then centrifuged in 14ml polycarbonate tubes overnight at 35000 rpm and 4°C in a MSE Prepspin centrifuge. The CsCl forms a gradient which separates intact phage from other contaminants. The intact phage was visible as a thin blue band. This was removed by puncturing the centrifuge tube and drawing out with a hypodermic syringe in a volume of about 1-2ml. This small volume containing the intact recombinant phage was dialysed overnight at 4°C in one litre of dialysis buffer (50mM Tris.HCl, pH8.0, 10mM $MgCl_2$ and 10mM NaCl). To each 0.5ml of phage dialysate was added 25µl 10% SDS / 40µl 250mM EDTA and 30µg of proteinase K and the mix incubated at 65°C for one hour. This was then extracted once with phenol, twice with phenol/chloroform, once with chloroform and finally preciptated from 0.3M NaOAc (pH7.0) with ethanol. The recovered phage DNA was then quantified and analysed for purity by spectrophotometry at 220-300nm and by gel electrophoresis.

2.21 cDNA library construction

cDNA synthesis was carried out by the RNAseH method as described by Gubler and Hoffman (1983), but in a single reaction tube. 1/25 of each reaction was removed for scintillation counting to asess effficiency of synthesis reaction. First strand cDNA synthesis was caried out in a 50µl volume using 6µg $PolyA^+$ RNA in:

50 mM Tris HCl (pH 8.1)

140 mM KCl
6 mM MgCl
30 mM ß-mercaptoethanol
1 mM of each of dATP, dCTP, dGTP, dTTP
100 µg/ml oligo dT_{12-18}
50 units Placental RNAse Inhibitor
10 µCi [α-^{32}P]dCTP (800 Ci/mMol)
1u/µl AMV Reverse Transcriptase
React 1 hr 41°C

20 units of RNAseH was then added and incubated for 5 min at 44 C. The volume was then increased to 250µl with compoments of the second strand cDNA synthesis reaction:

20 mM TrisHCl(pH 7.5)
5 mM $MgCl_2$
10 mM $(NH_4)_2SO_4$
100 mM KCl
0.15 mM ß-NAD
0.1u/µl DNA polymerase I
0.05u/µl *E.coli* ligase
incubate 1 hr 12°C
1 hr 22°C

Following second strand cDNA synthesis, the reaction mixture was heat inactivated at 70°C for 15 min. The ends of the cDNA were blunted by addition of 10 units Klenow fragment of DNA polymerase plus 4 units T4 DNA polymerase and incubation for 20 min at 22°C. Two phenol/chloroform and one chloroform extractions were carried out before ethanol precipitation.

For the *Eco*RI methylation protection reaction the cDNA was resuspended in 20µl 100mM Tris HCl (pH 8.0), 10mM EDTA, 1.5mM S-adenosyl methionine and 20 units *Eco*RI methylase and incubated 5 hr 37°C. The reaction was heat inactivated for 15 min 65°C and ethanol precipitated.

For attition of linkers the cDNA was resuspended with 2 µg of kinased *Eco*RI linkers in a 10µl volume of:

50 mM Tris HCl pH7.8

10 mM $MgCl_2$
20 mM dithiothreitol
1 mM ATP
50 ug/ml BSA (nuclease free)
3 weiss units of T4 DNA ligase.

incubate 16 hr 10°C

Following ligation, the reaction was heat killed 15 min 65°C and the volume was made up to 50 µl with *Eco*RI restriction buffer. 200 units *Eco*RI was added and incubated 37°C 3 hr. The volume was increased to 150 ul with TE and the digested cDNA/linkers were then extracted twice with phenol/chloroform, twice with chloroform and twice with ether and the ether driven off by heating to 65°C. Immediately after heating, digested linkers were separated from cDNA on a sephadex G50 spun column, previously equilibrated with TE buffer. A second separation step was carried out using a "NACS prepac" (GIBCO-BRL) ion exchange column. The eluate from the spun column was mixed with 0.5 ml 0.2M NaCl/TE solution and applied to the column. The column was rinsed once with 1 ml 0.2M NaCl/TE and 150 µl aliquots of 0.3M, 0.4M, 0.5M and 0.75M NaCl/TE applied to the column. Elution of cDNA from the column was asessed using a hand monitor. A large rise in eluted material was detected with the 0.75m NaCl wash. this fraction was saved and a further 2 X 150 µl aliquots of 2 M NaCl/TE applied to the column and pooled with the 0.75 M NaCl/TE eluate.

1µg of dephosphorylated λgt11 arms (Vector Cloning Systems) was mixed with the eluted cDNA and co-precipitated with ethanol overnight.

Following co-precipitation, vector/cDNA was resuspended in 5µl ligation mix:

50 mM Tris HCl pH7.8
10 mM $MgCl_2$
20 mM dithiothreitol

1 mM ATP
50 ug/ml BSA (nuclease free)
1 weiss unit of T4 DNA ligase.
incubate 16 hr 10°C

In vitro packaging was carried out using a "Gigapack" from Vector Cloning Systems according to the manufacturers instructions. 2.5µl of ligation mix was added to contents of the freeze/thaw extract tube on ice. 15 µl of sonicated extract was then added, mixed gently and incubated for 2 hr at 22°C. 0.5ml of phage storage media (SM) was added followed by 20µl chloroform and mixed gently. Dilutions were plated out on *E.coli* Y1090 as in section 2.17 to titre library.

An amount equivalent to 1/4 of the original 6 µg of PolyA$^+$ was finally used in the packaging reaction, and this produced a total of 3.3 X 10^5 recombinants, with approximately 7% non-recombinants as assessed by blue/white colony screening.

2.22 Hybridisation to DNA in gel membranes

Hybridisation of labelled oligonucleotides (prepared as in Section 2.9), to plasmid and to λ phage DNA fractionated on agarose gels, was performed on dried gel membranes as described in Singer-Sam *et al* (1983). Following electrophoresis the gel was stained, photographed, trimmed of excess agarose and soaked in 0.5M NaOH / 0.15M NaCl for 30 min. The gel was then neutralised with 0.5M Tris.HCl (pH8.0) and 0.15M NaCl for 30 min, placed on two sheets of 3MM paper and dried at 60°C. The dried gel, resting on a single sheet of 3MM paper, was floated in DDW till the dried gel membrane could be carefully peeled off the 3MM paper and placed in a sandwich box for hybridisation. The DNA in the gel membrane was then hybridised, without pre-hybridisation, to the labelled oligonucleotide probe at the appropriate temperature for 6-7 hours (10 degrees below Td). The hybridisation solution contained 6 x SSC, 0.1% SDS and 100

µg/ml tRNA or denatured salmon sperm DNA. The gel membrane was subsequently washed twice in 6 X SSC, 0.1% SDS for 15 min at the hybridisation temperature before autoradiography at -70°C.

2.23 Rapid restriction mapping of DNA cloned in λ phage vectors

^{32}P labelled oligonucleotides complementary to the cohesive (cos) ends of λ phage (ON-L and ON-R) can be used to rapid restriction map DNA cloned in λ phage vectors. They are annealled to a partial digest of the recombinant and run out on an agarose gel, dried and autoradiographed, as described in Rackwitz *et al* (1984).

The λ DNA to be mapped was partially digested in a solution containing 1µg of λ DNA, 1 µl of 10 X restriction buffer, 0.5u of the enzyme diluted in DDW and the volume adjusted to 10 µl with DDW. This mix was incubated at 37°C. After 7 min two 2µl aliquots were removed and added to 1µl of 150mM EDTA on ice to stop further digestion.

The cos oligonucleoties, ON-L and ON-R (New England Biolabs) were 5'-end labelled with ^{32}P. ON-L is complementary to the left cohesive terminus of λ DNA and ON-R to the right terminus. 0.5ng of labelled ON-L or ON-R was added to the aliquot of partial digest and DDW and NaCl added to make the solution 150mM NaCl. The mixture was heated to 75°C for 2 min and immediately transferred to a 45°C waterbath for 30 min to anneal. 5µl of gel loading buffer (36mM Tris.HCl, pH7.7, 30mM NaH_2PO_4, 60mM EDTA, 50% glycerol and 0.1% bromophenol blue) were added and the mixture was loaded onto a 0.5% agarose gel. Electrophoresis was carried out at 1.5V/cm for 24 hours in 36mM Tris.HCl, pH7.7, 30mM NaH_2PO_4 and 1mM EDTA. Relevant ^{32}P labelled markers (lambda C^{I857} partial digests) were also run. The gel was washed briefly and dried onto Whatman DE-81 cellulose paper at 60 C before autoradiography at -70°C.

2.24 DNA sequencing in M13 phage by the dideoxy chain-

termination method

The ligation mixture of M13 vector and insert DNA was transformed into *E. Coli* JM101, made competent by the $CaCl_2$ procedure (Section 2.6) and plated with IPTG and X-gal. Under these conditions recombinant and non recombinant phage can be distinguished due to insertional inactivation of the ß-galactosidase gene of M13 mp18 or mp19; recombinants appear as white plaques on the chromogenic (X-gal) substrate.

Single-stranded template was prepared from 1.5ml cultures of single white plaques picked into 1:100 dilutions of *E.Coli* JM101 and grown for 6-7 hours at 37°C. Cells were removed from the culture medium and the phage precipitated in 3.5% PEG 6000, 500mM NaCl for 15 min at RT. The protein coats of the particles were removed by one phenol extraction and one phenol chloroform extraction and the single-stranded DNA recovered by ethanol precipitation.

Aliqouts of single-stranded DNA were annealed to 2ng of appropriate primer in 10µl of 10mM Tris.HCl, pH8.5, 10mM $MgCl_2$ at 50-55°C for one hour. Nucleotide mixtures were prepared, containing the appropriate dideoxynucleotide triphosphate so that the reaction could be terminated at each of the four bases. These contained dTTP, dGTP and dCTP at 125µM or, if the corresponding dideoxy base was present, at 6µM. The dideoxy bases were present at 67µM for ddTTP and 33µM for ddATP, ddCTP and ddGTP. The annealed primer-template was mixed with 10µCi of ^{35}S-α-dATP (400 Ci/mmol) and 2u of Klenow DNA polymerase. This mix was divided between four tubes (2.5µl each), and 2.5µl of the nucleotide mix added. After 20 min at RT the reactions were chased by addition 1µl of a nucleotide mix containing each of the dNTPs at 100µM and given an additional 20 min incubation at RT. Samples were prepared for loading onto the 8% acrylamide sequencing gels by adding 4µl of formamide dye mix (80% formamide, 10mM EDTA, 0.1% xylene cyanol and 0.1% bromphenol blue) and boiling for 3 min. After running the gels were dried and exposed to X-ray film at RT for 24-48 hours.

2.25 Tissue culture and cell lines

Tissue culture practice was standard as described in Freshney, 1983. Media (Gibco-BRL or Sigma) were either reconstituted from powder and sterilised by filtration through a 0.22µm filter or, bought ready-made. Cells were passaged in 175cm^2 culture flasks, either Falcon (Becton-Dickinson) or Nunc (Gibco-BRL), at 37°C and 5% CO_2. As far as possible cells were maintained in logarithmic growth phase at 2-4 X 10^5 cells/ml. Total cell and viable cell counts were done using a Neubauer haemocytometer and Trypan Blue (Sigma) staining. All media were supplemented with 10% (FCS) foetal calf serum (Flow), 100 international u/ml penicillin, 100µg/ml streptomycin, 2.5µg/ml amphotericin B and 2mM glutamine.

A wide variety of cell lines were used in the work in this thesis. HeLa cells are a line of human cervical carcinoma cells (Gey *et al.*, 1952). SW480 cells are a human colorectal adenocarcinoma cell line (Leibovitz *et al.*, 1976). Both these cell lines were maintained in Dulbeccos modified Eagles medium (DMEM). K562 cells are human erythroleukaemic cells with an embryonic/foetal phenotype (Andersson *et al.*, 1979; Villeval *et al.*, 1983). K562 SAI cells are a semi-adherent subclone of K562 (Spandidos, 1984). HEL (92.1.7) cells are human erythroleukaemic cells (Martin & Papayannopoulou, 1982). These erythroid cell lines were maintained in Iscoves modified Dulbeccos medium (IMDM). Finally, CEM (pre-T-cell, REF) and H9 (T-cell, REF) were maintained in RPMI 1040 medium.

CHAPTER 3: CAI cDNA ISOLATION

Two human reticulocyte cDNA libraries were used in this work both of which used the vector λgt11. This vector uses an *E.coli lacZ* promoter to drive expression of inserted cDNA sequences allowing antibody screening of recombinants (Young and Davis, 1983). The first of these libraries had been constructed and screened prior to the work described here (using a CAI specific polyclonal antibody) by J. Barlow and Y. H. Edwards (personal communication). Three positive clones had been identified and designated λCAI.3, λCAI.9 and λCAI.11. This chapter describes the analysis of these recombinants and the construction and screening of a second cDNA library in order to find sequences not found in the first set of recombinants. The sequence and structure of the various recombinants referred to throughout the text is shown in Figs. 3.9 and 3.10.

3.1 Analysis of CAI recombinants from the first cDNA library.

The first isolate to be analysed was λCAI.3. *Eco*RI digestion of DNA from this recombinant released a single band of about 550 bp. Despite the fact that this band size was smaller than that expected for *CA-1* message (based on protein size) this band was isolated and ligated into *Eco*RI digested M13mp18 and 19. Following transformation, several white plaques were picked and the DNA extracted. The single-stranded DNA from these recombinants were annealed to one another prior to running on an agarose gel to determine the orientation of the insert in each recombinant. Several recombinants were then sequenced using the dideoxy chain termination method (Sanger 1977).

The results of this (Fig. 3.1) showed the sequence expected from the carboxy-half of the known human CAI protein sequence, commencing at an *Eco*RI site within the coding sequence at Asn residue no 124. At the other end of

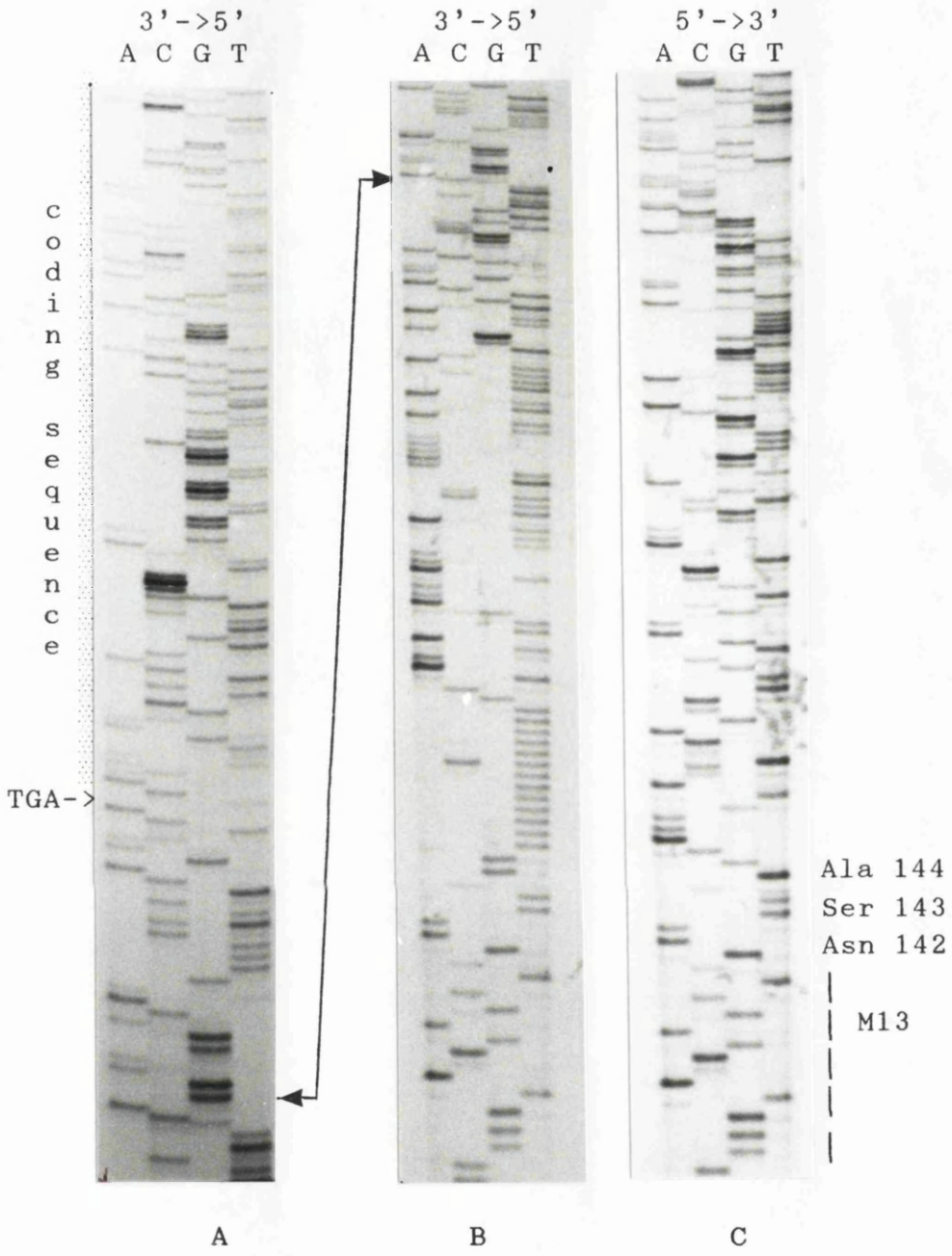


Fig. 3.1 Sequencing of insert from the cDNA isolate λCAI.3. A and B: Sequencing in the 3'->5' direction showing sequence corresponding to the carboxy-terminal end of the protein (position of TGA stop codon indicated), 3'-untranslated sequence and poly(A) tail. C: 5'->3' sequence showing protein-coding sequence from residue 142 of the protein. The full sequence of this cDNA recombinant is shown in Fig. 3.9 together with the sequence of other recombinants described in this chapter.

this fragment a poly (A) tail sequence (14 A's) was found 106 nt downstream from the stop codon close to a consensus polyadenylation signal (AATAAA). Restriction enzyme analysis allowed a *Pvu*II-*Eco*RIsubfragment to be subcloned for confirmatory sequence determination. Since no other band was seen in *Eco*RI digests of λCAI.3 it was assumed that the 5'-end of the CAI cDNA was absent from this recombinant.

Recombinant clones λCAI.9 and λCAI.11 were also grown and purified as potential sources of a full-length cDNA. *Eco*RI digestion showed that, like λCAI.3, only a single insert band was released from the vector (Fig. 3.2). These fragments were of a different size to the λCAI.3 insert being smaller in the case of λCAI.9 and larger in the case of λCAI.11. This difference in size, it was hoped, signified that different sequences (perhaps the 5' end) existed in these recombinants.

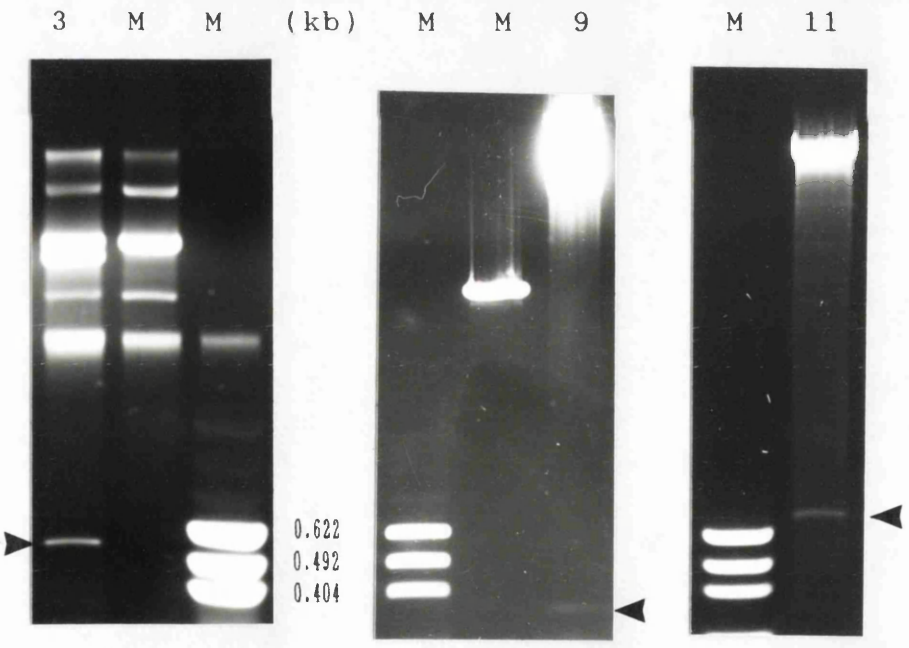


Fig. 3.2 Agarose gels of cDNA recombinant digests showing sizes of inserts (arrowed) in isolates λCAI.3, 9 and 11. M= molecular weight markers (sizes in kb).

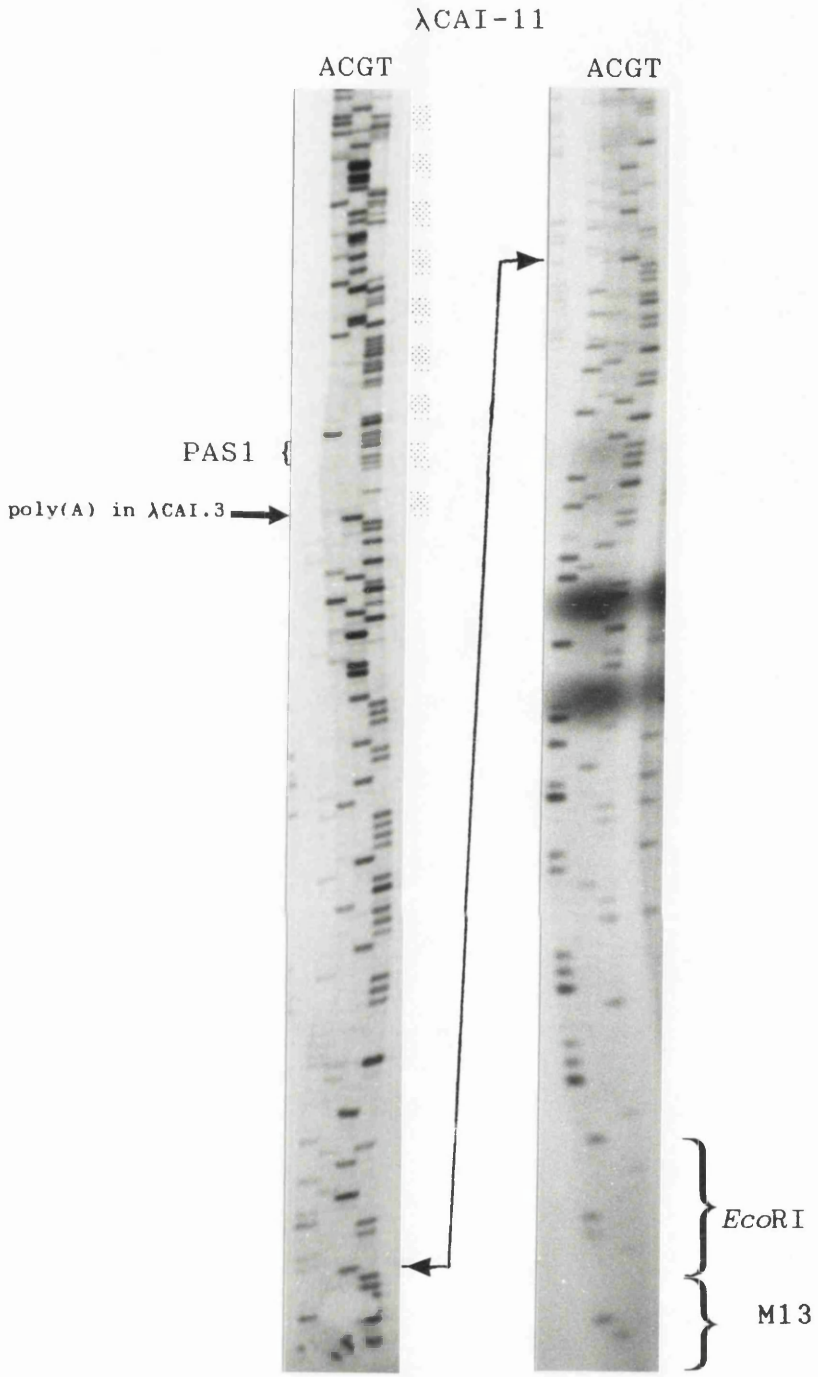


Fig. 3.3 Sequencing of the 3'-end of the insert from λCAI.11. an additional 150 nt of 3' untranslated sequence is found which is not present in λCAI.3. No Poly(A) tail sequence is seen in this recombinant or in λCAI.9. The stippled region indicates sequence common to λCAI.3 with the position of the poly(A)tail and polyadenylation signal (PAS1) (AATAAA) found in that recombinant indicated.

Sequence analysis however showed that both these fragments contained only the 3' end of the cDNA. Unlike isolate λCAI.3 no poly (A) tail sequences were found in these recombinants (Fig. 3.3) and instead a longer 3'-untranslated region was present implying use of a second polyadenylation downstream of the *Eco*RI site in the untranslated sequence. Isolate 9 appeared to be the product of incomplete cDNA synthesis or degradation, containing about 70 bp of coding sequence (commencing from amino acid residue 237), while the 5'-end of isolate 11 was the same as λCAI.3. starting at residue 124 of the protein. The structure of these clones is shown in Fig. 3.10.

In an attempt to find a full-length cDNA, the DNA fragment from λCAI.3 was used to re-screen the reticulocyte library. Several recombinants were found, but none of these contained more than one insert fragment. It was concluded that the *Eco*RI methylase protection step of the library construction had not worked.

The insert from λCAI.3 was also used in the screening of a genomic library in order to find *CA-1* genomic recombinants (see Chapter 4) which could be used as a source of probes for the 5'-end cDNA (see below).

3.2 Construction and screening of a second human reticulocyte library

Since the first library screened had not produced clones containing the 5' end of human carbonic anhydrase I, a second library construction was undertaken. Prior to construction of the library, a trial *in vitro* *Eco*RI methylase protection reaction was carried out since the inadequacy of this step was responsible for the failure to isolate a complete cDNA in the first library screening. This was apparently successful although a higher than recommended concentration of the methyl-donor, (S-Adenosylmethionine) was required (Fig. 3.4).

The source material for constructing this library was RNA

isolated from the blood of an individual suffering from pyruvate kinase deficiency-induced haemolytic anaemia in which the reticulocyte level is elevated from its normal value of 2% to approximately 20%. 660µg of RNA was isolated from 50 ml of blood and poly A^+ selected to give 20 µg $polyA^+$ RNA.

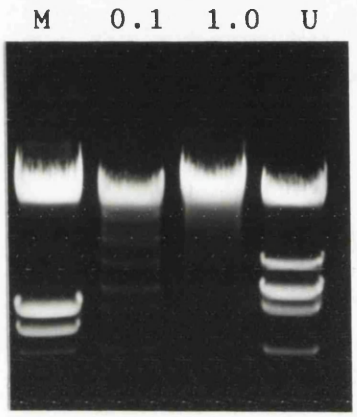


Fig. 3.4 *In vitro Eco*RI methylase protection. 2µg of λDNA was treated with *Eco*RI methylase prior to digestion with *Eco*RI. The two middle lanes are of DNA treated with 0.1 and 1.0 mM S-adenosyl methionine next to untreated (U) DNA. M : markers (λ *Hind*III *Eco*RI).

6 µg of $PolyA^+$ RNA was used in the cDNA synthesis reactions and 1/4 of this was used in the final packaging reaction. This produced 3.3×10^5 clones of which 7% were non-recombinant as assessed by blue/white plaque screening (about 2×10^5 recombinants per µg input RNA).

A total of 150,000 plaques were screened using *E.coli* strain Y1090 as a host. Primary screening was carried out using the insert from recombinant λCAI.11 (the larger of the three recombinants isolated from the first library). This first screening produced 19 positive clones. Subsequent secondary and tertiary screening produced 12 definitive positive clones. DNA was isolated from these recombinants and digested with *Eco*RI. This revealed that, as with the previously isolated clones, only single fragments were released from the vector. These were all the same size and

appeared to be identical to the insert found in λCAI.11 (Fig. 3.5) implying that despite increasing the concentration of the methyl donor the methylation protection of the *Eco*RI site internal to the cDNA had been incomplete.

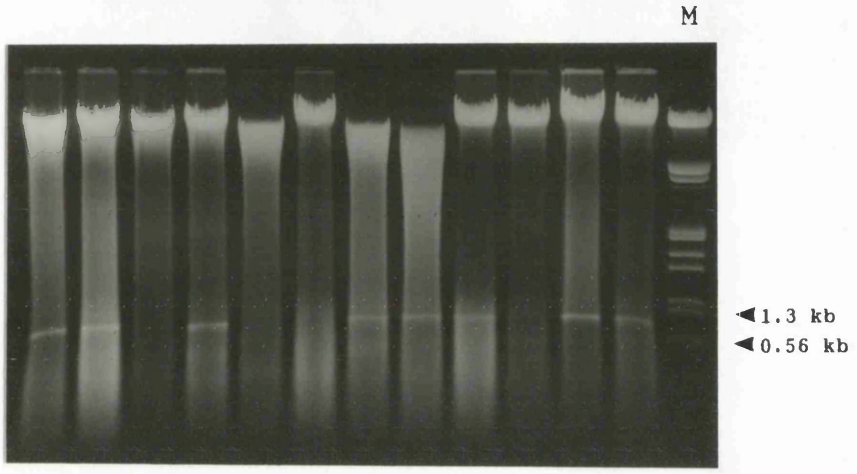


Fig. 3.5 *Eco*RI digestion of 12 cDNA recombinants isolated from the second cDNA library using the insert from λCAI.11 as a probe. All DNA samples which digested released a single insert fragment of the same size as seen in λCAI.11. M = markers.

In order to find recombinants containing the 5' end of the CAI cDNA, a genomic fragment probe was utilised (see Chapter 4 for details of the isolation and characterisation of genomic recombinants). The probe was a 1.2 kb *Eco*RI-*Hind*III subclone containing exon 3 of the *CA-1* gene and this identified 10 potential cDNA clones after screening approximately 200,000 plaques. Seven of these were finally used for DNA preparation.

Analysis of these recombinants showed that only a single insert was released upon *Eco*RI digestion. These inserts were all of approximately the same size except for one isolate - λ5'CAI.6 which was larger than the rest by about 50 bp (Fig. 3.6). The insert from this isolate and one other isolate (λ5'CAI.1) was purified and subcloned into M13 for sequencing.

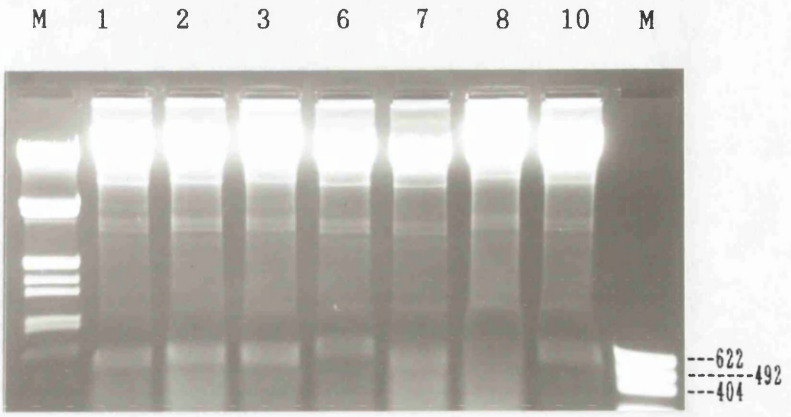


Fig. 3.6 *Eco*RI digests of isolates containing the 5'-end of the cDNA sequence (λ5'CAI.1 to -.10). Sequencing of isolates 1 and 6 showed that the larger size of λ5'CAI.6 is due to a small insert in the 5' leader sequence (Fig. 3.7).

Sequencing of isolates 1 (λ5'CAI.1) and 6 (λ5'CAI.6) showed that both these inserts contained sequences upstream (5') of the *Eco*RI site in the coding region (which formed the 5'-end of the 3'-cDNA recombinants described in Section 3.1) and extended into the 5'-non-coding region of the message. There was however an unexpected feature of the sequences in that the longer of the two isolates (λ5'CAI.6) contained an insert of 54 nt in the leader sequence compared to the shorter isolate (Fig. 3.8). This 54 bp region was termed the 1b element while the adjacent regions common to both cDNAs being designated 1a and 1c. Sequence analysis of a third isolate (λ5'CAI.10) showed it to be of the same type as λ5'CAI.1, lacking 1b.

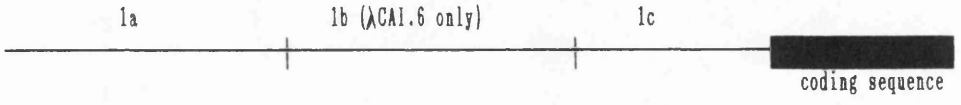


Fig. 3.7 Diagram of the elements found in the 5'-leader sequence of the CAI cDNA isolates

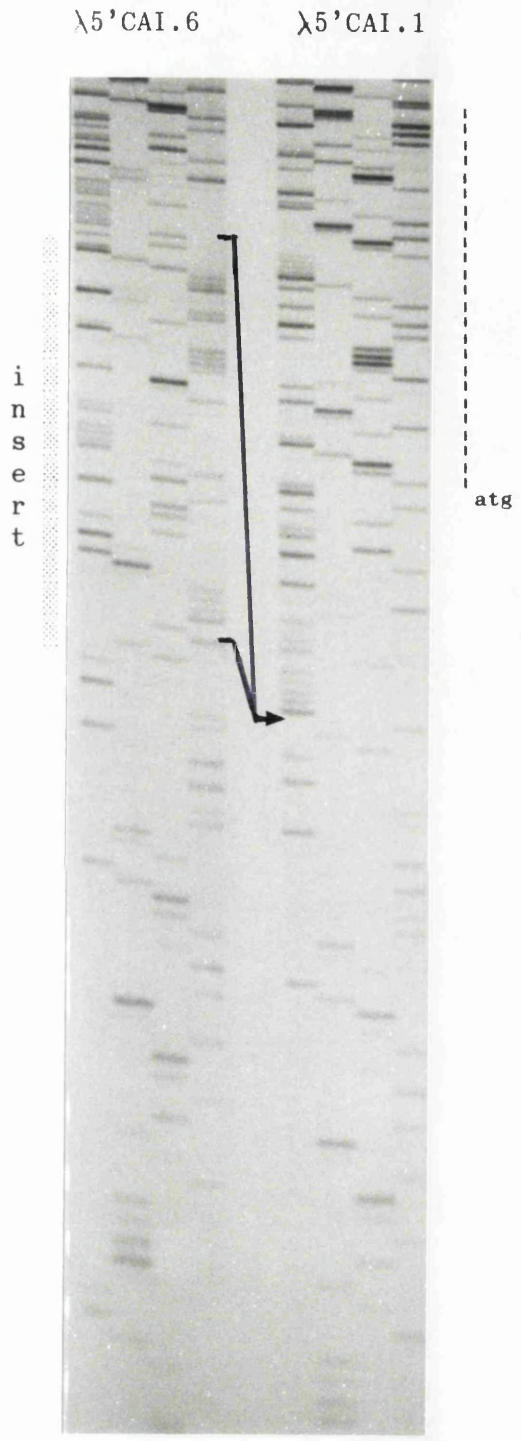


Fig. 3.8 5'->3' sequencing of the 5'-leader region of λ5'CAI.6 and -.1 showing the insert (stippled region) in λ5'CAI.6. Coding sequence is indicated by a broken line together with the position of the initiation codon (atg).

3.3 Identification of a second more distal polyadenylation site

Section 3.1 has already described the finding of 3'-end cDNA recombinants which did not contain any polyadenylation sequences. These recombinants (λCAI.9 and λCAI.11) appeared to be derived from a message which was truncated at an *Eco*RI site downstream of the polyadenylation site seen in recombinant λCAI.3. To identify cDNA recombinants containing sequences distal to this *Eco*RI site the cDNA library was screened with a genomic probe containing 3' flanking sequence. The genomic fragment used as a probe for these 3' untranslated cDNAs was a 2.2 kb *Hind*III-*Xba*I fragment containing downstream flanking sequence 3' of the *Eco*RI site which formed the 3' boundary of isolates λCAI.9 and λCAI.11. Sixteen potential positive clones were detected (although hybridisation was weak) following screening and one of these was sequenced (Fig. 3.9). This isolate contained a long Poly (A) region at one end and was 81 nt in length (from the centre of the *Eco*RI site and not including poly A tract) making the total 3'-untranslated message length 334 nt. The poly (A) tract started 18 nt from a consensus AATAAA polyadenylation signal. Sequencing of the 3'-flanking region from genomic recombinants (Chapter 4) confirmed the location of this sequence at the 3' end of the cDNA

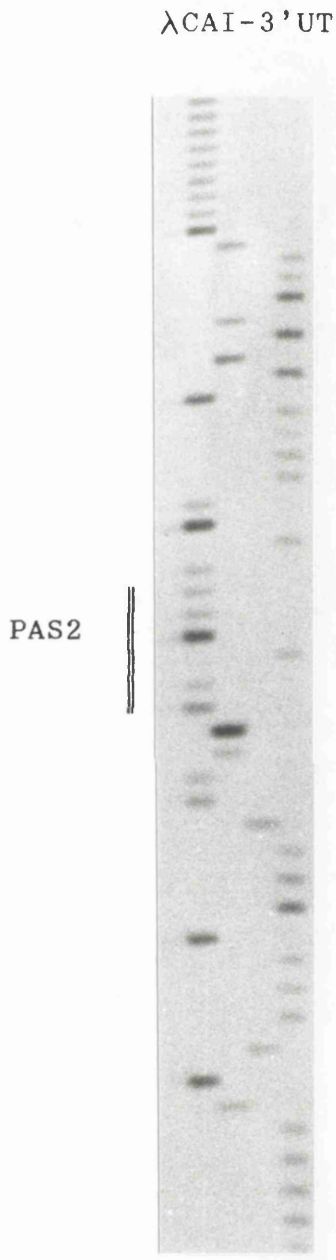
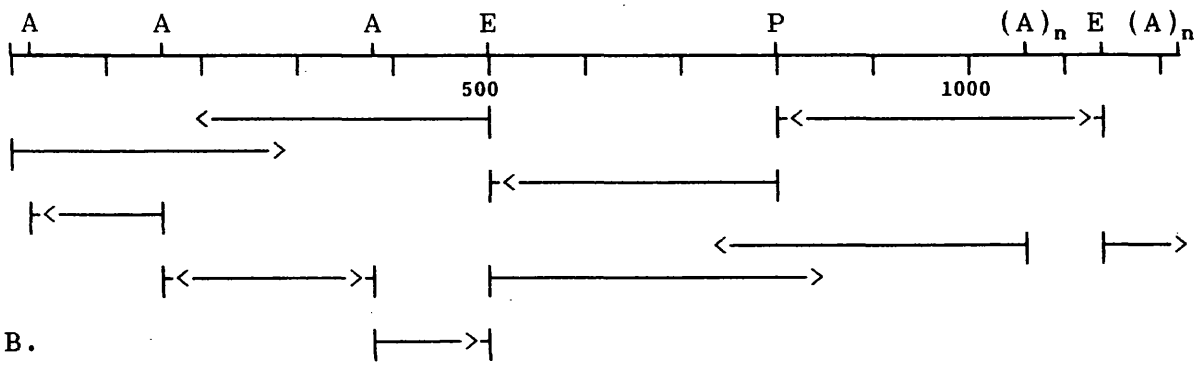


Fig. 3.9 Sequencing of the insert from cDNA isolate λCAI3'UT containing sequence downstream of the *Eco*RI site in the 3'-untranslated region. Poly(A) tail sequence is at the top of the gel and the position of the second polyadenylation signal (AATAAA) is marked (PAS2).

```
                    1a element                          50   <------
caggtgcaacccccctgcgtggtcctctgtggcagccttctctcattcagagctgttttcc

------------------- 1b element ----- 100 ----->
acagaggtagtgaaaagaactggattttcaagttcactttgcaagagaaaaagaaaactc

          1c element        150
agtagaagataATGGCAAGTCCAGACTGGGGATATGATGACAAAAATGGTCCTGAACAAT
             MetAlaSerProAspTrpGlyTyrAspAspLysAsnGlyProGluGln>
                                                  10
                    200
 GGAGCAAGCTGTATCCCATTGCCAATGGAAATAACCAATCCCCTGTTGATATTAAAACCA
TrpSerLysLeuTyrProIleAlaAsnGlyAsnAsnGlnSerProValAspIleLysThr>
             20                                   30
        250                                                 300
 GTGAAACCAAACATGACACCTCTCTGAAACCTATTAGTGTCTCCTACAACCCAGCCACAG
SerGluThrLysHisAspThrSerLeuLysProIleSerValSerTyrAsnProAlaThr>
             40                                   50
                                                        350
 CCAAAGAAATTATCAATGTGGGGCATTCTTTCCATGTAAATTTTGAGGACAACGATAACC
AlaLysGluIleIleAsnValGlyHisSerPheHisValAsnPheGluAspAsnAspAsn>
             60                                   70
                                   400
 GATCAGTGCTGAAAGGTGGTCCTTTCTCTGACAGCTACAGGCTCTTTCAGTTTCATTTTC
ArgSerValLeuLysGlyGlyProPheSerAspSerTyrArgLeuPheGlnPheHisPhe>
             80                                   90
                         450
 ACTGGGGCAGTACAAATGAGCATGGTTCAGAACATACAGTGGATGGAGTCAAATATTCTG
HisTrpGlySerThrAsnGluHisGlySerGluHisThrValAspGlyValLysTyrSer>
            100                                  110
                   500  EcoRI
 CCGAGCTTCACGTAGCTCACTGGAATTCTGCAAAGTACTCCAGCCTTGCTGAAGCTGCCT
AlaGluLeuHisValAlaHisTrpAsnSerAlaLysTyrSerSerLeuAlaGluAlaAla>
            120                                  130
        550                                                 600
 CAAAGGCTGATGGTTTGGCAGTTATTGGTGTTTTGATGAAGGTTGGTGAGGCCAACCCAA
SerLysAlaAspGlyLeuAlaValIleGlyValLeuMetLysValGlyGluAlaAsnPro>
            140                                  150
                                                        650
 AGCTGCAGAAAGTACTTGATGCCCTCCAAGCAATTAAAACCAAGGGCAAACGAGCCCCAT
LysLeuGlnLysValLeuAspAlaLeuGlnAlaIleLysThrLysGlyLysArgAlaPro>
            160                                  170
                                   700
 TCACAAATTTTGACCCCTCTACTCTCCTTCCTTCATCCCTGGATTTCTGGACCTACCCTG
PheThrAsnPheAspProSerThrLeuLeuProSerSerLeuAspPheTrpThrTyrPro>
            180                                  190
                         750
 GCTCTCTGACTCATCCTCCTCTTTATGAGAGTGTAACTTGGATCATCTGTAAGGAGAGCA
GlySerLeuThrHisProProLeuTyrGluSerValThrTrpIleIleCysLysGluSer>
            200                                  210
                   800                          >λCAI.9
 TCAGTGTCAGCTCAGAGCAGCTGGCACAATTCCGCAGCCTTCTATCAAATGTTGAAGGTG
IleSerValSerSerGluGlnLeuAlaGlnPheArgSerLeuLeuSerAsnValGluGly>
            220                                  230
```

```
        850                                                        900
 ATAACGCTGTCCCCATGCAGCACAACAACCGCCCAACCCAACCTCTGAAGGGCAGAACAG
AspAsnAlaValProMetGlnHisAsnAsnArgProThrGlnProLeuLysGlyArgThr>
             240                                  250

                                                          950
 TGAGAGCTTCATTTTGAtgattctgagaagaaacttgtccttcctcaagaacacagccct
ValArgAlaSerPhe***
             260

                                          1000
 gcttctgacataatccagttaaaataataatttttaagaaataaatttatttcaatatta
                                           PAS1
   .p(A)I                    1050
 gcaagacagcatgccttcaaatcaatctgtaaaactaagaaacttaaattttagttctta
  (a)n in λCAI.3
                  1100
 ctgcttaattcaaataataattagtaagctagcaaatagtaatctgtaagcataagctta

      1150              EcoRI                                1200
 tcttaaattcaagtttagtttgaggaattctttaaaattacaactaagtgatttgtatgt

                                                        p(A)II
 ctatttttttcagtttatttgaaccaataaaataatttttatctctttc(a)n
                          PAS2
```

A.



B.

Fig. 3.10 Sequence obtained from the human CAI cDNA recombinants described in this Chapter. Numbering of the protein sequence does not include the first methionine. 5'-leader sequence (designated 1b) found only in λ5'CAI.6 is underlined. The *Eco*RI sites (bold) which formed the boundaries of the recombinants are shown as are the two polyadenylation signals (PAS1 and 2) B: Sequencing strategy. *Eco*RI (E) fragments were subdivided with *Ava*II (A) or *Pvu*II (P) before subcloning into M13. Arrows indicate direction of sequencing.

3.4 Summary of results of cDNA cloning

The structure of the different CAI cDNA recombinants isolated is shown in Fig. 3.10. There were variations at both the 3'- and 5'-untranslated ends of the cDNA. For the 3'-end, this appeared to be the result of variable use of two polyadenylation sites (p(A)I and p(A)II), both of which used consensus polyadenylation signals. These signals have been designated PAS1 and PAS2 and give rise to a 3'-untranslated message length of 109nt and 334nt respectively.

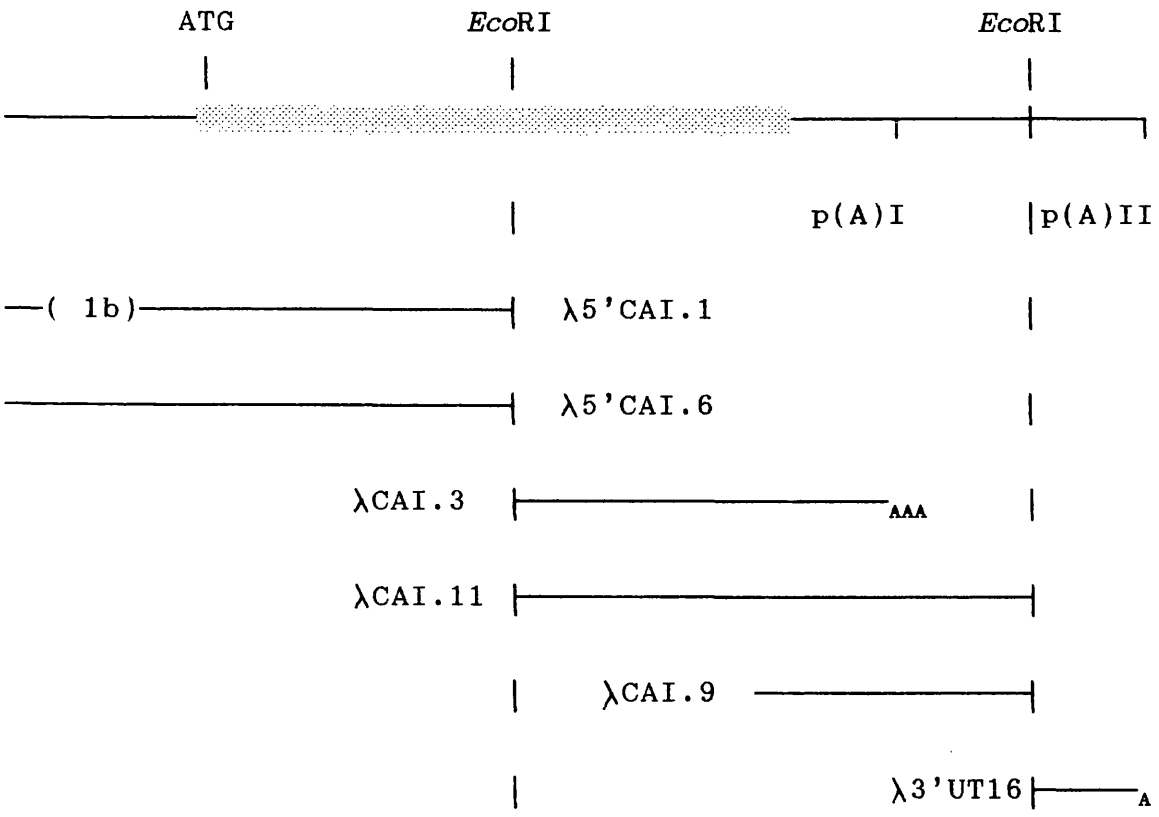


Fig. 3.11 Diagrammatic representation of the cDNA recombinants isolated and their position relative to a hypothetical message containing the longer of the two 5'-leader sequences found in the cDNA clones and polyadenylated at the more distal poly A site. The coding sequence (stippled), *Eco*RI sites and polyadenylation sites (AAA) are indicated. The brackets in λ5'CAI.1 indicate the region absent in this cDNA relative to λ5'CAI.6.

The untranslated leader sequence heterogeneity presented a less straightforward picture. All three sequenced recombinants shared a common sequence at the extreme 5' end and immediately 5' of the coding sequence, but one of these also contained a 54 bp insert within the leader sequence. These three regions were designated 1a (the first 50 nt or so of transcript), 1b (the insert found in λ5'CAI.6) and 1c (including the first part of the coding sequence. This divergence of sequence could not be adequately explained by either a heterogenous transcript initiation site or as a result of a cloning artifact. One explanation for this complexity would be that this insert is the result of a complex splicing pattern (and therefore exon structure) in the 5' flanking sequences. This was thought to be unlikely at the time since no such feature had been found in the closely related genes for CAI and CAII (Venta *et al.*, 1985a; Lloyd *et al.*, 1986). Further work however revealed that *CA-1* had an exon structure which differed from that of the other carbonic anhydrase genes (described in Chapter 4).

3.5 Assessment of relative use of polyadenylation sites I and II

Of the three cDNA recombinants sequenced containing the 3'-half of the coding sequence, one was a product of polyadenylation at p(A)I, while the other two (λCAI.9 and 11) extended further 3' to an *Eco*RI site. These were presumed to be products of transcripts which were processed at the second identified polyadenylation site p(A)II. Other clones were isolated but not sequenced and appeared, from the sizes of insert, to be identical to isolate λCAI.11. This finding suggested a preferential use of p(A)II rather than p(A)I. To test this, an oligonucleotide probe for the region between the two polyadenylation site was synthesised. This probe (Oligo 3'UT2, sequence $^{5'}$GATTACTATTTGCTAGCTTAC$^{3'}$) was hybridised to Northern blots of total RNA from reticulocytes (Fig. 3.11). For comparison a second oligonucleotide probe

(Oligo CAI#3, sequence $^{5'}$GTCATCATATCCCCAGTCTGG$^{3'}$) for the protein coding region was hybridised to a duplicate blot. Dot blots of samples of recombinant plasmids containing CAI sequences were used as a control to assess the relative efficiency of hybridisation. Similar levels of hybridisation were obtained from each probe by this method. Unfortunately differences in levels of background hybridisation and variation in signal intensity depending on temperature made accurate quantitation impossible. It could however be concluded that the majority of transcripts were not processed at p(A)I since this would result in a low signal level using Oligo 3'UT2.

It should also be borne in mind that the distance between the two poly(A) sites is about 200 nt. A difference in size of this magnitude should easily be visible on a northern blot given that the total message length is only 1,400 or so (including poly(A) tail). Since only one band has ever been detected using a variety of CAI specific probes (this work and data not shown, H. Brady personal communication) it was concluded that the majority of message terminated at the second polyadenylation site. No other consensus polyadenylation signal is found in the genomic sequence for several hundred nt downstream of poly (A) II.

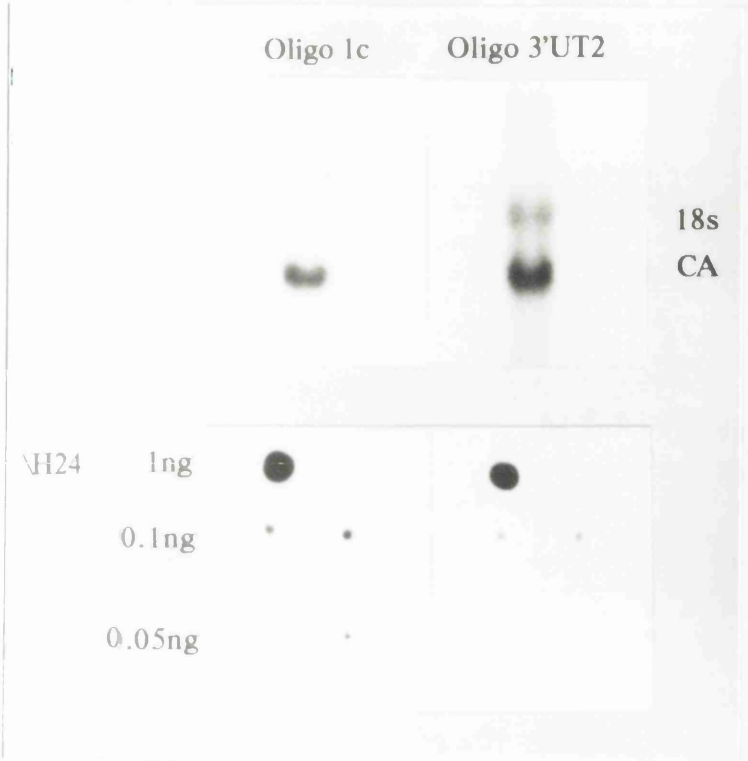


Fig. 3.12 Assessment of relative usage of p(A)I and p(A)II. Autoradiograph of northern blots of reticulocyte RNA probed with oligonucleotide probes for either the protein coding region (Oligo. 1c) or sequence between the two polyadenylation sites (Oligo. 3'UT2). Both probes were hybridised at 10^0 below their Td (64^0 and 45^0C respectively). Background hybridisation seen with 3'UT2 is probably due to the relatively low hybridisation temperature used and is not found using more stringent hybridisation conditions. Use of a higher temperature however also changes the absolute level of signal obtained, and level relative to DNA controls. For this reason this method was not considered useful except as a rough guide to transcript level.

CHAPTER 4: ISOLATION OF THE *CA1* GENE

Since the long term aim of the group was the dissection of the genetic elements involved in controlling *CA1* expression, the first step was the isolation of *CA1* containing recombinants. This chapter describes the work undertaken to isolate and physically characterise the *CA1* gene. Due to the unexpectedly large size of the *CA1* gene in comparison to other members of this gene family the process was more complicated than initially envisaged. This section has therefore been divided up into three parts. These describe: 1) the isolation of the clones containing protein coding sequence, 2) the isolation of clones, not contiguous with the first recombinants, containing sequence from the extreme 5'-end of the cDNA and 3) the isolation of overlapping clones linking these two sets of recombinants. This work was carried out in close collaboration with Jon Barlow who isolated the first *CA1* recombinants and Hugh Brady.

When the large size of the *CA1* gene became apparent, parallel screening of cosmid libraries were also carried out in the hope of isolating recombinants containing either the whole gene or larger sections than those found in the lambda recombinants. These experiments are described in Section 4.7.

4.1 Genomic clones containing *CA1* coding sequence.

The genomic recombinants containing the *CA1* gene were isolated from a λ2001 vector library kindly supplied by Dr T. Rabbits (M.R.C. Laboratory of molecular biology, Cambridge). This replacement vector contains multiple cloning sites for the restriction enzymes *Eco*RI, *Hind*III, *Bam*HI, *Xba*I, *Sst*I and *Xho*I (Karn *et al.*, 1984). The library was constructed by cloning *Sau*3A partially-digested DNA into the *Bam*HI sites of the vector, removing the *Eco*RI and *Hind*III sites (Le Franc *et al.*, 1986).

4.1.1 Initial recombinant isolation

The library was initially screened by J. Barlow using the 650 bp *Eco*RI fragment from the cDNA recombinant λCAI.11 (Chapter 3). Two recombinants were isolated from this screening and designated λHGCAI.JB#2 and JB#5 (referred to subsequently as λJB#2 and λJB#5). Following initial restriction site mapping, several subfragments were isolated which hybridized to the cDNA recombinant. These recombinants were partially restriction mapped and the approximate position of several exons identified. The exon/intron junctions were sequenced - after subcloning fragments which hybridized to the cDNA into M13 - using oligonucleotide primers designed to read outwards from the coding sequence. The opposite strand was then sequenced using oligonucleotides synthesised on the basis of this sequence and designed to read back into the coding sequence (J. Barlow, personal communication). A map of these recombinants is shown in Fig. 4.1.

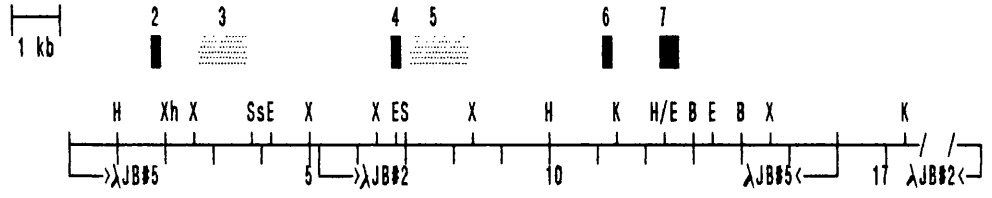


Fig. 4.1. Map of the recombinants λHGCAI.JB#2, and JB#5. The positions of the exons are indicated above the restriction map. Exons shown as solid boxes were positioned by their proximity to restriction sites from which they were sequenced. Shaded areas indicate those regions thought to contain exons, 3 and 5. The *Eco*RI site which formed the boundaries of the cDNA recombinants described in Chapter 3 lie in exons 4 and 7. B: *Bam*HI, E: *Eco*RI, H: *Hind*III, K: *Kpn*I, S: *Sma*I, Ss: *Sst*I, X: *Xba*I, Xh: *Xho*I

Since one of the main aims of the group was dissection of the promoter sequences responsible for expression, attempts were made to locate this region. It was found however that neither of these two recombinants would hybridize to an oligonucleotide probe (Oligo #3) for the first part of the protein coding sequence (amino acid residues 3-9 (Pro-Asp), sequence: $^{5'}$GACATCATATCCCCAGTCTGG$^{3'}$) (Fig 4.3).

4.1.2 Mapping of exons 3 and 5

Fig. 4.2 shows the experiments carried out to localise exons 3 and 5. Work carried out while subcloning fragments for sequencing (J. Barlow, personal communication) together with data supplied with recombinant H24 (see Section 4.1.4) indicated that exon 3 lay within a 1.2 kb *Xba*I-*Eco*RI region (part of a 3.2 kb *Xho*I fragment which could be isolated from λJB#5), while exon 5 lay within a 1.6 kb *Eco*RI-*Xba*I fragment. Examination of the cDNA sequence showed that exon 3 contained single sites for *Ava*II and *Ssp*I, while exon 5 contained a *Pst*I site. The fragments were isolated and digested with the appropriate enzymes, allowing positioning of the exons.

4.1.3 Isolation of recombinants containing the 5' end of the coding sequence

In order to find recombinants containing the first exon and the 5'-flanking sequence, the λ2001 library was screened using the insert from cDNA clone λ5'CAI.6 containing the 5'-half of the cDNA sequence (Chapter 3). Seven clones were identified and designated λ101-λ107 (the λ100 series). These clones were then screened with the Oligo #3 (Fig. 4.5). One of these - λ104 - hybridized to this probe but not the 3'-half of the cDNA and was therefore thought likely to extend further upstream of JB#5 (Fig. 4.3). This recombinant was restriction mapped and shown to extend some 10 kb upstream of the protein coding sequence (Fig. 4.4).

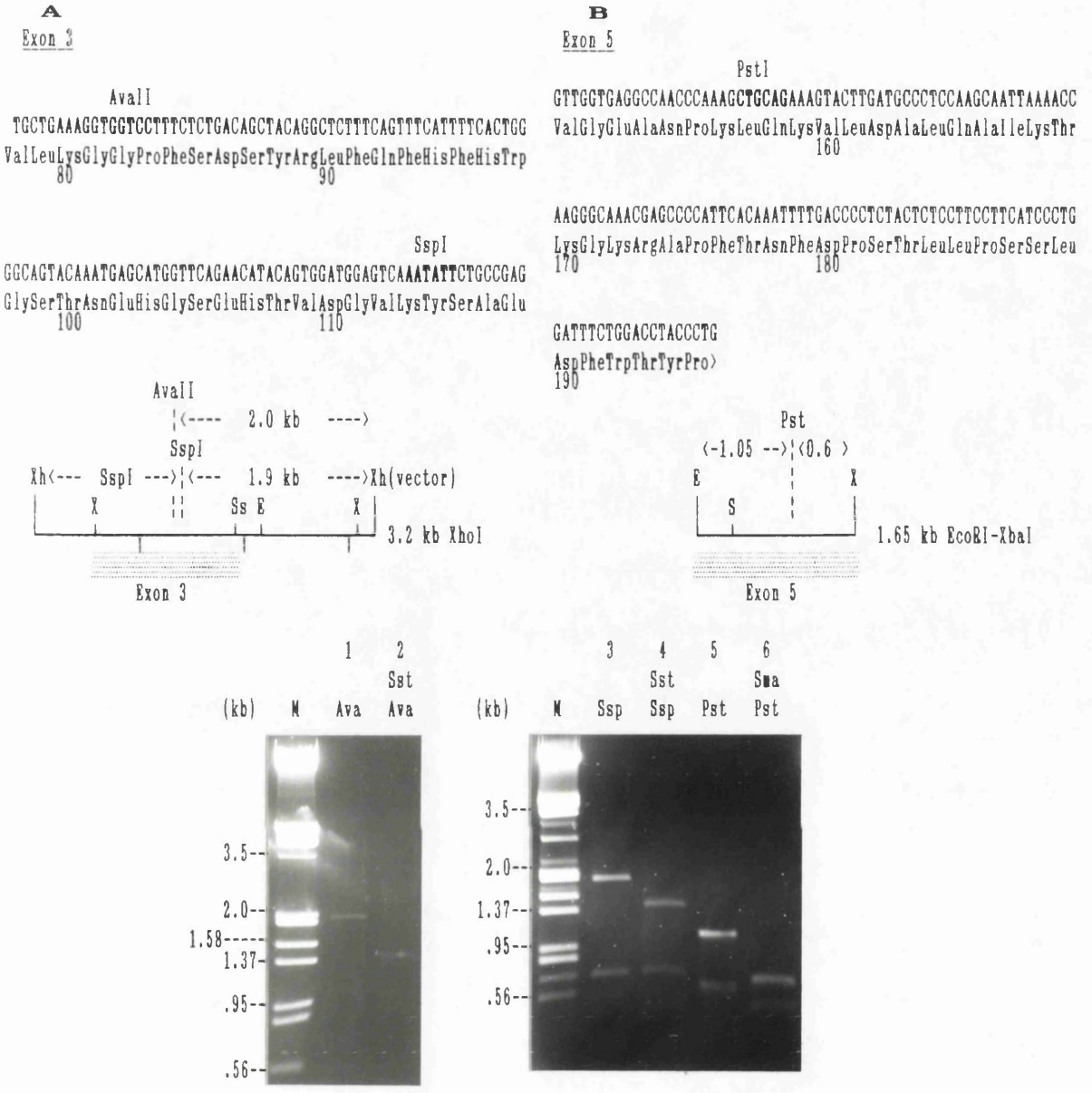


Fig. 4.2 Localisation of exons 3 and 5. A: Exon 3 was known to lie within the shaded *Xba*I-*Sst*I fragment. A 3.2 kb *Xho*I fragment containing this region was digested with *Ava*II or *Ssp*I (lanes 1 and 3). The 2.0 and 1.9 kb fragments produced by these digests were both cut with *Sst*I (lanes 2 and 4) showing these sites to be 2.0 and 1.9 kb from the 3'*Xho*I. B: The 1.6 kb *Eco*RI-*Xba*I fragment known to contain exon 5 was digested with *Pst*I (found in DNA sequence) to give 1.05 kb fragment. The previously mapped *Sma*I site was then used for orientation. DNA fragments in the *Pst*I/*Sma*I digest have run with a slightly altered mobility probably due to the acetate buffer used in the digest. E: *Eco*RI, S: *Sma*I, Ss: *Sst*I, X: *Xba*I, Xh: *Xho*I.

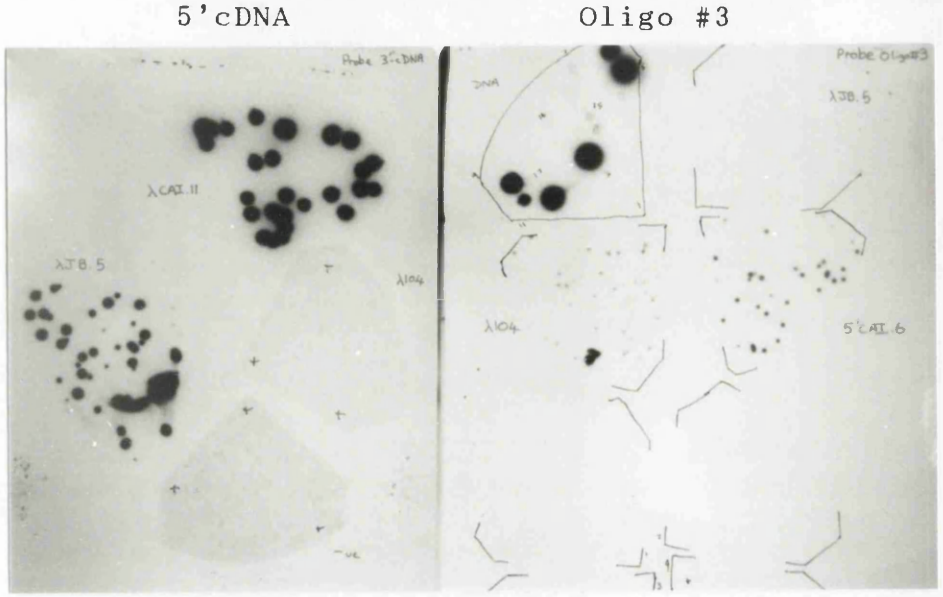


Fig. 4.3 Hybridisation of the 3'-cDNA (left) and Oligo #3 (right) to plaque lifts of λ104 and other genomic recombinants containing protein coding sequence. In the left panel λJB#5 and λCAI.11 are detected while λ104 does not cross hybridize. In the right panel, Oligo #3 hybridizes to λ104, λ5'CAI.6 (containing the 5' half of the cDNA) and a number of DNA controls (top left). λJB#5 however is not detected. At the bottom of the right panel are filters containing DNA from the other λ100 series of clones which did not cross hybridize to Oligo #3.

At this time the laboratory received a recombinant designated H24 from Dr Richard Tashian (University of Michigan) which also contained the human *CA1* gene. H24 had been isolated from a λCharon 4A genomic library by low stringency screening with a human *CA2* cDNA probe (the isozyme on which his laboratory is working). The clone had been restriction enzyme mapped, and the position of exons 4, 6 and 7 determined and partially sequenced. The data supplied with this recombinant matched the restriction enzyme mapping and sequence data for λJB#2, λJB#5. and λ104.

In summary, λJB#5 contained all but exon 1 of the protein coding sequence, while \JB#2 extended further 3' to these isolates and did not contain the first 3 exons. λ104 contained

only exons 1-3 and extended some 10 kb 5' to the coding sequence. H24 extended about 2.5 kb 5' of the coding sequence at its 5'-end and ended within the 3'-untranslated region. A map of the region covered by these recombinants is shown in Fig 4.4. The coding region is spread over 14 kb and as with the other carbonic anhydrase genes the coding sequence is interrupted by six introns. These were found to lie at the same position in the protein sequence as other *CA*s (Venta *et al* 1985a, Lloyd *et al.* 1987). The sequence at the exon intron junctions is shown in Fig. 4.19 (sequence determined by J. Barlow, U.C.L. and P. J. Venta/R. E. Tashian University of Michigan, personal communication)

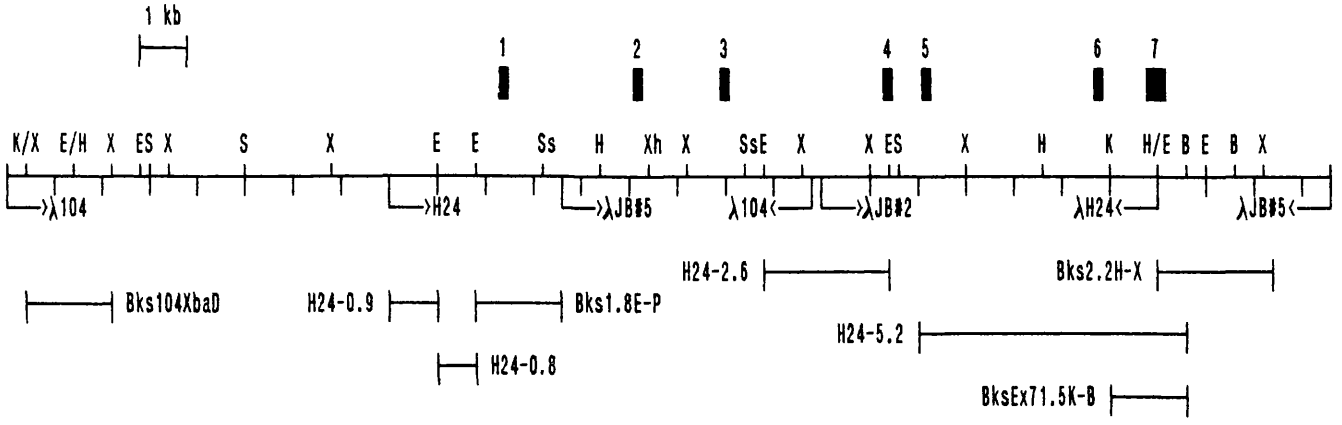


Fig. 4.4 Map of the region covered by recombinants containing *CA1* coding sequence, λ104, λJB#2, JB#5 and H24. Positions of exons (1-7) are shown above the restriction map. Below the map are shown those regions subcloned into plasmids and refered to elsewhere in the text as sources of probes or sequencing material. B: *Bam*HI, E: *Eco*RI, K: *Kpn*I, S: *Sma*I, Ss: *Sst*I, X: *Xba*I, Xh: *Xho*I,

4.1.4 The 5'-end of the cDNA is not found upstream of the coding sequence

Chapter 3 has described the finding of cDNA clones with two different types of 5'-leader sequence, one of which contained an apparent insert of 54 nt relative to the other sequences - the 1b element (Fig. 4.5) It was assumed initially that this may have been some kind of cloning artifact and that sequencing of the corresponding region from genomic clones would help resolve the origin of 1b. To this end the region upstream of the first coding exon was isolated for sequencing.

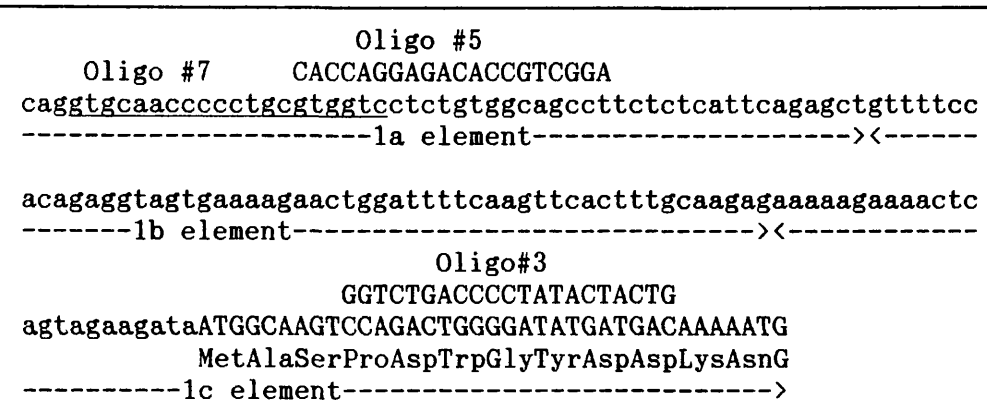


Fig. 4.5 The 5'-end of the cDNA showing the three different elements of the leader sequence. The 1b element is found only in a minority of cDNA species (see Chapter 3, Section 3.2). This sequence was used to generate several oligonucleotide probes used in cloning and sequencing and the position of these is indicated. Oligo sequence written above the cDNA sequence shows oligos complementary to the message, while the regions underlined are from the other strand.

Oligo #3 which would hybridize to sequence coding for amino acid residues 3-9 was used as a probe to localise the region containing exon 1. Recombinant clones were digested, separated on agarose gels and southern blotted prior to hybridisation with this probe.

A 1.8 kb *Eco*RI-*Pst*I fragment from clone λ104 was identified in this way and subcloned into M13 for

sequencing. Sequencing using oligo #3 as a primer showed that the genomic sequence upstream of exon 1 did not correspond to either of the two sequences found in the 5'-leader of the cDNA clones and suggested that this fragment did not contain the promoter of the gene. To check that this was not specific to recombinant λ104 , the same fragment was also sequenced from clone H24 with the same result. Fig. 4.7 shows the sequence obtained upstream of the protein coding sequence compared with the cDNA sequences.

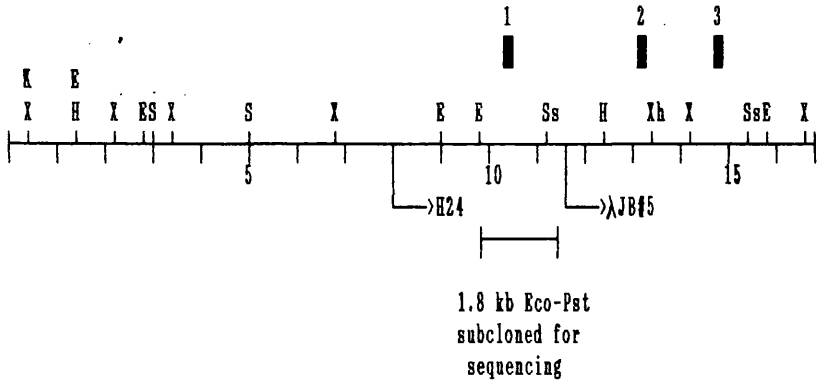


Fig. 4.6 Map of recombinant λHGCAI.104 showing exon position and region cloned for sequencing

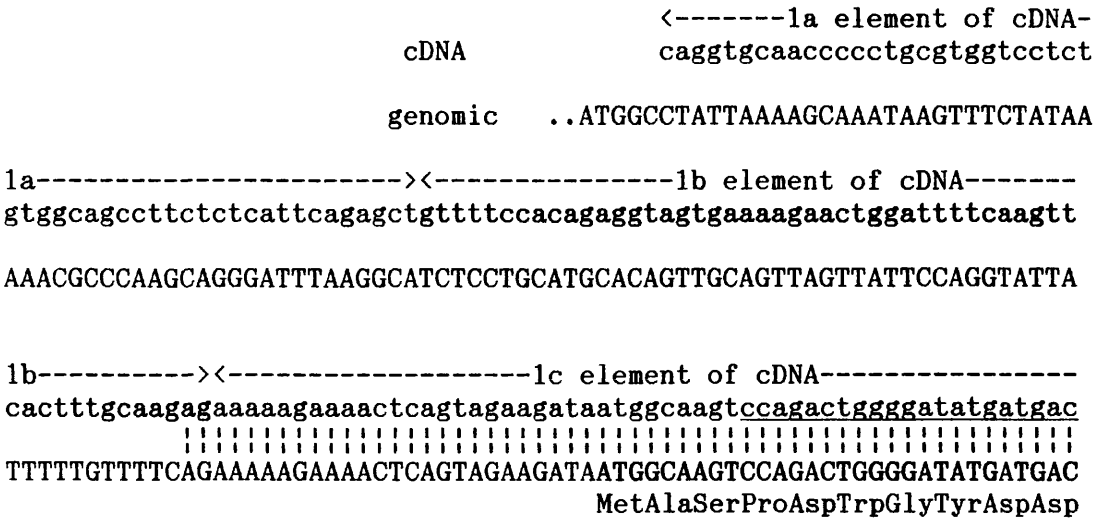


Fig. 4.7 Comparison between genomic sequence upstream of the first coding exon and corresponding sequence from the cDNA clones. Genomic sequence is in upper case, cDNA in lower case. The sequence of the cDNA and genomic recombinants diverges 5' of the 1c region found in the cDNA. Underlined region in the coding sequence indicates the complementary sequence to oligo #3 used as a probe to isolate λ104.

4.1.5 Sequencing of the 3'-flanking region

The cDNA recombinants λCAI.9 and λCAI.11 both contained 3' non-coding sequence which extended 100 nt downstream of the polyadenylation site found in recombinant λCAI.3 but did not contain any poly(A) tract (Chapter 3). It was assumed that since both these recombinants ended at the same *Eco*RI site, that this was not a cloning artifact (i.e. one of the linkers used in cloning) but represented a genuine *Eco*RI site in the 3' flanking region of the gene. Restriction mapping of λHGCAI.JB#5 and information supplied with recombinant H24 allowed identification of a 2.2 kb *Hind*III-*Xba*I fragment which contained an *Eco*RI site close to the *Hind*III site, as found in the cDNA sequence, and should therefore contain *CA1* 3'-flanking sequence (see Chapter 3). This fragment was subcloned and used for two purposes a) as a source of

fragments for sequencing, and b) as a probe for cDNA clones containing 3' untranslated sequence and a more distal polyadenylation site (described in chapter 3). Fig. 4.8 shows a map of the region at the 3'-end of the gene.

Sequencing of this region showed the expected untranslated sequence found in the cDNA clones together with a second polyadenylation signal 150 nt downstream of the *Eco*RI site which formed the 3'-end of λCAI.9 and λCAI.11. Altogether some 700 bp of flanking sequence was obtained as shown in Fig. 4.9.

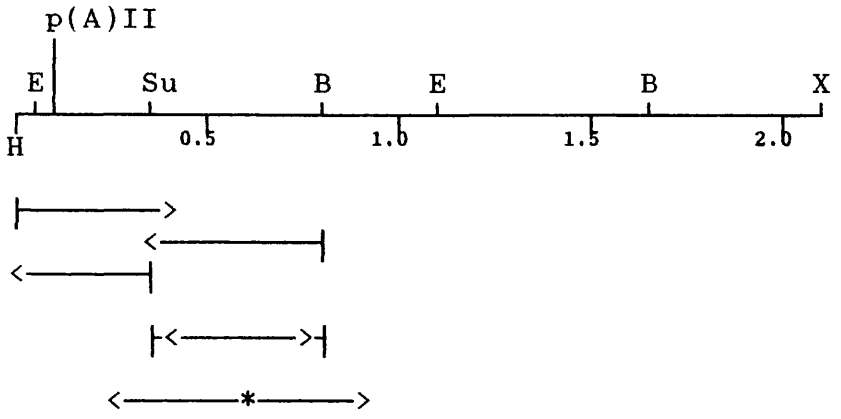


Fig. 4.8 Restriction map of the subcloned 2.2 kb *Hind*III-*Xba*I fragment containing 3'-flanking sequence of the *CA1* gene. B: *Bam*HI, E: *Eco*RI, Su: *Sau*3A, X: *Xba*I. The *Eco*RI site close to the left-hand end of this map forms the 3'-end of cDNA recombinant clones λCAI.9 and λCAI.11. The position of the more distal polyadenylation site (p(A)II) is also shown. The more proximal polyadenylation site (p(A)I) is not found in this clone and lies about 100 bp 5' of the *Hind*III site. Some 700 bp of sequence downstream of the *Hind*III site was determined (Fig 4.9.) as shown in the sequencing strategy below the restriction map. Sequence commencing at a "*" indicates use of an oligonucleotide primer specific for that region.

```
HindIII
   GATA-1                              EcoRI                50
AAGCTTATCT TAAATTCAAG TTTAGTTTGA GGAATTCTTT AAAATTACAA CTAAGTGATT
---------- λCAI.9/11 ---------------><--------------------------
                                   PAS2 100                   P(A)II
TGTATGTCTA TTTTTTTCAG TTTATTTGAA CCAATAAAAT AATTTTATCT CTTTCTTTCT
------------ λCAI.3'UT ------------------------------------->
                                150
GTTGTGCATT CAGTTTCTAA AACCATTAAG TTTCTACTCC ATTTACATTC AAAAATCTTA

                    200
AATACTTTAC TTGCAAGAGT ATTTTGCTTC AAATACAACA ACCTAAGAGC AGCTGGAGAT

       250                                                      300
GAAATATTGG GAAATTCATT TGCTTACTCC TGAAGACAAA AATATAGCTG AGATGACCAC

                                                      350     Sau3A
TGGATTTAAT ATCGTTATGC TGGCCCAACA TTGCTACCAT TTGTGTTGTC TGTGATCAAA

  GATA-1                                    400
ATGATTATCT TTTATATAGG AAGATGACGC TTCTGGATAT TGCTTTCACT TCTTCTCCCC

                                450
ACGTTAGCAA GGACAATGCT TCTCTGCCAT TATTACAACT AGTTAGTTTG CATGGAGAAT

                    500
CTTTACTTTA AAATTGGAAG AAAAGTCACA AGTGAATGGT TTATAAAAAT GCTAAAGAAG

       550                                          Ap1           600
TCATTCTTGC TTAGAATCAT ATAGAAACAT CATGCAATCT TTTAGTCAGA TGTGCGCTTC

              GATA-1                                    650
ACCTTATGCT ATTTTTATCT TTAATTGACA CACAATAATT GTACATGTTT ATGGAGTATA

                                        700
GTGTGGTGTT TTCTGTTTGT TTGTTTGTTT TTTGAGACAA GGTCTCACTC TGCCAGTCAG

       730
GGTGGAGTGC GATGGT
```

Fig. 4.9 Sequence of the 3'-flanking region of the *CA1* gene. The sequence in bold is found in the cDNA clones (Chapter 3). The position of the more distal polyadenylation signal (PAS2) is underlined as well as the second polyadenylation site p(A)II. Potential transcription factor binding sites are also indicated including the erythroid-specific factor GATA-1 (see Section 4.6)

4.2 Isolation of recombinants containing the *CA1* promoter (the λ200 recombinants)

Since the 5'-end of the cDNA sequence had not been found adjacent to the protein coding sequence in λ104 and H24, a second oligonucleotide (Oligo #5, sequence: $^{5'}$AGGCTGCCACAGAGGACCAC $^{3'}$) was synthesised as a probe for this region. This probe was designed to hybridize to the 1a element of the cDNA sequence (see Fig. 4.5). Oligo #5 identified four recombinants from the λ2001 library designated λ201-λ204. These recombinants were restriction mapped by a combination of conventional restriction digests and "COS" mapping in which ^{32}P-labelled oligonucleotides complemetary to either the right or left cohesive end of lambda are annealed to partially digested λDNA.

This revealed that λ201 and λ202 were identical and that there was extensive overlap amongst the other clones. λ204 was found to extend some 15 kb 5' to the region detected by Oligo #5 while λ203 extended 12 kb 3' to this region (Fig. 4.11).

Restriction digests of the λ200 series were probed with Oligo #5 (Fig. 4.5) to identify fragments containing the promoter region (Fig. 4.19) of λ and from this mapping 4.2 kb *Hind*III-*Sst*I fragment was selected for subcloning into Bluescript KS vector (plasmid pBks204HS4.2, Fig. 4.12). This plasmid provided the source material for the sequencing of approximately 900 bp of upstream sequence. An additional 1.4 kb of upstream sequence was obtained from an *Ava*II fragment lying adjacent to this region. See Figs. 4.12 to 4.14 for sequencing strategy and sequence obtained from this region.

This sequencing showed that these genomic clones contained sequence identical to that of the 1a element of the cDNA recombinants but did not contain 1b or 1c (see Fig. 4.5)

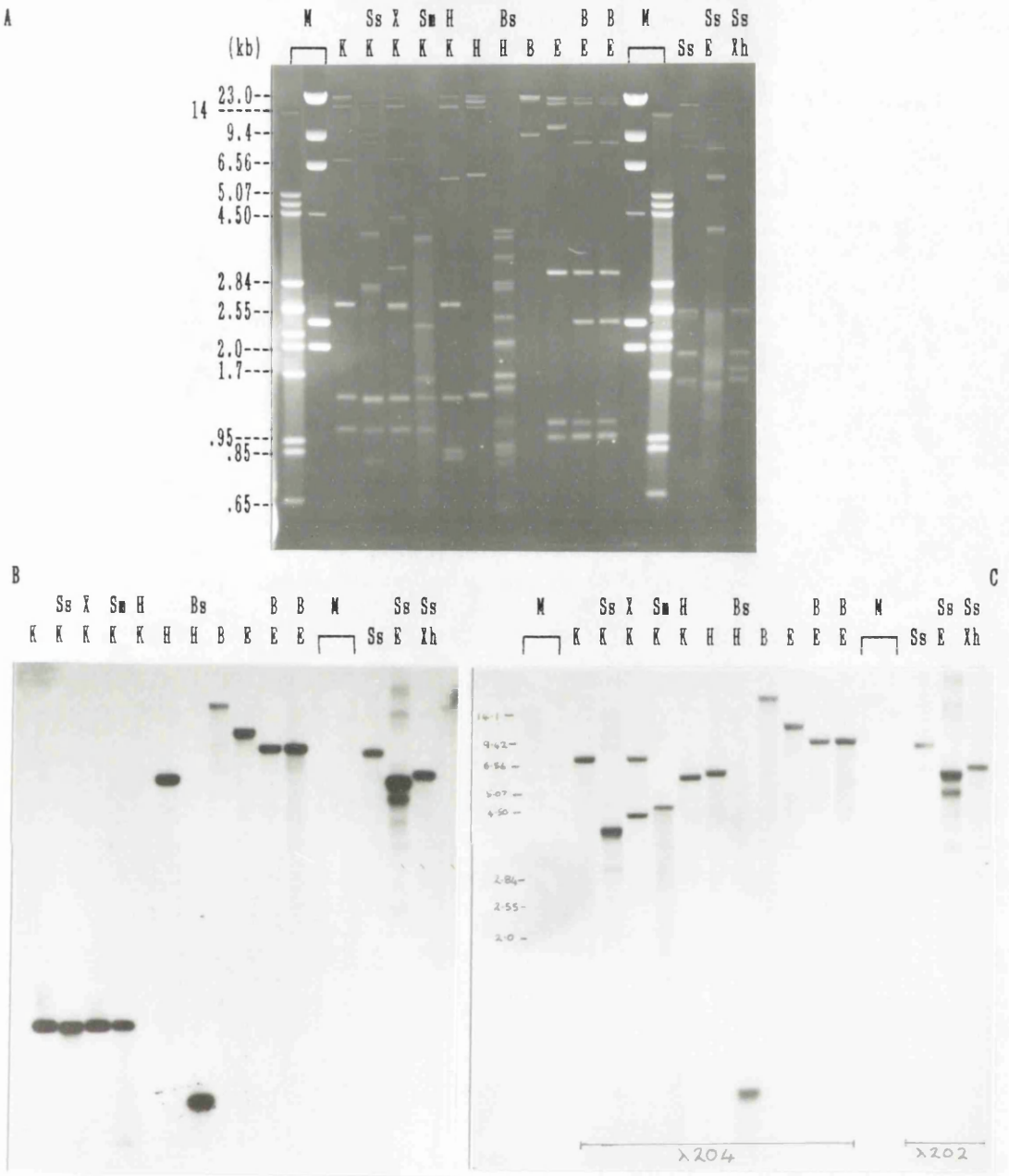


Fig. 4.10 Restriction mapping of recombinants containing the promoter region. A: Agarose gel of two recombinants digested with various enzymes. B and C: The gel shown in A has been southern blotted and probed with the Oligo #5 which hybridizes to the end of the cDNA sequence (B) and a 1.4 kb *Ava*II fragment lying upstream of the promoter (C). A map of the promoter region is shown in Fig. 4.11. B: *Bam*HI, Bs: *Bst*XI, E: *Eco*RI, K: *Kpn*I, Sm: *Sma*I, Ss: *Sst*I, X: *Xba*I, Xh: *Xho*I. M: size markers (kb)

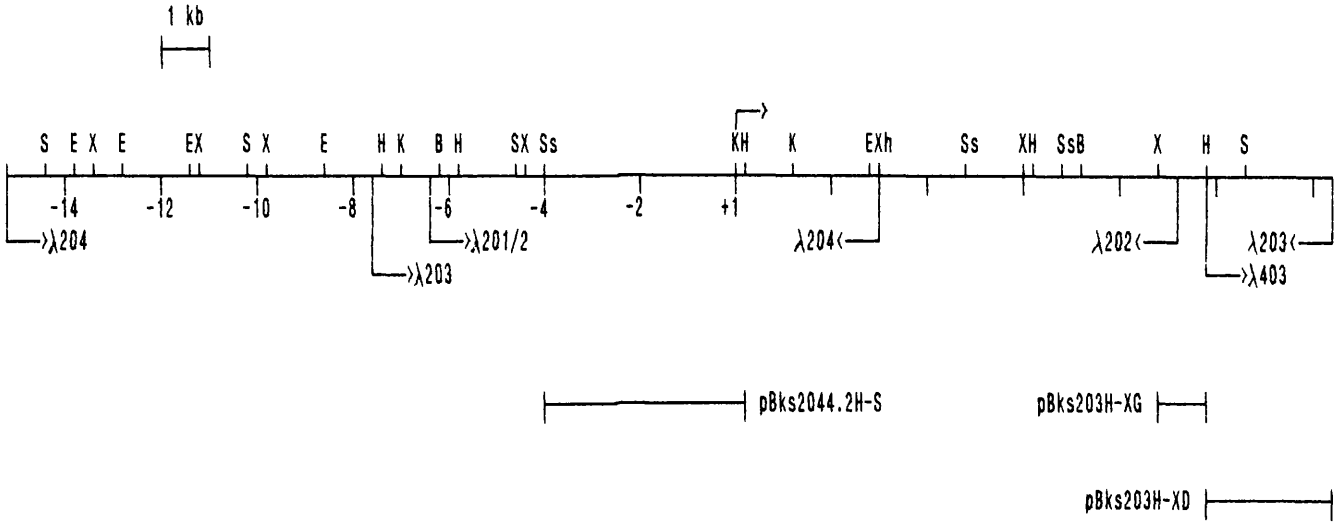


Fig 4.11. Map of the region covered by the λ200 series of recombinants. The transcription start site is found adjacent to the *Kpn*I site (+1). The probe (Oligo #5), used to isolate these recombinants, hybridises between this *Kpn*I site and the adjacent *Hind*III site.

The transcription start point for reticulocyte mRNA was mapped by primer extension analysis using both Oligo #3 and Oligo #5 as well as by S1 nuclease analysis and was found some 20 nt upstream of the 5'-end of the cDNA (Brady *et al.* 1989).

λ201-204 did not show any overlap with the previously isolated genomic recombinants as far as could be determined by restriction mapping and no cross-hybridisation between these recombinants and a radiolabelled 2.0 kb *Xba*I fragment from the 5' end of λ104 could be detected. Thus at this stage there appeared to be two sets of unlinked genomic recombinants, one of which contained coding sequence while the other appeared to contain the promoter. The determination of the physical relationship between these two sets of recombinants is described in the next section.

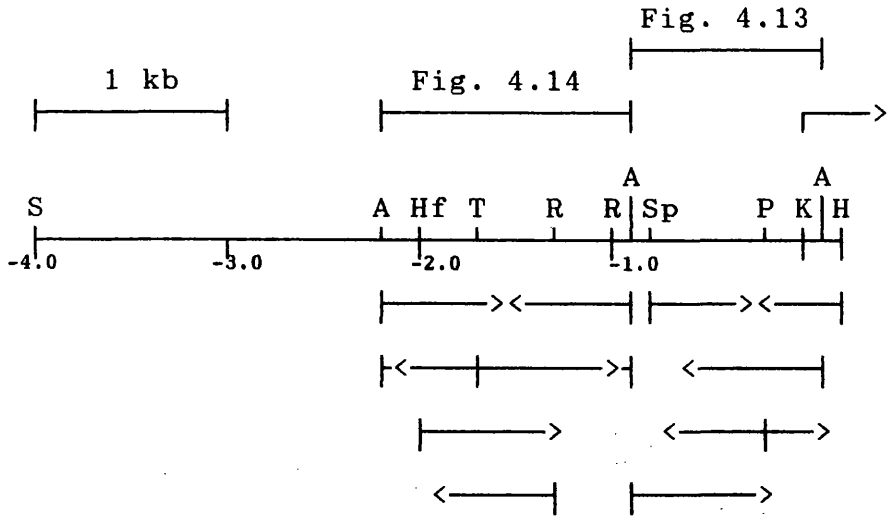


Fig. 4.12 Restriction map of 4.2 Kb *Hind*III-*Sst*I fragment containing the promoter region of the *CA1* gene. Transcription start point is shown by an arrow over the *Kpn*I (K) site. A: *Ava*II, K: *Kpn*I, H: *Hind*III, Hf: *Hin*fI, P: *Pvu*II, R: *Rsa*I S: *Sst*I, Sp: *Ssp*I.

```
1CTTTAGCCCA ACAGTCAAAA ATAATTGATG CTACCCTACA AATGTCCAAA ACTCTAGTAT

                    SspI                                  -800
61ATCATATTTC TAAGTTACAG CAAATATTAG TCCTGCTAAA CCAGGGAGCT TTGGCAAAAA

121TGTTTTTTGA CAGTAAATTT GTCCTTGATT ATATATTAAC TAGTCAAAGA GGTGTTTGTA

                    -700
181ACATTATTAG AGCTTCTTGT TGTAGGTGGG TTAACACCAC CAATCAAGAG GTCATTCTAA

241CAGAAAGCCT GGATCAGAAA ACCATCACCC TAAAAAAACA TGCCTTACAT ATTTAACACA

    -600                                                  -500
301CTCTGAAATC CAGTCAAAAT ATGACTAAAG GCCCTTGCCA TGACTGATGT ATTCTCCTGG

361CCAACGCCAA ACAAATGGGA GCCTGGTTAC GAGTCAGCCT TCAGGGACTT GTCACATTTC

421TACTTGGTTT CTTCCTTGTT ATTGTCATAA TAAAATGTTT TCTATGCTGT TTAGTGCAAC

                    -400                                  Ap1
481TTAGGCCCTA TTCTGTAGAA GTCTCCTCTA CTATTCAGGC CACTCAAACA CCCCAAATAA

541TTGAGTTCAA AATCGACATC AAGATATAAA GGAATCAGTG ACTAAATATA TTTCATATAT

    -300          GATA-1                                  GATA-1
601GGTATTTTTA TTGATTATTG TGCTGTCTTG ACCTAGTATG GAGGCCTTGG CTAGAGGCTG

                    PvuII                 CACCC   -200    GATA-1
661GTCAGTTTCC TCTCTTGAGC AGCTGATTAA ATCCACACCC CAACCACTTC CCTTATCAGG

                                        GATA-1
721TTCTCACACT CTGGGGCCAC TATGTACCCA CTCTAATCAC CACAGGGCCA GACATCAGAC

                    -100      Sp1           Oct-1
781AATTAAGGAC AGCGCCCATG CCCCAAAGCC CGCCAAAATT ATGCAAATTA TTCAAAATTA

                                        TATA
841TTCAACCTAG CTAACCCCAC CCTTTTTGCT GTACATAAGC TGCCCATTCC CCCTCCAGCC

                                                      Oligo #5
    +1  KpnI                                  AvaII
901TGTGGTACCC AGTCCTCAGG TGCAACCCCC TGCGTGGTCC TCTGTGGCAG CCTTCTCTCA
   <---------------------------- Exon 1a -------------------------

961TTCAGAGCTG TAAGTAACAA AGACTTCTGC CTTTCATCTA ->  100 bases to HindIII
  -------->------ Intron 5'UTI (25 kb) -----
```

Fig. 4.13 Human *CA1* promoter sequence. The transcription start point (+1) is at residue 902 with the region highlighted in bold text the first exon (Exon 1a, 68 bp). This region is found at the 5'-end of the cDNA sequence. Sequence complementary to Oligo #5 used in isolation of the λ200 series of recombinants is underlined. The position of several sequence motifs associated with transcriptional control are also underlined. These include the TATA box and potential binding sites for several transcription factors including the erythroid specific factor GATA-1. (See Section 4.6 and Fig. 4.21).

```
   1 GGTCCAATTA TAGGAGTATT CTGGATGCCA GCATCTTTGC TAGCTGTGCA CCCACAAAGC
                                                                -2100
  61 CTTTGACTAA AAAGACATTT GTGAATCAAG CTCCCAGGAC CCAACATTGT ATTCAAATTC
                                                  HinfI
 121 TACAAATGAC TCAGCCCCAT TATAACACCA CAGAGTCACA AAAATACAGC CCACATCAAC
                                                                -2000
 181 CCTAGTGGGT ATCCATGTGG GAAGGATACC CTAGTCCGAC ACAGTGCCCT AGCTCAAGAC

 241 CATGCCATTA AAGATGACAT GCCCCCTGCC TCCCAACCAA CTATGGGCAG TGCCAACTAC
                                                -1900
 301 ACTAGCACAT GGAATGTCTT GCCATTCATA GGAGACCAGA GGCTCTTGGT TTTCCAGTAG

 361 CTGGTTACAC TGCTGACTTC TGTTTTAAAA AAGCAAGACC CACAACACTG GAGTAACTTT
                                                                TaqI
-1800
 421 CCAGACTTTA CAGACACCTA TAAAACTCGA AGTTCTGCTC TCTGTGATTG TAATTCCACT

 481 CTATACCTGG CTTAGCCCTT ATGAGACCAA GTCCAATACC TCCCTAATGC TTCTAAGATT
                GATA-1      CACCC                                -1700
 541 CAGGTTCCTG CTAATCCCAC CCACTTCTCC TGCCTAAAGG GTAACCTGAC TGAAGTATTC

 601 CACCCTGTGT GCTCTCCTAG GAGTAAAAGG AGCATCTGCT GGTCTCCGAC TGTCAAACCA
                                                -1600
 661 TTTATTCCCT TCCATCCTCC TCGCCTTCCC TTTCTCATTG TCAGTATCCC TACTCCCACC

 721 CCTATCTTTT CCGTCCTAAC AGCATCTTCT CCTCTTGCCC TCTAAAAGAC TACGTAAGTA
-1500
 781 AAATTCAAAC CCACTGCCCC AGCTTCTCTG ATATTTATAC AGACATCCAA AATACAAACA
                                                          CACCC
 841 CATTTGACTG AAGTAGGCTC TTCTGCACTG ACTATCCATA ACTCATCACC CTGCTTCCAC
GATA-1                                  GATA-1          -1400
 901 ATATCCCACA TCACACATTT ACCCAACTGA TAATTGCCCC AGAGATGAAA AATACTGAGA

 961 AATGAAAAAC ACTTCCAAAC GGCCTTTTCT TCTGTATTGG GGTGCCCCCT CATGGCAAGT
                                                -1300
1021 GGCAACTGCA CTCGTATTTC CTCAATACAT ACCACAGATG GGCTACCCAA TGAACAACTG

1081 CATATTCTCA CCATATGCCC AACAAAAGCC ACTGTCTTCC TCCAGTCCCA TTCTTTAGTA
-1200                                   RsaI                 b3/c2
1141 ATTTCCCTTT ACCTAAATCA AGGTACAATT TTCCAGTCTA ATGGTAATGC CACAAAACTA
                                                               -1100
1201 TATTGTGTCC CAACCAACAA GGTGTTTATC CCCTCTCCTT TGCCCTTCCA TTACTCTGAA

1261 ACTGGAAGCC ACGAGATGCA TTCTGGCCTC TCTTGGCATT CCTATTTATA TAAATGTTAA
                                                -1000           RsaI
1321 TTATTCTTTC CCTACCCAGC CAAATGTCAT TGATGTGCCA CATTTGTACC TATAATACTG

1381 GGACTCACCA CTGCTCTGGG ACTTGCTGCG ATGGCCACAA GGACC
```

Fig. 4.14 Sequence of 1.4 kb *Ava*II fragment upstream of the promoter region. Potential transcription factor binding sites are underlined. (See Section 4.6 and Fig. 4.21). Positions (relative to +1) are approximate as sequence around -900 is not well defined.

4.3 A chromosome walk to isolate recombinants lying between λ104 and λ203

In order to test whether the genomic sequences isolated in \104 and λ204 lay on a contiguous region of the chromosome, the linkage of these two clones was analysed using pulsed-field gel electrophoresis (PFGE). Fig. 4.15 shows the result of a PFGE experiment in which K562 DNA has been digested with *Sal*I, *Cla*I, *Sst*II and *Nae*I, and probed with one probe from the 3'-end of λ203 (λ203H-XG) and another from the 5'-end of λ104 (λ104*Xba*D). Both probes hybridized to the same *Sal*I, *Sst*II and *Nae*I bands showing that the recombinants isolated were in fact linked and less than 150 kb apart.

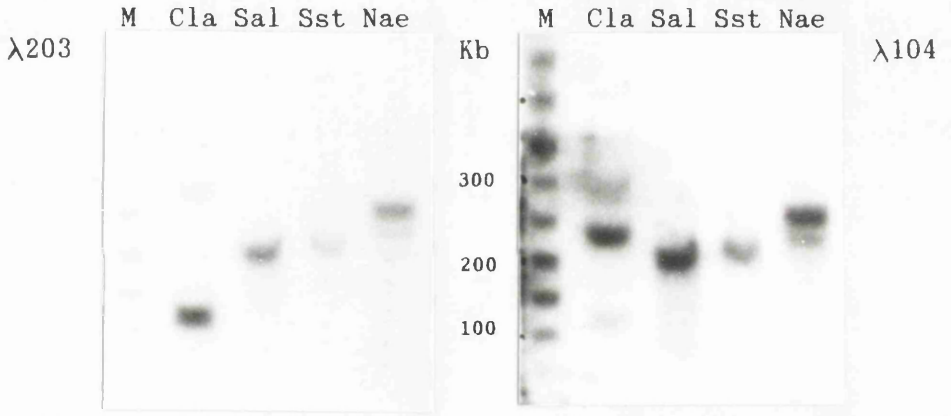


Fig. 4.15 Demonstration of linkage of λ104 and λ203 using pulsed-field gel electrophoresis (PFGE). K562 DNA was digested with *Sal*I, *Cla*I, *Sst*II and *Nae*I, southern blotted and sequentially probed with the 3'-end of λ203 (203*Hind*III-*Xba*IG fragment)and the 5'-end of λ104 (104*Xba*ID fragment). Both probes detect the same *Sst*II and *Sal*I and *Nae*I bands.

To isolate recombinants upstream of λ104, the λ2001 genomic library was re-screened with the 2.0 kb *Xba*I probe from the 5'-end of λ104. Five potentialy useful recombinants were identified and designated λ301-λ305. Mapping of these showed that λ301 extended 6.5 kb 5' of λ104 while λ303 extended another 5 kb beyond this i.e. 11.5 kb 5' of λ104. Hybridisation of these recombinants with the *Hind*III-*Xba*I fragment from the 3'-end of λ203 showed no cross

hybridisation.

A 2.5 kb *Eco*RI fragment from the 5'-end of λ303 was identified as a suitable probe for further genomic library screening, and used to isolate several recombinants designated λ401-λ403. λ403 was found to extend some 5 kb upstream of λ303 and a 3.6 kb *Xba*I fragment (fragment C) from the 5'-end of this recombinant was subcloned into the vector Bluescript for use as probes (pBKS403XbaC).

4.3.1 λ203 and λ403 share overlapping sequence

With the isolation of λ303 and λ403 over 25 kb of sequence had been cloned upstream of the protein coding sequence. In order to test how close these recombinants might be to the promoter containing recombinants, duplicate southern blots were prepared of human placental DNA digested with various enzymes. These were then hybridized with probes from the 3'-end of λ203 (*Hind*III-*Xba*I,G) and the 5'-end of λ403 (subfragments of λ403 *Xba*I,C). Several bands were seen to be common to both probes including fragments for *Bam*HI, *Eco*RI, *Sph*I and *Tth*III (Fig. 4.17,A). The smallest of these shared fragments (*Sph*I) was 7 kb in size indicating close proximity of the two clones. Because restriction mapping had shown that a *Sma*I site was found at the 3'end of λ203 and the 5'end of λ403 it was suspected that this region overlapped. To test this, detailed restriction mapping of the subcloned end-fragments of these recombinants was carried out (Fig. 4.17,B). This showed that these recombinants overlapped by about 2 kb.

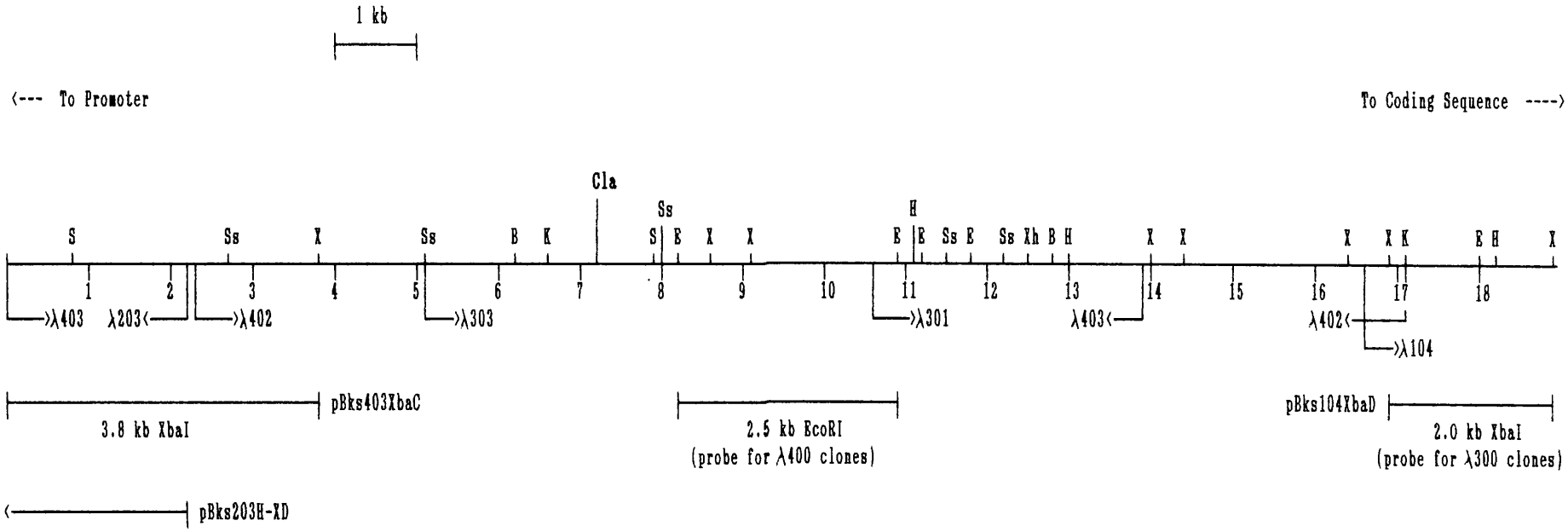
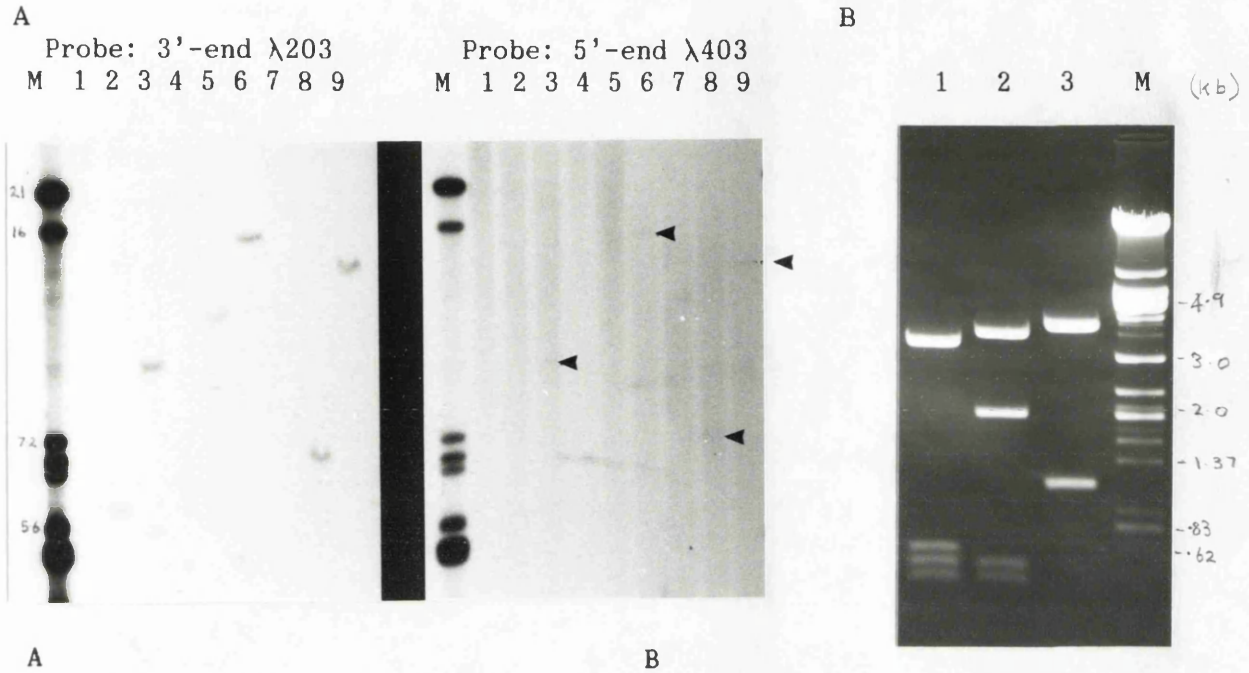


Fig 4.16. Map of the region covered by the λ300 and λ400 series of recombinants. Below the map is shown those regions used as probes for the chromosome walking. The 2.0 kb *Xba*I fragment was used to isolate the λ300 recombinants while the 2.5 kb *Eco*RI fragment from λ303 was used to isolate the λ400 recombinants. The *Cla*I site which separates the two probes shown in the PFGE experiment shown in Fig 4.18 is shown 7.2 kb from the 5'-end of λ403.



A
1. *Apa*I 4. *Bgl*I 7. *Hind*III
2. *Ava*I 5. *Bst*EII 8. *Sph*I
3. *Bam*HI 6. *Eco*RI 9. *Tth*III

B
1. pBks 203H-XD, *Sma*I+*Acc*I
2. pBks 403XbaD, *Sma*I+*Acc*I
3. pBks 203H-XD, *Sma*I+*Eco*RVI

Fig. 4.17 λ203 and λ403 share overlapping sequence. A: Southern blot of human genomic DNA digested with various enzymes and probed with a probe from the 3'-end of λ203 (λ203H-XG, left panel) and from the 5'-end of λ403 (λ403XbaD, right panel). Both probes detect the same *Bam*HI, *Eco*RI, *Sph*I and *Tth*III fragments (arrowed in right-hand autoradiograph), the smallest of which is the 7 kb *Sph*I band (lane 8). B: Lanes 1 and 2 show two different plasmids containing the subcloned ends of λ203 (Lane 1) and λ403 (Lane 2) cut with *Sma*I+*Acc*I. Both give the same fragments of 600 bp and 550 bp

4.4 Summary of *CA1* recombinant isolation

Altogether five sets of recombinants were isolated covering the *CA1* transcription unit. Of these JB#2, JB#5 and λ104 contained the protein coding sequence and was isolated using a cDNA probe, as did recombinant H24 which was isolated in the laboratory of Dr. R. Tashian. Another set - the λ200 series - containing the promoter sequences was isolated using the 5'-cDNA oligo probe. The other isolates (the λ300 and λ400 series of recombinants) were isolated by walking along the chromosome from the coding sequence to the upstream promoter.

In all 65 kb of genomic sequence was cloned with the coding sequence found in a region of 14 kb, separated from the promoter by about 36 kb (Fig. 4.18).

In order to retain some consistency of nomenclature between the different *CA* genes the numbering of the coding exons is the same as that given to *CA2* and *CA3*, with the 5' non-coding exons being designated 1a and 1b and the first coding exon designated 1c. This nomenclature however differs from that given to the mouse gene where the first coding exon has been designated Exon 2.

The small insert found in the untranslated leader sequence of the 5'cDNA recombinant λ5'CAI.6 was found to lie some 10 kb upstream of the coding sequence. This was determined by carrying out hybridisations using an oligonucleotide probe specific to this region. The sequence round this exon was determined by J. Sowden (personal communication).

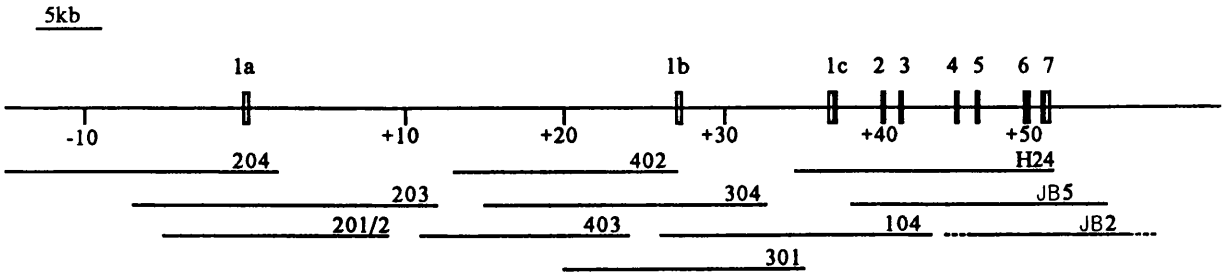


Fig. 4.18 The structure of the *CA1* gene. The scale is numbered in kb from the transcription start point. Exons, marked 1a-7 are indicated by black bars (coding sequence) or open boxes (non-coding sequence) and are not drawn to scale. The annotated lines below indicate the extent of the cloned regions in the various lambda recombinants isolated.

Exon 1a	27kb	Exon 1b
ATTCAGAGCTgtaagtaaca..	intron 5'UTA.	aatcctgtgcttgcctctagGTTTTTCCAC
Exon 1b	9.5kb	Exon 1c
TTTGCAAGAGgcagtaggaa..	intron 5'UTB.	ggtattatttttgttttcagAAAAAGAAAA
Exon 1c	2.8kb	Exon 2
GACAAAAATGgtaacacttc...	intron 1....	acacgtgtttgtcctggtagGTCCTGAACA
Exon 2	1.0kb	Exon 3
AACCGATCAGgtgagctgaa...	intron 2....	tcggttccctttttcttccagTGCTGAAAGG
Exon 3	3.6kb	Exon 4
TTCTGCCGAGgtaatgtaat...	intron 3....	aaccatagtatcatttttagCTTCACGTAG
Exon 4	0.9kb	Exon 5
TTTGATGAAGgtgagttaca...	intron 4....	ttattttcttaaatctccagGTTGGTGAGG
Exon 5	3.0kb	Exon 6
TAAAACCAAGgtaaacacac...	intron 5....	gatgtattcttttcttccagGGCAAACGAG
Exon 6	0.8kb	Exon 7
CTCAGAGCAGgtagagttgt...	intron 6....	aaaatattttatccttctagCTGGCACAAT

Fig. 4.19 Sequence of the *CA1* exon/intron junctions. Exon sequence is shown in capitals, introns in lower case. Coding sequence is shown in bold. The nonconsensus donor splice site of intron 5'UTB is underlined.

4.5 CA1 contains a second promoter adjacent to the coding sequence

While this work was in progress, Fraser and Curtis (1989) demonstrated that a second promoter is found in mouse *CA1* adjacent to the coding sequence. This promoter is responsible for *CA1* expression in colon epithelial tissue and probably the other secondary sites of *CA1* expression.

Comparison of the mouse colon promoter and the human sequence upstream of exon 1c shows considerable sequence homology in this region (Fig. 4.20). Transcript analysis (primer extension and S1 mapping) of human colon RNA confirmed the presence of a colon specific leader sequence (H. Brady personal communication and manuscript in press).

```
Hum  ataattctctttacaatataagaaaattaagcaa-tgaaactaacatagcccttgtagtattttt
       *   **     * *  *  *  *  *  *     * **  *  *      ***     **
Mus  atatttcatttttatattacaataacataaaacaattcatgctgaccta-ccccaat-agacattt

                                        GATA-1
Hum  tacaacacctttttttagatatgtgtacttcctgataagcagagatgatgaaataatggcctatt
         *   *    *  **  *   *           *    *  *        * *  **
Mus  tacaataccatttatttatgtgtatacttcctgatgagcaaagttgatgaaacagtgaactatt

                                                    •+1
Hum  aaaagcaa-ataagtttctataaaaacgcccaagcagggatttaagGCATCTCCTGCATGCACAGT
      * * *    *   *          **     * ** ** *     ***   *     * *** **
Mus  acacggaagacaagcttctataaaa--gcccatggtgtcacttaagatctctgCTGCAGGTGTAAG
                                                          •+1
                                                             GATA-1
Hum  TGCAGTTAGTTATTCCAGGTATTATTTTTGTTTTCAGAAAAAGAAAACTCAGTAGAAGATAATG
               *   *  **   *    * *   *   * ****  *  * *  **   * Met..
Mus  TGCAGTTAGTCATTTCACATATCATTTCTCTTTACAGGAATCACAACCTAAATAAGAGAAAATG
```

Fig. 4.20 Comparison of mouse and human DNA sequence upstream of the coding sequence. Upper case letters indicate transcribed sequence with transcription start sites being those of Fraser *et al.* (1989) for the mouse and H. Brady personal communication for the human. The proposed TATA boxes are underlined. Also shown are the potential binding sites for the transcription factor GATA-1 discussed in section 4.6.2.

4.6 Analysis of the DNA sequence flanking the *CA1* gene

4.6.1 General features associated with control of gene expression

Examination of the sequence obtained from the regions flanking the *CA1* transcription unit (Figs. 4.9,13,14,20) shows there to be a number of potential *cis*-acting sequences known to be associated with control of gene expression.

Close to the (erythroid) transcription start point at -28 a "TATA" box motif is found. This sequence -$^{5'}$CATAAG$^{3'}$- diverges from the canonical TATA box sequence ($^{5'}$TAT$^{A}/_{T}$A$^{3'}$) but is similar to the TATA box of several globin genes in having the first T of the motif replaced with a C residue (Konkel *et al.*, 1979 [human ß-globin], Lawn *et al.*, 1980 [mouse ß^{maj}-globin], Leung *et al.*, 1987 [human θ-globin]). There are no very good candidate sequences for a CCAAT box motif, the other commonly found motif in promoter sequences, upstream of this TATA box although there are three CCAAT like regions between -60 and -90. The CCAAT box is a less highly conserved element in promoters, and is often replaced by G/C rich region which may bind the factor Sp1. A potential binding site for this factor, identical to that found in the HSV-1 intermediate early gene (Jones and Tijan, 1985) is found at -93. This region also contains a sequence which may be recognised by the Oct-1/Oct-2 transcription fators ($^{5'}$ATGCAAAT$^{3'}$) (Fletcher *et al.*, 1987). While Oct-1 is ubiquitous and could possibly influence the activity if this promoter, Oct-2 is unlikely to play a role in the expression of *CA1* since it has been shown to be a lymphoid specific factor (a cell type in which CAI is absent).

Also found at both the 5'- and 3'-flanking sequence of the gene (at -342 and 801 bases downstream from the end of the coding sequence respectively) are binding sites for the factor AP-1 (Lee *et al.*, 1987, Jones *et al.*, 1988).

4.6.2 Sequence motifs associated with erythroid-specific gene expression

Potential binding sites for the GATA-1 erythroid specific factor (consensus sequence $Pu^{T}/_{A}GAT^{A}/_{T}Pu^{T}/_{G}$, Wall *et al.*, 1988) are found at three places in the promoter region at positions -190, -149 and -290, at the 3'-end of the gene 223, 581 and 833 bases downstream of the coding sequence and at the 5' end of the coding sequence, one overlapping the Met codon, and one 158 bp upstream of the coding sequence (position -77 of the colon promoter start site).

The erythroid promoter of the gene also contains a binding site for the CACCC box factor about 20 bp from the most consensus GATA-1 site and a similar clustering of potential GATA-1 and CACCC binding sites are also seen in the sequence of the upstream 1.4 kb *Ava*II fragment. These sites lie about 1.5 kb upstream of the promoter at a position defined as a DNaseI hypersensitive site (J. Sowden personal communication, see also Discussion, Chapter 7). This region also contains a binding site termed b3/c2 (A/T N A/G TAATNNN A/G) which is found in the erythroid promoter of the human porphobilinogen deaminase gene and the 3'-enhancer of the human ß-globin gene (Wall *et al.*, 1988; Mignotte *et al.*, 1989).

The Ap1 site at -320 has already been mentioned as a generalised transcription factor binding site, but this motif may also bind an erythroid specific factor NF-E2 as found in the PBGD gene (Mignotte *et al.*, 1989).

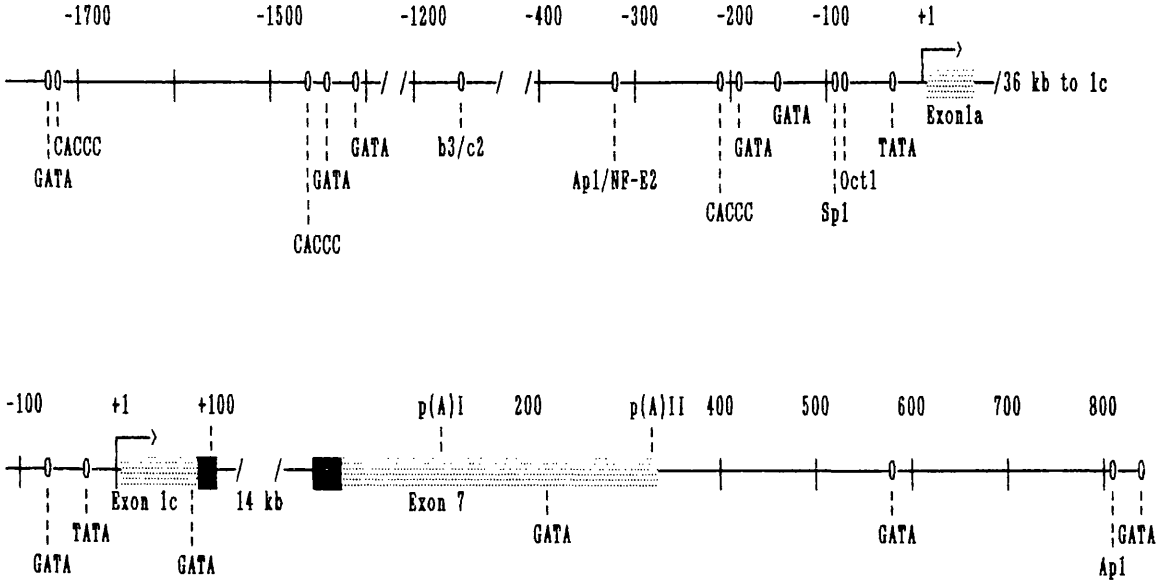


Fig. 4.21 Diagram of potential transcription factor binding sites in the *CA1* gene. The top line shows the region upstream of the erythroid promoter while the lower line shows the region flanking the coding sequence. Potential factor binding sites are indicated by circles. Light shading indicates non-coding transcribed sequence, dark shading indicates coding sequence. The numbering at the 3' end of the gene is distance from the stop codon.

4.7 Cosmid library screening for *CA1*

Following the isolation and characterisation of λ203 it became apparent that due to the large intron in the 5' leader large numbers of recombinants would have to be isolated in order to clone the entire gene. Several cosmid libraries were therefore screened since the larger inserts in this type of vector would reduce the work required for a walking exercise.

Three different cosmid libraries were screened (kind gifts of Dr P. Little, Department of biochemistry, Imperial College, Dr T. H. Rabbitts, M.R.C. Laboratory of Molecular Biology, Cambridge and Dr P. Brickell, Department of biochemistry, University College London). In all a total of approximately 2 X 10^6 recombinant clones were analysed and multiple hybridisations on duplicate lifts were carried out using probes from the large intron of *CA1* (1 kb *Hind*III-*Xba*I from λ203 and 1 kb *Eco*RI-*Xba*I from λ104). Despite the large number of hybridisations carried out no positive clones were detected.

To test the screening procedure, one of these libraries was screened using a 1.8 kb c-*myc* genomic probe. The c-*myc* region of the chromosome in the cell line from which this library was prepared (COLO 320 HSR) was highly amplified, providing a good positive control for the screening process. The results of this screening is shown in Fig. 4.22. Several strong positives are seen indicating that the screening methodology was not responsible for the failure to isolate recombinant clones from this library.

The library was screened again using a 1.4 kb *Ava*II fragment from close to the promoter. This screening revealed 15 positive recombinant clones, but following stripping of the filters no hybridisation was obtained using probe 203 *Hind*III-*Xba*I.G despite the fact that this probe lay only 10 kb from the first probe.

This failure to find cosmid clones using probes for the large intron, despite the apparent presence of several

recombinants for the region adjacent to the promoter (1.4 *Ava*II fragment), should also be set against two other observations. These are that firstly, while carrying out the chromosome walking exercise between λ203 and λ104 the library was re-screened with λ203*Hind*III-*Xba*IC fragment but no other recombinants apart from λ203 were found containing this region. Secondly an earlier screening exercise carried out on a cosmid library using a CAI specific oligonucleotide probe produced isolates which appeared to be unstable i.e. positive colonies could not be purified to homogeneity (J. Barlow personal communication).

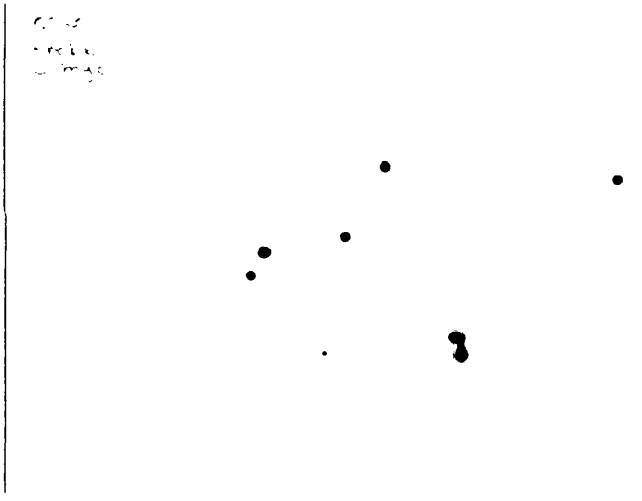


Fig. 4.22 Hybridisation of COLO320 cosmid library with a probe for the oncogene c-*myc* which is amplified in the cell line from which this library was constructed.

CHAPTER 5: PHYSICAL MAPPING OF *CA1, CA2, and CA3*

As described in the introduction, *CA1* has been shown to be closely linked to two other carbonic anhydrase genes *CA2* and *CA3*, lying on the long arm of chromosome 8 (Edwards *et al.*, 1986a,b; Davis *et al.*,1987; Nakai *et al.*, 1987). It has also been shown using pulsed-field gel electrophoresis that the genes lie less than 200 kb apart (Kearney *et al.*, 1987; Venta *et al* 1987).

This chapter describes PFGE experiments carried out to map this cluster of genes.

5.1 Notes on methodology and probes used

The PFGE system used for this work was principally the field inversion (FIGE). This uses a standard agarose gel electrophoresis system in which the polarity of the electrodes is periodically altered (Carle *et al.*, 1986). This system produces straight lanes and allows full use of the gel width without the need for complicated electrode arrays.

Initially a double inhomogeneous field system (Carle and Olson, 1984) was used in which two separate sets of electrodes lie at right angles. This system produces curved lanes leading to two problems: difficulty in sizing and a relatively small useable area of the gel. Examples of these two systems are shown in Fig. 5.1

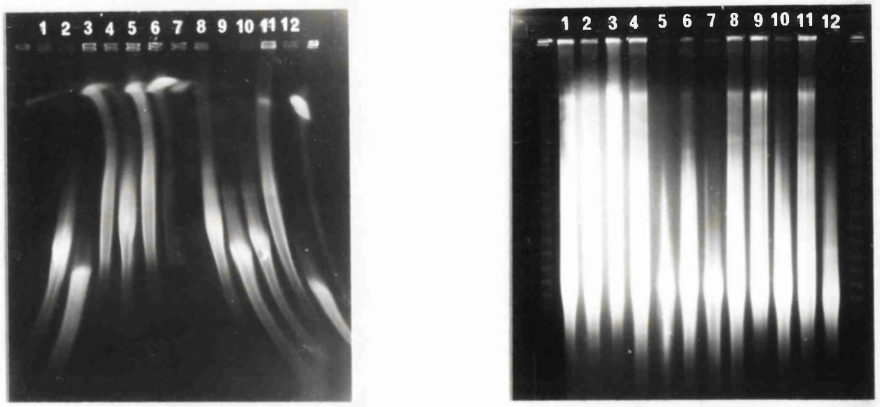


Fig. 5.1 Examples of pulsed field gel electrophoresis using double inhomogeneous field (left) or field inversion (right).

Figure 5.2 shows diagrams of the *CA1*, *CA2* and *CA3* genes and indicates the positions of the probes used in this work relative to the genes. Similar maps are also shown in the fold out Appendix Fig. A2 and A3 which can be used as reference. The 5' 1.4 kb *Ava*II fragment was isolated from a plasmid (pBks204HS4.2) containing the promoter region of *CA1*, a subclone of the lambda recombinant HGCA204. The 1kb *Eco*RI-*Xba*I fragment was isolated from plasmid pBks104XbaD containing the untranslated exon 1b, a subclone of λHGCA104 (see Chapter 4 and Fig A1). The *CA2* probes (2.3 kb *Eco*RI-*Cla*I and 1.5 kb *Eco*RI-*Cla*I) were prepared by *Cla*I, *Eco*RI digestion of plasmid H25-3.8 containing a 3.8 kb *Eco*RI fragment containing exons 1 and 2 of *CA2* together with 1.4 kb of upstream sequence (kindly provided by Richard Tashian, University of Michigan and originally subcloned from the lambda recombinant H25, Venta *et al.*, 1984). The *CA3* probe used was a 2.8kb *Eco*RI-*Hind*III fragment isolated from a plasmid containing the promoter region of the gene, a subclone of the lambda recombinant CA2.1 (Lloyd *et al.*, 1987).

The human erythroleukaemic K562-SA1 cell line (Spandidos *et al.*, 1984) was used for all the PFGE work described here. Some experiments have been repeated with other cell lines including: CEM (T-Lymphoblastoid), H/9 (T-Lymphoblastoid), HeLa (fibroblast) and HEL (erythroleukaemic). Although differences in restriction fragment sizes could be seen between these cell lines, the results were consistent with those found in K562. Differences in banding patterns were all explicable as changes in the methylation state of the recognition sites for the enzymes used in mapping.

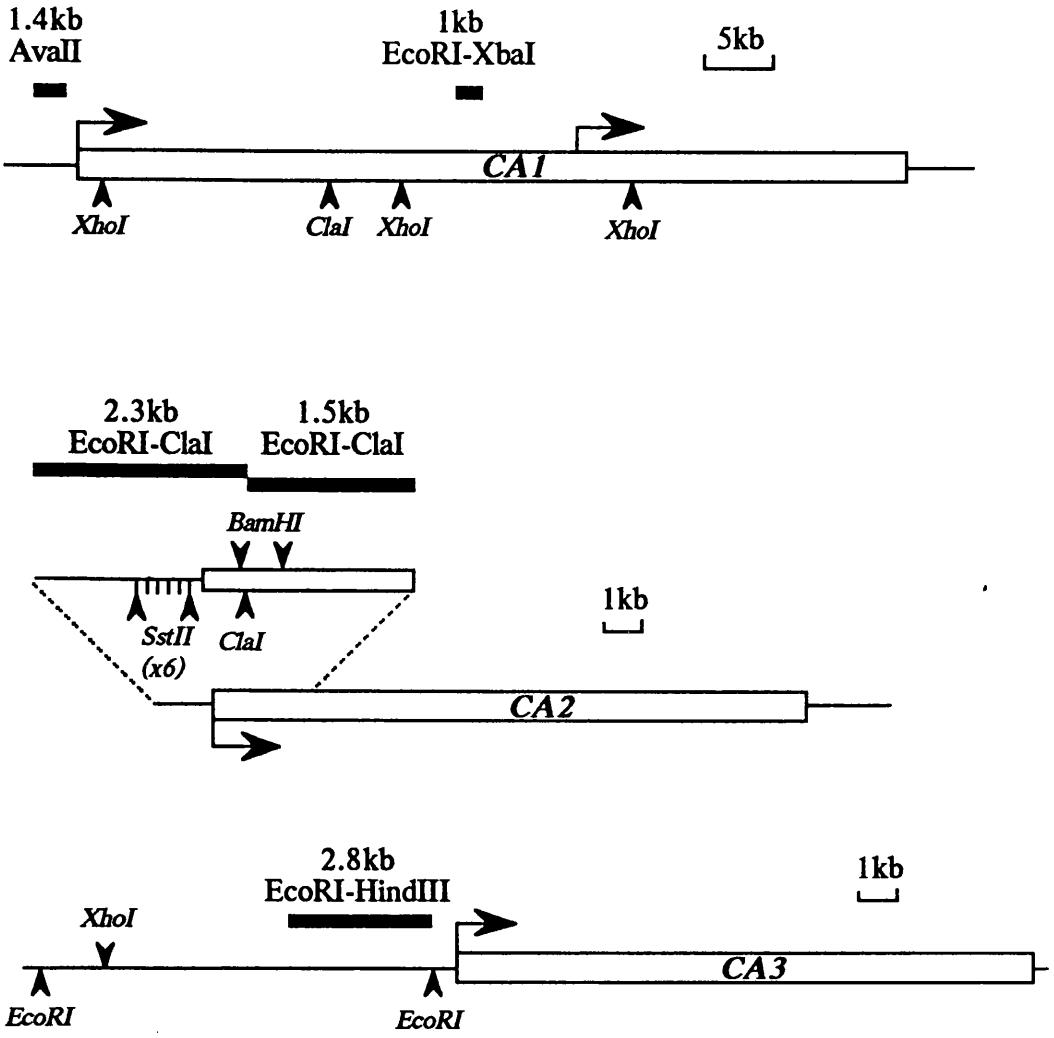


Fig. 5.2 Diagram of probes used for PFGE mapping. The *CA1*, *CA2*, *CA3* genes together with the relative positions of the nucleic acid probes (denoted by bars above the genes) used in the mapping experiments shown in Figs. 5.3-5.5. The extent of the transcribed regions (position of exons not shown) is shown by open boxes with arrows indicating the direction of transcription. All the restriction enzyme sites shown have been mapped in recombinants, and only those sites relevant for the mapping work presented here have been shown.

5.2 PFGE to map *CA1*, *CA2* and *CA3*

5.2.1 Determining the order of the genes: CA3 lies between CA1 and CA2.

Fig. 5.3 shows the results of sequential Southern blot hybridisations to *Sal*I, *Xho*I and *Cla*I digested DNA with genomic probes specific for *CA1*, *CA2* and *CA3*. The *CA1* 1.4 kb *Ava*II and *CA3* 2.8 kb *Eco*RI-*Hin*dIII probes detect a common *Sal*I fragment (170 kb) not detected by the *CA2* 1.5 kb *Eco*RI-*Cla*I probe, while the *CA2* and *CA3* probes detect an *Xho*I fragment (110 kb) which is not detected by the *CA1* probe. Taken together these findings indicate that *CA3* lies between *CA1* and *CA2*, separated from *CA2* by one or more *Sal*I sites and from *CA1* by one or more *Xho*I sites. In addition each gene probe detects a different *Cla*I fragment.

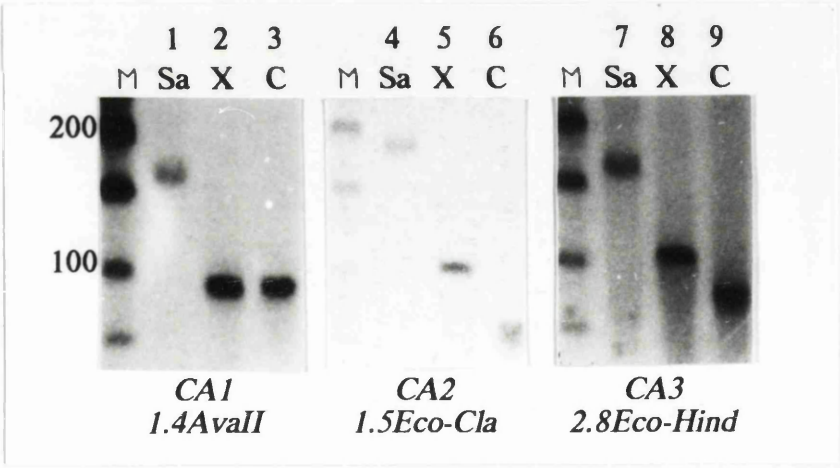


Fig. 5.3 Determining the order of the CA genes. Southern transfer of a pulsed-field gel sequentially hybridised with probes for *CA1*, *CA2* and *CA3* (see Fig. 5.2 for probe details). DNA was digested with *Cla*I (C), *Sal*I (Sa), or *Xho*I (X). Lambda concatamer markers (M) are sized in kb. *CA3* shares a common *Sal*I fragment with *CA1* (lanes 1 and 7) and a common *Xho*I fragment with *CA2* (lanes 5 and 8).

5.2.2 The relative orientation of the genes: *CA1* is transcribed away from, and in the opposite direction to *CA2* and *CA3*.

Mapping of recombinant clones (Chapter 4) shows that a single *Cla*I site exists in the *CA1* gene (see Fig. 5.2). A probe (1.4kb *Ava*II) for the 5' end of the gene detects a *Cla*I fragment of about 80 kb (Fig. 5.3, lane 3) while a probe (1kb *Eco*RI-*Xba*I) which lies 3' to this site detects a band of over 200 kb in size (Fig. 5.4A., lane 1). Since neither of these fragments are detected by the *CA2* or *3* probes, and the *Cla*I fragment extending 3' to the *Cla*I site in *CA1* is larger than the maximum distance apart of these genes, *CA2* and *3* must be located upstream (5') of *CA1* and separated from it by 80 kb or more.

Orientation of the *CA2* gene relative to *CA1* and *CA3* was made possible by the identification of a *Sal*I site between *CA2* and *3* (see above) and a *Cla*I site at the 5' end of the *CA2* gene. This *Cla*I site lies within the region cloned and used for probe preparation (see Fig. 5.2). The *CA2* probe lying 5' to this *Cla*I site (2.3 kb *Eco*RI-*Cla*I) detects a *Sal*I site 10 kb upstream of the promoter (Fig. 5.4B) in double digests using *Sal*I together with *Cla*I (or *Bam*HI, or *Sst*II, sites for which lie close to the *Cla*I site in the recombinant). This probe also detects the same *Cla*I-*Xho*I fragment as the *CA3* probe (see Fig. 5.5). The probe (1.5 kb *Eco*RI-*Cla*I) lying 3' of this *Cla*I site on the other hand fails to detect a *Sal*I site downstream of the gene, and hybridises to a *Cla*I-*Xho*I fragment not detected by the *CA3* probe. These data confirm the observation of Venta *et al.* (1987) that *CA1* (and from this study *CA3*) lies 5' to *CA2*.

The relative orientation of *CA3* was determined using an *Xho*I site lying 5 kb upstream of the transcription start site which had been mapped in recombinant clones (Y. Edwards, personal communication). Since the ability of the

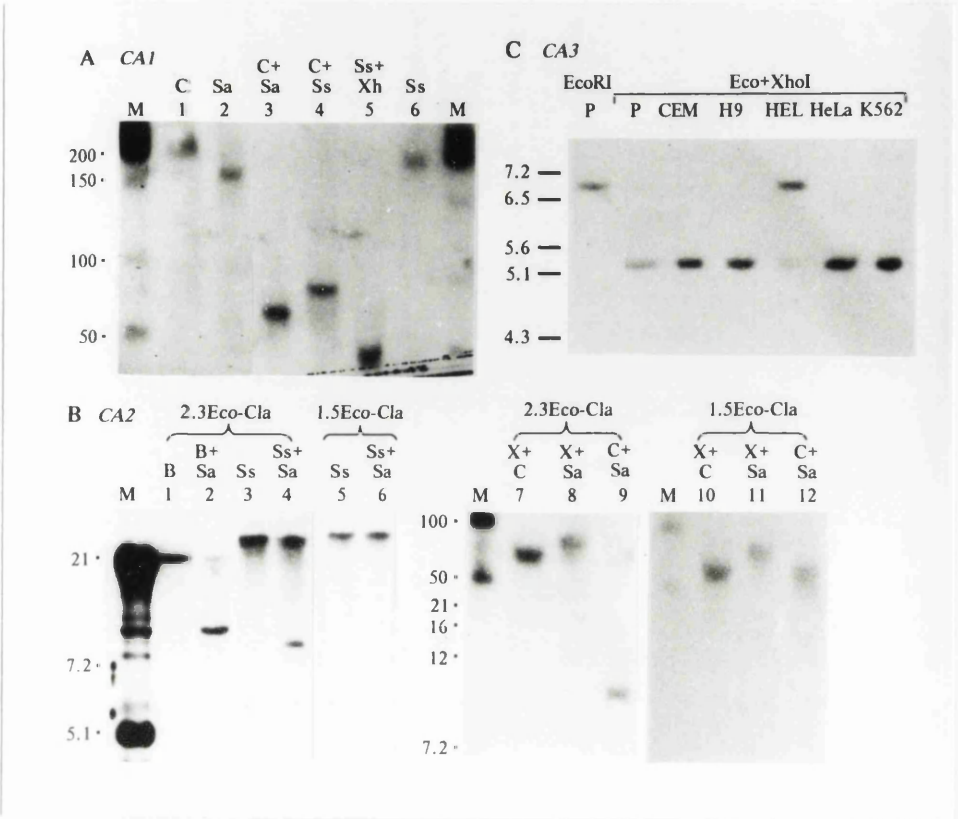


Fig. 5.4 Determining the orientation of the genes. A: Southern transfer of pulsed-field gel hybridised with a *CA1* probe (1 kb *Eco*RI-*Xba*I) lying 3' to the single *Cla*I site in the gene. The 200 kb *Cla*I fragment seen in lane 1 is larger than the maximum separation between the genes and does not contain either *CA2* or *CA3* which lie on smaller fragments (see Fig. 5.3). Lanes 3 and 4 contain double digests of *Cla*I together with *Sal*I or *Sst*II, and show the position of these sites downstream of *CA1*. Panel B: Southern transfer of standard and pulsed-field gels hybridised with *CA2* probes lying 5' (2.3 kb *Eco*RI-*Cla*I) or 3' (1.5 kb *Eco*RI-*Cla*I) to the single *Cla*I site in *CA2*. Lanes 1-4 and 7-9 probed with 5' probe, lanes 5, 6 and 10-12 probed with 3' probe. Double digests with *Sal*I together with either *Bam*HI (lane 2), *Sst*II (lane 4) or *Cla*I (lane 9) using the 5' probe detects a *Sal*I site 10 kb upstream of the gene. This site lies between *CA2* and *CA3* (see Fig. 5.3). No site is detected using the 3' probe (lanes 6 and 12). Panel C: Southern transfer, hybridised with a *CA3* probe (2.8 kb *Eco*RI-*Hin*dIII), of DNA from several cell lines digested with *Eco*RI and *Xho*I. The parental *Eco*RI fragment of 7 kb (lane 1) is reduced in size by *Xho*I, indicating that the single *Xho*I site (5kb upstream of the gene) identified in *CA3* recombinant clones is susceptible to digestion

so-called rare-cutter restriction enzymes used in this work to digest DNA is affected by DNA methylation state, digests were carried out to show this site was in fact being cut in the cell line used in this work (K562). This is shown in Fig. 5.4C in which a number of different cell lines have been digested with *Eco*RI and *Xho*I, *Xho*I reducing the *Eco*RI band from 7 to 5.2 kb. Having established that this site is indeed susceptible to digestion, and that *CA2* and *CA3* share a common *Xho*I fragment, it becomes apparent that *CA2* lies downstream (3') of *CA3*.

5.2.3 Distance separating the genes.

The distance between the *CA1* and *CA2* genes can be determined from the size and termini positions of the *Sal*I fragment on which both *CA1* and *CA3* lie. One end of this fragment has been shown to lie between *CA2* and *CA3* 10 kb 5' to *CA2* (see above). The other end of this fragment, lying approximately 70 kb 3' to the *Cla*I site in the *CA1* gene, can be detected with a *Cla*I/*Sal*I double digest using probe *CA1* 1kb *Eco*RI-*Xba*I (Fig. 5.4A, lane 3). This data provides an estimate of about 110 kb as the distance between the two genes. A similar figure is arrived at from calculations based on the size of a 200 kb *Sst*II fragment which also contains both *CA1* and *3* sequences (Fig. 5.4A, lane 6 and data not shown). This fragment is roughly co-linear with the *Sal*I fragment, one end lying slightly further 3' to *CA1* than the *Sal*I site (compare lanes 3 and 4, Fig. 5.4A), while the other end lies near the promoter of *CA2* (within the region cloned in H25-3.8).

The distance between *CA2* and *CA3*, can be determined by making use of the *Cla*I site in *CA2* and the *Xho*I site in *CA3* to fix the position of these genes using a *Cla*I/*Xho*I double digest. However while carrying out this work it became apparent that there were certain anomalies in the sizing of these fragments. For example the *Xho*I fragment detected by *CA2* and *CA3* probes has an apparent size of about 110 kb (Fig.

5.3, lanes 5 and 8). This fragment can be subdivided with *Cla*I and the fragments produced hybridised with the two *CA2* probes (2.3 kb *Eco*RI-*Cla*I and 1.5 kb *Eco*RI-*Cla*I) separated by a *Cla*I site (see for example Fig. 5.4B, lanes 7 and 10). The apparent size of both these bands is greater than 60 kb which, when summed together, exceeds the size of the parental *Xho*I fragment. Similar problems arose when considering the sizes of sub-fragments of the *Sal*I band containing *CA1* and *CA3* sequences. This suggested that either the smaller fragments have an aberrantly low mobility, or the larger fragments have an aberrantly high mobility in this electrophoretic system. Since the size of the *Sal*I fragment was in good agreement with estimates from other workers (R. Tashian, personal communication.) it was suspected that the mobility of the smaller DNA fragments differed from that of the molecular size markers. To test whether this was the case, a conventional (unpulsed) gel was used to size the *Cla*I-*Xho*I fragment containing *CA3* and the 5' end of *CA2* and showed that indeed this fragment should be sized below 50 kb (Fig. 5.5).

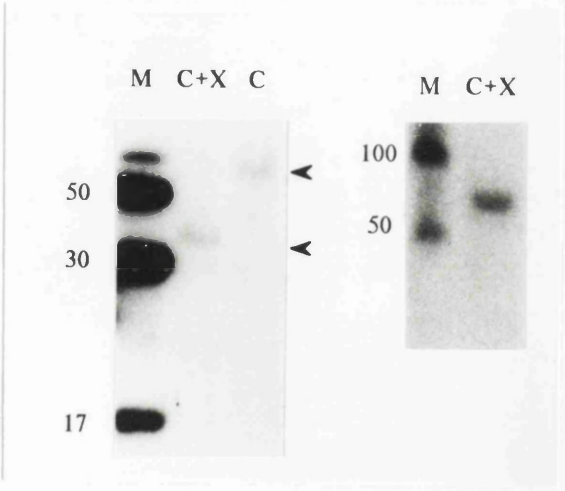


Fig. 5.5 Comparison of fragment sizes in PFGE and unpulsed gels. Southern transfer, probed with *CA3* 2.8 kb *Eco*RI-*Hind*III. The gel on the left shows a *Cla*I and *Cla*I,*Xho*I digest separated on an unpulsed 0.5% agarose gel. The right hand gel shows a similar *Cla*I/*Xho*I digest separated by FIGE. The *Cla*I,*Xho*I band which appears to be 60kb using the field inversion system is sized at about 40 kb using an unpulsed gel.

In pulsed field gels, migration of restriction enzyme fragments has previously been reported to be slower than expected in regions of high local DNA concentrations (Michiels *et al.*, 1987). This phenomenon probably accounts for the anomalous running positions of the smaller fragments seen using our field inversion system.

The map shown in Fig. 5 has been drawn taking the above considerations into account. It is assumed that there are no additional sites for either *Cla*I or *Xho*I between *CA1* and *CA3* i.e. that the *Xho*I and *Cla*I fragments containing the 5' end of *CA1* (Fig. 2, lanes 2 and 3) abuts the fragments containing *CA3* (Fig. 2, lanes 8 and 9). Although the possibility of extra sites between these genes cannot be discounted, the size constraints within which identified fragments have to be fitted and the relative rarity of these sites would seem to make this unlikely.

5.3. Summary of PFGE mapping data

The genes lie in the order *CA2*, *CA3*, *CA1*, with *CA2* and *CA3* being transcribed in the same direction, away from *CA1*, which is transcribed in the opposite direction. *CA2* and *3* are relatively close together, with a gap of about 20 kb between the 3' end of *CA3* and the 5' end of *CA2*. *CA1* lies further away with the 5' end of *CA1* lying about 80 kb from the 5' end of *CA3* (the distance between the coding regions of *CA1* and *CA3* is in fact over 110 kb because of the large intron in the 5' leader of *CA1*). It is not known how these genes lie on the chromosome in terms of centromere/telomere orientation.

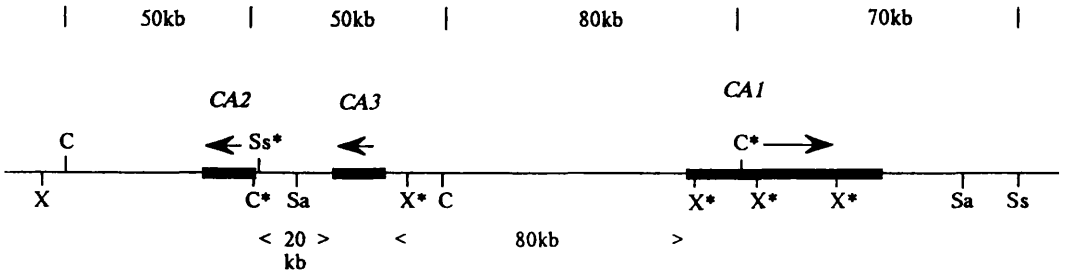


Fig. 5.6 Map of the human carbonic anhydrase locus located on the long arm of chromosome 8 (8q22). The three genes found at this locus - *CA1*, *2* and *3* - are indicated by black boxes, with arrows indicating the direction of transcription. The distances between the genes are shown below the map (sizes in kb). Letters indicating the sites for various enzymes are: C, *Cla*I; X, *Xho*I; Sa, *Sal*I; Ss, *Sst*II. All the sites shown are those susceptible to digestion using K562 cell-line DNA, but may be resistant to digestion in other cell types. Those sites which have been mapped in recombinant clones are marked with an asterisk. The distances between selected sites are shown above the map.

CHAPTER 6: METHYLATION ANALYSIS OF THE *CA* GENES

6.1 Notes on cell lines and methodology.

This chapter presents data on methylation states of CpG dinucleotides in and around the *CA1* gene. As described in Chapter 1, the methylation state of cytosine in CpG dinucleotides can vary from cell type to cell type, usually with a lower level of methylation found in the cell type in which the gene is active (Cedar, 1988). Analysis of the methylation state of DNA in human erythroid cells is hampered by the nature of the source material. Nucleated erythroid cells are found only in small numbers in the bone marrow of adults, making studies on in vivo material difficult. A number of workers have however used erythroleukaemic cell lines to study methylation patterns in the vicinity of the globin genes. These studies have largely supported the notion of hypomethylation being associated with gene expression (Bird *et al.*, 1987b; Enver *et al.*, 1988a).

Two cell lines in particular are of interest with regard to analysis of erythroid *CA1* expression, the HEL-92 cell line (Martin and Papayannopoulou, 1982) and the K562 (SA1) cell line (Spandidos 1984). These cell lines appear to mimic different developmental stages of erythroid tissue. HEL cells express both *CA1* and low levels of ß-globin and for this reason may be considered to have a foetal/adult phenotype (Enver *et al.*, 1988b). K562 cells on the other hand express neither *CA1* or ß-globin, but do express foetal and embryonic globins and are considered to represent an earlier foetal or foetal/embryonic stage of development (Benz *et al.*, 1980; Enver *et al.*, 1988b). It has been shown that the lack of ß-globin in K562 cells can be overcome by fusion with the MEL C88 mouse erythroleukaemic cell line which has an "adult" phenotype (Wright *et al.*, 1983; Baron and Maniatis, 1986) and, subsequent to the isolation of *CA1* recombinant clones, the same phenomenon has been shown for *CA1* (Brady *et al.*, 1990). In these short term fusion experiments (transient

heterokaryons), nuclei do not fuse, and this trans-activation suggests that *CA1* and ß-globin genes in K562 are responding to developmental stage specific trans-acting factors. By contrast the non-erythroid HeLa cell line (Gey *et al.*, 1952) does not show this type of trans-activation (Brady *et al.*, 1990; Baron and Maniatis, 1986) and so it might be considered that in some way the *CA1* and ß-globin genes are poised for expression in K562 cells, perhaps by having a suitable chromatin conformation.

The experiments below were therefore largely concerned with comparing the methylation state of *CA1* in HEL and K562 cell lines. Other cell lines have also been examined including the HeLa cell line and other non-erythroid cells, although not in the same detail.

Methylation state was determined by restriction digests and Southern blot analysis using the enzyme *Hpa*II which recognises the sequence CCGG. Methylation of the cytosine in the central CG doublet renders this site insensitive to digestion by *Hpa*II, making the site "invisible" on Southern analysis. The *Hpa*II isoschizomer *Msp*I is insensitive to methylation and can be used to disclose the position of *Hpa*II sites independent of methylation state.

In addition to this examination of *CA1* some information regarding methylation states of the whole region round the *CA1, 2* and *3* genes became available as a by-product of the PFGE mapping experiments described in Chapter 5. This data is described in Section 6.3.

6.2 Methylation analysis of *CA1*

Figs 6.1 to 6.4 show the results of experiments carried out to determine the methylation states of *Hpa*II sites along the length of the *CA1* gene. These experiments used *Hpa*II single digests and/or *Hpa*II+*Kpn*I and *Hpa*II+*Sst*I double digests. In each figure is a map of the region examined with relevant restriction sites shown. The map positions (in kb) in these figures are the same as those shown in the Appendix figure A1. Each of the *Hpa*II sites analysed has, for

convenience, been designated a letter code (A to O) and the positions of these within the *CA1* gene is shown in the fold out Appendix Fig. A2. This figure also shows the location of the probes used in this work and should be used in conjunction with figures 6.1 to 6.4.

6.2.1 Methylation patterns at the 5' end of *CA1*.

Methylation at the 5' end of the *CA1* gene was studied by probing Southern blots of *Hpa*II digests with a 1.4 kb *Ava*II genomic fragment from 1 kb upstream of the erythroid promoter. Four *Hpa*II sites lie in this region A1 and A2 at -4.75 and -4.5 and B1 and B2 at +1.25 and +1.5 (see Fig. 6.1). Analysis was carried out on DNA from two erythroleukaemic cell lines, K562 (non-*CA1*-expressing) and HEL (*CA1*-expressing), together with a number of non-erythroid lines: CEM (pre T-cell, Foley *et al.*, 1965), H9 (T-cell, L. Goff, Department of Haematology, University College Hospital, personal communication), HeLa (fibroblast, Gey *et al.*, 1952) and SW480 (colon carcinoma, Leibovitz *et al.*, 1976).

In *Hpa*II single digests (lanes 1-6) it was found that HEL DNA appeared to have the highest level of methylation, producing high molecular weight bands (>25 kb, lane 2). DNA from K562 and CEM showed multiple bands indicating partial methylation while the other cell lines (lanes 3, 4, and 6) appeared to be completely demethylated producing a band size the same as that produced by *Msp*I (lane 7). These single digests were not informative as to the pattern of methylation at individual sites and more information was generated by double digests.

Digestion with *Hpa*II+*Kpn*I allows analysis of sites B1 and B2 downstream of the probe while *Hpa*II+*Sst*I digestion allows analysis of sites A1 and A2 upstream. Only K562 and HEL were analysed in this way. In K562 DNA all four sites were partially methylated since bands corresponding to digestion at each *Hpa*II site can be detected as well as the parental *Kpn*I and *Sst*I bands (lanes 10 and 11).

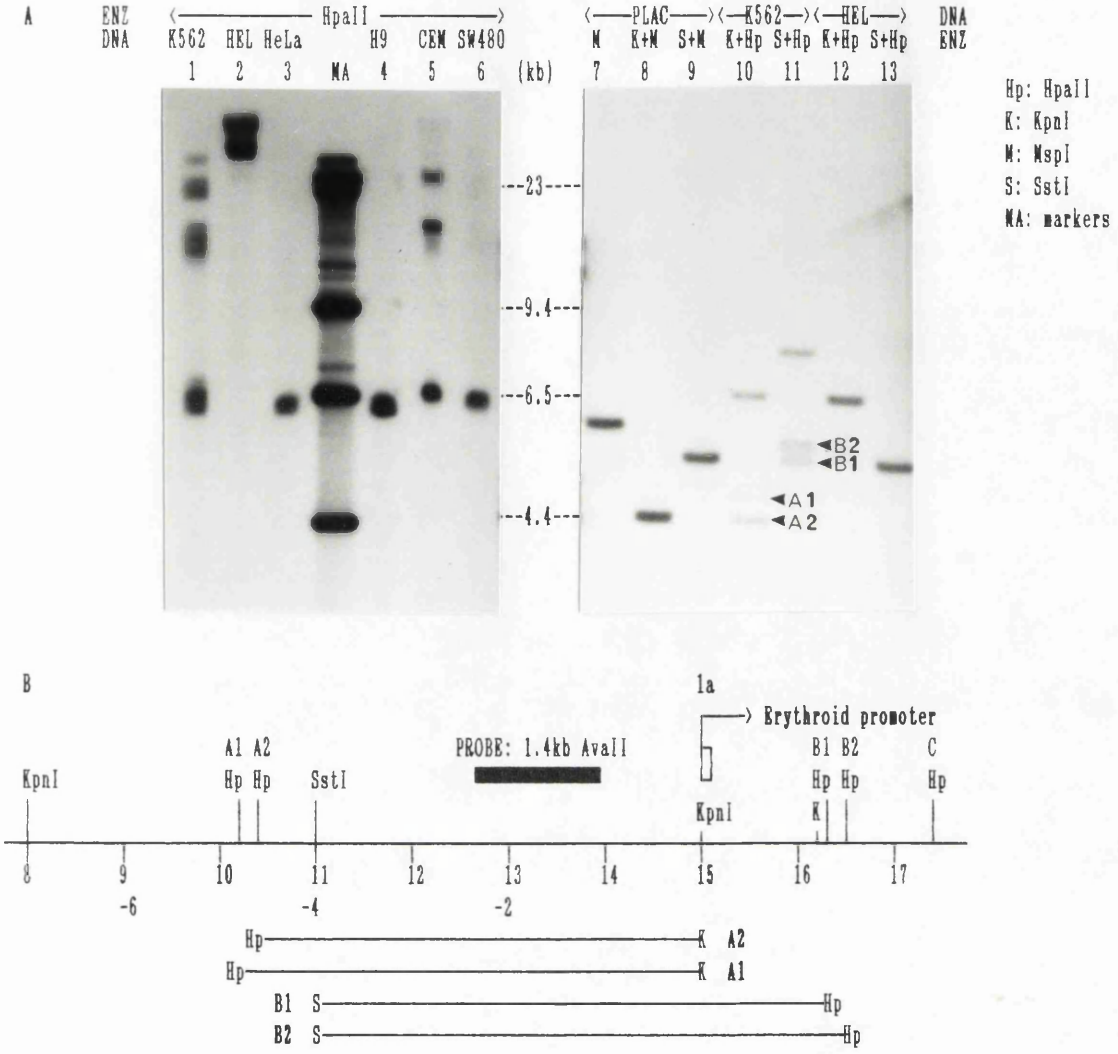


Fig. 6.1 *Hpa*II analysis of methylation states at the erythroid promoter.

A: Southern blot of DNA digested with *Hpa*II (Hp), *Msp*I (M), *Kpn*I (K) or *Sst*I (S) as indicated and hybridised with the 1.4kb *Ava*II fragment probe. Lanes 1-6 show *Hpa*II digests of DNA from the K562, HEL, HeLa, H9, CEM, and SW480 cell lines. Lanes 10-13 double digests of K562 and HEL DNA. Lanes 7-9 show placental DNA digested with the methylation insensitive *Hpa*II isoschizomer *Msp*I (together with *Kpn*I or *Msp*I as indicated). The arrowed bands in lanes 10 and 11 correspond to the fragments labelled in B. The highest bands in these two tracks are the *Kpn*I (7.0kb) and *Sst*I (8.5kb) fragments.

B: Map of the region around the erythroid promoter with particular fragments produced in the digestions indicated by lines below the map (not all fragments are indicated). The map positions are those of Fig. A1. Only *Hpa*II, *Kpn*I and *Sst*I sites are shown.

In HEL DNA the upstream sites A1 and A2 are completely methylated and onlythe *Kpn*I fragment is detected in *Hpa*II+*Kpn*I digests (lane 12) while the downstream site B1 is completely unmethylated, producing a fragment size in *Hpa*II+*Sst*I digests (lane 13) identical to that of *Msp*I+*Sst*I (lane 9)

6.2.2 Methylation patterns within the large intron and at exon 1c.

*Hpa*II sites C, D and E lie 2.3, 9.3 and 10.4 kb downstream of the erythroid promoter, within the large intron. Methylation at these sites was assessed using a 1 kb *Hind*III-*Xba*I probe which covered site D (Fig. 6.2). As in Fig. 6.1 *Hpa*II single digests of K562, HEL, HeLa, CEM, H9 and SW480 were analysed as well as double digests of K562 and HEL.

In single digests HEL DNA appears to be highly methylated producing a high molecular weight band, CEM DNA is partially methylated producing multiple bands, while HeLa, H9 and SW480 DNA are unmethylated producing a band size identical to that seen in *Msp*I digests (lane 7). K562 DNA which was partially methylated at sites A1, A2, B1 and B2 also appeared to be largely demethylated although some faint high molecular weight bands could be seen (lane 1). Double digests of HEL DNA (lanes 12 and 13) failed to detect demethylation at any of sites C, D or E or at any sites downstream of the probe up to 16 kb 3' to exon 1a.

The region around the colon promoter was analysed using a 0.8 kb *Eco*RI probe from 0.3 kb upstream of exon 1c. *Hpa*II site J lies within this probe with site I positioned 0.8 kb upstream of this and site K 3.5 kb downstream, close to exon 2 (Fig. 6.3). K562 DNA shows no sign of methylation at any of these sites and *Hpa*II digestion (lane 1) gives the same fragment sizes as those seen in *Msp*I digests (lane 3). HEL DNA, as found previously, is highly methylated and in *Hpa*II+*Kpn*I and *Hpa*II+*Sst*I digests only bands corresponding to *Kpn*I or *Sst*I fragments were detected (lanes 7 and 8). These *Kpn*I and *Sst*I fragments are large (22.5 and 16 kb) and show that in DNA from this cell line no *Hpa*II sites are cut in the region from 15 kb 5' of exon 1c to close to exon 6. Other data (not shown) also indicated, as found in analysis of other sites, that CEM DNA was partially methylated in this region while DNA from HeLa and H9 was completely digested. The methylation state of SW480 could not be determined.

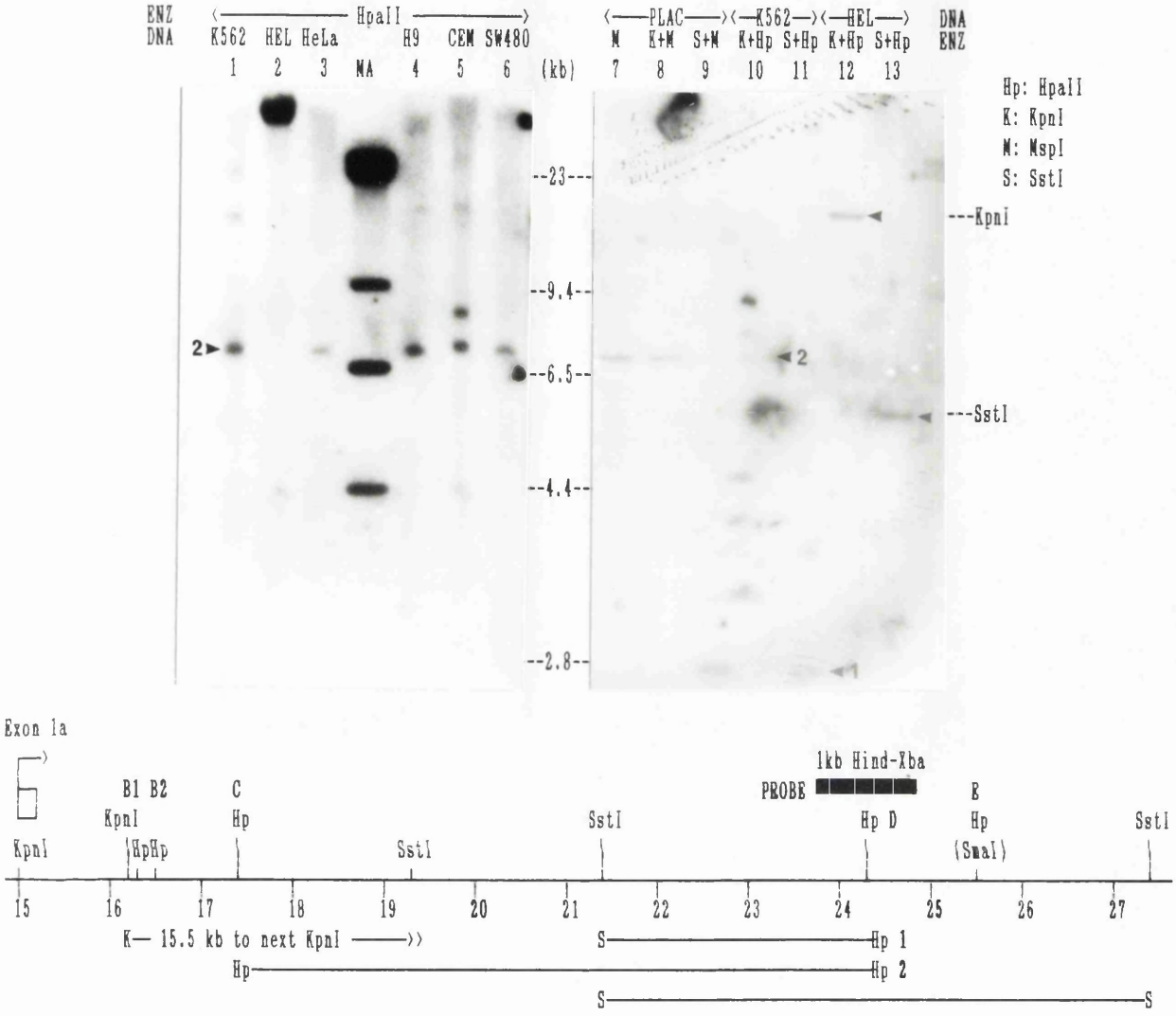


Fig. 6.2 Methylation analysis within the large intron of *CA1*. Sites C, D and E analysed using the 1kb *Hind*III-*Xba*I probe. Details of the figure are the same as those in Fig. 6.1. Lanes 1-6: *Hpa*II digested DNA from the cell lines K562, HEL, HeLa, H9, CEM and SW480. Lanes 7-13: Double digests of K562 and HEL DNA using *Hpa*II together with *Kpn*I or *Sst*I. The position of the *Kpn*I and *Sst*I fragments is shown to the side of the figure. The arrowed bands are those shown in the map beneath. The small fragment created by digestion at sites D and E cannot be seen on this blot, but is detected in experiments where the gel is not run as far as shown here (data not shown). The strong band of 8 kb in size seen in the CEM digest (lane 5) may indicate partial digestion at site D.

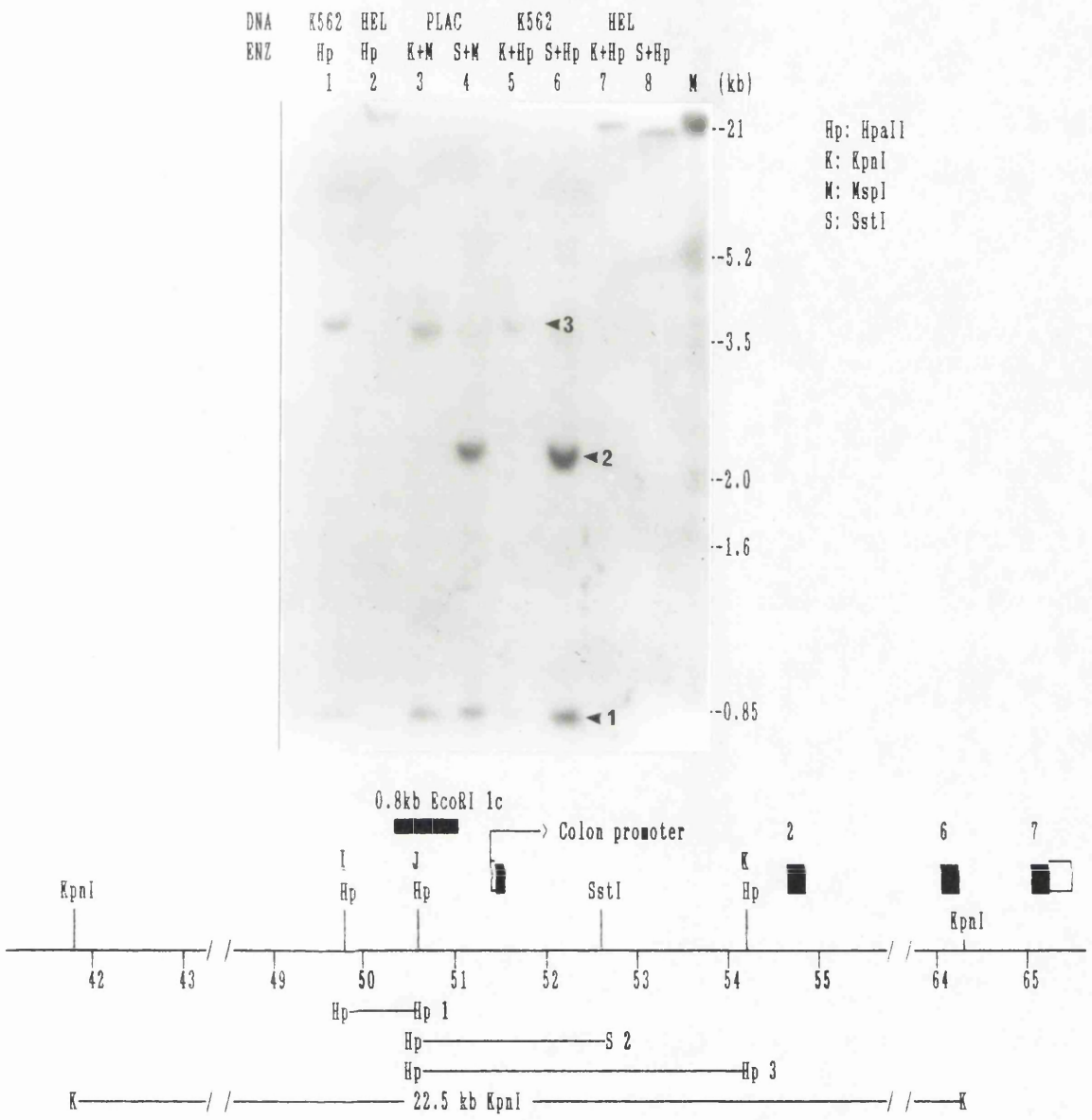


Fig. 6.3 Methylation analysis of the region around exon 1c. K562 and HEL cell DNA digested with *Hpa*II (lanes 1 and 2) alone or with *Kpn*I (lanes 5 and 7) and *Sst*I (lanes 6 and 8). Lanes 3 and 4 are *Kpn*I+*Msp*I and *Sst*I+*Msp*I controls. The probe used was a 0.8 kb *Eco*RI fragment lying just 5' to exon 1c allowing analysis of sites I, J and K. All other details as Fig. 6.1. The large fragments detected in lanes 7 and 8 close to the 21 kb marker are the sizes expected (from recombinant mapping) from *Kpn*I and *Sst*I digests (22 and 16 kb).

6.2.3 Methylation patterns at the 3'-end of *CA1*.

*Hpa*II+*Kpn*I and *Hpa*II+*Sst*I digests of DNA from K562 and HEL cells were hybridised using a 0.7 kb *Hind*III-*Rsa*I probe covering exon 7 and extending into 3'-flanking sequence. Three *Hpa*II sites were analysed in this way, sites M and N lying between exons 5 and 6 and site O lying 4.2 kb downstream of exon 7 (see Fig. 6.4).

K562 DNA is completely unmethylated at site O and produces a 5.2 kb fragment (band 1) in *Hpa*II+*Kpn*I digests (Fig. 6.4, lane 3). Site N upstream of exon 6 is partially methylated and allows the fragment produced from digestion at sites M and O to be seen in the *Hpa*II+*Sst*I digest (band 4, lane 4). Site M appears to be completely digested, i.e. unmethylated, and no fragments can be detected above band 4.

In HEL DNA site O is partially methylated and the *Kpn*I fragment (band 2 in lane 4) can still be detected in *Hpa*II+*Kpn*I digestion. Sites M and N are completely methylated and the 14 kb *Sst*I band (extending almost to exon 3) is seen in *Hpa*II+*Sst*I digests as well as a slightly smaller band formed by partial digestion at site O.

6.3 Summary of *CA1* methylation analysis

The pattern of methylation in the cell lines examined is shown in Fig. 6.5. A wide variation in methylation level has been found in the various cell types examined. HEL cells express *CA1* and it was anticipated that these cells would have a generally lower level of methylation in the *CA1* gene. The results shown above however demonstrate that, in general, the *CA1* gene is very highly methylated in this cell line and that most *Hpa*II sites cannot be digested. Only two exceptions to this pattern were found. One site (B1) lying about 1.5 kb downstream of exon 1a is completely demethylated (the neighbouring site 0.25 kb 3' to this may also be demethylated but this could not be determined).

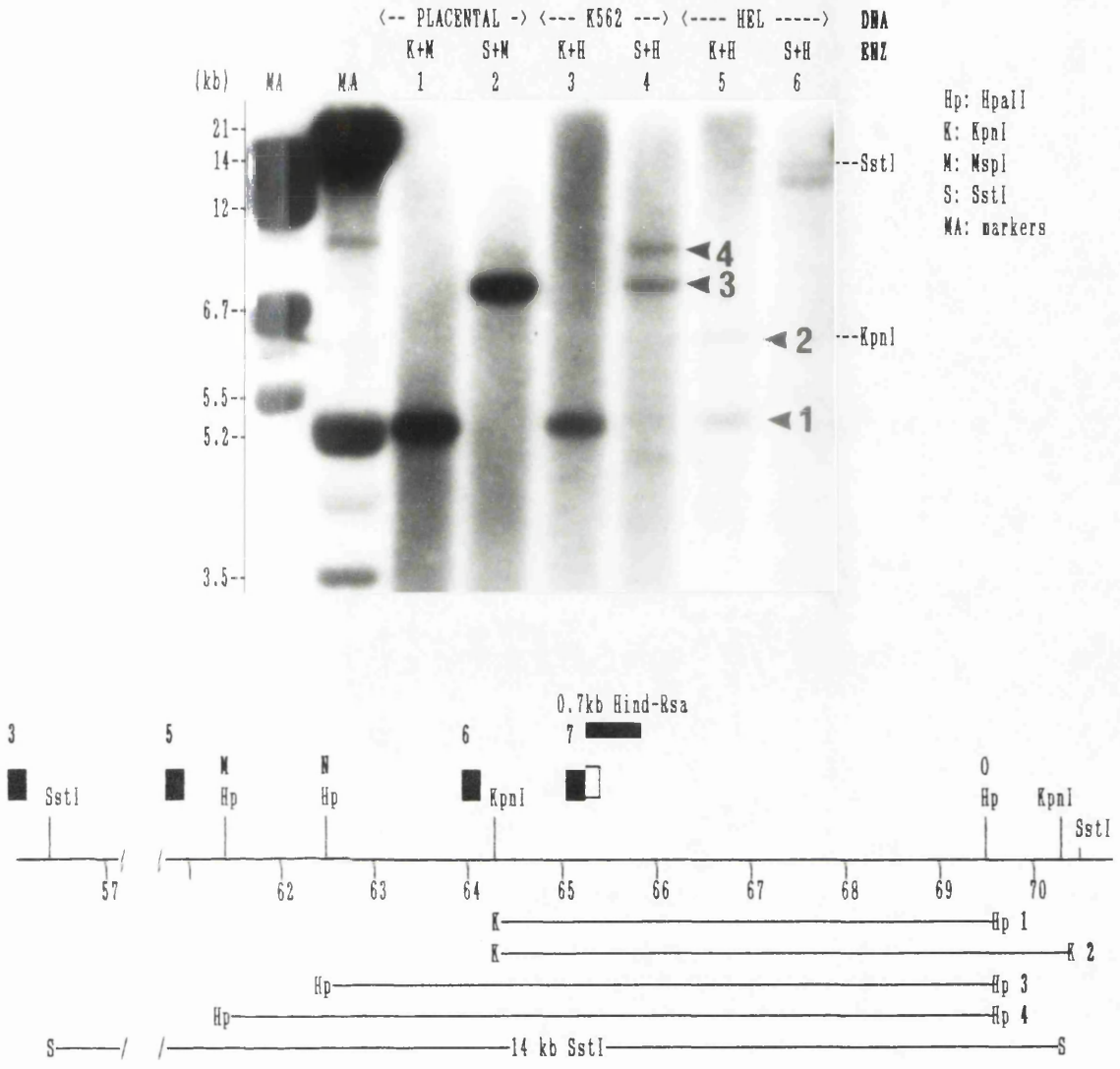


Fig. 6.4 Methylation at the 3'-end of the *CA1* gene. Southern blot of K562 and HEL cell DNA digested with *Kpn*I+*Hpa*II or *Sst*I+*Hpa*II and probed with a 700 bp *Hind*III-*Rsa*I fragment from the 3'-end of the gene. All other details as in Fig. 6.1. Lanes 1 and 2 are digests using *Kpn*I+*Msp*I and *Sst*I+*MspI* to show products of complete digestion at *Hpa*II sites. Scale (kb) is that shown in Fig. A1.

The second site which shows demethylation (showing partial digestion) in HEL cells is site O at the 3'-end of the gene 4kb from exon 7. Unfortunately the lack of *Hpa*II sites in critical areas round the promoter and 3'-flanking region of the gene did not allow more detailed examination of what may be demethylated *regions* rather than single sites. Indeed the general paucity of *Hpa*II sites in *CA1* may well have prevented disclosure of other demethylated areas.

K562 cells which are considered to have a foetal phenotype show partial DNA methylation at many of the *Hpa*II sites examined, with multiple bands produced on Southern blots. There was no obvious relationship between the methylation patterns seen in K562 and HEL cells i.e. sites which showed lower methylation levels in one cell line were not those with lower methylation levels in the other.

All the non-erythroid cell lines showed little or no methylation of the *Hpa*II sites close to the *CA1* probes used. The exception to this was the pre-T-cell CEM cell line which showed partial methylation at most sites, and appeared to show a higher level of methylation than K562 DNA.

Since most of these cell lines showed digestion of *Hpa*II sites in a way which implied complete demethylation over tens of kb, it was clearly of interest to know whether this was a general phenomenon or localised to the *CA1* gene. This is investigated in Section 6.3.

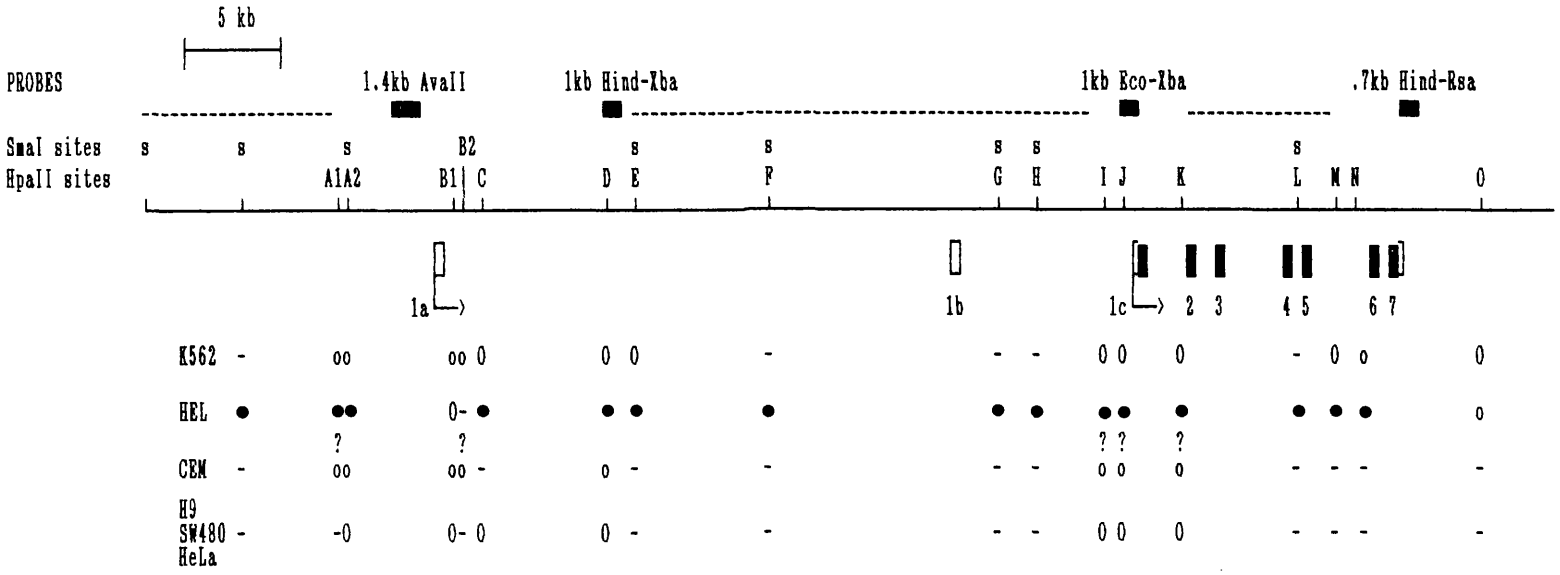


Fig. 6.5 Methylation states of *Hpa*II sites in the *CA1* gene. *Hpa*II sites designated letters A1 to O. Those sites with "s" above them are the central four bases of *Sma*I sites. Black bars above the map show positions of probes used. The broken lines indicate those regions where *Hpa*II sites have not been mapped and those *Hpa*II sites indicated in this region are known to exist only from *Sma*I site mapping. Question marks (?) indicate where methylation state could not be definitively assessed due to complex banding patterns.

6.3 Methylation states of the other *CA* genes and in non-cell line DNA.

Although no comprehensive survey of the methylation state of the other *CA* genes has been carried out in these cell lines, some information can be derived from the physical mapping experiments described in Chapter 5. The restriction enzymes used for PFGE analysis (of mammalian cells) contain CG dinucleotides within their recognition sequences and are susceptible to inhibition by methylation of cytosine. Because of this, PFGE analysis, of necessity often has to address the question of DNA methylation since changes in methylation state can complicate analysis of such data. Chapter 5 has described the location of rare cutter sites in the *CA* locus and Appendix Fig. A3 shows a slightly more comprehensive map of these sites. These sites have been designated numbers and are referred to as RC1 to RC18.

Difficulties in obtaining accurate sizes of fragments in PFGE analysis have already been mentioned in Chapter 5 and, from other reports, appears to be a common problem (Chen *et al.*, 1988; Michiels *et al.*, 1987). For this reason the fragment sizes given in this section should always be regarded as apparent and may differ considerably from the actual fragment size. For example the *Cla*I band detected with a 5'-*CA1* probe seen in lane 3 of Fig. 6.8 appears to be about 140 kb, but has been shown in other experiments, for example those shown in Chapter 5, to run at around 100 kb and is probably slightly smaller than this. The position of some of the sites shown in Fig. A3, which may only have been detected in one or two experiments, should therefore be regarded as provisional. This is especially the case for those sites lying downstream of *CA2*.

6.3.1 Variation in methylation patterns between cell lines

All of the cell lines examined in section 6.2 have been analysed, to a greater or lesser extent, with pulsed-field gel analysis. These experiments have been carried out principally with the enzymes *Cla*I, *Sal*I, *Sst*II and *Xho*I although the enzymes *Sma*I, *Mlu*I and *Nae*I have also been used. The results shown in Figs 6.6 to 6.11 illustrate particular features of this work, and the overall results are summarised diagrammatically in Fig. 6.12.

One of the first indications that methylation patterns at rare cutter sites varied dramatically from cell line to cell line came from a comparison of fragment sizes obtained by digestion of HEL DNA and K562 DNA. K562 DNA shows fragment sizes of 80 to 280 kb with single digests of *Cla*I, *Nae*I, *Sal*I, *Xho*I, *Sst*II (Chapter 5) and *Mlu*I (data not shown) when probed with a fragment from the 5'-end of *CA1*.

The results obtained with HEL DNA were very different (Fig. 6.6). No bands were produced from either the *Sal*I or *Mlu*I digests, while the *Cla*I fragment of 300 kb (lane 1) was much larger than that detected in K562 *Cla*I digests (80-100 kb). In K562 DNA, 5'- and 3'-end *CA1* probes detect different *Cla*I fragments whereas re-probing of the blot of HEL DNA shown in Fig. 6.8 with a 3'-*CA1* probe did not detect any new bands, indicating that the *Cla*I site in *CA1* (RC12) was not being digested.

*Cla*I digests were also carried out on CEM, H9, HeLa, K562 and HEL DNA and hybridised with probes for *CA1* and *CA3*. As can be seen in Fig. 6.7A, HEL DNA stands out as producing a higher molecular weight band than the other cell lines when probed with *CA1* (lane 2). By re-probing this filter, it was determined that the smaller bands, seen in all cell lines but HEL, were the products of digestion at RC10 and RC12.

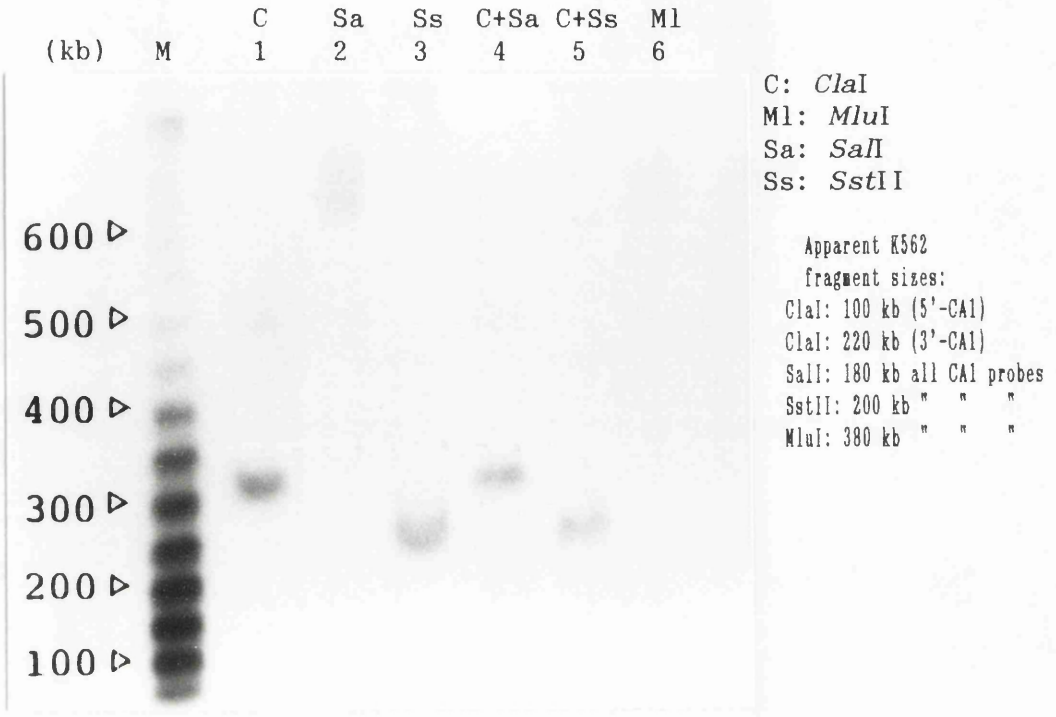


Fig. 6.6 PFGE analysis of HEL DNA. DNA was digested with *Cla*I, *Sal*I, *Sst*II, *Mlu*I, or double digests and hybridised with the 1 kb *Hind*III-*Xba*I fragment from *CA1*. Only high molecular weight smears are seen for either *Sal*I or *Mlu*I digests while the *Cla*I fragment is substantially larger than that produced in K562 digests. All the fragments seen here are also seen when probed with a 5'-end probe for *CA1* whereas different *Cla*I fragments are detected in K562 DNA using 5' and 3'-end *CA1* probes.

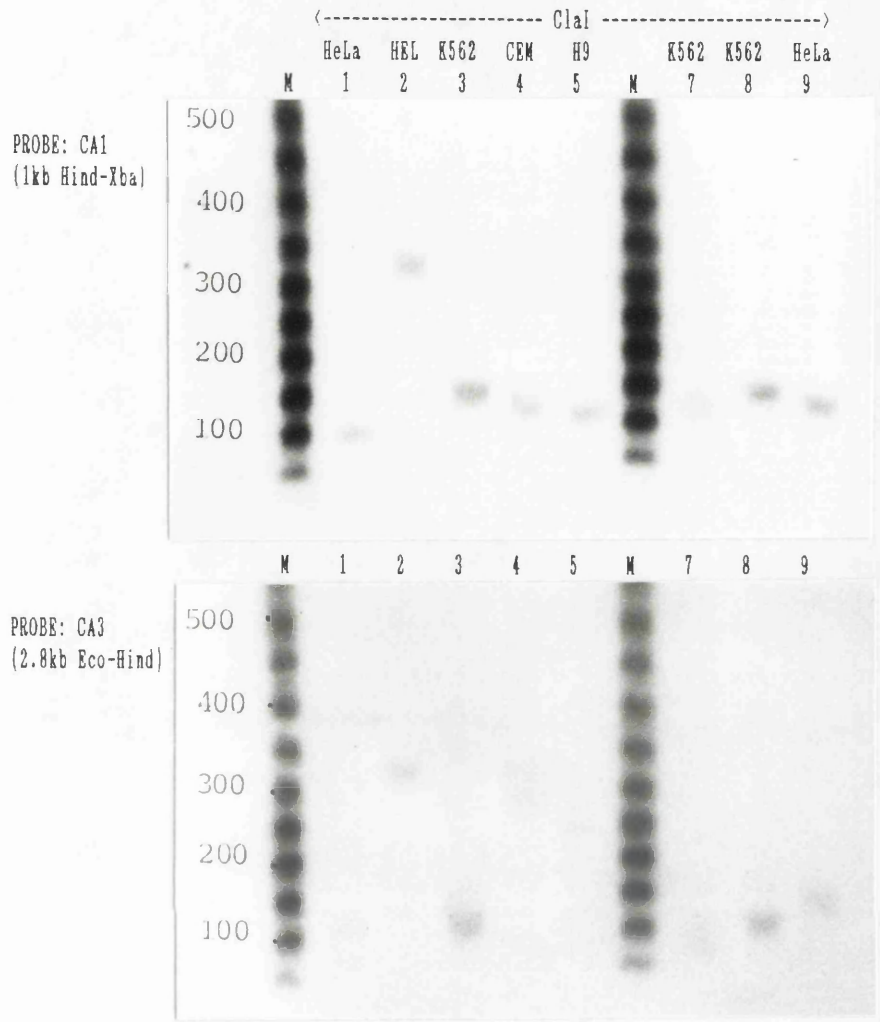


Fig. 6.7 *Cla*I digests of CEM, HEL, HeLa, H9 and K562 cell line DNA hybridised with *CA1* and *CA3* probes. A: blot was probed with *CA1* 1kb *Hind*III-*Xba*I probe lying 5'-to the *Cla*I site in the *CA1* gene (RC12). HEL DNA shows a clear difference to the other cell lines producing a band size of over 300 kb (lane 2) while all other cell lines give apparent fragment sizes of between 100 and 150 kb.

B: The same blot has been hybridised with a probe for *CA3* and shows that although the same HEL band (lane 2) is detected with this probe, all other bands are different, with CEM and H9 giving notably higher fragment sizes than K562 (280 and 250 kb faint bands in lanes 4 and 5). This gel also demonstrates the variation in mobility seen between different DNA samples of the same cell line for example lanes 6, 7 and 8 both contain K562 DNA while lanes 1 and 9 are HeLa samples. In each case samples with lower DNA concentration show greater mobility.

When this filter was re-hybridised with a probe for the 5'-end of *CA3* (2.8 *Eco*RI-*Hind*III) an interesting difference was noted from the results obtained using the *CA1* probe. Whereas all the cell lines (HEL excepted) appeared to give similar digestion patterns with the *CA1* probe, the *CA3* probe showed a more heterogenous fragment pattern, with CEM and H9 (lanes 4 and 5) in particular showing larger *Cla*I fragments than K562.

CEM and H9 were explored further to see how the methylation state of the rare-cutter sites round the *CA2* gene may vary. A *Xho*I site (RC9) lies 5 kb upstream of the *CA3* promoter and it was known that this site was digested in all cell lines except HEL (see Chapter 5, Fig.5.4c). Because of this, *Xho*I digests could provide a defined fragment end, allowing analysis of the region downstream of the *CA3* gene.

Fig. 6.8 and 6.9 show comparative digests of K562, H9 and CEM DNA probed with *CA1* and *CA3* probes. In Fig. 6.8 *Xho*I, *Xho*I+*Cla*I, and *Xho*I+*Sal*I digests of K562 and H9 DNA show identical bands when probed with the *CA1* 1 kb *Hind*III-*Xba*I probe, suggesting identical methylation patterns in this region (compare K562 lanes 1, 2, and 3 and their corresponding digests using H9 DNA in lanes 5, 6 and 7). Hybridisation with *CA2* (arrowed bands in Fig. 6.8B) produces larger fragment sizes from the *Xho*I digest (lanes 1 and 5) and the *Xho*I+*Cla*I digest of H9 DNA (lanes 2 and 6) indicating that the *Xho*I site downstream of *CA2* (RC3) and the *Cla*I site close to the *CA2* promoter (RC5) are poorly digested. The faint multiple banding pattern seen in the *Xho*I+*Cla*I digest (also seen with *CA3* probes see Fig. 6.9) is probably caused by partial *Cla*I digestion at RC4 and RC5.

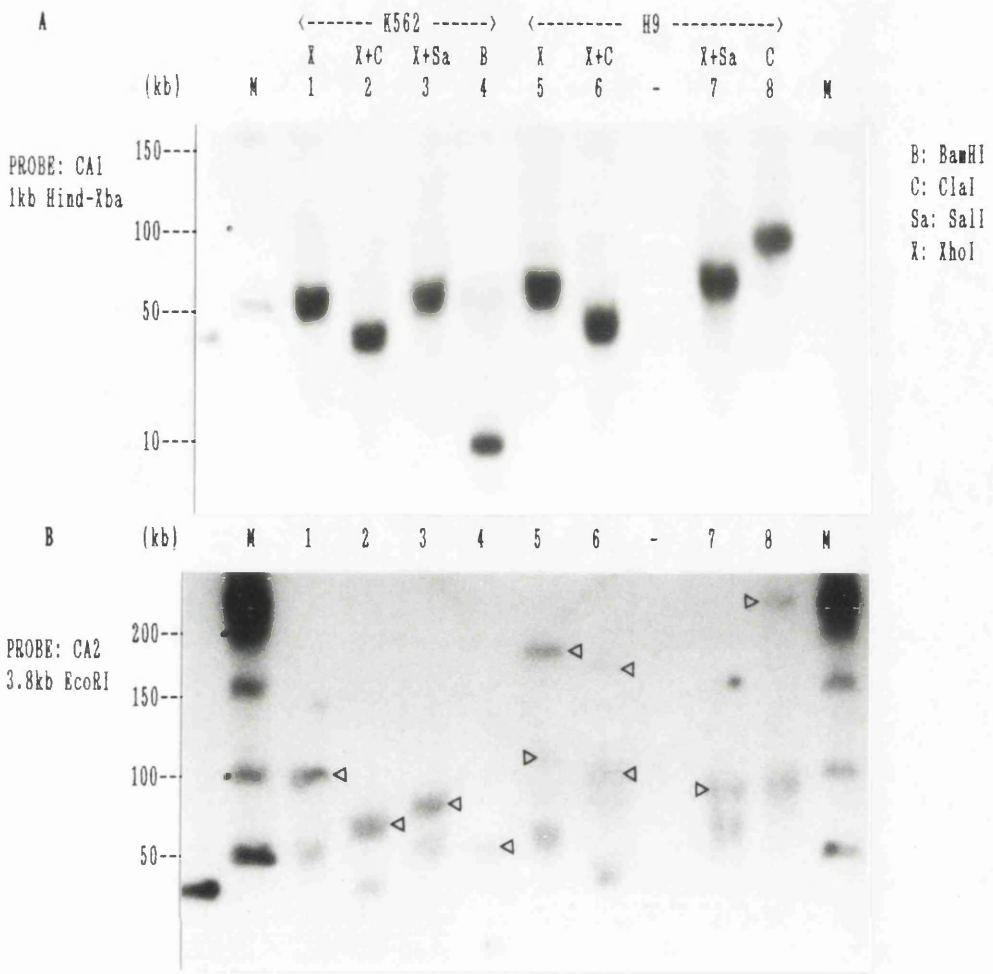


Fig. 6.8 Comparative digests of K562 and H9 DNA1 digested with *Xho*I, *Xho*I+*Cla*I, *Xho*I+*Sal*I. A: blot hybridised with *CA1* 1 kb *Hind*III-*Xba*I fragment showing the same band size produced in each cell line. Lanes 4 and 8 are not comparable (*Bam*HI and *Cla*I digests). B: The blot has been re-hybridised with *CA2* 3.8 kb *Eco*RI fragment without removing the first probe and novel bands are indicated by arrows. Comparison of the *Xho*I digests (lanes 1 and 5) shows that RC3 is scarcely digested in H9 DNA producing a faint band of 100 kb (lower arrow lane 5) while the major band detected is 175 kb in size. *Xho*I+*Cla*I digest of H9 DNA (lane 6) produces several faint bands most of which are larger than that seen in the K562 digests showing partial digestion at RC5.

That these alterations in banding pattern are due to changes in methylation state (rather than for example deletions or insertions in the different cell lines) is supported by the finding that the *Xho*I+*Sal*I digest gives the same band size (80 kb) in both cell lines and that a faint band can be detected in the H9/*Xho*I digest (lower arrow, lane 5) corresponding to a low level of digestion at RC3.

A similar experiment to that shown in Fig. 6.8 is shown in Fig. 6.9, with CEM, H9 and K562 DNA digested with *Xho*I and *Xho*I plus *Cla*I, *Sal*I or *Sst*II. When hybridised with a probe for the 5'-end of *CA1*, all three cell types give the same size *Xho*I fragment (90 kb). Since the mapping data showed there to be a *Cla*I site 5' of the *CA3* gene it would be expected that the *Xho*I+*Cla*I digest should be slightly smaller than the *Xho*I fragment. This can be seen as a slight increase in mobility for the K562 and H9 digests (lanes 6 and 10). This was not seen in lane 2 due to aberrations in the gel).

Hybridising the same blot with a *CA3* probe (Fig. 6.9B) demonstrates that CEM DNA, like H9 DNA is methylated at RC3, producing a larger *Xho*I fragment than K562 (about 175 kb in size). In CEM DNA both the *Cla*I and *Sst*II sites close to *CA2* (RC5 and RC6) are methylated (lanes 2 and 4) while the *Sal*I site is unmethylated, producing the same fragment size as the K562 digest (lanes 3 and 10). H9 DNA is unmethylated at the *Sst*II and *Sal*I sites (RC6 and RC8) close to *CA2* and as in Fig. 6.8 produces faint multiple bands from the *Xho*I+*Cla*I digest (lane 6).

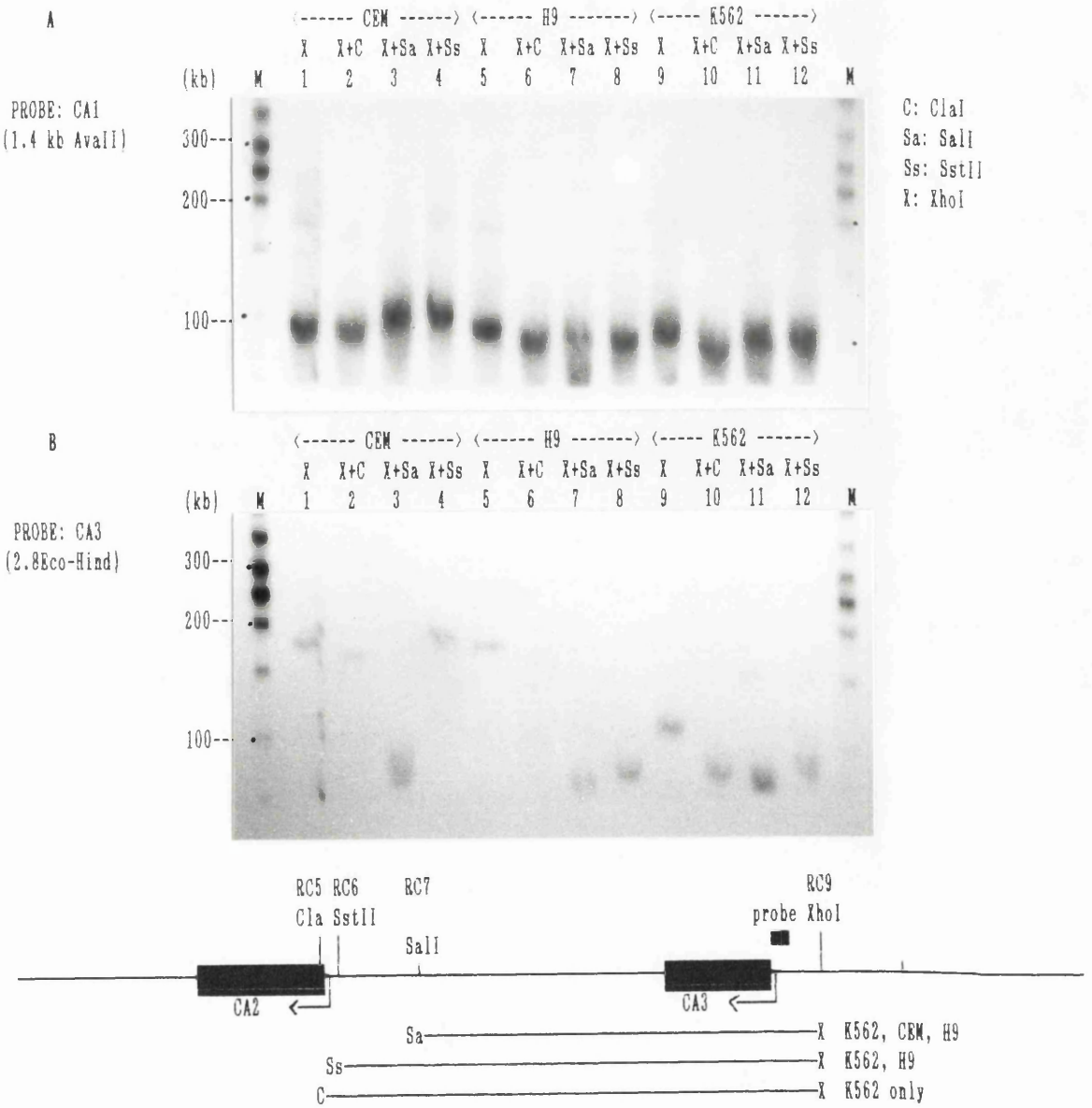


Fig. 6.9 Comparative digests of CEM, H9 and K562 DNA digested with *Xho*I, *Xho*I+*Cla*I, *Xho*I+*Sal*I and *Xho*I+*Sst*II. A: Hybridisation carried out using probe *CA1* 1.4 *Ava*II. All three cell lines demonstrate the same digestion pattern but note aberration in gel around lane 3 and 4. B: The same blot hybridised with *CA3* 2.8 *Eco*RI-*Hind*III. Several faint bands are detected in the *Xho*I+*Cla*I digests of H9 DNA (lane 6, not visible in this figure) and one of these is the same as that seen in the *Xho*I+*Cla*I digest of CEM DNA (lane 2) but no band was detected of the same size as that produced in K562 *Xho*I+*Cla*I digests.

6.3.2. Methylation of the *CA* genes in DNA from non transformed cells

Figs. 6.10-6.11 shows some of the experiments carried out using DNA from sperm and white blood cells (W.B.C.). Although detailed mapping was not carried out, it was clear that both sperm and white blood cell DNA demonstrate a generally high level of methylation.

In Fig. 6.10 DNA from sperm, K562 and HEL cells has been digested with *Cla*I, *Bsc*I (an exact isoschizomer of *Cla*I), *Sal*I and *Sst*II. As described for HEL digests, *no* fragments could be detected from *Sal*I digestion of the sperm DNA. This suggests that no two demethylated *Sal*I sites exist in this region within the resolving power of the PFGE system used (700-800 kb). Again, as with HEL digests, probes for *CA1* and *CA3* both detect the same 300 kb *Cla*I fragment (Fig. 6.10A and 6.10C lanes 1 and 2), indicating that the *Cla*I site within *CA1* (RC12) and close to *CA3* (RC10) are resistant to digestion. That these 300 kb *Cla*I fragments from both sperm and HEL terminate at RC5 in the promoter of *CA2* is demonstrated in Fig. 6.10B. In this panel the blot has been hybridised with the 3.8 kb *Eco*RI fragment containing the *Cla*I site in the promoter region of *CA2* (RC5) and two bands are seen in both sperm and HEL *Cla*I digests (lanes 1,2 and 9,10). All three cell types generate *Sst*II fragments of approximately 200 kb which terminate in the promoter of *CA2* (RC6).

The only other non-cell line DNA examined was that of peripheral white blood cells. It is particularly pertinent to compare these cells to HEL and K562 since the various white blood cell types and erythroid cells are derived from a common haemopoietic progenitor cell pool, and indeed HEL and K562 cells share some characteristics of these non-erythroid cell types (Gootenberg *et al.*, 1981; Villeval *et al.*, 1986).

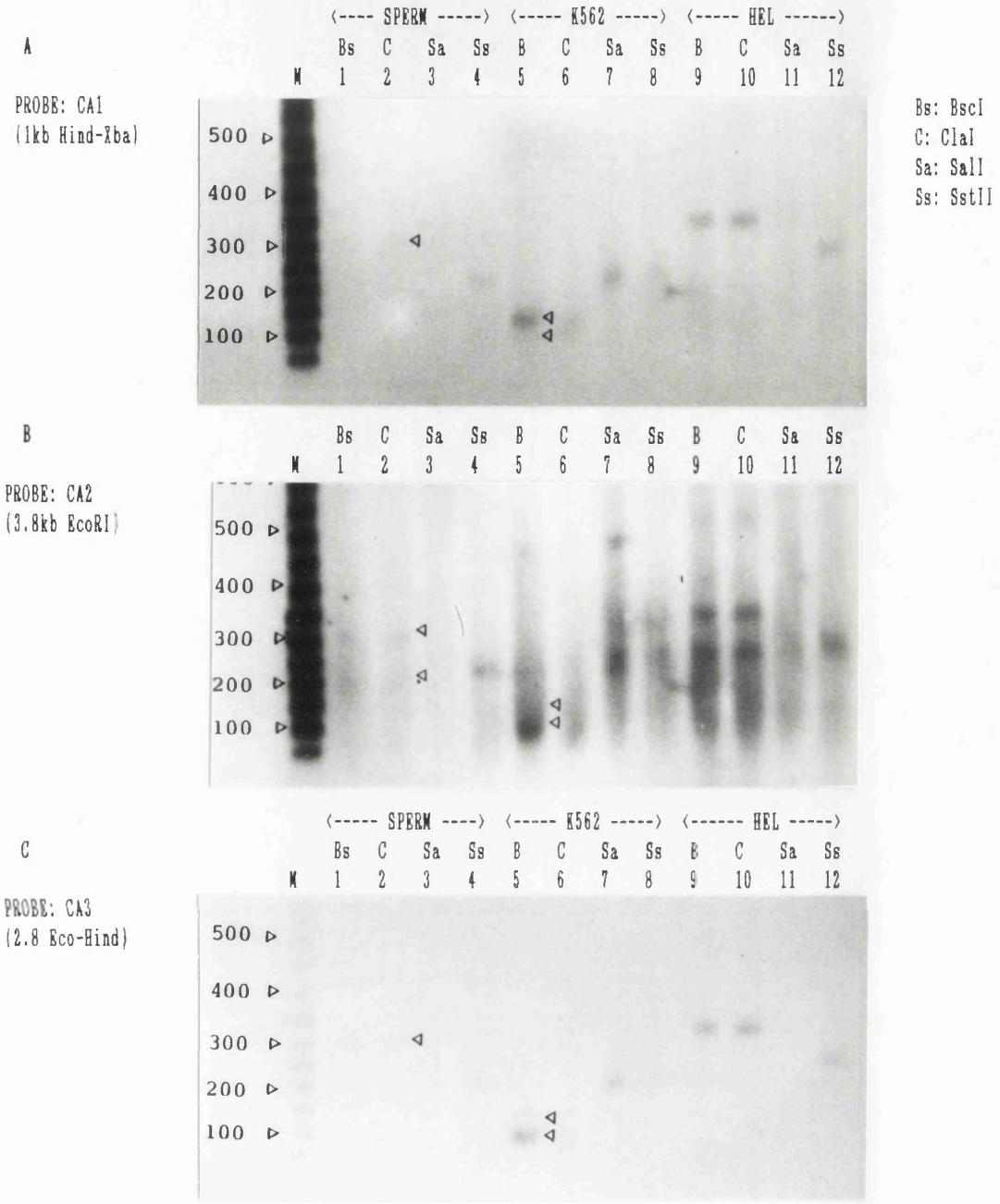


Fig. 6.10 PFGE analysis of sperm K562 and HEL DNA. DNA was digested with *Cla*I (or its isoschizomer *Bsc*I), *Sal*I and *Sst*II and hybridised with probes from A, *CA1* (1.0 kb *Hind*III-*Xba*I); B, *CA2* (3.8 kb *Eco*RI) or C, *CA3* (2.8 kb *Eco*RI-*Hind*III). Arrows show the position of the faint *Cla*I bands produced in the sperm lanes, while the two arrows in the K562 digests show the position of the two bands identified by the different probes. No *Sal*I band is seen in sperm or HEL DNA. K562 DNA digested with *Cla*I shows only a single band with the *CA2* probe due to the similarity in size of the fragments produced.

Results from PFGE analysis, such as that shown in Fig. 6.11 showed that white blood cell DNA displayed, if anything, a higher level of methylation than sperm. No fragments could be detected in *Sal*I (lane 4) or *Cla*I (lanes 3 and 8) digests using a *CA1* probe. Multiple bands were produced from *Xho*I digests but it is not known which sites might have been digested to produce these bands.

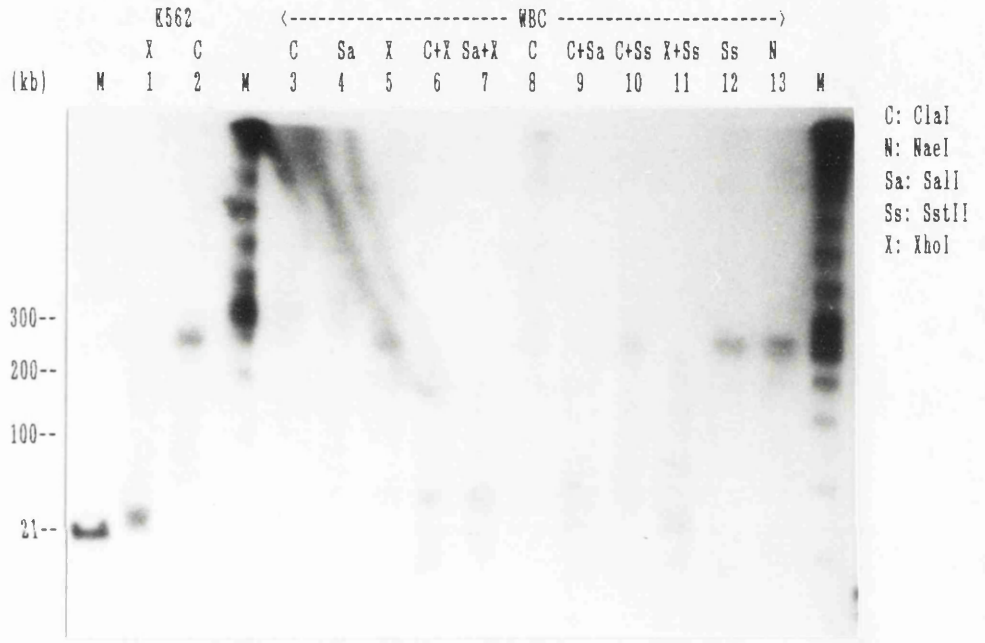


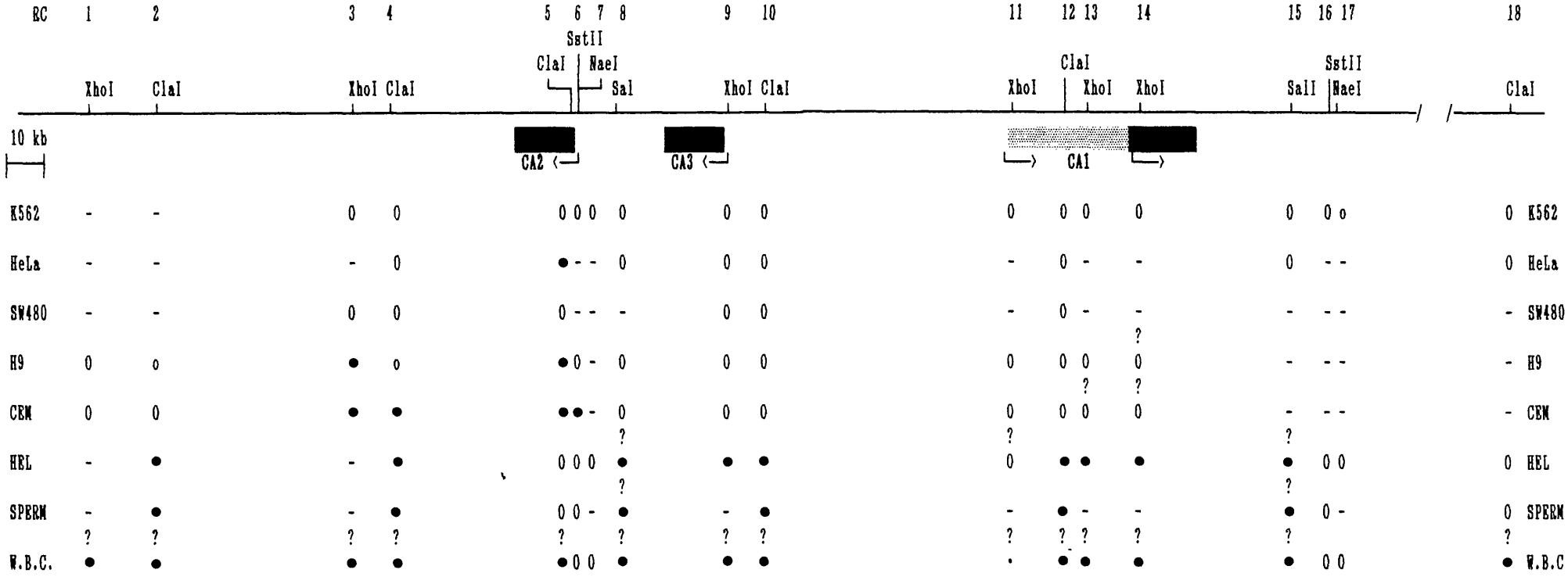
Fig. 6.11 PFGE analysis of white blood cell DNA. Peripheral blood white cell DNA was digested with *Xho*I, *Cla*I, *Sal*I, *Nae*I and various double digests and hybridised with a *CA1* probe (1 kb *Hind*III-*Xba*I). Lanes 1 and 2 show K562 DNA digested with *Cla*I and *Xho*I. All fragments detected are larger than those found in K562 digests with the exception of the *Nae*I and *Sst*II digests which appear to be the same size. This and other data (not shown) suggests that these two fragments are more or less co-linear and it is proposed that one end of these fragments lie in or near the *CA2* promoter, while the other lies less than 40 kb downstream of *CA1*. No bands can be detected in either *Cla*I (lanes 3 and 8) or *Sal*I (lane 4) digests.

6.4. Summary of results of methylation analysis.

Results from the pulsed-field analysis (summarised in Fig. 6.12) appear at first sight to suggest that the methylation patterns described in Section 5.2 for *CA1* extend beyond this gene and into the area around *CA2* and *CA3*, with K562 and HeLa appearing to be completely demethylated but with HEL cells selectively demethylated at particular sites. The CEM and H9 digests however show that these cell lines may have low levels of methylation in one region while other regions are more highly methylated. Despite this variation between the cell lines it would appear that most of the cell lines have a much lower level of methylation around this gene cluster than either white blood cell or sperm DNA.

Unfortunately little is known about the expression profiles of the various cell lines examined (apart from what has already been mentioned regarding *CA1* in HeLa, HEL and K562 cells). However, since the distribution of CAI and CAIII is restricted to a relatively small number of cell types it seems unlikely that either CAI or CAIII would be found in any of these cells. It is more likely that the *CA2* gene is active in some of these cells since this isozyme is far more widely distributed and indeed the CpG-rich promoter of this gene appears more typical of a housekeeping gene. However it is known that CAII is not found in HeLa cells (Shapiro *et al.*, 1987).

Both sperm and W.B.C. appear highly methylated and show demethylation at a few specific sites, resembling HEL DNA in this respect. These sites are probably only those which lie within CpG-rich demethylated regions. The *CA2* promoter is one such region and examination of its sequence (Venta, *et al.*, 1985b) shows that several rare cutter sites, including *Sst*II and *NaeI* and multiple *Hpa*II sites are located here. No other *Sst*II or *Nae*I sites have ever been detected in this gene cluster apart from those in the *CA2* promoter so it is almost certain that the two fragments seen in the W.B.C. digests are co-linear and probably mark another CpG-rich



0 : unmethylated
o : partial methylation
● : complete methylation
- : not tested

Fig. 6.12 Methylation states of the rare cutter sites at the *CA* gene cluster on chromososome 8. Sites marked with a question mark could not be definitively designated due either to sizing difficulties, or because highly methylated DNAs (HEL, sperm, W.B.C.) did not produce restriction fragments resolvable on the gels used.

region 3'-of the *CA1* gene. Similarly sized *Sst*II and *Nae*I fragments have also been found in K562 and HEL digests and mapping experements using K562 DNA places these sites about 50 kb from the 3'-end of *CA1*.

CHAPTER 7: DISCUSSION

This thesis has described several findings. Firstly the isolation and characterisation of reticulocyte cDNA clones encoding human carbonic anhydrase I in which a heterogeneous 5'-leader was found. This was followed by the cloning and characterisation of the human *CA1* gene during which it was discovered that the 5'-ends of the cDNA clones were derived from an erythroid specific promoter located tens of kilobases upstream of the protein-coding sequence. The physical relationship of this gene to two other members of the carbonic anhydrase gene family, *CA2* and *CA3*, was also determined using PFGE. Finally the methylation state of CpG dinucleotides in and around the *CA1* gene was determined in a number of cell types including those cell lines which had been used for *CA1* expression studies by other members of the research group.

7.1 Structure and transcription of the *CA1* gene

One of the most striking findings made during the course of this work was the difference between *CA1* and the other carbonic anhydrase genes characterised to date (*CA2* and *CA3* Venta *et al.*, 1985; Lloyd *et al.*, 1987), namely the presence of a large intron (36.5 kb in size) which separates the erythroid-specific promoter of the gene from the coding sequence (Chapter 4, Fig. 4.18). A number of other genes have also been shown to have introns in the 5'-non-coding region (see Leff *et al.*, for review). Although in most cases such introns are relatively small the *c-abl* gene possesses an intron of some 200 kb in its leader sequence (Bernards *et al.*, 1987) and so clearly *CA1* cannot be considered unique in this respect.

Within this large intron, next to the coding sequence is found a second promoter. This more conventionally positioned promoter is active in colon epithelia, and is probably also responsible for *CA1* transcription in the other non erythroid

tissues where CAI is found.

Many genes are required to be active in more than one, although not all cell types. This limited pattern of gene activity has been shown to be adequately coped with by use of a single promoter active in different cell types. In this instance the promoter may for example contain binding sites for a number of different tissue-specific transcription factors, different sets of which would be important for expression in different cell types (Dynan, 1989). The occurrence of two or more promoters in a single gene is not however unique (see Schibler and Sierra 1987 for review) although the large distance separating the two *CA1* promoters is unusual. The porphobilinogen deaminase gene (PBGD) for example possesses two promoters, one of which has a housekeeping function, while the other, like *CA1*, is active in erythroid cells (Chretien *et al.*, 1988). The two promoters in this instance are relatively close together separated by only 3 kb with the erythroid promoter (unlike *CA1*) adjacent to the coding sequence.

It should be noted that although the exact distance of the erythroid promoter from the colon promoter/coding sequence in the mouse gene for CAI (*car-1*) has not been determined it is known to be at least 10 kb (Fraser *et al.*, 1989). In general it has been observed that intron size is quite flexible, and since the mouse and human lineages have been separated for some 100 million years the maintenance of a large intron may reflect some functional significance. For example the erythroid promoter may lie within another chromatin domain from that of the colon promoter. Investigation of this possibility, or any other aspect of non-erythroid expression would be facilitated by the availability of a non-eythroid CAI expressing cell line.

It has already been mentioned in Chapter 1 that individuals deficient in erythrocyte CAI appear to have no overt clinical symptoms (Kendall and Tashian, 1977) and it would be of interest to see whether such individuals lack CAI in other sites of expression such as colon. The distance

of the erythroid promoter from the body of the gene would clearly allow for a range of mutations in and around exon 1a which would not affect the colon promoter. Lack of *CA1* expression in the red blood cell where CAII would be capable of substituting for CAI activity would probably be of less consequence than loss of activity in non-CA2 containing sites.

The evolutionary origin of the *CA1* erythroid promoter is a topic for speculation. It may of course be possible that this promoter results from a duplication or capture of a pre-existing erythroid-specific promoter from another gene rather than *de-novo* creation of a new promoter. If this were the case traces of a such a gene may be found within the large intron such as open reading frames or sequences with a higher than average G/C-content than normally found in introns. It might even be possible that another active gene could be located in this region. Experiments have not been carried out to see whether any other transcripts can be detected using probes from within this intron so this latter possibility has not been tested. Northern analysis using oligonucleotide probes specific for exon 1a detect only *CA1* message (H. Brady personal communication).

This large intron also contains a small 54 bp exon (exon 1b) 10 kb upstream of the coding sequence (Fig. 4.18) which only appears in a minority of the cDNA species and could not be detected in northern blot analysis using exon 1b-specific probes(Chapter 3, Fig. 3.8 to Fig. 3.10 and J. Sowden personal communication). The gene which most resembles *CA1* in the possession of such an "optional" exon in the 5'-leader sequence is the 3-hydroxy-3-methylglutaryl coenzyme A synthase gene which gives rise to a small 5'-untranslated exon in approximately 50% of its transcripts. In this instance it has been suggested that the exon has a functional role on the basis that it is conserved between hamsters and humans (Gill *et al.*, 1987) but no such occasional exons have been reported in the mouse gene encoding CAI (*car-1*) (Fraser *et al.*, 1989) The production of occasional exons

may simply be a reflection of the fact that, within an intron of this size, more than one sequence capable of acting (albeit inefficiently) as splice junctions would be found. The non-consensus splice site at the 3'-end of exon 1b may be responsible for the rarity of this exon since the adjacent intron starts with a GC, rather than the almost universal GT (Fig. 4.19). (In a survey of over 400 intron/exon boundaries, a GT at the 5' end of the intron was found in over 99% of the sites examined; Padgett *et al.*, 1986). Although the 5'-GT is almost invariant, suggesting that this sequence is required for efficient splicing, it may well be the case that other features of a splice junction may reduce any stringent requirement for this sequence. Thus the cox 5b gene from *S. cerevisiae*, which contains a 5'-GC in its most 5' intron, is unaffected in terms of its splicing by being mutagenised to the consensus GT sequence (Hodge and Cumsky, 1989), while chicken and duck α-globin genes which are required to be efficiently expressed at high levels also possess 5'-GC-containing introns (Padgett *et al.*, 1986).

7.2 Evolution of the *CA1, CA2, CA3* gene cluster.

The PFGE mapping work described in Chapter 5 has shown the *CA* genes lying on the long arm of chromosome 8 to be in the order *CA2, CA3, CA1*. *CA2* and *CA3* are relatively close together, being separated by about 20 kb and are transcribed in the same direction, away from *CA1*. These two genes are separated from *CA1* by some 80 kb and *CA1* is transcribed in the opposite direction to the other two genes i.e. away from *CA2* and *CA3* (Fig. 5.6). This is the first time this gene cluster has been mapped in any species.

The three genes mapped in this study have existed as distinct forms for over 300 million years, being formed some time between the divergence of the elasmobranchs (450 million years ago (mya)) and the divergence of the amniotes (300 mya). Comparison of the sequences of *CA1*, *CA2* and *CA3* at

both the protein and nucleic acid levels suggests that *CA2* and *CA3* are slightly more closely related to each other than either is to *CA1* (Lloyd et al. 1986., Hewett-Emmett and Tashian, 1990). This suggests that the duplication event which gave rise to *CA2* and *3* (estimated to be 300-320 mya; Fraser and Curtis 1986) post-dated the duplication giving *CA1* and the *CA2/3* ancestral gene although Hewett-Emmett and Tashian (1990) suggests that till further evidence is available it would be safer to regard these genes as being formed by a trifurcation event. Since genes tend to become separated through evolution, the fact that the *CA2* and *CA3* lie relatively close together physically compared with the distance between *CA3* and *CA1* would lend support to the later duplication event to form *CA2* and *CA3*, though clearly no firm conclusion can be drawn from this line of evidence. Against this interpretation, it should be noted that models of gene duplication based on unequal crossing over between sister chromatids at meiosis assume that once an initial duplication event has taken place, secondary duplications should be facilitated by slippage and mis-pairing between the two adjacent genes (Maeda and Smithies, 1986). This process would be less likely to take place once the *CA1* gene was inverted and so suggests that either the inversion of *CA1* relative to *CA2* and *CA3* took place some time after the second duplication event producing *CA2* and *CA3*, or that *CA1* was produced (with an inversion) after *CA2* and *CA3*. If duplication and inversion events were separate, species may be found with all three genes in the same orientation. Although as yet this gene cluster has not been definitively mapped using PFGE in any other species some data ~~is~~ are available for the mouse. In this organism interspecific backcrosses between *Mus musculus* and *Mus spretus* showed recombination between *car-3* and the other two genes in this cluster (located on chromosome 3) indicating that *car-3* is unlikely to be located between *car-1* and *car-2* as is found in humans (Beechey *et al.*, 1990). This indication that the genes may be arranged differently in the mouse is supported by

preliminary PFGE data placing *car-2* downstream of *car-1* (Peter Curtis, Wistar Institute, Philadelphia, personal communication).

A very similar arrangement of genes is seen in the cluster of human protease inhibitor genes α_1-antitrypsin (PI), α_1-antichymotrypsin (AACT) and the PI-like gene PIL found on chromosome 14. In this cluster the two closely related genes PI and PIL are found close together and are transcribed in the same direction away from the more distant (220 kb) AACT gene which is, like *CA1*, transcribed in the opposite direction (Sefton *et al.*, 1990). Inversion of genes within multigene families does not appear to be a particularly rare event, and have been found in human HOX gene clusters (Acampora *et al.*, 1989), Drosophila α-amylase (Boer and Hickey, 1986), human CD1 human thymocyte differentiation antigen (Yu and Milstein, 1989), MHC class III (Sargent *et al.*, 1989) and the mouse major urinary protein genes (Ghazal *et al.*, 1985).

Whatever the exact arrangement of the carbonic anhydrase genes may be in the mouse, it is clear that they have remained closely linked for some considerable time (*CA1* and *CA2* are also known to be tightly linked in the guinea-pig and pigtail macaque; Carter, 1972; DeSimone *et al.*, 1973a). It could be hypothesised that the maintenance of such a close relationship is indicative of a functional relationship between these genes and in this regard it is interesting to note that a partial deficiency of CAI in the pigtail macaque is always associated with a reduction in the level of CA2 produced from the *cis-CA2* gene (DeSimone *et al.*, 1973a and b), while Carter *et al.* (1984) have reported an increase in CAIII levels in individuals showing a reduction in the amounts of red cell CAII.

The different patterns of expression seen in these three genes is reflected in the different nature of the promoters of these genes. It has already been noted in Chapter 6 that *CA1* is particularly deficient in sites for the restriction enzyme *Hpa*II, and examination of the erythroid promoter

sequence reveals that only 3 CpG dinucleotides are found between -300 to +100. The colon promoter is similarly lacking in *Hpa*II sites and contains only one CpG dinucleotide in the region from -175 to +80. By contrast the *CA2* promoter appears to resemble the promoter of an idealised housekeeping gene being G+C rich and containing 50 CpGs between -300 to +100. This CpG richness is also reflected in the large number of rare cutter restriction sites in the promoter of *CA2* - sites which are demethylated even in the highly methylated DNA from sperm, white blood cells and HEL cells. *CA3* lies somewhere between these two extremes with 30 CpGs in the -300 to +100 region. Despite the fact that expression of this gene is restricted to only a few tissues, several of the *Hpa*II sites in the promoter of *CA3* are demethylated in all tissues examined (Edwards *et al.*, 1988), a characteristic typical of HTF-island containing housekeeping genes. It has been suggested (Edwards, 1990) that this property of the *CA3* promoter may be a reflection of this gene's evolution away from a more ubiquitously active, *CA2*-like gene.

The tentative identification of a demethylated CpG-rich region downstream of *CA1* (Section 6.3, Fig. 6.11) prompts further investigation and the first step in examining this region would be to carry out a more extensive search using rare cutter restriction enzymes. If the limited mapping data is correct this region should be about 50 kb beyond the 3'-end of the *CA1* gene. This is a comparatively short intergenic distance for the mammalian genome and, if a gene does reside in this area, the possibility that another member of the CA family is located here should be considered.

7.3 Gene expression

Analysis of the sequence obtained at both the 5'- and 3'-ends of the gene showed that these regions contained a number of potential transcription factor binding sites.

Of great interest is the finding that the *CA1* gene has features in common with other genes expressed in erythroid tissues. The most obvious of these is the presence of several binding sites for the transcription factor GATA-1 (Section 4.6.2, Fig. 4.21). As described in the introduction, binding sites for this factor are found to be associated not only with the globin genes, but have also been found upstream of the erythroid-specific promoter of the porphobilinogen deaminase gene (Mignotte *et al.*, 1989).

Potential binding sites for this factor (consensus sequence $Pu^{T}/_{A}GAT^{A}/_{T}Pu^{T}/_{G}$, Wall *et al.*, 1988 or a close variant thereof, Plumb *et al.*, 1989) are found at three places in the erythroid promoter region at positions -190, -149 and -290, at the 3'-end of the gene 223, 581 and 833 bases downstream of the coding sequence and at the 5' end of the coding sequence, one overlapping the Met codon, and one 158 bp upstream of the coding sequence (position -77 of the colon promoter start site). Several of these sites have been analysed by the methods of DNAseI footprinting and gel mobility shift assays and shown to bind GATA-1 (Brady *et al.*, 1989 and J. Sowden, personal communication) and are almost certainly of importance for expression of *CA1* in erythroid tissues. A similar arrangement of GATA-1 sites at the 5'- and 3'-ends of the gene has also been found in the human ß-globin gene (Wall *et al.*, 1988).

The erythroid promoter of the gene also contains a binding site for the CACCC box factor about 20 bp 5'-of the GATA-1 site at -190. Close proximity of the CACCC box and the GATA-1 site has been found previously, and it has been suggested that the factors binding these motifs act co-operatively (Mantovani *et al.*, 1988; Mignotte *et al.*, 1989). Interestingly, a similar clustering of potential GATA-1 and CACCC binding sites are also seen in the sequence of the upstream 1.4 kb *Ava*II fragment (Fig. 4.14). DNaseI hypersensitive sites have been identified at both -200 and -1500 bp upstream of the erythroid promoter (J. Sowden personal communication and PhD thesis, 1991) and it would

seem likely that these sites are of critical importance for *in vivo* gene expression. This region also contains a binding site termed b3/c2 (A/T N A/G TAATNNN A/G) which is found in the erythroid promoter of the human porphobilinogen deaminase gene and the 3'-enhancer of the human ß-globin gene (Wall *et al.*, 1988; Mignotte *et al.*, 1989).

Following the work described in this thesis, attempts have been made to analyse the properties of the *CA1* erythroid promoter by means of reporter gene constructs. Chimeric genes were created by linking CA1 promoter fragments extending 5 kb upstream of the erythroid promoter to a reporter gene and these were stably transfected into erythroid and non-erythroid cells (J, Sowden personal communication and PhD thesis, 1991).

These constructs have been introduced into the three cell lines described in Chapter 6; 1) The *CA1*-expressing mouse MEL cell line; 2) The K562 cell line which is non-*CA1*-expressing, but can be trans-activated by fusion with MEL cells to produce *CA1*; 3) The non-erythroid, non-transactivatable HeLa cell line. In brief, these constructs have produced low levels of expression (<5% of endogenous *CA1*) in both erythroid cell lines despite the fact that the foetal/embryonic K562 cells do not normally produce *CA1*, while no expression is seen in HeLa cells. This result implies that while the promoter appears to confer erythroid tissue specificity, sequences other than those upstream of the erythroid start site are responsible for mediating the developmental regulation of this gene. The location of possible regulatory regions can only be speculated on at present. Clearly attention might be paid to the 3'-end of the gene which appears to show demethylation (Section 6.2.2), while DNAseI hypersensitive sites discovered by J. Sowden (personal communication and PhD thesis 1991) may indicate other regions of interest. As well as HSS sites within the vicinity of the promoter at -200 bp and -1.5 kb, other sites have been found within the large intron. These sites lie approximately 6.5 kb and 10.5 kb downstream of

exon 1a and are found in HEL and K562 but are absent in HeLa cells.

Work has been carried out defining HSS in the mouse *car-1* gene. In this gene two sites have been found in similar positions to the two sites upstream of the erythroid promoter in the human gene (Peter Curtis, Wistar Institute, personal communication). The same study has also identified two sites close to the colon promoter which may bind a factor(s) associated with the differentiation process (Ma *et al.*, 1991 and P. Curtis personal communication). This region has not yet been studied for hypersensitive sites in the human gene.

If more work using expression constructs is to be carried out in order to define those sequences involved in the temporal and developmental regulation of this gene certain considerations should be borne in mind. It is now known that important contributions are made to gene regulation by remote DNA sequences such as locus controlling regions and nuclear scaffold attachment regions. Such considerations and the size of *CA1* make it probable that large expression constructs will be required to take this aspect of the work further. Since even cosmid clones could not accommodate the entire gene, alternative strategies such as the use of phage P1 vectors (Sternberg, 1990) which are capable of accomodating 100 kb insertions may be needed.

The difficulties experienced in trying to isolate cosmid clones containing this gene should also be noted if this work is to be undertaken. Other workers have reported difficulties in isolating particular regions of eukaryotic DNA in bacterial cloning vectors (Wyman *et al.*, 1985, Coulson *et al.*, 1986) and it would appear that some sequences will remain resistant to this type of cloning. If parts of *CA1* are unclonable an alternative approach to expression work may be needed. One such approach may be to use yeast artificial chromosomes (YACs; see Schlessinger, 1990 for review). YACs are capable of accommodating hundreds or even thousands of kb of sequence and are thought to be less prone

to the problems of sequence rejection than bacteria.

Acquisition of yeast artificial chromosome clones would also potentially be of use in other lines of investigation. For example allowing investigation of any cis-acting co-regulation effects as suggested in Section 7.2. Characterisation of *CA1*-containing YACs would also have the by-product of allowing isolation of the CpG-rich, demethylated region hypothesised to exist downstream of *CA1* (Section 6.4 and Fig. 6.11).

Such expression work would obviously be complemented by more detailed analysis of *in-vivo* chromatin conformation, hypersensitive sites and a clearer picture of methylation states of the *CA1* gene.

7.4 Methylation Patterns of *CA1*

The finding of regions of demethylated CpGs in a generally highly methylated background is a fairly common observation, and so the results seen in the *Hpa*II digestion of HEL cell DNA were not unexpected (Fig. 6.5). Such demethylated sites may lie in or close to those regions important for control of gene expression and so the fact that demethylated sites are found at the 3'- as well as the 5'-end of the gene may reflect a need for the 3'-end of the *CA1* gene to be incorporated into expression constructs in order to reproduce the expression patterns seen *in-vivo*.

Although only two demethylated *Hpa*II sites were found in the *CA1* gene it is quite likely that other CpGs which do not lie within *Hpa*II recognition sequences would also be demethylated. Unfortunately the general paucity of *Hpa*II sites in this gene has made a more precise analysis impossible by this method. This work could in future be extended using other methylation sensitive restriction enzymes such as *Hha*I.

The lower level of methylation seen in the *CA1* gene in the non-*CA1*-expressing cells was somewhat less expected and is at variance with results reported in studies of other

genes. The fact that K562 cells show partial methylation particularly at the sites closest to the erythroid promoter (Section 6.2.1) is especially interesting since it has been shown that the *CA1* can be trans-activated in this cell line following fusion with the adult phenotype mouse erythroleukaemia cell line, MEL (Brady *et al.*, 1989). By contrast, the completely demethylated *CA1* gene of the HeLa cell line cannot be activated in this manner, and it is tempting therefore to suggest that in the case of *CA1* the relationship between methylation and gene expression is the reverse of that found for most other genes.

This interpretation however should be treated with caution since DNA methylation states of transformed cells may well have a different significance from that of normal, untransformed tissue. A number of studies have demonstrated that transformed cells and tumour cell lines exhibit an overall reduction of around 10% in their 5-methylcytosine content (Hoffman, 1984; Feinberg *et al.*, 1988; Gama-Sosa *et al.*, 1983), while genes fully methylated in normal tissues may be substantially hypomethylated in malignant cells (Feinberg and Vogelstein, 1983a; Goelz *et al.*, 1985). Such observations have lead some workers to propose that the lower levels of 5-methylcytosine found in transformed cells may be linked to the demethylation and subsequent aberrant expression of oncogenes, implicating demethylation as an early event in neoplastic transformation (see for example Feinberg and Ranier 1990). In apparent contradiction, other workers point to the findings that many genes are apparently inactivated during transformation (Holliday, 1990) and that 5-azacytidine treatment of cell lines can reactivate inactive genes in transformed cells (Jones, 1985), suggesting that "established cell lines frequently inactivate genes by *de-novo* methylation" (Holliday, 1990).

These apparently conflicting views of the relationship of the transformed state to DNA methylation may be reconciled if it is assumed that a qualitatively different type of demethylation is taking place in these cells and that this

is what has been detected in and around the *CA* genes (Section 6.3 and Fig. 6.13). This aberrant type of demethylation may be characterised by being a) more complete than that found in normal tissues, where demethylation along the length of the gene would usually be *reduced* rather than absent and b) extensive, covering tens or even hundreds of kilobases. Within such regions of demethylation the relationship between gene activity and methylation may be completely different from that found in normal somatic tissues and it seems likely that the maintenance of a partially or wholly methylated gene would be a requirement for activation. Such a notion would be supported by the observation that genes not normally active in colonic mucosa such as γ-crystalin and γ-globin are hypomethylated in colorectal carcinomas and premalignant adenomas (Feinberg and Vogelstein, 1983; Goelz *et al.*, 1985). Most studies of methylation states in particular genes examine relatively small regions of DNA and would not therefore disclose any long range demethylation such as that seen here.

The relationship between DNA methylation and chromatin structure is particularly relevant here. The association between hypomethylation and chromatin decondensation or nuclease sensitivity has been well documented (Schmidt *et al.*, 1984; Buschhausen *et al.*, 1987; Cedar 1986, also see discussion in Selker 1990). If it is a general rule that demethylated DNA is coincident with a decondensed chromatin state, it might be expected that the demethylated regions demonstrated here would show sensitivity to nucleases. However, since it has been found that DNAse treatment of nuclei has detected DNAse hypersensitive sites only in the HEL and K562 cell lines (J. Sowden personal communication and PhD thesis, 1991), it could be tentatively suggested that the HeLa DNA in this region is in a condensed, but demethylated form. Unfortunately these hypersensitivity studies could not determine the general level of DNaseI sensitivity of this region. Clearly further investigation is needed in this area, but it would seem likely that such work

on the demethylation seen here (which appears to have taken place in most of the cell lines), is more likely to throw light on the relationship between methylation and the transformed state than that between methylation and *CA1* expression. The difficulty of obtaining nucleated erythroid tissue from bone marrow has unfortunately prevented study of more normal *CA1*-expressing cells.

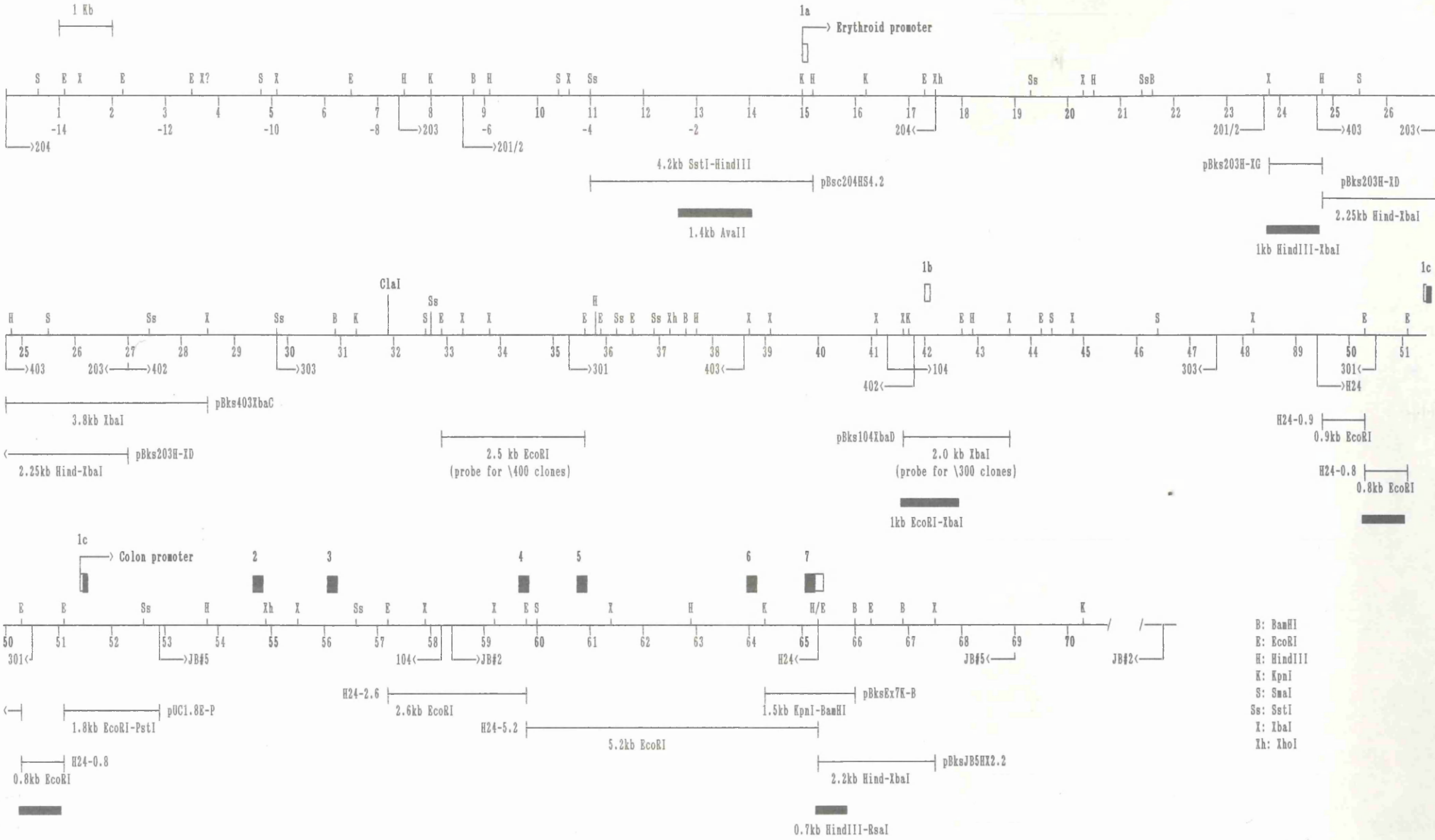


Fig. A1. MAP OF THE HUMAN *CA1* GENE. Distances are marked in kb from the 5'-end of lambda recombinant 204. The erythroid promoter lies at 15kb and the colon promoter at 51 kb. Exons are shown above the line as black bars (coding sequence) or open boxes (untranslated sequence). Below the line, the positions of the ends of the various lambda recombinants are indicated. Lines below the map indicate those regions subcloned into plasmids or used as probes. Bars beneath these lines indicate repetitive-sequence-free regions most commonly used as probes.

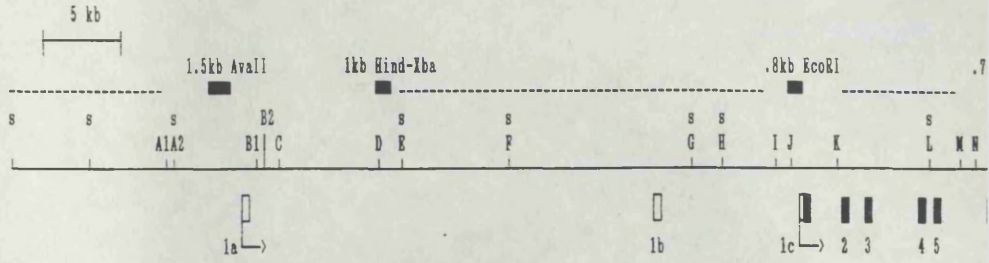


Fig. A2 Map of the *CA1* gene showing positions of *Hpa*II sites A1 to O. Th
sites with "s" above them are the central four bases of *Sma*I sites.
broken lines indicate those regions where *Hpa*II sites have not been map
and those *Hpa*II sites indicated in this region are known to exist only f
*Sma*I site mapping.

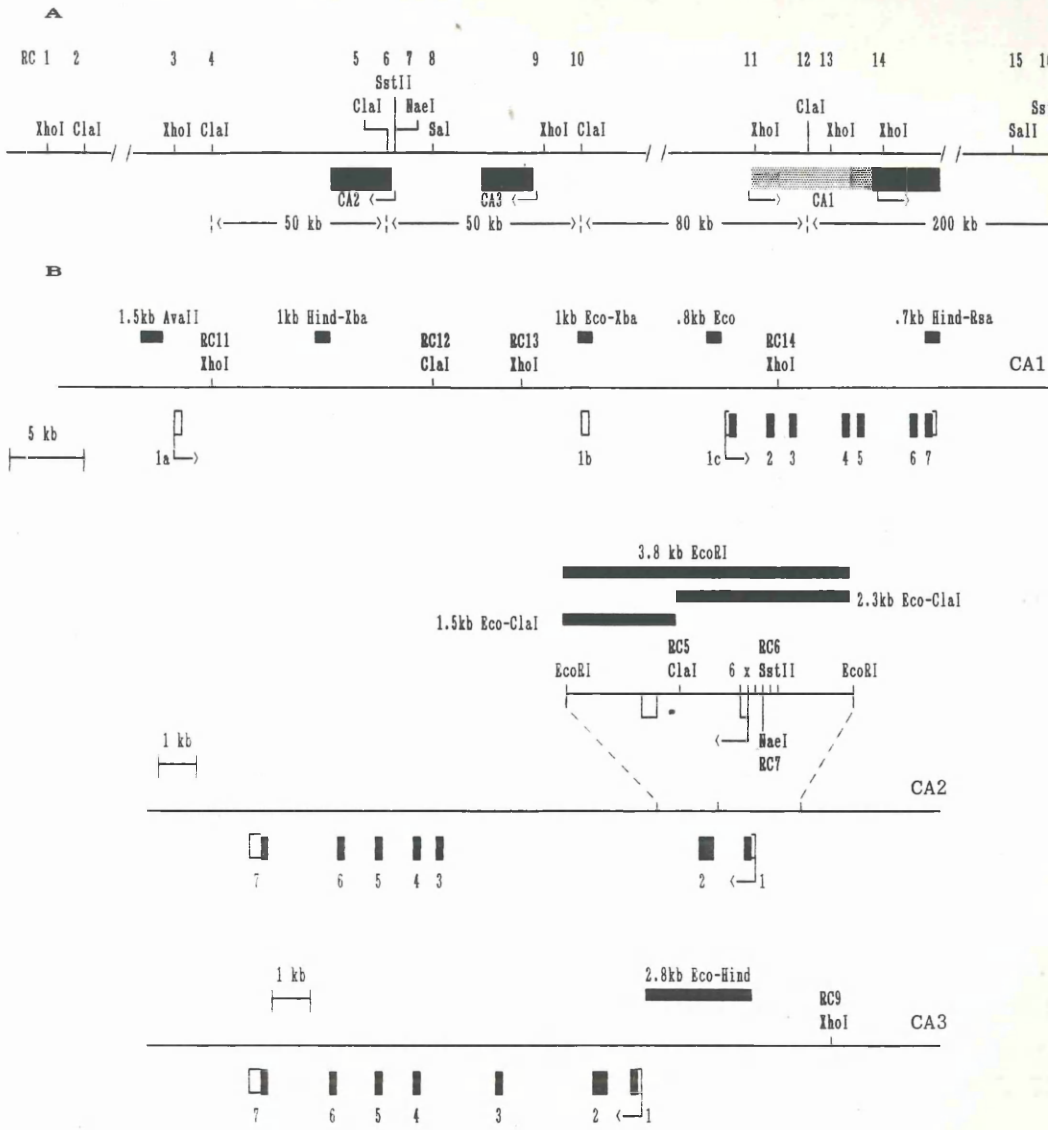


Fig. A3 Map of rare-cutter restriction sites in the *CA* gene cluster on chromoson
8 and probes used in methylation analysis. A: Map of the whole CA cluster wit
*Cla*I, *Nae*I, *Sal*I, *Sst*II and *Xho*I sites shown. Numbers over each site are thos
reffered to in the text (i.e. RC1 to RC18). Genes are shown by boxes beneath th
map with arrows indicating direction of transcription. B: Maps of each gene i
the cluster. Exons are shown as filled (coding sequence) or open boxes (non-codin
sequence). Bars above each gene map indicate probes used in the methylatio
analysis.

REFERENCES

Acampora, D., D'Esposito, M., Faiella, A., Pannese, M., Migliaccio, E., Morelli, F., Stornaiuolio, A., Nigro, V., Simeone, A., and Boncinelli, E. (1989) The human HOX gene family. Nucl. Acids Res. 17: 10385-10403.

Alter, B. P., Rappeport, J. M., Huisman, T. H. J., Shroeder, W. A. and Nathan, D. G. (1976) Fetal erythropoiesis following bone marrow transplantation. Blood 48: 843-853.

Andersson, L. C., Nillson, K. and Gahmberg, C. G. (1979) K562- a human erythroleukemic cell line. Int. J. Cancer. 23: 143-147.

Antoniou, M., deBoer, E., Habets, G. and Grosveld, F. (1988) The human ß-globin gene contains multiple regulatory regions: identification of one promoter and two downstream enhancers. EMBO J. 7: 377-384.

Auffray, C. and Rougeon, F. (1980). Purification of mouse immunoglobulin heavy-chain messenger RNAs from total myeloma tumor RNA. Eur. J. Biochem. 107: 303-314.

Banerji, J., Rusconi, S. and Schaffner, W. (1981). Expression of a ß-globin gene is enhanced by remote SV40 DNA sequences. Cell 27: 299-308.

Barnhart, K. M., Kim, C. G. and Sheffery, M. (1989). Purification and charachterisation of an erythroid cell-specific factor that binds the murine α- and ß-globin genes. Mol. Cell. Biol. 9: 2606-2614.

Baron, M. H. and Maniatis, T. (1986) Rapid reprogramming of globin gene expression in transient heterokaryons. Cell 46: 591-602.

Becker, P. B., Ruppert, S. and Schütz, G. (1987). Genomic footprinting reveals cell type-specific DNA binding of ubiquitous factors. Cell 51: 435-443.

Beechey, C., Tweedie, S., Spurr, N., Ball, S., Peters, J. and Edwards, Y. (1990). Mapping of mouse carbonic anhydrase-3, *Car-3:* another locus in the

homologous region of mouse chromosome 3 and human chromosome 8. Genomics 6: 692-696.

Behringer, R. R., Hammer, R. E., Brinster, R. L., Palmiter, R. D. and Townes, T. M. (1987) Two 3' sequences direct erythroid specific expression of human ß-globin genes in transgenic mice. Proc. Natl. Acad. Sci. USA. 84: 7056-7060.

Benz, E. J., Murnane, M. J., Tonkonow, B. L., Berman, B. W., Mazur, E. M., Cavallesco, C., Jenko, T., Snyder, E. L., Forget, B. G. and Hoffman, R. (1980) Embryonic-fetal characeristics of a human erythroleukaemic cell line. Proc. Natl. Acad. Sci. USA. 77: 3509-3513.

Bernards, A., Rubin, C. M., Westbrook, C. A., Paskind, M. and Baltimore, D. (1987) The first intron in the human c-*abl* gene is at least 200 kilobases long and is a target for translocations in chronic myelogenous leukemia. Mol. Cell. Biol. 7: 3231-3236.

Bird, A. P., Taggart, M., Frommer, M., Miller, O. J. and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell 40: 91-99.

Bird, A. P. (1987). CpG islands as gene markers in the vertibrate genome. Trends Genet. 3: 342-247.

Bird, A. P., Taggart, M. H., Nicholls R. D. and Higgs, D. R. (1987b). Non methylated CpG-rich islands at the human α-globin locus: implications for evolution of the α-globin pseudogene. EMBO. J. 6: 999-1004.

Birnboim, H.C. & Doly, J. (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucl. Acids Res. 7: 1513-1523.

Boer, H. and Hickey, D. A. (1986) The alpha-amylase gene in *Drosophila melanogaster*: nucleotide sequence, gene structure and expression motifs. Nucl. Acids Res. 14: 8399-8411.

Boyer, S. H., Siegel, S. and Noyes, A. N. (1983). Developmental changes in

human erythrocyte carbonic anhydrase: Coordinate expression with adult hemoglobin. Dev. Biol. 97: 250-253.

Brady, H. J. M., Sowden, J. C., Edwards, M., Lowe, N., and Butterworth, P. H. W. (1989). Multiple GF-1 binding sites flank the erythroid specific transcription unit of the human carbonic anhydrase I gene. FEBS. Letters 257: 451-456.

Brady, H. J. M., Lowe. N., Sowden, J. C., Edwards, M. and Butterworth, P. H. W. (1991) The human carbonic anhydrase I gene has two promoters with different tissue specificities. Biochem. J. In Press.

Brady, H. J. M., Edwards, M., Linch, D. C., Barlow, J. H. and Butterworth, P. H. W. (1990). Expression of the human carbonic anhydrase I gene is activated late in fetal erythroid development and regulated by stage-specific *trans*-acting factors. Brit. J. Haematol. 76: 135-142.

Breathnach, R. and Chambon, P. (1981) Organisation and expression of eukaryotic split genes coding for proteins. Ann. Rev. Biochem. 52: 441-466.

Brent, R. and Ptashne, M. (1985) A eukaryotic transcription activator bearing the DNA specificity of a prokaryotic repressor. Cell 43: 729-736.

Briggman, J. V., Tashian, R. E. and Spicer, S. S. (1983). Immunohistochemichal localisation of carbonic anhydrasae I and II in eccrine sweat glands from control subjects and patients with cystic fibrosis. Am. J. Pathol. 112: 250-257.

Buschhausen, G., Wittig, B., Graessmann, M. and Graessmann, A. (1987). Chromatin structure is required to block transcription of the methylated herpes simplex virus thymidine kinase gene. Proc. Nat. Acad. Sci. USA. 84: 1177-1181.

Busslinger, M., Hurst, J. and Flavell, R. A. (1983). DNA methylation and the regulation of globin gene expression. Cell 34: 197-206.

Carle, G. F. and Olson, M. V. (1984) Separation of chromosomal DNA molecules from yeast by orthogonal-field-alternation gel electrophoresis. Nuc. Acids Res. 12: 5647-5664.

Carle, G. F., Frank, M. and Olson, M. V. (1986) Electrophoretic separation of large DNA molecules by periodic inversion of the electric field. Science 232: 65-68.

Carter, M. J. and Parsons, D. S. (1971). The isoenzymes of carbonic anhhydrase: tissue, subcellular distribution and functional significance, with particular reference to the intestinal tract. J. Physiol. 215: 71-94.

Carter, M. J. (1972). Carbonic anhydrase: Isoenzymes, properties, distribution, and functional significance. Biol. Rev. 47: 465-513.

Carter, N. D. (1972). Carbonic anhydrase isozymes in *Cavia porcellus, Cavia aperea* and their hybrids. Comp. Biochem. Physiol. 43b: 743-747.

Carter, N. D., Hewett-Emmett, D. Jeffery, S., and Tashian, R. E. (1981). Testosterone-induced sulfonamide-resistant carbonic anhydrase of rat liver is indistinguishable from skeletal muscle carbonic anhydrase III. FEBS Lett. 128: 114-118

Carter, N. D., Heath, R., Welty, J. R., Hewett-Emmett, D., Jeffery, S., Shiels, A. and Tashian, R. E. (1984) Red cells genetically deficient in carbonic anhydrase II have elevated levels of a carbonic anhydrase indistinguishable from muscle CAIII. Ann. N.Y. Acad. Sci. 429: 284-286.

Carle, G. F., Frank, M., and Olson, M. V. (1986). Electrophoretic separation of large DNA molecules by periodic inversion of the electric field. Science 232: 65-68.

Cedar, H. (1988). DNA methylation and gene activity. Cell 53: 3-4.

Chen, J.-d., Denton, M. J., Morgan, G., Pearn, J. H. and Mackinlay, A. G. (1988) The use of field inversion gel electrophoresis for deletion detection in Duchenne muscular dystrophy. Ann. Hum. Genet. 42: 777-780.

Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. and Rutter, W. J. (1979). Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry. 18: 5294-5299.

Chretien, S., Dubart, A., Beaupain, D., Raich, N., Grandchamp, B., Rosa, J., Goosens, M. and Romeo, P-H. (1988). Alternative transcription and splicing of the human porphobilinogen deaminase gene result either in tissue-specific or in housekeeping expression. Proc. Natl. Acad. Sci. USA. 85: 6-10.

Cockerill, P. N. and Garrard, W. T. (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.

Comb, M. and Goodman, H. M. (1990). CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. Nucl. Acids Res. 18: 3975-3982.

Coulson, A., Sulston, J., Brenner, S. and Korn, J. (1986) Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA 83: 7821-7825.

Cross, S. H. and Little, P. F. R. (1986). A cosmid vector for systematic chromosome walking. Gene 49: 9-22

Darmon, M., Nicolas, J. F. and Lamblin, D. (1984) 5-Azacytidine is able to induce the conversion of teratocarcinoma derived mesechymal cells into epithelial cells. EMBO J. 3: 961-967.

Davenport, H. W. and Fisher, R. B. (1938). Carbonic anhydrase in the gastrointestinal mucosa. J. Physiol. 94: 16*P*.

Davenport, H. W. and Wilhelmi, A. E. (1941). Renal carbonic anhydrase. Proc. Soc. Exp. Biol. Med. 48: 53-56.

Davidson, E. H., Jacobs, H. T. and Britten, R. T. (1983) Very short repeats

and coordinate induction of genes. Nature 301: 468-470.

Davis, M. B., West, L. F., Barlow, J. H., Butterworth, P. H. W. B., Lloyd, J. C. and Edwards, Y. H. (1987). Regional localisation of carbonic anhydrase genes *CA1* and *CA3* on human chromosome 8. Somat. Cell Mol. Genet. 13: 173-178.

DeSimone, J., Linde, M. and Tashian, R. E. (1973a). Evidence for linkage of carbonic anhydrase isozyme genes in the pig-tailed macaque, *Macaca nemestrina*. Nature New Biol. 242: 55-56.

DeSimone, J., Magid, E. and Tashian, R. E. (1973b). Genetic variation in the carbonic anhydrase isozymes of macaque monkeys. II. Inheritence of red cell carbonic anhydrase levels in different carbonic anhydrase I genotypes of the pig-tailed macaque, *Macaca nemestrina*. Biochem. Genet. 8: 165-174.

Dexter, T. M. (1979) Cell interactions in vitro. Clin. Haematol. 8: 453-458

Dierks, P., Van Ooyen, A., Cochran, M. D., Dobkin, C., Reiser, J. and Weissmann, C. (1983). Three regions upstream of the cap site are required for efficient and accurate transcription of the rabbit ß- globin gene mouse 3T6 cells. Cell 31: 695-706.

Dodgson, S. J., Forster, R. E. Schwed, D. A. and Storey, B. T. (1983). Contribution of matrix carbonic anhydrase to citrulline synthesis in isolated guinea pig liver mitochondria. J. Biol. Chem. 258: 7696-7701

Dodgson, S. J., Forster, R. E. and Storey, B. T. (1984). The role of carbonic anhydrase in hepatocyte metabolism. Ann. N. Y. Acad. Sci. 429: 516-524.

Dodgson, S. J. and Contino, L. C. (1988). Rat kidney mitochondrial carbonic anhydrase. Arch. Biochem. Biophys. 260: 334-341.

Dynan, W. S. (1989). Modularity in promoters and enhancers. Cell 58: 1-4.

Edwards, Y. H., Barlow, J. H., Konialis, C. P., Povey, S. and Butterworth, P. H. W. B. (1986a). Assignment of the gene determining human carbonic anhydrase, CAI, to chromosome 8. Ann. Hum. Genet. 50: 123-129

Edwards, Y. H., Lloyd, J., Parkar, M. and Povey, S. (1986b). The gene for human muscle specific carbonic anhydrase III (CAIII) is assigned to chromosome 8. Ann. Hum. Genet. 50: 41-47.

Edwards, Y. H., Charlton, J. and Brownson, C. (1988) A non-methylated CpG-rich island associated with the human muscle-specific carbonic anhydrase III gene. Gene 17: 473-481.

Edwards, Y. H. (1990) CpG islands in genes showing tissue-specific expression. Phil. Trans. R. Soc. Lond. B. 326: 207-215.

Eicher, E. M., Stern, R. H., Womack, J. E., Davisson, M. T., Roderick, T. H. and Reynolds, S. C. (1976). Evolution of mammalian carbonic anhydrase loci by tandem duplication: Close linkage of *Car-1* and *Car-2* to the cerntrome reregion of chromosome 3 of the mouse. Biochem. Genet. 14: 651-660.

Elgin, S. C. R. (1981). DNAaseI-hypersensitive sites of chromatin. Cell 27: 414-415

Emerson, B. M., Nickol, J. M. and Fong, T. C. (1989). Erythroid-specific activation and derepression of the chick ß-globin promoter in vitro. Cell 57: 1189-1200.

Eng, L.-I. L. and Tarail, R. (1966). Carbonic anhydrase deficiency with persistence of foetal haemoglobin: a new sindrome. Nature. 211: 47-49.

Enver, T., Zhang, J-W., Papayannopoulou, T. and Stamatoyannopoulos, G. (1988a). DNA methylation: a secondary event in globin gene switching ?. Genes Dev. 2: 698-706.

Enver, T., Zhang, J-W., Anagnou, N. P., Stamatoyannopoulos, G. and

Papayannopoulou, T. (1988b). Developmental programs of human erythroleukemia cells: globin gene expression and methylation. Mol. Cell Biol. 8: 4917-4926.

Eriksson, A. E. (1988). Structural differences between high and low activity forms of carbonic anhydrases. Ph.D. thesis, University of Uppsala, Sweden.

Evans, R. M. and Hollenberg, S. M. (1988) Zinc Fingers: Gilt by association. Cell 52: 1-3.

Evans, R. M. (1988) The steroid and thyroid hormone receptor superfamily. Science 240: 889-895

Evans, T., Reitman, M., and Felsenfeld, G. (1988). An erythrocyte-specific DNA binding factor recognises a regulatory sequence common to all chicken globin genes. Procl. Natl. Acad. Sci. USA 85: 5978-5980.

Evans, T. and Felsenfeld, G. (1989). The erythroid-specific transcription factor Eryf1: A new finger protein. Cell 58: 877-885.

Feinberg, A. P., Gehrke, C. W., Kuo, K. C. and Ehrlich, M. (1988) Reduced genomic 5-Methylcytosine content in human colonic neoplasia. Cancer Res. 48: 1159-1161

Feinberg, A. P. and Ranier, S. (1990) A novel strategy for identifying potential targets of altered DNA methylation in neoplastic transformation. *in* "Nucleic acid methylation" UCLA Symposia on molecular and cellular biology. 128 (Eds. Clawson, G. A., Willis, D. B., Weissbach, A. and Jones, P. A.). Wiley-Liss N.Y. pp221-228.

Feinberg, A. P., and Vogelstein, B. (1983a) Hypomethylation distinguishes genes of some human cancer from their normal counterparts. Nature 301: 89-91.

Feinberg, A. P., and Vogelstein, B. (1983b). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal.

Biochem. 132: 6-13.

Fernley, R. T., Coghlan, J. P., and Wright, R. D. (1988a) Purification and charachterisation of a high-M_r carbonic anhydrase from sheep parotid gland. Biochem. J. 249: 201-207.

Fernley, R. T., Wright, R. D. and Coghlan, J. P. (1988b). Complete amino acid sequence of ovine salivary carbonic anhydrase. Biochemistry. 27: 2815-2820.

Fernley, R. T. (1988) Non cytoplasmic carbonic anhydrases. Trends in Biochem. Sci. 13: 356-359.

Ferrell, R. E., Osborne, W. R. A. and Tashian, R. E. (1981). Effect of metabolic acidosis on hydrogen ion excretion in a pigtail macaque with erythrocyte carbonic anhydrase I deficiency. Proc. Soc. Exp. Biol. Med. 168: 155-158.

Fletcher, C., Heintz, N. and Roeder, R. G. (1987) Purification and characterisation of OTF-1, a transcription factor regulating cell cycle expression of a human histone H2b gene. Cell 51: 773-781.

Foe, V. E., Wilkinson, L. E. and Laird, C. D. (1976) Comparative organisation of active transcription units in Oncopeltus fasciatus. Cell. 9: 131-146.

Foley, G. E., Lazarus, H., Farber, S. (1965) Continuous culture of human lymphoblasts from peripheral blood of a child with acute leukemia. Cancer 18: 522-529.

Fourney, R. M., Miakoshi, J., Day, R. S. and Paterson, M. C. (1988). Northern Blotting: Efficient staining and transfer. BRL. Focus. 10: 5-7.

Fraser, P. J. and Curtis, P. J. (1986). Molecular evolution of the carbonic anhydrase genes: Calculation of divergence time for mouse carbonic anhydrase I and II. J. Mol. Evol. 23: 294-299.

Fraser, P. J. and Curtis, P. J. (1987). Specific pattern of gene expression during induction of mouse erythroleukemia cells. Genes Dev.

Fraser, P. J., Cummings, P. and Curtis, P. J. (1989). The mouse carbonic anhydrase I gene contains two tissue-specific promoters. Mol. Cell Biol. 9: 3308-3313.

Freshney, R.I. (1983) Culture of animal cells - a manual of b a s i c techniques, Alan R. Liss Inc., New York.

Fried, M. and Crothers, D. M. (1981) Equilibria kinetics of lac repressor-operator interactions ,b polyacrylamide gel electrophoresis. Nucl. Acids Res. 9: 6505-6523

Funakoshi, S. and Deutsch, H. F. (1971). Human carbonic anhydrases. V. Levels in erythrocytes in various states. J. Lab. Clin. Med. 77: 39-45.

Galas, D. and Shimitz, A. (1978) DNase footprinting: a simple method for the detection of protein-DNA binding specificity. Nucl. Acids Res. 5: 3157-3170.

Gallarda, J. L., Foley, K. P., Yang, Z. Y. and Engel, J. D. (1989) The ß-globin stage selector element factor is erythroid-specific promoter/enhancer binding protein NF-E4. Genes Dev. 3: 1845-1859.

Gama-Sosa, M. A., Slagel, V. A., Trewin, R. W. and Ehrlich, M. (1983). The 5-methylcytosine content of DNA from human tumors. Nucl. Acids Res. 11: 6683-6894.

Garel and Axel (1976). Selective digestion of transcriptionally active ovalbumin genes from oviduct nuclei. Proc. Natl. Acad. Sci. USA. 73: 3966-3977.

Gazit, B Panet, A., and Cedar, H. (1980) reconstetution of DNasel sensitive structure on active genes. Proc. Natl. Acad. Sci. USA 77: 1787-1780.

Gey, G. O., Coffman, W. D. and Kubicek, M. T. (1952) Tissue culture studies

of proliferative capacity of cervical carcinoma and normal epithelium. Cancer Res. 12: 264-265.

Ghazal, P., Clark, J. and Bishop, J. O. (1985) Evolutionary amplification of a pseudogene. Proc. Natl. Acad. Sci. USA 82: 4182-4185

Gibson T. J., Rosenthal, A. and Waterston, R. H. (1978). Lorist 6, a cosmid vector with *Bam*HI, *Not*I, *Sca*I and *Hin*dIII cloning sites and altered neomycin phosphotransferase gene expression. Gene 53: 283-286.

Gill, G., Smith, J. R., Goldstein, J. L. and Brown, M. S. (1987) Optional exon in the 5'-untranslated region of 3-hydroxy-3-methylglutaryl coenzyme A synthase gene: Conserved sequence and splicing pattern in humans and hamsters. Proc. Natl. Acad. Sci. USA 84: 1863-1866.

Goelz, S. E., Vogelstein, B. Hamilton, S. R. and Feinberg, A. P. (1985) Hypomethyylation of DNA from benign and malignant human colon neoplasms. Science 228: 187-190.

Gootenberg, S. E., Ruscetti, F. W., Mier, J. W., Gazdar, A. and Gallo, R. C. (1981) Human cutaneous T cell lymphoma and leukemia cell lines produce and respond to T cell growth factor. J. Exp. Med. 154: 1403-1418.

Gordon, J. W. and Ruddle, F. H. (1985). DNA-mediated genetic transformation of mouse embryos and bone marrow - a review. Gene 33: 121.

Greaves, D. R., Wilson, F. D., Lang, G. and Kioussis, D. (1989) Human CD2 3'-flanking sequences confer high-level T cell-specific, position-independent gene expression in transgenic mice. Cell 56: 979-986.

Green, P. M., Montandon, A. J., Bently, D. R., Ljung, R., Nilsson, I. M. and Giannelli, F. (1990). The incidence and distribution of CpG - TpG transition in the coagulation factor IX gene. A fresh look at CpG mutational hotspots. Nucl. Acids Res. 18: 3227-3231.

Green, S. and Chambon, P. (1987). Oestradiol induction of a glucocorticoid-responsive gene by a chimaeric receptor. Nature 325: 75-78.

Grosfeld, F., Blom van Assendelft, G., Greaves, D. R. and Kollias, G. (1987). Position-independent, high-level expression of the human ß-globin gene in transgenic mice. Cell 51: 975-985.

Gross, D. S. and Garrard, W. T. (1988). Nuclease hypersensitive sites in chromatin. Ann. Rev. Biochem. 57: 159-197.

Groudine, M., Kohwi-Shigematsu, T., Gelinas, R., Stamatoyannopoulous, G. and Papayannopoulo, T. (1983). Human fetal to adult hemoglobin switcing: Changes in chromatin structure of the ß-globin gene locus. Proc. Natl. Acad. Sci. USA. 80: 7551-7555.

Gubler, U. and Hoffman, B. J. (1983) A simple and very efficient method for generating cDNA libraries. Gene 25: 263-269.

Gumucio, D. L., Rood, K. L., Gray, T. A., Riordan, M. F., Sartor, C. I. and Collins, F. S. (1988) Nuclear protein that binds the human γ-globin gene promoter: alteration in binding produced by point mutations associated with hereditary persistence of fetal hemoglobin. MolCell Biol 8: 5310-5322.

Henriques, O. M. (1928). Die Bindungsweise des Kohlendioxids im Blute. Biochem. Z. 200: 1-24.

Hewett-Emmett, D., Hopkins, P. J., Tashian, R. E. and Czelusniak. (1984). Origins and molecular evolution of the carbonic anhydrase isozymes. N. Y. Acad. Sci. 249: 338-358

Hewett-Emmett, D., Cook, R. G. and Dodgson, S. J. (1986). Novel gene encodes mitochondrial carbonic anhydrase (CA V) that lacks the active-site tyrosine of the amino-terminal region of CA I, II and III. Fed. Proc. 45: 1661, (abstract #1055).

Hewett-Emmett, D. and Tashian, R. (1990). Structure and evolutionary origins of the carbonic anhydrase multigene family. *In* "The carbonic anhydrases: cellular physiology and molecular genetics" (S. Dodgson, G. Gross and R. E. Tashian, Eds.), Plenum Publishing Corp., N. Y. (in press).

Higgs, D. R., Wood, W. G., Jarman, A. P., Sharpe, J., Pretorius, I.-M. and Ayyub, H. (1990) A major positive regulatory region located far upstream of the human α-globin locus. Genes Dev. 4: 1588-1601

Hodge, M. R. and Cumsky, M. G. (1989) Splicing of a yeast intron containing an unusual 5' junction sequence. Mol. Cell Biol. 9: 2765-2770.

Hoffman, R. M. (1984) Altered methionine metabolism, DNA methylation and oncogene expression in carcinogenesis. A review and synthesis. Biochim. Biophys. Acta 738: 49-87

Höller, M., Westin, G., Jiricny, J. and Schaffner, W. (1988). Sp1 transcription factor binds DNA and activates transcription even when the binding site is CpG methylated. Genes Dev. 2: 1127-1135.

Holliday, R. (1990) DNA methylation and epigenetic inheritance. Phil. Trans. Royal. Soc. Lond. B 326: 329-338

Holmes, R. S. (1976). Mammalian carbonic anhydrase isozymes: evidence for a third locus. J. Exp. Zool. 197: 289-295.

Huang, H. C. and Cole, R. D. (1984) The distribution of H1 histone is non-uniform in chromatin and correlates with different degrees of condensation. J. Biol. Chem. 259: 14237-14242.

Hutcheon, T., Dixon, G. H. and Levy Wilson, B. (1980). Transcriptionally active mononucleosomes from trout testis are heterogeneous in composition. J. Biol. Chem. 255: 681-685.

Iguchi-Ariga, S. M. M. and Schaffner, W. (1989). CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. Genes Dev. 3: 612-619.

Ish-Horowicz, D. & Burke, J.F. (1981) Rapid and efficient cosmid cloning. Nucl. Acids Res. 9, 2989-2998.

Jones, P. A. (1985). Altering gene expression with 5-azacytidine. Cell. 40: 485-486.

Jones, N. C. Rigby, P. W. J. and Ziff, E. B. (1988) *Trans*-acting protein factors and the regulationof eukaryotic transcription: lessons from studies on DNA tumor viruses. Genes Dev. 2: 267-281

Jones, K. A. and Tijan, R. (1985) Spl Binds to promoter sequences and activates HSV "immediate-early" gene transcription in vitro. Nature 317: 179-182.

Kadonga, J. T., Carner, K. R., Masiars, F. R. and Tijan, R. (1987) Isolation of cDNA encoding transcription factor Spl and functional analysis of the DNA binding domain. Cell 51: 1079-1090

Kannan, K. K. (1980). Crystal structure of carbonic anhydrase. *in* "Biophysics and biochemistry of carbon dioxide. (C. Bauer, G. Gros and H. Bartels, Eds.) Springer Verlag. Berlin. pp 184-205.

Karlsson, S. and Nienhuis, A. W. (1985) Developmental regulation of human globin genes. Ann. Rev. Biochem. 54: 1071-1108

Karn, J.S., Brenner, S., Barnett, L. & Cesareni, G. (1980) Novel bacteriophage \ cloning vector. Proc. Natl. Acad. Sci. USA 77: 5172-5176.

Karn, J., Matthes, H. W. D., Gait, M. J. and Brenner, S. (1984). A new selective phage cloning vector, \2001, with sites for *Xba*I, *Bam*HI, *Hind*III, *Eco*RI, *Sst*I and *Xho*I. Gene 32: 217-224.

Kearney, P., Barlow, J., Wolfe, J. and Edwards, Y. (1987). Physical linkage of CA1 and CA3: Human gene mapping 9. Cytogenet. Cell Genet. 46: 637-638.

Kendall, A. G. and Tashian, R. E. (1977). Erythrocyte carbonic anhydrase I: inherited deficiency in humans. Science. 197: 471-472.

Keshet, I., Yisraeli, J. and Cedar, H. (1985). Effect of regional DNA

methylation on gene expression. Proc. Natl. Acad. Sci. USA. 82: 2560-2564.

Keshet, I., Lieman-Hurwitz, J. and Cedar, H. (1986). DNA methylation affects the formation of active chromatin. Cell 44: 535-543.

Khalifa, R. G. (1971). The carbon dioxode hydration activity of carbonic anhydrase: Stop-flow kinetic studies on the native human isozymes B and C. J. Biol. Chem. 246: 2561-2573.

Kiouissis, D., Wilson, F., Daniels, C., Leveton, C., Taverne, J. and Playfair, J. H. L. (1987). Expression and rescuing of a cloned human tumor necrosis factor gene using an EBV-based shuttle cosmid vector. EMBO J. 6: 355-361.

Koester, M. K., Register, A. M. and Noltman, E. A. (1977). Basic muscle protein, a third genetic locus isoenzyme of carbonic anhydrase? Biochem. Biophys. Res. Comm. 76: 196-204.

Kollias, G., Hurst, J., deBoyer, E. and Grosveld, F. (1987) Atissue and developmental specific enhancer is located downstream from the human ß-globin gene. Nucl. Acids Res. 15, 5739-5747.

Konkel, D. A., Maizel, J. V., and Leder, P. (1979) The evolution and sequence comparison of two recently diverged mouse chromosomal ß-globin genes. Cell 18: 865-873

Kondo, T., Taniguchi, N., Taniguchi, K., Matsuda, I. and Murao, M. (1978). Inactive form of erythrocyte carbonic anhydrase B in patients with primary renal tubular acidosis. J. Clin. Invest. 62: 610-617

Kondo, T., Tanaguchi, N., Hirano, T., and Kawakami, Y. (1984). Inactive form of carbonic anhydrase I in erythrocytes from primary aldosteronism. N. Y. Acad. Sci. 429: 302-305.

Kovesdi, I. Reichel, R. and Nevins, J. R. (1987). Role of an adenovirus E2 promoter binding factor in E1A-mediated coordinate gene control. Proc. Natl. Acad. Sci. USA. 84: 2180-2184.

Kumar, V., Green, S., Stack, G., Berry, M., Jin, J.-R. and Chambon, P. (1987) Functional domains of the human estrogen receptor. Cell 51: 941-951

Kumar, V. and Chambon, P. (1988) The estrogen receptor binds tightly to its responsive element as a ligand-induced homodimer. Cell 55: 145-156

Kumpulainen, T. (1981) Human carbonic anhydrase isoenzyme C. Histochemistry 72: 425-431

Kumpulainen, T. and Korhonen, L. K. (1978). Immunohistochemichal demonstration of carbonic anhydrase. Histochemistry 58: 83-192

Lajtha, L. G. (1979) Stem cell concepts. Differentiation, 14: 23-30

Landschulz, W. H., Johnson, P. F. and McKnight, S. L. (1988) The leucine zipper: A hypothetical structure common to a new class of DNA binding proteins. Science. 240: 1759-1764

Latchman, D. S. (1990) Eukaryotic transcription factors. Biochem. J. 270: 281-289.

LaThange, N. B. and Rigby, P. W. J. (1988) *Trans*-acting protein factors and the regulation of eukaryotic transcription. In "Transcription and Splicing", B. D. Hames and D. Glover Eds. 3-42. Oxford: IRL Press.

Lawn, R., Efstradis, A., O'Connell, C. and Maniatis, T. (1980) The nucleotide sequence of the human ß-globin gene. Cell 21: 647-651.

Lee, W., Haslinger, A., Karin, M. and Tijan, R. (1987). Activation of transcription by two factors that bind promoter and enhancer sequences of the human metallothionein gene and SV40. Nature, 235: 368-372

Leff, S. E., Rosenfeld, M. G. and Evans, R. M. (1986) Complex transcriptional units: Diversity in gene expression by alternative RNA processing. Ann. Rev. Biochem. 55: 1091-1117

LeFranc, M.-P., Foster, A., Baer, R., Stinson, M. A. and Rabbitts, T. H. (1986). Diversity and rearrangement of the human T cell rearranging genes: Nine germ-line variable genes belonging to two subgroups. Cell. 45: 237-246.

Leibovitz, A., Stinson, J. C., McCombs III, W. B., McCoy, C. E., Mazur, K. C. and Mabry, N. D. (1976) Classification of human colorectal adenocarcinoma cell lines. Cancer Res. 36: 4562- 4569.

Leung, S., Proudfoot, N. J. and Whitelaw, E. (1987) The gene for O-globin is transcribed in human fetal erythroid tissue. Nature 329: 551-554.

Lewis, C. D., Clark, S. P., Felsenfeld, G. and Gould, H. (1988a). An erythrocyte-specific protein that binds to the poly(dG) region of the chicken ß-globin gene promoter. Genes Dev. 2: 863-873.

Lewis, S. E., Erickson, R. P., Barnett, L. B., Venta, P. J. and Tashian, R. E. (1988b). *N*-Ethyl-*N*-nitrosourea-induced null mutation at the mouse *Car-2* locus: An animal model for human carbonic anhydrase II deficiency syndrome. Proc. Nat. Acad. Sci. USA 85: 1962-1966.

Lindskog, S. (1960). Purification and properties of bovine erythrocyte carbonic anhydrase. Biochem. Biophys. Acta. 39: 218-226.

Little, P. F. R. (1987) Choice and use of cosmid vectors. *in* "DNA cloning: a practical approach" Vol. 3. (D. M. Glover, Ed.). IRL Press Ltd, Oxford. pp 19-42.

Lloyd, J., McMillan, S., Hopkinson, D. and Edwards, Y. H. (1986). Nucleotide sequence and derived amino acid sequence of a cDNA encoding human muscle carbonic anhydrase. Gene, 41: 233-239.

Lloyd, J., Brownson, C., Tweedie, S., Charlton, J., and Edwards, Y. H. (1987). Human muscle carbonic anhydrase gene: Structural and DNA methylation patterns in fetal and adult tissue. Gene. Dev. 1: 594-602.

Loc, P.-V. and Strätling, W. H. (1988). The matrix attatchment regions of

the chicken lysozyme gene co-map with the boundaries of the chromatin domain. EMBO J. 7: 655-664

Lonnerholm, G. (1984) Histochemichal localisation of carbonic anhydrase in mammalian tissue. Ann. N. Y. Acad. Sci. 429: 369-381.

Lyons, G. E., Buckingham, M. E., Tweedie, S. and Edwards, Y. H. (1990) Carbonic anhydrase III, an early mesodermal marker, is expressed in embryonic mouse skeletal muscle and notochord. Development 111: 233-244

Ma, X., Fraser, P. and Curtis, P. J. (1991). A differentiation stage-specific factor interacts with mouse carbonic anhydrase form I gene and a conserved sequence in mammalian ß-globin genes. In press

Maeda, N. and Smithies, O. (1986). The evolution of multigene families: human haptoglobin genes. Ann. Rev. Genet. 20: 81-108.

Maniatis, T., Fritsch, E. F. and Sambrook, J. (1982). Molecular Cloning. A Laboratory Manual. Cold Spring Harbour Laboratory Press, Cold Spring Harbour, NY.

Mann, T. and Keilin, D. (1940). Sulphanilamide as a specific inhibitor of carbonic anhydrase. Nature 146: 164-165.

Mantovani, R., Malgaretti, N., Nicolis, S., Giglioni, B., Comi, P., Cappelini, N., Bertero, M. T., Caligaris-Cappio, F. and Ottolenghi, S. (1988). An erythroid specific nuclear factor binding to the proximal CACCC box of the ß-globin gene. Nucl. Acids Res. 16: 4299-4313.

Maren, T. H. (1984) The general physiology of reactions catalysed by carbonic anhydrase and their inhibition by sulfonamides. Ann. New York Acad. Sci 249: 568-579.

Maren, T. H. (1988). The kinetics of HCO_3^- synthesis related to fluid secretion, pH control, and CO_2 elimination. Ann. Rev. Physiol. 50: 695-717

Martin and Papayannopoulou (1982) HEL cells: a new human erythroleukemia

cell line with spontaneous and induced globin expression. Science 216: 1233-1235

Martin, D. I. K., Tsai, S.-F. and Orkin, S. H. (1989). Increased γ- globin expression in a nondeletion HPFH mediated by an erythroid-specific DNA-binding protein. Nature 338: 435-438.

Martin, D. I. K., Zon, L. I., Mutter, G. and Orkin, S. H. (1990). Expression of an erythroid transcription factor in megakaryocytic and mast cell lineages. Nature 344: 444-446.

Mavilio, F., Giampaolo, A., Care, A., Migliaccio, G., Calandrini, M., Russo, G., Pagliardi, G. L., Mastroberardino, G., Marinucci, M., and Peschle, C. (1983). Molecular mechanisms of human hemoglobin switching: selective undermethylation and expression of globin genes in embryonic, fetal, and adult erythroblasts. Proc. Natl. Acad. Sci. USA 80: 6907-6911.

Meldrum, N. U. and Roughton, F. J. W. (1933). Carbonic anhydrase. Its preparation and properties. J. Physiol. 80: 113-142.

Melis, M., Demopulos, G., Najfeld, V., Zhang, J.-W., Brice, M., Papayanopoulu and Stamatoyannopoulos, G. (1987). A chromosome 11-linked determinant controls fetal globin expression and the fetal to adult globin switch. Proc. Natl. Acad. Sci. USA. 84: 8105-8109

Michiels, F., Burmeister, M. and Lehrach, H. (1987). Derivation of clones close to *met* by preparative field inversion-gel electrophoresis. Science 236: 1305-1308.

Mignotte, V., Wall, L., deBoer, E., Grosveld, F. and Romeo, P.-H. (1989). Two tissue-specific factors bind the erythroid promoter of the human porphobilinogen deaminase gene. Nucl. Acids Res. 17: 37-58.

Mitchell, P. J. and Tijan, R. (1989) Transcriptional activation by sequence-specific DNA binding proteins. Science 245: 371-378.

Montgomery, J. C., Venta, P. J., and Tashian, R. E. (1986). Isolation and

characterisation of a human genomic clone containing a portion of a carbonic anhydrase-like gene. Isozyme Bull. 19: 12.

Montgomery, J. C., Shows, T. B., Venta, P. J. and Tashian, R. E. (1987). Gene for novel human carbonic anhydrase (CA) isozyme on chromosome 16 is unlinked to the *CAI/CAII/CAIII* gene cluster. Amer. J. Hum. Genet. 41: A229 (abstract #680)

Murakami, H. and Sly, W. S. (1987). Purification and characterization of human salivary carbonic anhydrase. J. Biol. Chem. 262: 1382-1388.

Murray, E. J. and Grosveld, F. (1987) Site-specific demethylation in the promoter of human γ-globin gene does not alleviate methylation mediated supression. EMBO J. 6: 2329-2335.

Nakai, H., Byers, M. G., Venta, P. J., Tashian, R. E. and Shows, T. B. (1987). The gene for human carbonic anhydrase II (CA2) is located at chromosome 8q22. Cytogenet. Cell. Genet. 44: 234-235.

Nostrand, B., Vaara, I. and Kannan, K. K. (1974). Structural relationship of human erythrocyte carbonic anhydrase isozymes B and C. *in* "Isozymes: Molecular Structure." Vol. 1 (C. L. Market, Ed). Academic Press, New York. pp. 575-599

Nyman, P. O. (1961). Purification and properties of carbonic anhydrase from human erythrocytes. Biochem. Biophys. Acta 52: 1-12

Orkin, S. H., Kazazian, H. H., Antonorakis, S. E., Goff, S., Boehm, C. D., Sexton, J., Waber, P. and Giardina, P. (1982). Linkage of ß-thalassaemia mutations and ß-globin gene polymorphisms with DNA polymorphisms in the human ß-globin gene cluster. Nature 296: 627-631.

Orkin, S. H., Antonorakis, S. E. and Kazazian, H. H. (1984). Base substitution at position -88 in a ß-thalassemic globin gene. Further evidence for the role of distal promoter element ACACCC. J. Biol. Chem. 259: 8679-8681.

Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S. and Sharp, P. A.: Splicing of messenger RNA precursors. Ann. Rev. Biochem. 55 (1986) 1119-1150.

Palmiter, R. D. and Brinster, R. L. (1985). Transgenic mice. Cell 41: 343.

Perbal, B. V. (1988) A practical guide to molecular cloning. John Wiley and Sons, New York (Second edition).

Perkins, N. D., Orchard, K. H., Collins, M. L. K., Latchman, D. S. and Goodwin, G. H. (1990). Detection in non-erythroid cells of a factor with the binding characteristics of the erythroid cell transcription factor EF1. Biochem. J. 269: 543-545

Pevny, L., Simon. M. C., Robertson, E., Klein, W. H., Tsai, S.-F., D'Agati, V., Orkin, S. H. and Constantini, F. (1991) Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. Nature 349: 257-260.

Plumb, M., Frampton, J., Wainwright, H., Walker, M., Macleod, K., Goodwin, G. and Harrison, P. (1989). GATAAG; a *cis*-control region binding an erythroid-specific nuclear factor with a role in globin and non-globin gene expression. Nucl. Acids Res. 17: 73-92.

Ptashne (1988) How eukaryotic transcriptional activators work. Nature 335: 683-689.

Pukkila, P. J. (1987). Telling right from wrong: a role for DNA methylation. TIG. 3: 1-2.

Rackwitz, H. R., Zehtner, G., Frischauf, A-M. and Lehrach, H. (1984). Rapid restriction mapping of DNA cloned in lambda phage vectors. Gene 30: 195-200.

Register, A. M,. Koester, M. K. and Noltman, E. A. (1978). Discovery of carbonic anhydrase in rabbit skeletal muscle and evidence for its identity with "basic muscle protein". J. Biol. Chem. 253: 4143-4152.

Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) Labelling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. J. Mol. Biol. 113: 237-251.

Rogers, J. H. (1987). Sequence of carbonic anhydrase II cDNA from chick retina. Eur. J. Biochem. 162: 119-122.

Romeo, P.-H., Prandini, M.-H., Joulin, V., Mignotte, V., Prenant, M., Vainchenker, W., Marguerite, G. and Uzan, G. (1990). Megakaryocytic and erythrocytic lineages share specific transcription factors. Nature 344: 447-449.

Russell, G. J., Walker, P. M. B., Elton, R. A. and Subak-Sharpe, J. H. (1976). Doublet frequency analysis of fractionated vertebrate nuclear DNA. J. Mol. Biol. 108: 1-23.

Ryan, U. S., Whitney, P. L. and Ryan, J. W. (1982). Localisation of carbonic anhydrase on pulmonary artery endothelial cells in culture. J. Appl. Physiol. 53: 914-919.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) Molecular cloning: A laboratory manual. Second Edition. Cold Spring Harbour Laboratory Press, Cold Spring Harbour, New York.

Sandeen, G., Wood, W. and Felsenfeld, G. (1980). The interaction of high mobility proteins HMG 14 and 17 with nucleosomes. Nucl. Acids Res. 8: 3757-3778.

Sanger, F., Nicklen, S. and Coulson, A. R.: DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA 74 (1977) 5463-5467.

Sanyal, G. and Maren, T. H. (1981). Thermodynamics of carbonic anhydrase catalysis: A compparison between the human isoenzymes B and C. J. Biol. Chem. 256: 608-612

Sargeant, C. A., Dunham, I. and Campbell, R. D. (1989) Identification of

multiple HTF-island associated genes in the human major histocompatability complex class III region. EMBO J. 8: 2305-2312.

Sato, S., Zhu, X. L. and Sly, W. S. (1990). Carbonic anhydrase isozymes IV and II in urinary membranes from carbonic anhydrase II-deficient patients. Proc. Natl. Acad. Sci. USA. 87: 6063-6076

Schäfer, A. and Dietsch, P. (1984). A 54,000 molecular weight protein with carbonic anhydrase activity in rabbit erythrocytes. Ann. N. Y. Acad. Sci. 249: 241-242.

Schibler, U. and Sierra, F. (1987) Alternative promoters in developmental gene expression. Ann. Rev. Genet. 21: 237-257

Schlessinger, D., (1990) Yeast artificial chromosomes: tools for mapping complex genomes. Trends Genet. 6: 248-258.

Schlissel, M. S. and Brown, D. D. (1984). The transcriptional regulation of Xenopus 5S RNA genes in chromatin: the roles of active stable transcription complexes and histone H1. Cell 37: 903-913.

Schmid, M., Haaf, T. and Grunert, D. (1984) 5-azacytidine-induced undercondensations in human chromosomes. Hum. Genet. 67: 257-263

Sefton, L., Kelsey, G., Kearney, P. and Wolfe, J. (1990) A physical map of the human *PI* and *AACT* genes. Genomics 7: 382-388

Selker, E. U. (1990) DNA methylation and chromatin structure: A view from below. Trends in Biochem. Sci. 15: 103-107.

Sell, J. E. and Petering, H. G. (1974). Carbonic anhydrase from human neonatal erythrocytes. J. Lab. Clin. Med. 84: 369-377.

Serfling, E., Jasin, M. and Schaffner, W. (1985) Enhancers and eukaryotic gene transcription. Trends in Genet. 1: 224-230.

Shapiro, L. H., Venta, P. J. and Tashian, R. E. (1987). Molecular analysis

of G+C-rich upstream sequences regulating transcription of the human carbonic anhydrase II gene.

Shapiro, L. H., Venta, P. J., Yu, Y.-S. L. and Tashian, R. E. (1989). Carbonic anhydrase II is induced in HL-60 cells by 1,25-dihydroxyvitamin D_3: a model for osteoclast gene regulation. FEBS lett. 249: 307-310

Sheridan, B. L., Weatherall, D. J., Clegg, J. B., Pritchard, J., Wood, W. G, Callender, S. T., Durrant, I. J., McWhirter, W. R., Ali, M., Partridge, J. W., and Thompson, E. N. (1976). Br. J. Haematol. The patterns of fetal haemoglobin production in leukaemia. 32, 487-506.

Sigler, P. B. (1988) Acid blobs and negative noodles. Nature 333: 210-212

Singer-Sam, J., Simmer, R.L., Keith, D.H., Shively, L., Teplitz, M., Itakura, K., Gartler, S.M. and Riggs, A.D. (1983) Isolation of a cDNA clone for human X-linked 3-phospholycerate kinase by use of a mixture of synthetic oligodeoxyribonicleotides as a detection probe. Proc. Natl. Acad. Sci. USA 80: 802-806.

Sly, W. S., Hewett-Emmett, D., Whyte, M. P., Yu, Y.-S. and Tashian, R. E. (1983). Carbonic anhydrase II deficiency identified as the primary defect in the autosomal recessive syndrome of osteopetrosis with renal tubular acidosis and cerebral calcification. Proc. Nat. Acad. Sci. USA. 80: 2752-2756.

Southern, E. M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98: 503-517.

Spandidos, D. A. (1984). Transfer of human globin genes to human erythroleukemia cells. Mol. Biol. Med. 2: 167-175.

Spicer, S. S., Sens, M. A. and Tashian, R. E. (1982). Immunocytochemichal demonstration of carbonic anhydrase in human epithelial cells. J. Histochem. Cytochem. 30: 864-873.

Spicer, S. S., Sens, M. A., Hennigar, R. A. and Stoward, P. J. (1984).

Implications of the immunohistochemichal localisation of the carbonic anhydrase isozymes for their function in normal and pathalogic cells. Ann. N. Y. Acad. Sci. 429: 382-397.

Stamatoyannopoulos, G., Constantoulakis, P., Brice, M., Kurachi, S. and Papayannopoulou, T. (1987). Coexpression of embryonic, fetal and adult globins in erythroid cells of human embryos: relevance to the cell-lineage models of globin switching. Dev. Biol. 123: 191-197.

Stein, R., Razin, A. and Cedar, H. (1982). *In vitro* methylation of the hamster adeninephosphoribosyl transferase gene inhibits its expression in mouse L cells. Proc. Natl. Acad. Sci. USA. 79: 3418-3422.

Sternberg, N. (1990) Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. Proc. Natl. Acad. Sci. USA 87: 103-107.

Stief, A., Winter, D. M., Stratling, W. H. and Sippel, A. E. (1989). A nuclear DNA attatchment element mediates elevated and position-independent gene activity. Nature 341: 343-345.

Strauss, F. and Varshavsky, A. (1984) Aprotein binds to a satellite DNA repeat at three specific sites that would be brought into mutual proximity. Cell 37: 889-901.

Struhl, K. (1989) Helix-turn-helix, zinc-finger and leucine zipper motifs for eukaryotic transcriptional regulatory proteins. Trends in Biochem. Sci. 14: 137-140.

Suggs, S. V., Wallace, R. B., Hirose, T., Kawashima, E. H. and Itakura, K. (1981) Proc. Natl. Acad. Sci. USA 78: 6613-6617.

Talbot, D., Collis, P., Antoniou, M., Vidal, M. Grosveld, F. and Greaves, D. R. (1989). A dominant controll region from the human ß-globin locus conferring integration site-independant gene expression. Nature 338: 352-353.

Tashian, R. E., Hewett-Emmett, D., Dodgson, S. J., Forster, R. E. and Sly, W. S. (1984). The value of inherited deficiencies of humancarbonic anhydrase isozymes in understanding their cellular role. Ann. N. Y. Acad. Sci. 429: 262-275.

Tashian, R. E. (1989). The carbonic anhydrases: Widening perspectives on their evolution, expression and function. BioEssays 10: 186-192.

Taylor, S. M. and Jones, P. A. (1979) Multiple new phenotypes induced in 10T1/2 and 3T3 cels treated with 5-azacytidine. Cell 17: 771-779.

Townes, T. M. and Behringer, R. R. (1990). Human globin locus activation region (LAR): role in temporal control. TIG. 6: 219-223.

Trainor, C. D., Evans, T., Felsenfeld, G. and Boguski, M. S. (1990). Structure and evolution of a human erythroid transcription factor. Nature. 343: 92-96.

Tsai, S-F., Martin, D. I. K., Zon, L. I., D'Andrea, A. D., Wong, G. G., and Orkin, S. H. (1989). Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. Nature 339: 446-451.

Tu, C. K., Sanyal, G., Wynns, G. S. and Silverman, D. N. (1983) The pH dependenceof the hydration of CO_2 catalyzed by carbonic anhydrase III from skeletal muscle of the cat: Steady-state and equilibrium studies. J. Biol. Chem. 258: 8867-8871.

Tuan, D., Solomon, W., Li, Q. and London, I. M. (1985). The "ß-like-globin" gene domain in human erythroid cells. Proc. Natl. Acad. Sci. USA. 82: 6384-6388.

van der Ploeg, L. H. T. and Flavell. (1984). DNA methylation in the human ß-globin locus in erythroid and nonerythroid tissues. Cell 19: 947-958.

van Slyke, D. D. and Hawkins, J. A. (1930). Studies of gas and electrolyte

equilibria in blood XVI. The evolution of carbon dioxide from blood and buffer solutions. J. Biol. Chem. 87: 265-279.

Venta, P. J., Montgomery, J. C., Wiebauer, K., Hewett-Emmett, D. and Tashian, R. E. (1984). Organisation of the mouse and human carbonic anhydrase II genes. Ann. N. Y. Acad. Sci. 429: 309-323.

Venta, P. J., Montgomery, J. C., Hewett-Emmett, D., Weibauer, K. and Tashian, R. E. (1985a). Structure and exon to protein domain relationships of the mouse carbonic anhydrase II gene. J. Biol. Chem. 260: 12130-12135.

Venta, P. J., Montgomery, J. C., Hewett-Emmett, D. and Tashian, R. E. (1985b). Comparison of the 5' regions of the human and mouse carbonic anhydrase II genes and identification of possible regulatory elements. Biochim. Biophys. Acta 826: 195-201.

Venta, P. J., Montgomery, J. C. and Tahsian, R. E. (1987). Molecular genetics of carbonic anhydrase isozymes. *In* "Isozymes: Current topics in biological and medical research" (M. C. Razatti, J. G. Scandalios and G. S. Whitt, Eds.). Alan R. Liss. Inc., New York. pp. 58-72.

Vidali, G., Boffa, L. C., Bradbury, E. M. and Allfrey, V. G. (1978). Butyrate supression of histone deacetylation leads to accumulation of multiavetylated forms of histones H3 and H4 and increased DNase1 sensitivity of the associated DNA sequences. Proc. Natl. Acad. Sci. USA 75: 2239-2243.

Villeval, J. L., Pellicci, P. G., Tabilio, A., Titeux, M., Henri, A., Louache, F., Thomopuolos, P., Vainchenker, W., Garbaz, M., Rochant, H., Breton-Gorius, J. Edwards, P. A. W. and Testa, U. (1983) Erythroid properties of K562 cells. Effects of hemin, butyrate and TPA induction. Exp. Cell. Res. 146: 423-431.

Villeval, J. L., Testa, U. Vinci, G., Tonthat, H., Bettaieb, A., Titeux, M., Cramer, P., Edelman, L., Rochant,. Breton-Gorius, J. and Vainchenker, W. (1985). Carbonic anhydrase I is an early specific marker of normal human

erythroid differentiation. Blood. 66: 1162-1170.

Villeval, J. L., Cramer, P., Lemoine, F., Henri, A. Bettaieb, A., Bernaudin, F., Beuzard, R., Berger, R., Flandrin, G., Breton-Gorius, J. and Vainchenker, W. (1986). Phenotype of early erythroblastic leukemias. Blood 68: 1167-1174.

Wall, L., deBoer, E., and Grosveld, F. (1988). The human ß-globin gene 3' enhancer contains multiple binding sites for an erythroid-specific protein. Gene. Dev. 2: 1089-1100.

Wallace, B.R., Johnson, M.J., Hirose, T., Miyake, T., Kawashima, E. & Itakura, K. (1981) The use of synthetic oligonucleotides as hybridization probes. II. Hybridization of oligonucleotides of mixed sequence to rabbit ß-globin DNA. Nucl. Acids Res. 9, 879-894.

Wang, R. Y.-H., Kuo, K. C., Gherke, C. W., Huang, L.-H. and Ehrlich, M. (1982). Heat- and alkali-induced deamination of 5-methylcytosine and cytosine. Biochim. Biophys. Acta. 697: 371-377.

Weatherall, D. J. and McIntyre, P. A. (1967). Developmental and acquired variations in erythrocyte carbonic anhydrase isozymes. Br. J. Haematol. 13: 106-114.

Weatherall, D. J. and Clegg (1981) The thalassaemia syndromes (Blackwell Scientific, Oxford).

Weintraub, H. and Groudine, M. (1976). Chromosomal subunits in active genes have an altered conformation. Science 193: 848-858.

Weisbrod, S. and Weintraub, H. (1979). Isolation of a subclass of nuclear proteins responsible for conferring a DNase1 sensitive structure on globin chromatin. Proc. Natl. Acad. Sci. USA 76: 630-634.

Whitney, P. L. and Briggle, T. V. (1982). Membrane-associated carbonic anhydrase purified from bovine lung. J. Biol. Chem. 257: 12056-12059.

Wigler, M., Levy, D. and Perucho, M. (1981). The somatic replication of DNA methylation. Cell 24: 33-40.

Wistrand, P. J. (1984a). Properties of membrane-bound carbonic anhydrase. Ann. N. Y. Acad. Sci. 429: 195-209.

Wistrand, P. J. (1984b). The use of carbonic anhydrase inhibitors in opthalmology and clinical medicine. N. Y. Acad. Sci. 429: 609-619.

Wistrand, P. J. and Knuuttila, K.-G. (1989). Renal membrane-bound carbonic anhydrase. Purification and properties. Kidney Int. 35: 851-859.

Wood, W. G., Clegg, J. B. and Weatherall, D. J. (1977). Developmental biology of human hemoglobins. *in* "Progress in hematology X". (Ed. E. B. Brown). Grune and Stratton, New York. pp43-90.

Wood, W. G., Bunch, C., Kelly, S., Gunn, Y. and Breckon, G. (1985a). Control of haemoglobin switching by a developmental clock? Nature. 313: 320-323.

Wood, W. I., Gitshier, J., Lasky, L. A. and Lawn, R. M. (1985b). Base composition-independent hybridisation in tetramethylammonium chloride: A method for oligonucleotide screening of highly complex gene libraries. Proc. Natl. Acad. Sci. USA 82: 1585-1588.

Wood, W. G., Clegg, J. B. and Weatherall, D. J. (1979) Hereditary persistence of fetal haemoglobin (HPFH) and ß thalassemia Br. J. Haematol. 43: 509-512.

Woolley, P. (1975). Models for metal ion function in carbonic anhydrase. Nature 258: 677-682.

Wright, S., de Boer, E., Grosveld, F., and Flavell, R. A. (1983) Regulated expression of the human ß-globin gene family in murine erythrolukaemia cells. Nature 339: 446-451.

Wyman, A. R., Wolfe, L. B. and Botstein, D. (1985) Propogation of some

human DNA sequences in bacteriophage \ vectors requires mutant *Escherichia coli* hosts. Proc. Natl. Acad. Sci. USA 82: 2880-2884.

Yamamoto, M., Ko, L. J., Leonard, M. W., Beug, H., Orkin, S. H., and Engel, J. D. (1990) Activity and tissue-specific expression of the transcription factor NF-E1 multigene family. Genes Dev. 4: 1650-1662.

Yisraeli, J., Adelstein, R. S., Melloul, D., Nudel, U., Yaffe, D. and Cedar, H. (1986). Muscle-specific activation of a methylated chimeric actin gene. Cell 46: 409-416.

Young, R. A. and Davis, R. W. (1983) Efficient isolation of genes by using antibody probes. Proc. Natl. Acad. Sci. USA 80: 1194-1198.

Young, R. A., Bloom, B. R., Grosskinsky, C. M., Iranyi, J., Thomas, D. and Davies, R.W. (1985) Dissection of *Mycobacterium tuberculosis* antigens using recombinant DNA Proc. Natl. Acad. Sci. USA 82, 2583-2587.

Yu, C. Y. and Milstein, C. (1989) A physical map linking the five human thymocyte differentiation antigen genes.

Zhu, X. L. and Sly, W. S. (1988). Carbonic anhydrase IV from human lung and kidney: Purification,characterisation and amino acid sequence. J. Cell Biol. 107: 852a (abstract #4851).

Zon, L. I., Tsai, S.-F., Burgess, S., Matsudiara, P., Bruns, G. A. P. and Orkin, S. H. (1990). The major erythroid ,DNA-binding protein (GF-1): Primary sequence and localisation of the gene to the X chromosome. Proc. Nat. Acad. Sci. USA. 87: 668-672.

GENE 03648

Structure and methylation patterns of the gene encoding human carbonic anhydrase I

(Recombinant DNA; nucleotide sequence; alternative splicing; polyadenylation; erythroid expression; transcription factors; erythroleukemic cell lines)

Nicholas Lowe, Hugh J.M. Brady, Jonathan H. Barlow, Jane C. Sowden, Mina Edwards and Peter H.W. Butterworth

Biochemistry Department, University College London, London WC1E 6BT (U.K.)

Received by R.W. Davies: 20 December 1989
Revised: 2 May 1990
Accepted: 3 May and 15 May 1990

SUMMARY

The gene (*CAI*) encoding human carbonic anhydrase I (CAI) has been isolated and shown to have a total length of 50 kb. Some 36 kb of this consists of a large intron separating the erythroid-specific promoter from the coding region. A small (54 bp) noncoding exon from within this intron is occasionally found in transcripts. Two different polyadenylation sites have been found, the most distal of which is the most commonly used. Methylation levels near the promoter differ widely in cell lines. In *CAI*-expressing cells, a region of DNA near the promoter is demethylated in a generally highly methylated background. Surprisingly, non-*CAI*-expressing cell lines show much lower levels of methylation.

INTRODUCTION

The carbonic anhydrases are a family of zinc metalloenzymes which catalyse the reversible hydration of CO_2. They are found in almost all organisms and in mammals exist in a number of isoforms each of which exhibits a characteristic pattern of tissue distribution. Of these isozymes, the cytosolic carbonic anhydrases CAI, II and III are probably the best characterised. CAI is found at its highest levels in red blood cells, CAIII is characteristic of muscle and male rat liver while CAII is distributed in a wide range of tissues and cell types. Distinct isozymes have been shown to exist in mitochondria (CAV), as a membrane-bound form in lung and kidney (CAIV) and as a secreted salivary form (CAVI). A seventh isozyme (designated CAVII or CAZ) has been proposed to exist based on the analysis of genomic clones containing a carbonic anhydrase-like gene. For recent reviews see Fernley (1988) and Tashian (1989).

CAI is the second most abundant protein in human erythrocytes after globin and has been shown to be a very early marker for erythroid differentiation. It appears before most other proteins considered to be characteristic of the erythroid lineage, being found in cells of the BFU-E stage or earlier (Villeval et al., 1985). At lower levels, CAI has been shown to be present in the intestinal epithelium, vascular endothelium, corneal epithelium and lens of the eye. In addition to its tissue specificity, CAI is also regulated at a developmental level: during fetal development CAI levels in the blood change, being virtually absent early in gestation, rising just before birth and reaching their adult levels several years after birth (Boyer et al., 1983). Developmental changes in gene expression have received much attention in the case of the globins and the molecular mechanisms underlying these changes are just starting to be understood. Similar molecular analysis of nonglobin genes such as *CAI* will be critical in determining whether these mechanisms are

Correspondence to: Dr. P.H.W. Butterworth, Biochemistry Dept., University College London, Gower St., London WC1E 6BT (U.K.) Tel. (071)387-7050, ext. 2234/7; Fax (071)380-7193.

Abbreviations: aa, amino acid(s); bp, base pair(s); CAI, carbonic anhydrase I; *CAI*, gene (DNA) Rencoding CAI; cDNA, DNA complementary to RNA; EBV, Epstein–Barr virus; kb, kilobase(s) or 1000 bp; nt, nucleotide(s); oligo, oligodeoxyribonucleotide; *tsp*, transcription start point.

gene-specific or are examples of a more general form of developmental control.

As a family, the carbonic anhydrases present evolutionary biologists with a closely related set of genes which have, through evolution, acquired diverse patterns of expression. The molecular analysis of these genes should be a useful complement to studies of other multigene families such as the globins in which all members of the family are expressed in the same tissue type.

The aim of the present study was to isolate and characterise recombinants containing the human *CAI* gene, as a prelude to the elucidation of the molecular mechanisms controlling its regulation.

RESULTS AND DISCUSSION

(a) Isolation and characterisation of recombinants encoding the entire *CAI* gene

Initial recombinants (λHGCAI5.1, λHGCAI2.1) containing most of the protein coding sequence were isolated by screening a human genomic library in λ2001 with the human *CAI* cDNA (Barlow et al., 1987). Recombinants (λHGCAI201-204) containing exon 1a at the extreme 5′ end of the transcript were isolated using an oligonucleotide probe designed to hybridise to the 5′ end of the cDNA, while clones spanning the large intron (λHGCAI104, 301-303, 402-403) were isolated by chromosome walking. In addition to these recombinants we were also provided with a λCharon4a recombinant (λH24) by Richard Tashian (University of Michigan), isolated by low stringency probing of a library with a human *CAII* cDNA. The position of these recombinants is shown in Fig. 1.

In total, over 70 kb of DNA was cloned with the protein coding region lying within a stretch of 14 kb. Exon 1a lies approximately 36 kb 5′ to the translation start, separated by a large intron giving a total gene length of approximately 50 kb. Within this large intron lies a small (54 bp) exon (exon 1b) which is found in a minority of cDNAs.

Restriction fragments containing cDNA sequence were subcloned into M13 and sequenced, either using M13 sequencing primers where the intron/exon junctions lay close to a convenient restriction site or using oligos which flank the junctions. As with the other *CA* genes, we found that, apart from the unusual exon structure in the 5′ leader, the gene is divided into seven exons (Fig. 1). To retain some consistency of nomenclature within the *CA* family, those exons containing protein-coding sequence have been termed 1c-7, and are equivalent to exons 1–7 in *CAII and CAIII* (Venta et al., 1985; Lloyd et al., 1987). The two noncoding exons lying upstream from these have been designated 1a and 1b (Fig. 1). The boundaries of exons 1c-7 lie in the same positions within the protein as those in other *CA* genes, except for mouse *CAII* where the junction between exons 4 and 5 interrupts a Gly codon 14 nt 5′ to its position in *CAI* (Yoshihara et al., 1987). All the intron/exon junctions have the form Exon/GT..intron..AG/Exon except for intron 5′ UTB which has a GC at its 5′ terminus (the donor splice site; Fig. 3).

(b) cDNA sequences suggest the possibility of minor alternative RNA processing events

Sequencing and restriction enzyme analysis showed that, at the 5′ end of the message, only one out of seven cDNA recombinants contained exon 1b suggesting that message containing this exon is a minority species. To support this hypothesis, Northern blots of reticulocyte RNA were hybridised with oligo probes specific for either exon 1b or exon 1c (see Fig. 2 for position of oligos). Strong signals were obtained using the exon 1c probe, while no signal could be seen with the 1b probe (data not shown). In addition primer extension experiments, carried out to map

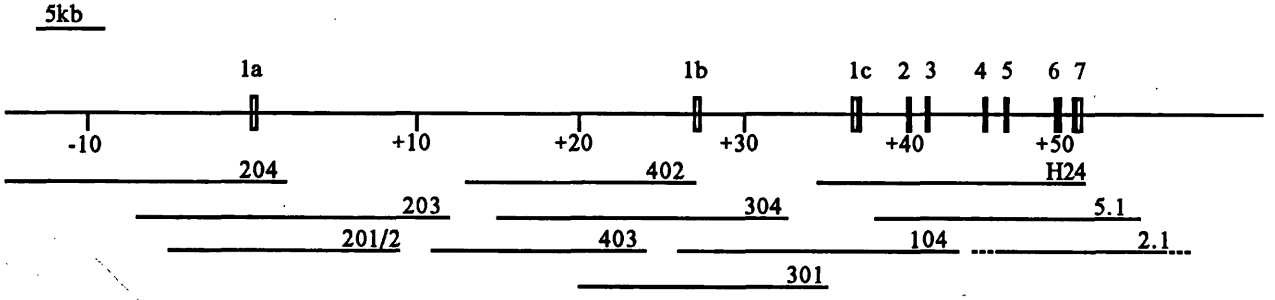


Fig. 1. The structure of the human carbonic anhydrase I gene. The scale is numbered in kb from the *tsp* identified in Brady et al. (1989). Exons, marked 1a–7, are indicated by black bars (coding sequence) or open boxes (noncoding sequence) and are not drawn to scale due to their small size. The annotated lines below indicate the extent of the cloned regions in the various λ recombinants isolated. The methylation analysis shown in Fig. 4 was carried out on the region around 1a. The λ2001 library used (a kind gift of Dr T. Rabbits, M.R.C. Laboratory of Molecular Biology, Hills Rd., Cambridge, U.K.) was generated by insertion of *Sau*3A partially-digested EBV-transformed SH-cell (a human B-lymphocyte cell line) DNA into the *Bam*HI site of the vector (LeFranc et al., 1986). The library was screened on PALL Biodyne A membranes using nick translated or random oligo-labelled probes by the method of Maniatis et al. (1982). For the isolation of λ HGCAI201-204, a ^{32}P-labelled oligo probe '5′UT18-38' (see Fig. 2) was used with a tetramethylammonium chloride wash following hybridisation (Wood et al., 1985). 'Cos' mapping of the recombinants was carried out essentially by the method of Rackwitz et al. (1984).

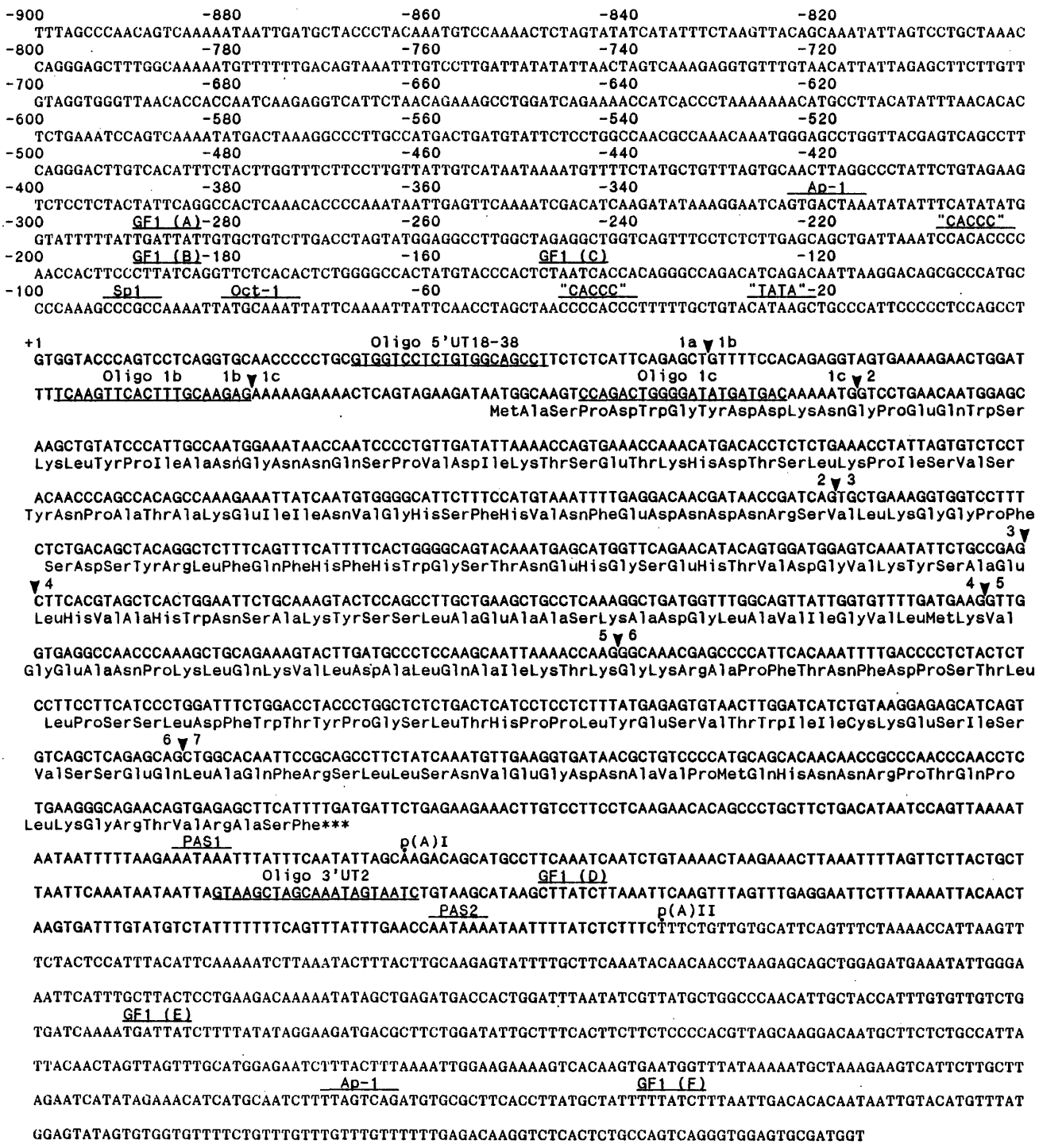


Fig. 2. Sequence of human *CAI* and flanking nt sequence. The transcribed sequence is shown in bold type, with the CAI aa sequence shown below the translated region. The points where introns interrupt the sequence are shown by arrowheads, with the exon numbers given on either side. The putative 'TATA' box (−28) is shown. Overlined are potential binding sites for several transcription factors — Ap1, Sp1, Oct1, 'CACCC' and the erythroid specific factor GF1. The signals for polyadenylation are overlined (PAS1 and PAS2) and the sites of polyadenylation (p(A)) are dotted over the nt where the first A of the poly(A) tail is found. Three oligos were used as probes to assess transcript levels of various parts of the message (see text); these have been designated as oligos 1b, 1c and 3′UT2. The sequence to which they would hybridise is underlined and labelled. Oligo 5′UT18-38 was used in the isolation of genomic recombinants λHGCAI201-4. Sequencing was carried out using the chain termination method of Sanger et al. (1977) following subcloning of overlapping restriction fragments isolated from genomic λ recombinants (see Fig. 1). This sequence has been deposited with Genbank and carries the accession number M33987. Three asterisks: stop codon.

the *tsp*, also failed to produce an extension product corresponding to that expected if exon 1b is present (Brady et al., 1989). Mouse *CAI* has also recently been shown to possess a large intron in the leader sequence but 'optional' exons have not been reported. In this instance it has been shown that a second promoter, responsible for expression of *CAI* in colon, lies just upstream of the protein-coding region within this large intron (Fraser et al., 1989). Given the

```
 Exon 1a                    27kb                              Exon 1b
ATTCAGAGCTgtaagtaaca..intron 5'UTA.aatcctgtgcttgcctctagGTTTTTCCAC

 Exon 1b                    9.5kb                             Exon 1c
TTTGCAAGAGgcagtaggaa..intron 5'UTB.gtattattctctgttttcagAAAAAGAAAA

 Exon 1c                    2.8kb                             Exon 2
GACAAAAATGgtaacacttc...intron 1....acacgtgtttgtcctggtagGTCCTGAACA

 Exon 2                     1.0kb                             Exon 3
AACCGATCAGgtgagctgaa...intron 2....tcggttccctttttcttccagTGCTGAAAGG

 Exon 3                     3.6kb                             Exon 4
TTCTGCCGAGgtaatgtaat...intron 3....aaccatagtatcatttttagCTTCACGTAG

 Exon 4                     0.9kb                             Exon 5
TTTGATGAAGgtgagttaca...intron 4....ttattttcttaaatctccagGTTGGTGAGG

 Exon 5                     3.0kb                             Exon 6
TAAAACCAAGgtaaacacac...intron 5....gatgtattcttttcttccagGGCAAACGAG

 Exon 6                     0.8kb                             Exon 7
CTCAGAGCAGgtagagttgt...intron 6....aaaatattttatccttctagCTGGCACAAT
```

Fig. 3. Sequence of human *CAI* intron/exon junctions. Exon sequence is shown in capitals, introns in lower case. Coding sequence shown in bold. The nonconsensus donor splice site of intron 5′UTB is underlined. Restriction fragments from genomic λ recombinants which had been shown, by hybridisation to the cDNA or oligo probes, to contain cDNA sequence were subcloned into M13 and sequenced using the chain termination method of Sanger et al. (1977). Those exon/intron junctions which did not lie close enough to a convenient cloning site to be sequenced with the M13 universal primer were sequenced using primers from within the exon and the sequence obtained used to make primers for sequencing the other strand.

similarity of structure between the mouse and human gene, it seems likely that the same arrangement could exist in the human *CAI* gene.

At the 3′ end of the gene, two different cDNA species have been found, terminating at two different polyadenylation sites. The proximal site (p(A)I) gives rise to a 3′-untranslated message length of 109 nt, while the distal site (p(A)II) produces a length of 334 nt. Although significantly different in size, only one transcript has ever been detected by Northern blot analysis, showing a predominant use of only one site. This site would appear to be p(A)II, based on two lines of evidence. Firstly, sequencing and restriction analysis of cDNA recombinants showed that only one out of eight clones were of the shorter type. Secondly, an oligo specific for the transcript sequence between the two polyadenylation sites produces a comparable signal to that produced with a protein-coding region probe (data not shown). Although we have no direct evidence for the use of the proximal site, both types of cDNAs terminate in a string of A's not found in the genomic sequence which lie 18 nt beyond a consensus polyadenylation signal (AATAAA) and therefore probably indicates genuine message rather than a cloning artifact.

The mouse *CAI* cDNA sequence reported also shows a polyadenylation signal 60 nt beyond the termination codon. The reported sequence however extends beyond this 320 nt downstream from the termination codon. Whether this first signal is utilised is not reported (Fraser and Curtis, 1986).

(c) The gene contains sequences associated with erythroid expression

Analysis of sequence flanking the 5′ and 3′ ends of the transcript shows the presence of a number of motifs associated with gene expression (Fig. 2). The promoter has a rather poor TATA box (−28), and no obvious CAAT box in a suitable position suggesting that expression from this promoter may be low in the absence of other elements regulating transcription. Potential binding sites for several transcription factors (including Ap-1, CACCC, Oct1, and Sp1) which may increase transcription are found flanking the gene and are shown in Fig. 2.

It has been shown that for many genes which are expressed specifically in erythroid tissues, high levels of expression are dependent on the presence of certain sequence motifs (RWGATWR$^{T}_{G}$; Wall et al., 1988), which act as the binding site for the erythroid-specific transcription factor GF1 (see Tsai et al., 1989, for a description of the cloning of this factor). Several of these sequences are found at both the 5′ and 3′ ends of the gene (Fig. 2) and are likely to be an important determinant of the expression of this gene. These sites have been shown to bind a protein using bandshift assays and DNase I footprinting, while a 191 bp DNA fragment containing GF1 sites A and B has been shown to up-regulate a heterologous gene promoter in erythroid cells (Brady et al., 1989).

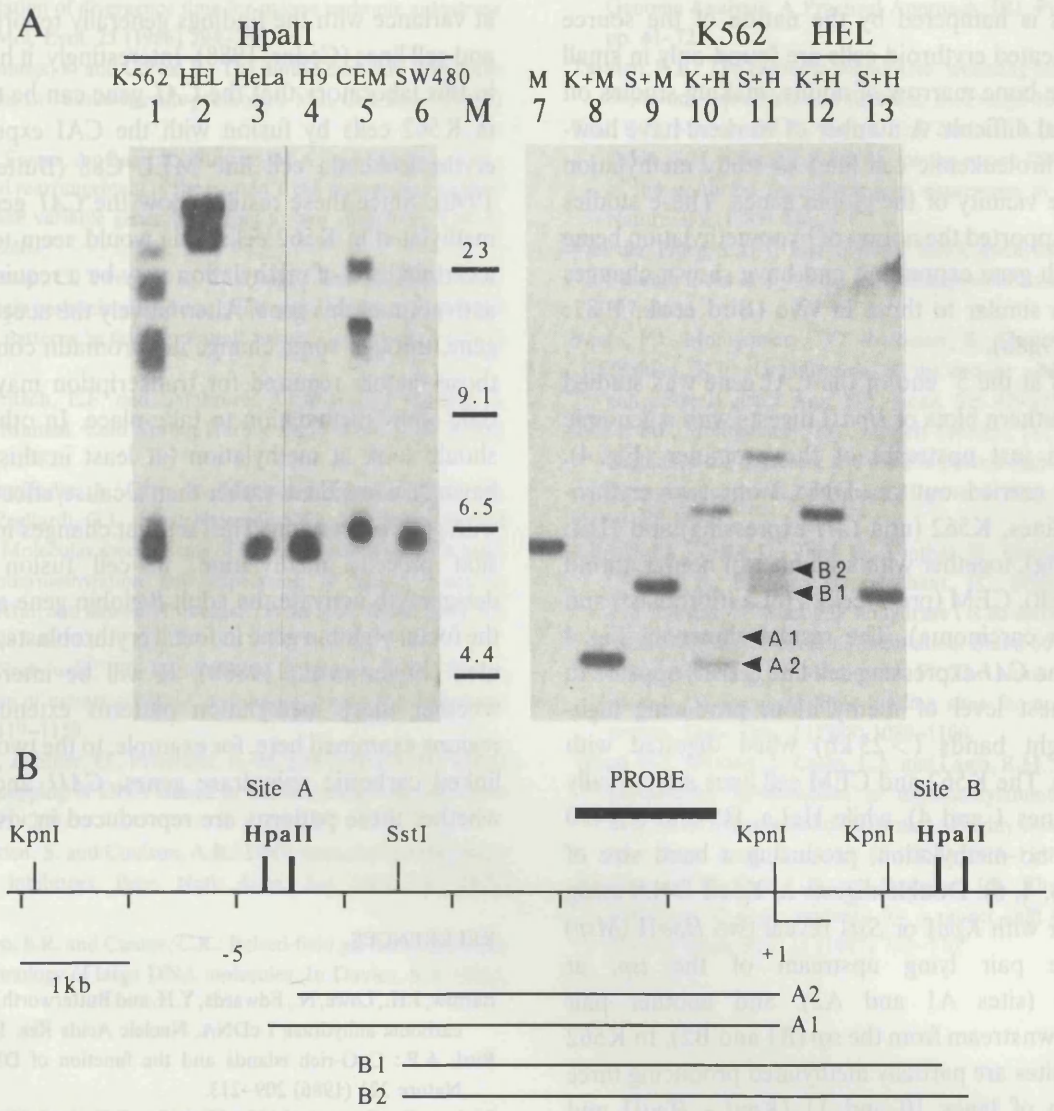


Fig. 4. *Hpa*II analysis of methylation states at the promoter region. (Panel **A**) DNA digested with *Hpa*II (H), *Msp*I (M), *Kpn*I (K) or *Sst*I (S) as indicated and hybridised with the 1.4-kb *Ava*II fragment probe shown in map **B**. M indicates position of size markers (sizes in kb). Lanes 1–6 show *Hpa*II digests of DNA from K562, HEL, HeLa, H9, CEM, and SW480 cell lines. Lanes 7–9 show placental DNA digested with the methylation insensitive *Hpa*II isoschizomer *Msp*I (together with *Kpn*I or *Msp*I as indicated) to show the fragments produced by *Hpa*II in the absence of methylation. The arrowed bands in lanes 10 and 11 correspond to those fragments labelled in map **B**. The highest bands in these two lanes are the parental *Kpn*I (7.0 kb) and *Sst*I (8.5 kb) fragments. Lanes 10–13 are double digests, using *Kpn*I + *Hpa*II or *Sst*I + *Hpa*II, of K562 (lanes 10 and 11) and HEL (lanes 12 and 13) cell DNA. (Map **B**) Map of the erythroid specific promoter of *CAI* indicating the probe used in methylation analysis and relevant restriction sites. The *tsp* (at a *Kpn*I site) is indicated (+ 1). The lines A1, A2, B1, and B2 correspond to the bands labelled on the Southern blot. Neither of the bands A1 or A2 is seen in the HEL digests (lane 12) indicating high levels of methylation at site A; conversely, only B1 is seen in lane 13 indicating lack of methylation at site B. DNA was separated on 0.7% agarose gels and alkali blotted onto Hybond N + hybridisation membrane (Amersham) according to the manufacturer's instructions. Probes (50–100 ng) were labelled using a random oligo labelling kit (BCL) and hybridised according to the protocol recommended for Hybond N + , but with the inclusion of 6% polyethylene glycol (mol.wt. 6000) in the hybridisation/prehybridisation mixture. Methylation studies were carried out on DNA prepared on separate occasions by two different methods. This was done in order to reduce the risk of artifactual digestion patterns resulting from poor DNA preparations. DNA was prepared either by the method of Maniatis et al. (1982), producing DNA in solution, or by the method of Smith et al. (1988), producing DNA embedded in 0.5% agarose. For DNA in solution, 1–5 μg of DNA was digested with 20–30 units of restriction enzyme(s) for at least 4 h in a volume of 30 μl. For DNA embedded in agarose, a block of 40 μl volume containing approximately 5 μg DNA would be digested in 150 μl of the appropriate restriction buffer and 30 units of enzyme overnight.

(d) Methylation patterns at the *CAI* gene in expressing, and non-expressing cell lines

Methylation at CpG dinucleotides has been shown to be negatively correlated with gene expression. For housekeeping genes, this is usually seen as a permanent demethylation of CpG-rich 'islands' close to the 5′ end of genes (Bird et al., 1987), while in genes whose expression is limited to particular cell types, hypomethylation of specific regions of DNA (usually close to the promoter) is frequently found in those cells or tissues expressing the gene (Cedar, 1988).

Analysis of the methylation state of DNA in human

erythroid cells is hampered by the nature of the source material. Nucleated erythroid cells are found only in small numbers in the bone marrow of adults, making studies on in vivo material difficult. A number of workers have however used erythroleukemic cell lines to study methylation patterns in the vicinity of the globin genes. These studies have largely supported the notion of hypomethylation being associated with gene expression and have shown changes in methylation similar to those in vivo (Bird et al., 1987; Enver et al., 1988a).

Methylation at the 5′ end of the CAI gene was studied by probing Southern blots of *Hpa*II digests with a genomic fragment from just upstream of the promoter (Fig. 4). Analysis was carried out on DNA from two erythroleukemic cell lines, K562 (non-*CAI*-expressing) and HEL (*CAI*-expressing), together with a number of non-erythroid lines: H9 (T-cell), CEM (pre T-cell), HeLa (fibroblast) and SW480 (colon carcinoma). The results shown in Fig. 4 indicate that the *CAI*-expressing cell line (HEL) appears to have the highest level of methylation, producing high-molecular-weight bands (>25 kb) when digested with *Hpa*II (lane 2). The K562 and CEM cell lines are partially methylated (lanes 1 and 4), while HeLa, H9 and SW480 show little or no methylation, producing a band size of 6.5 kb (lanes 3, 4, 6). Double digests of K562 DNA using *Hpa*II together with *Kpn*I or *Sst*I reveal two *Hpa*II (*Msp*) doublets, one pair lying upstream of the *tsp*, at −4.75/−4.5 kb (sites A1 and A2), and another pair 1.25/1.5 kb downstream from the *tsp* (B1 and B2). In K562 cells all these sites are partially methylated producing three bands in each of lanes 10 and 11 (*Kpn*I + *Hpa*II and *Sst*I + *Hpa*II respectively). In HEL cells, the upstream site (site A) is completely methylated producing only the parental *Kpn*I band of 7 kb (lane 12: *Kpn*I + *Hpa*II), while the 3′ site is completely unmethylated producing a fragment size in the *Sst*I + *Hpa*II digest (lane 13), the same as that of the *Sst*I + *Msp*I digest (lane 9). Analysis of similar Southern blots using a probe lying within the first large intron (intron 5′ 1A) at + 10 kb again shows high levels of methylation in HEL cells, suggesting that the region of low methylation covers only a region round the promoter (data not shown). Unfortunately, the low G + C content of this region and the lack of *Hpa*II sites make the assessment of the extent of this demethylated region impossible by this method.

The specific hypomethylation of a region of DNA near the promoter of an active gene, as found here in the HEL cells, has been found in many other studies including the globins in erythroid tissue (Mavilio et al., 1983; van der Ploeg and Flavell, 1984), so the finding of this at the promoter of the *CAI* gene was not surprising. The complete lack of methylation, or low level of methylation, in cell lines which do not express *CAI* was however unexpected, being at variance with the findings generally reported for tissues and cell lines (Cedar, 1988). Interestingly, it has been found in this laboratory that the *CAI* gene can be transactivated in K562 cells by fusion with the CAI expressing-mouse erythroleukemia cell line MEL C88 (Butterworth et al., 1990). Since these results show the *CAI* gene is partially methylated in K562 cells, this would seem to suggest that a certain level of methylation may be a requirement for the activation of this gene. Alternatively the accessibility of the gene, through some change in chromatin conformation, to those factors required for transcription may at the same time allow methylation to take place. In other words one should look at methylation (at least in this instance) as having a coincident, rather than a cause/effect, relationship with gene expression. The fact that changes in gene expression precede methylation, in cell fusion experiments designed to activate the adult β-globin gene and inactivate the foetal γ-globin gene in foetal erythroblasts, supports this idea (Enver et al., 1988b). It will be interesting to see whether these methylation patterns extend beyond the regions examined here, for example, to the two other closely linked carbonic anhydrase genes, *CAII* and *CAIII*, and whether these patterns are reproduced in vivo.

REFERENCES

Barlow, J.H., Lowe, N., Edwards, Y.H. and Butterworth, P.H.W.: Human carbonic anhydrase I cDNA. Nucleic Acids Res. 15 (1987) 2386.

Bird, A.P.: CpG-rich islands and the function of DNA methylation. Nature 321 (1986) 209–213.

Bird, A.P., Taggart, M.H., Nicholls, R.D. and Higgs, D.R.: Non-methylated CpG-rich islands at the human α-globin locus: implications for evolution of the α-globin pseudogene. EMBO J. 6 (1987) 999–1004.

Boyer, S.H., Siegel, S. and Noyes, A.N.: Developmental changes in human erythrocyte carbonic anhydrase levels: coordinate expression with adult hemoglobin. Dev. Biol. 97 (1983) 250–253.

Brady, H.J.M., Sowden, J.C., Edwards, M., Lowe, N. and Butterworth, P.H.W.: Multiple GF-1 binding sites flank the erythroid specific transcription unit of the human carbonic anhydrase I gene. FEBS Lett. 257 (1989) 451–456.

Butterworth, P.H.W., Barlow, J.H., Brady, H.J.M., Edwards, M., Lowe, N. and Sowden, J.C.: The structure and regulation of the human carbonic anhydrase I gene. In Dodgson, S., Gross, G. and Tashian, R.E. (Eds.), The Carbonic Anhydrases: Cellular Physiology and Molecular Genetics. Plenum, New York, 1990, in press.

Cedar, H.: DNA methylation and gene activity. Cell 53 (1988) 3–4.

Enver, T., Zhang, J.-W., Anagnou, N.P., Stamatoyannopoulos, G. and Papayannopoulou, T.: Developmental programs of human erythroleukemia cells: globin gene expression and methylation. Mol. Cell. Biol. 8 (1988a) 4917–4926.

Enver, T., Zhang, J.-W., Papayannopoulou, T. and Stamatoyannopoulos, G.: DNA methylation: a secondary event in globin gene switching? Genes Dev. 2 (1988b) 698–706.

Fernley, R.T.: Non cytoplasmic carbonic anhydrases. Trends in Biochem. Sci. 13 (1988) 356–359.

Fraser, P. and Curtis, P.: Molecular evolution of the carbonic anhydrase

genes: calculation of divergence time for mouse carbonic anhydrase I and II. J. Mol. Evol. 23 (1986) 294–299.

Fraser, P., Cummings, P. and Curtis, P.: The mouse carbonic anhydrase I gene contains two tissue-specific promoters. Mol. Cell. Biol. 9 (1989) 3308–3313.

LeFranc, M.-P., Foster, A., Baer, R., Stinson, M.A. and Rabbitts, T.H.: Diversity and rearrangement of the human T cell rearranging γ genes: Nine germ-line variable genes belonging to two subgroups. Cell 45 (1986) 237–246.

Lloyd, J., Brownson, C., Tweedie, S., Charlton, J. and Edwards, Y.H.: Human muscle carbonic anhydrase gene: gene structure and DNA methylation patterns in fetal and adult tissue. Genes Dev. 1 (1987) 594–602.

Maniatis, T., Fritsch, E.F. and Sambrook, J.: Molecular Cloning. A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1982.

Mavilio, F., Giampaolo, A., Care, A., Migliaccio, G., Calandrini, M., Russo, G., Pagliardi, G.L., Mastroberardino, G., Marinucci, M. and Peschle, C.: Molecular mechanisms of human hemoglobin switching: selective undermethylation and expression of globin genes in embryonic, fetal, and adult erythroblasts. Proc. Natl. Acad. Sci. USA 80 (1983) 6907–6911.

Padgett, R.A., Grabowski, P.J., Konarska, M.M., Seiler, S. and Sharp, P.A.: Splicing of messenger RNA precursors. Annu. Rev. Biochem. 55 (1986) 1119–1150.

Rackwitz, H.R., Zehtner, G., Frischauf, A.-M. and Lehrach, H.: Rapid restriction mapping of DNA cloned in lambda phage vectors. Gene 30 (1984) 195–200.

Sanger, F., Nicklen, S. and Coulson, A.R.: DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74 (1977) 5463–5467.

Smith, C.L., Klco, S.R. and Cantor, C.R.: Pulsed-field gel electrophoresis and the technology of large DNA molecules. In Davies, K.E. (Ed.), Genome Analysis, A Practical Approach. IRL Press, Oxford, 1988, pp. 41–72.

Tashian, R.E.: The carbonic anhydrases: widening perspectives on their evolution, expression and function. BioEssays 10 (1989) 186–192.

Tsai, S-F., Martin, D.I.K., Zon, L.I., D'Andrea, A.D., Wong, G.G. and Orkin, S.H.: Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. Nature 339 (1989) 446–451.

Van der Ploeg, L.H.T. and Flavell,: DNA methylation in the human β-globin locus in erythroid and nonerythroid tissues. Cell 19 (1984) 947–958.

Venta, P.J., Montgomery, J.C., Wiebauer, K., Hewett-Emmett, D. and Tashian, R.E.: Organisation of the mouse and human carbonic anhydrase II genes. Ann. N.Y. Acad. Sci. 429 (1984) 309–323.

Venta, P.J., Montgomery, J.C., Hewett-Emmett, D., Wiebauer, K. and Tashian, R.E.: Structure and exon to protein domain relationships of the mouse carbonic anhydrase II gene. J. Biol. Chem. 260 (1985) 12130–12135.

Villeval, J.L., Testa, U., Vinci, G., Tonthat, H., Bettaieb, A., Titeux, M., Cramer, P., Edelman, L., Rochant, H., Breton-Gorius, J. and Vainchenker, W.: Carbonic anhydrase I is an early specific marker of normal human erythroid differentiation. Blood 66 (1985) 1162–1170.

Wall, L., deBoer, E. and Grosveld, F.: The human β-globin gene 3′ enhancer contains multiple binding sites for an erythroid-specific protein. Gene Dev. 2 (1988) 1089–1100.

Wood, W.I., Gitshier, J., Lasky, L.A. and Lawn, R.M.: Base composition-independent hybridisation in tetramethylammonium chloride: a method for oligonucleotide screening of highly complex gene libraries. Proc. Natl. Acad. Sci. USA 82 (1985) 1585–1588.

Yoshihara, C.M., Lee, J.-D. and Dodgson, J.B.: The chicken carbonic anhydrase II gene: evidence for a recent shift in intron position. Nucleic Acids Res. 15 (1987) 753–770.

Physical Mapping of the Human Carbonic Anhydrase Gene Cluster on Chromosome 8

NICK LOWE,*,[1] YVONNE H. EDWARDS,† MINA EDWARDS,* AND PETER H. W. BUTTERWORTH*

*Department of Biochemistry, University College London, Gower Street, London WC1E 6BT, United Kingdom; and †MRC Human Biochemical Genetics Unit, The Galton Laboratory, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, United Kingdom

Received September 11, 1990; revised January 23, 1991

A cluster of genes encoding the three cytoplasmic carbonic anhydrase isozymes CAI, CAII, and CAIII lie on the long arm of chromosome 8 (8q22) in humans. These genes have been mapped using pulsed-field gel electrophoresis. The genes lie in the order *CA2*, *CA3*, *CA1*. *CA2* and *CA3* are separated by 20 kb and are transcribed in the same direction, away from *CA1*. *CA1* is separated from *CA3* by over 80 kb and is transcribed in the direction opposite to *CA2* and *CA3*. The arrangement of the genes is consistent with proposals that the duplication event which gave rise to *CA1* predated the duplication which gave rise to *CA2* and *CA3*. The order of these three genes differs from that suggested for the mouse based on recombination frequency.

INTRODUCTION

The carbonic anhydrases (CA) (EC 4.2.1.1) are a family of enzymes characterized by their ability to catalyze the reversible hydration of carbon dioxide to carbonic acid. The various mammalian isozymes characterized to date include cytoplasmic (CAI, CAII and CAIII), membrane-bound (CAIV, CAVII), mitochondrial (CAV), and secreted (CAVI) forms that carry out diverse roles in pH balance, ion exchange, and CO_2 metabolism.

The three well-characterized cytoplasmic isozymes CAI, CAII, and CAIII with which this paper is concerned each exhibit specific patterns of tissue distribution. CAI is found at high levels in red blood cells and at lower levels in epithelial tissue, notably intestinal epithelia. CAII is a high activity form and is found in a wide variety of tissue types and CAIII is confined mainly to skeletal muscle (and in the rat male liver) (see Tashian, 1989; Fernley, 1988, for recent reviews of the carbonic anhydrases).

[1] To whom all correspondence should be addressed.

The genes encoding each of these isozymes have been isolated and shown by direct mapping using molecular probes and by classical genetic studies to be clustered on chromosome 8 (8q22) in humans (Edwards *et al.*, 1986a,b; Davis *et al.*, 1988; Nakai *et al.*, 1987) and chromosome 3 in mice (Eicher *et al.*, 1976; Beechey *et al.*, 1990) and to lie within about 200 kb of each other in humans (Kearney *et al.*, 1987; Venta *et al.*, 1987). Other members of the carbonic anhydrase gene family have been assigned to other locations, *CA6* being found on chromosome 1 and *CA7* on chromosome 16 in humans.

This paper presents the structure of the *CA1*, *CA2*, and *CA3* gene cluster as determined by pulsed-field gel electrophoresis (PFGE).

MATERIALS AND METHODS

Sources of Probes Used

Two recombinant clones containing the human *CA1* gene (Lowe *et al.*, 1990) were the source of the probes shown in Fig. 1. The 5′ 1.4-kb *Ava*II fragment was isolated from a plasmid—pBKS204HS4.2—containing the promoter region of *CA1*, a subclone of the λ recombinant HGCAI204. The 1-kb *Eco*RI–*Xba*I fragment was isolated from plasmid pBKS104XbaD containing the untranslated exon 1b, a subclone of HGCAI104. The *CA2* probes (2.3-kb *Eco*RI–*Cla*I and 1.5-kb *Eco*RI–*Cla*I) were prepared by *Cla*I/*Eco*RI digestion of plasmid H25-3.8 containing a 3.8-kb *Eco*RI fragment containing exons 1 and 2 of *CA2* together with 1.4 kb of upstream sequence (kindly provided by Richard Tashian, University of Michigan and originally subcloned from the λ recombinant H25 (Venta *et al.*, 1984)). The *CA3* probe used was a 2.8-kb *Eco*RI–*Hin*dIII fragment isolated from a plasmid containing the promoter region of the gene, a subclone of the λ recombinant CA2.1 (Lloyd *et al.*, 1987).

0888-7543/91 $3.00

Cell Lines

The human erythroleukemic K562-SA1 cell line (Spandidos *et al.*, 1984) was used for all the PFGE work described in this paper. Some experiments have been repeated with other cell lines (see, for example, Fig. 2C), including CEM (T-lymphoblastoid), H/9 (T-lymphoblastoid), HeLa (fibroblast), and HEL (erythroleukemic), and the results were consistent with those found in K562. Differences in banding patterns between the various cell lines (data not shown) were all explicable as changes in the methylation state of the recognition sites for the enzymes used in mapping.

DNA Preparation and Digestion

DNA for conventional agarose gel electrophoresis was prepared and digested using standard methods (Maniatis *et al.*, 1982). For pulsed-field gel electrophoresis, the cells used to prepare the DNA were spun down and resuspended in phosphate-buffered saline at a concentration of 5×10^6–1×10^7. An equal volume of molten 1% agarose (Ultrapure, Bio-Rad) was added, and the suspension was pipetted into 80-μl block formers (Pharmacia, LKB). Solidified blocks were digested in >10 vol of 0.5% SDS, 100 m*M* EDTA, 100 μg/ml proteinase K at 55°C for at least 6 h. Following proteinase digestion cells were washed (20 min 55°C) twice in >10 vol TE (10 m*M* Tris–HCl, pH 8.0, 1 m*M* EDTA), twice in TE plus 2.5 m*M* phenylmethylsulfonyl fluoride, and twice more in TE. Half an 80-μl block was used per restriction enzyme digestion. After preequilibration in 400 μl of the appropriate reaction buffer, 40 units of enzyme was added to the block in a reaction volume of 150 μl and incubated for at least 4 hr. Following digestion blocks were equilibrated with stop buffer/dye-mix (0.5× TBE, 50 m*M* EDTA 0.1% bromophenol blue, 0.1% xylene cyanol) and inserted into wells.

Electrophoretic Conditions and Southern Blotting

A field inversion system (Carle *et al.*, 1986) was used for PFGE gels with switching intervals linearly ramped from 3 s forward 1 s reversed at the start of the run to 60 s forward 20 s reverse after 24 h. Gels were 1.1% to 1.4% agarose in 0.5× TBE (0.05 *M* Tris–HCl, 0.05 *M* boric acid, 1 m*M* EDTA) and were run at 7.5 V/cm. If separation of smaller fragments was required, the switching pattern was restarted after 16 h and allowed to run for a further 4 h. Following electrophoresis, DNA was alkali blotted onto a Hybond N+ hybridization membrane (Amersham International) according to the manufacturer's instructions. Hybridization was in 4× SSC, 10× Denhardt's solution, 0.1% SDS, 10 m*M* sodium phosphate, pH 6.8, 6% polyethylene glycol (mol wt 6000), and 100 μg/ml denatured salmon sperm DNA at 65°C. Fifty to two-hundred nanograms of DNA probe was used per hybridization, labeled by the random oligo primer technique (Feinberg and Vogelstein, 1983) to a specific activity of >5 $\times 10^8$ cpm/μg using a Boehringer-Mannheim kit. After overnight hybridization, filters were washed for 20 min in 2× SSC, 0.1% SDS once at room temperature and once at 65°C and once in 0.1× SSC, 0.1% SDS for 20 min at 65°C.

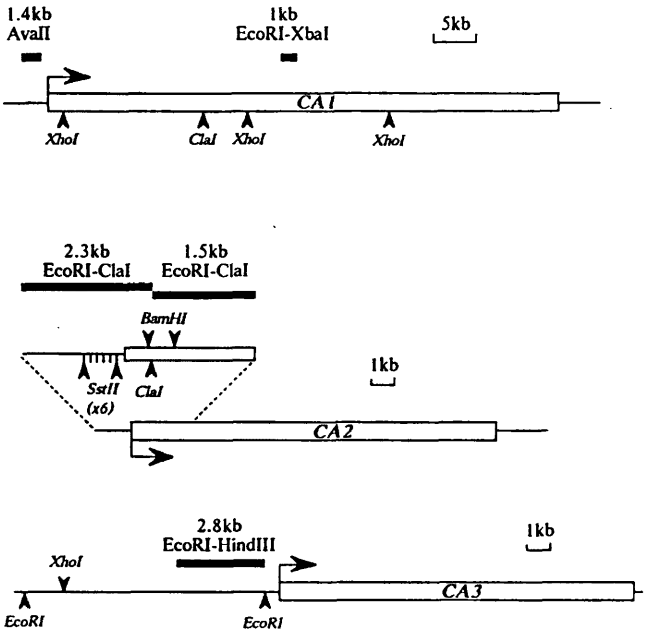


FIG. 1. The *CA1*, *CA2*, *CA3* genes together with the relative positions of the nucleic acid probes (denoted by bars above the genes) used in the mapping experiments shown in Fig. 2–4. The extent of the transcribed regions (position of exons not shown) is shown by open boxes with arrows indicating the direction of transcription. All the restriction enzyme sites shown have been mapped in recombinants, and only those sites relevant for the mapping work presented here have been shown.

RESULTS

We have carried out pulsed-field gel electrophoresis using a field inversion system (FIGE) with probes from three human carbonic anhydrase genes, *CA1*, *CA2*, and *CA3* (see Fig. 1 for the position of probes relative to each gene), principally using the human erythroleukemic cell line K562.

Determining the Order of the Genes: CA3 Lies between CA1 and CA2

Figure 2 shows the results of sequential Southern blot hybridizations to *Sal*I-, *Xho*I-, and *Cla*I-digested

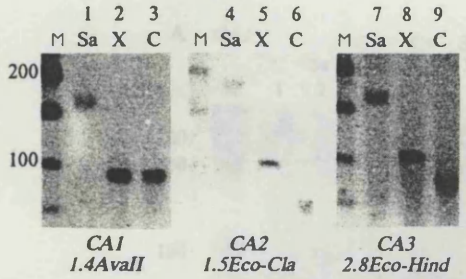


FIG. 2. The order of the CA genes. Southern transfer of a pulsed-field gel sequentially hybridized with probes for *CA1, CA2,* and *CA3* (see Fig. 1 for probe details). DNA was digested with *Cla*I (C), *Sal*I (Sa), or *Xho*I (X). The λ concatamer markers (M) are sized in kb. *CA3* shares a common *Sal*I fragment with *CA1* (lanes 1 and 7) and a common *Xho*I fragment with *CA2* (lanes 5 and 8).

DNA with genomic probes specific for *CA1, CA2,* and *CA3.* The *CA1* 1.4-kb *Ava*II and *CA3* 2.8-kb *Eco*RI–*Hin*dIII probes detect a common *Sal*I fragment (170 kb) not detected by the *CA2* 1.5-kb *Eco*RI–*Cla*I probe, while the *CA2* and *CA3* probes detect an *Xho*I fragment (110 kb) that is not detected by the *CA1* probe. Taken together these findings indicate that *CA3* lies between *CA1* and *CA2,* separated from *CA2* by one or more *Sal*I sites and from *CA1* by one or more *Xho*I sites. In addition each gene probe detects a different *Cla*I fragment.

The Relative Orientation of the Genes: CA1 Is Transcribed Away From and In the Opposite Direction to CA2 and CA3

Direct mapping of recombinant clones shows that a single *Cla*I site exists in the *CA1* gene (see Fig. 1). A probe (1.4-kb *Ava*II) for the 5′ end of the gene detects a *Cla*I fragment of about 80 kb (Fig. 2, lane 3), while a probe (1-kb *Eco*RI–*Xba*I) that lies 3′ to this site detects a band of over 200 kb in size (Fig. 3A, lane 1). Since neither of these fragments are detected by the *CA2* or *CA3* probes, and the *Cla*I fragment extending 3′ to the *Cla*I site in *CA1* is larger than the maximum distance apart of these genes, *CA2* and *CA3* must be located upstream (5′) of *CA1* and separated from it by 80 kb or more.

Orientation of the *CA2* gene relative to *CA1* and *CA3* was made possible by the identification of a *Sal*I site between *CA2* and *CA3* (see above), and a *Cla*I site at the 5′ end of the *CA2* gene, within the region cloned and used for probe preparation (see Fig. 1). The *CA2* probe lying 5′ to this *Cla*I site (2.3-kb *Eco*RI–*Cla*I) detects a *Sal*I site 10-kb upstream of the promoter (Fig. 3B) in double digests using *Sal*I together with *Cla*I (or *Bam*HI or *Sst*II, sites for which lie close to the *Cla*I site in the recombinant). This probe also detects the same *Cla*I–*Xho*I fragment as the *CA3* probe (see Fig. 4). The probe (1.5-kb *Eco*RI–*Cla*I) lying 3′ of this *Cla*I site on the other hand fails to detect a *Sal*I site downstream of the gene and hybridizes to a *Cla*I–*Xho*I fragment not detected by the *CA3* probe. These data confirm the observation of Venta *et al.* (1987) that *CA1* (and from this study *CA3*) lies 5′ to *CA2.*

The relative orientation of *CA3* was determined using an *Xho*I site lying 5 kb upstream of the transcription start site that had been mapped in recombinant clones. Since the ability of the so-called rare-cutter restriction enzymes used in this work to digest DNA is affected by DNA methylation state, digests were carried out to show that this site was in fact being cut in the cell line used in this work (K562). This is shown in Fig. 3C in which a number of different cell lines have been digested with *Eco*RI and *Xho*I, *Xho*I reducing the *Eco*RI band from 7 to 5.2 kb. Having established that this site is indeed susceptible to digestion and that *CA2* and *CA3* share a common *Xho*I fragment, it becomes apparent that *CA2* lies downstream (3′) of *CA3.*

Distance Separating the Genes

The distance between the *CA1* and the *CA2* genes can be determined from the size and termini positions of the *Sal*I fragment on which both *CA1* and *CA3* lie. One end of this fragment has been shown to lie between *CA2* and *CA3* 10 kb 5′ to *CA2* (see above). The other end of this fragment, lying approximately 70 kb 3′ to the *Cla*I site in the *CA1* gene, can be detected with a *Cla*I/*Sal*I double digest using probe *CA1* 1-kb *Eco*RI–*Xba*I (Fig. 3A, lane 3). These data provide an estimate of about 110 kb as the distance between the two genes. A similar figure is arrived at from calculations based on the size of a 200-kb *Sst*II fragment that also contains both *CA1* and *CA3* sequences (Fig. 3A, lane 6, and data not shown). This fragment is roughly colinear with the *Sal*I fragment, one end lying slightly further 3′ to *CA1* than the *Sal*I site (compare lanes 3 and 4, Fig. 3A), while the other end lies near the promoter of *CA2* (within the region cloned in H25-3.8).

The distance between *CA2* and *CA3* can be determined by making use of the *Cla*I site in *CA2* and the *Xho*I site in *CA3* to fix the position of these genes using a *Cla*I/*Xho*I double digest. While carrying out this work it became apparent that there were certain anomalies in the sizing of these fragments. For example, the *Xho*I fragment detected by *CA2* and *CA3* probes has an apparent size of about 110 kb (Fig. 2, lanes 5 and 8). This fragment can be subdivided with *Cla*I and the fragments produced hybridized with the two *CA2* probes (2.3-kb *Eco*RI–*Cla*I and 1.5-kb *Eco*RI–*Cla*I) separated by a *Cla*I site (see, for example, Fig. 3B, lanes 7 and 10). The apparent size of both

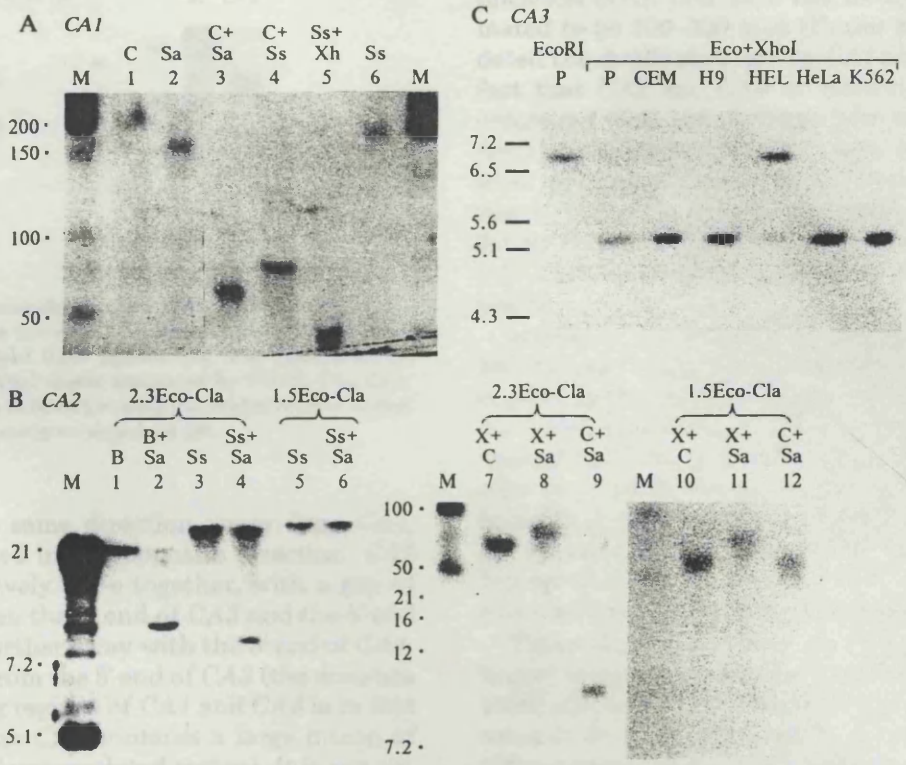


FIG. 3. Orientation of the genes. **(A)** Southern transfer of pulsed-field gel hybridized with a *CA1* probe (1-kb *Eco*RI–*Xba*I) lying 3′ to the single *Cla*I site in the gene. The 200-kb *Cla*I fragment seen in lane **1** is larger than the maximum separation between the genes and does not contain either *CA2* or *CA3* which lie on smaller fragments (see Fig. 2). Lanes **3** and **4** contain double digests of *Cla*I together with *Sal*I or *Sst*II and show the position of these sites downstream of *CA1*. **(B)** Southern transfer of standard and pulsed-field gels hybridized with *CA2* probes lying 5′ (2.3-kb *Eco*RI–*Cla*I) or 3′ (1.5-kb *Eco*RI–*Cla*I) to the single *Cla*I site in *CA2*. Lanes **1–4** and **7–9** were probed with 5′ probe, and lanes **5, 6,** and **10–12** were probed with 3′ probe. Double digests with *Sal*I together with *Bam*HI (lane **2**), *Sst*II (lane **4**), or *Cla*I (lane **9**) using the 5′ probe detect a *Sal*I site 10-kb upstream of the gene. This site lies between *CA2* and *CA3* (see Fig. 2). No site is detected using the 3′ probe (lanes **6** and **12**). **(C)** Southern transfer, hybridized with a *CA3* probe (2.8-kb *Eco*RI–*Hin*dIII), of DNA from several cell lines digested with *Eco*RI and *Xho*I. The parental *Eco*RI fragment of 7 kb (lane **1**) is reduced in size by *Xho*I, indicating that the single *Xho*I site (5 kb upstream of the gene) identified in *CA3* recombinant clones is susceptible to digestion.

these bands is greater than 60 kb which, when summed together, exceeds the size of the parental *Xho*I fragment. Similar problems arose when considering the sizes of subfragments of the *Sal*I band containing *CA1* and *CA3* sequences. This suggested that either the smaller fragments have an aberrantly low mobility or the larger fragments have an aberrantly high mobility in this electrophoretic system. Since the size of the *Sal*I fragment was in good agreement with estimates from other workers (R. Tashian, personal communication) it was suspected that the mobility of the smaller DNA fragments differed from that of the molecular size markers. To test whether this was the case, a conventional (unpulsed) gel was used to size the *Cla*I–*Xho*I fragment containing *CA3* and the 5′ end of *CA2* and showed that indeed this fragment should be sized below 50 kb (Fig. 4). In pulsed-field gels, migration of restriction enzyme fragments has previously been reported to be slower than expected in regions of high local DNA concentrations (Michiels *et al.*, 1987). This phenomenon probably accounts for the anomalous running positions of the smaller fragments seen using our field inversion system.

The map shown in Fig. 5 has been drawn taking the above considerations into account. It is assumed that there are no additional sites for either *Cla*I or *Xho*I between *CA1* and *CA3*, i.e., that the *Xho*I and *Cla*I fragments containing the 5′ end of *CA1* (Fig. 2, lanes 2 and 3) abuts the fragments containing *CA3* (Fig. 2, lanes 8 and 9). Although formally the possibility of extra sites between these genes cannot be discounted, the size constraints within which identified fragments have to be fitted and the relative rarity of these sites would seem to make this unlikely.

DISCUSSION

The three human carbonic anhydrase genes lie in the order *CA2, CA3, CA1,* with *CA2* and *CA3* being

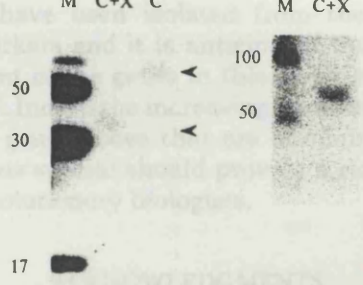


FIG. 4. Southern transfer, probed with *CA3* 2.8-kb *Eco*RI–*Hin*dIII. The gel on the left shows a *Cla*I and a *Cla*I/*Xho*I digest separated on an unpulsed 0.5% agarose gel. The right hand gel shows a similar *Cla*I/*Xho*I digest separated by FIGE. The *Cla*I/*Xho*I band that appears to be 60 kb using the field inversion system is sized at about 40 kb using an unpulsed gel.

transcribed in the same direction, away from *CA1*, which is transcribed in the opposite direction. *CA2* and *CA3* are relatively close together, with a gap of about 20 kb between the 3′ end of *CA3* and the 5′ end of *CA2*. *CA1* lies further away with the 5′ end of *CA1*, lying about 80 kb from the 5′ end of *CA3* (the distance between the coding regions of *CA1* and *CA3* is in fact over 110 kb because *CA1* contains a large intron of some 37 kb in its 5′ untranslated region). It is not yet known how these genes lie on the chromosome in terms of centromere/telomere orientation.

The three genes mapped in this study have existed as distinct forms for over 300 million years. The identification of homologous isozymes in different species indicates that the duplication events that gave rise to these genes occurred at some time between the divergence of the elasmobranchs (450 million years ago (mya)) and the divergence of the amniotes (300 mya). Comparison of the sequences of *CA1*, *CA2*, and *CA3* at both the protein and nucleic acid levels suggests that *CA2* and *CA3* are more closely related to each other than either is to *CA1* (Lloyd *et al.*, 1986, Hewett-Emmett and Tashian, 1990). This suggests that the duplication event that gave rise to *CA2* and *CA3* (estimated to be 300–320 mya (Fraser *et al.*, 1989)) postdated the duplication giving *CA1* and *CA2/CA3*. The fact that *CA2* and *CA3* lie relatively close together compared with the distance between *CA3* and *CA1* would lend support to this view, though clearly no firm conclusion can be drawn from this line of evidence. The other members of the carbonic anhydrase family diverged considerably earlier than the *CA1*, *CA2*, *CA3* divergence and lie on separate chromosomes.

Studies of other linked multigene families of similar antiquity have found many instances of gene conversion in the globins (Scott *et al.*, 1984; Slightom *et al.*, 1985), haptoglobins (Maeda and Smithies, 1986), immunoglobulins and histocompatability genes (Flanagan *et al.*, 1984; Weiss *et al.*, 1983), α-amylases (Weibauer *et al.*, 1985), and bovine vasopressin and oxytocin genes (Ruppert *et al.*, 1984). Despite their close linkage there are no signs of gene conversion taking place within the carbonic anhydrase genes.

These three genes are also known to be closely linked in the mouse (Eicher *et al.*, 1976; Beechey *et al.*, 1990) and have been assigned to band A2 on chromosome 3. In a linkage analysis using an interspecific (*Mus spretus*/*Mus mus domesticus*) backcross, no recombinants were found between *CA1* and *CA2*, whereas a recombination frequency of 2.4% was found between *CA3* and *CA1/CA2* (Beechey *et al.*, 1990). This implies that *CA3* does not lie centrally in the mouse gene cluster. If these linkage data are correct, we must suggest that the arrangement of these genes in the mouse genome differs from that in the human. Gene order between homologous genes within conserved segments in mouse and man is usually (Nadeau, 1989) but not always (Nadeau and Reiner, 1989) conserved. It is possible, for example, that in the evolutionary past a gene conversion event in a mouse or human ancestor switched positions of the CA genes. Genomic or cDNA clones for *CA1*, *CA2*,

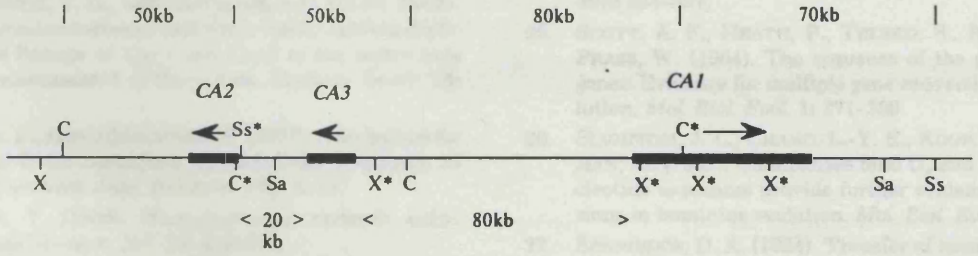


FIG. 5. Map of the human carbonic anhydrase locus located on the long arm of chromosome 8 (8q22). The three genes found at this locus—*CA1*, *CA2*, and *CA3*—are indicated by black boxes, with arrows indicating the direction of transcription. The distances between the genes are shown below the map (sizes in kb). Letters indicating the sites for various enzymes are C, *Cla*I; X, *Xho*I; Sa, *Sal*I; and Ss, *Sst*II. All the sites shown are those susceptible to digestion using K562 cell line DNA, but may be resistant to digestion in other cell types. Those sites that have been mapped in recombinant clones are marked with an asterisk. The distances between selected sites have been shown above the map.

and *CA3* have been isolated from the mouse by various workers and it is anticipated that the exact arrangement of the genes in this organism will soon be resolved. Indeed the increasing number of carbonic anhydrase gene probes that are becoming available from various species should provide a rich source of data for evolutionary biologists.

ACKNOWLEDGMENTS

Many thanks to Peter Rowe, Louise Sefton, and Phil Kearny for help and advice with the preliminary stage of this work, to Mike Larkum for building the field inversion system, and to Dr. Richard Tashian for providing *CA2* probes and unpublished results. This work was supported by a grant from the Wellcome Trust.

REFERENCES

1. BEECHEY, C., TWEEDIE, S., SPURR, N., BALL, S., PETERS, J., AND EDWARDS, Y. (1990). Mapping of mouse carbonic anhydrase-3, *Car-3:* Another locus in the homologous region of mouse chromosome 3 and human chromosome 8. *Genomics* **6:** 692–696.
2. CARTER, N. D., HEWETT-EMMETT, D., JEFFERY, S., AND TASHIAN, R. E. (1981). Testosterone-induced sulfonamide-resistant carbonic anhydrase of rat liver is indistinguishable from skeletal muscle carbonic anhydrase III. *FEBS Lett.* **128:** 114.
3. CARLE, G. F., FRANK, M., AND OLSON, M. V. (1986). Electrophoretic separation of large DNA molecules by periodic inversion of the electric field. *Science* **232:** 65–68.
4. DAVIS, M. B., WEST, L. F., BARLOW, J. H., BUTTERWORTH, P. H. W. B., LLOYD, J. C., AND EDWARDS, Y. H. (1987). Regional localisation of carbonic anhydrase genes *CA1* and *CA3* on human chromosome 8. *Somatic. Cell Mol. Genet.* **13:** 173–178.
5. EDWARDS, Y. H., BARLOW, J. H., KONIALIS, C. P., POVEY, S., AND BUTTERWORTH, P. H. W. B. (1986a). Assignment of the gene determining human carbonic anhydrase, CAI, to chromosome 8. *Ann. Hum. Genet.* **50:** 41–47.
6. EDWARDS, Y. H., LLOYD, J., PARKAR, M., AND POVEY, S. (1986b). The gene for human muscle specific carbonic anhydrase III (CAIII) is assigned to chromosome 8. *Ann. Hum. Genet.* **50:** 41–47.
7. EICHER, E. M., STERN, R. H., WOMACK, J. E., DAVISSON, M. T., RODERICK, T. H., AND REYNOLDS, S. C. (1976). Evolution of mammalian carbonic anhydrase loci by tandem duplication: Close linkage of *Car-1* and *Car-2* to the centromere region of chromosome 3 of the mouse. *Biochem. Genet.* **14:** 651–660.
8. FEINBERG, A. P., AND VOGELSTEIN, B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **132:** 6–13.
9. FERNLEY, R. T. (1988). Noncytoplasmic carbonic anhydrases. *Trends Biochem. Sci.* **13:** 356–359.
10. FLANAGAN, J. G., LEFRANC, M.-P., AND RABBITS, T. H. (1984). Mechanisms of divergence and convergence of the human immunoglobulin $\alpha 1$ and $\alpha 2$ constant region gene sequence. *Cell* **36:** 681–688.
11. FRASER, P., CUMMINGS, P., AND CURTIS, P. (1989). The mouse carbonic anhydrase I gene contains two tissue-specific promoters. *Mol. Cell. Biol.* **9:** 3308–3313.
12. HAMMER, M. F., SCHIMENTI, J., AND SILVER, L. M. (1989). Evolution of mouse chromosome 17 and the origin of inversions associated with the mouse *t* haplotypes. *Proc. Natl. Acad. Sci. USA* **86:** 3261–3265.
13. HEWETT-EMMETT, D., AND TASHIAN, R. (1990). Structure and evolutionary origins of the carbonic anhydrase multigene family. *In* "The Carbonic Anhydrases: Cellular Physiology and Molecular Genetics" (S. Dodgson, G. Gross, and R. E. Tashian, Eds.), Plenum, New York, in press.
14. KEARNEY, P., BARLOW, J., WOLFE, J., AND EDWARDS, Y. (1987). Physical linkage of CA1 and CA3: Human gene mapping 9. *Cytogenet. Cell Genet.* **46:** 637–638.
16. LLOYD, J., BROWNSON, C., TWEEDIE, S., CHARLTON, J., AND EDWARDS, Y. H. (1987). Human muscle carbonic anhydrase gene: Structural and DNA methylation patterns in fetal and adult tissue. *Gene. Dev.* **1:** 594–602.
15. LLOYD, J., MCMILLAN, S., HOPKINSON, D., AND EDWARDS, Y. H. (1986). Nucleotide sequence and derived amino acid sequence of a cDNA encoding human muscle carbonic anhydrase. *Gene* **41:** 233–239.
17. LOWE, N., BRADY, H. J. M., BARLOW, J. H., SOWDEN, J. C., EDWARDS, M., AND BUTTERWORTH, P. H. W. (1990). Structure and methylation patterns of the gene encoding human carbonic anhydrase I. *Gene,* in press.
18. MAEDA, N., AND SMITHIES, O. (1986). The evolution of multigene families: Human haptoglobin genes. *Annu. Rev. Genet.* **20:** 81–108.
19. MANIATIS, T., FRITSCH, E. F., AND SAMBROOK, J. (1982). "Molecular Cloning: A Laboratory Manual," Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
20. MICHIELS, F., BURMEISTER, M., AND LEHRACH, H. (1987). Derivation of clones close to *met* by preparative field inversion gel electrophoresis. *Science* **236:** 1305–1308.
21. NADEAU, J. H. (1989). Maps of linkage and synteny homologies between mouse and man. *Trends Genet.* **5:** 82–86.
22. NADEAU, J. H., AND REINER, A. H. (1989). Linkage and synteny homologies in mouse and man. *In* "Genetic Variants and Strains of the Laboratory Mouse" (M. F. Lyon and A. G. Searle, Eds.), 2nd ed., Oxford Univ. Press, Oxford.
23. NAKAI, H., BYERS, M. G., VENTA, P. J., TASHIAN, R. E., AND SHOWS, T. B. (1987). The gene for human carbonic anhydrase II (CA2) is located at chromosome 8q22. *Cytogenet. Cell Genet.* **44:** 234–235.
24. RUPPERT, S., SCHERER, G., AND SCHUTZ, G. (1984). Recent gene conversion involving bovine vasopressin and oxytocin precursor genes suggested by nucleotide sequence. *Nature* **308:** 554–557.
25. SCOTT, A. F., HEATH, P., TRUSCO, S., BOYER, S. H., AND PRASS, W. (1984). The sequence of the gorilla fetal globin genes: Evidence for multiple gene conversions in human evolution. *Mol. Biol. Evol.* **1:** 371–389.
26. SLIGHTOM, J. L., CHANG, L.-Y. E., KOOP, B. F., AND GOODMAN, M. (1985). Chimpanzee fetal G_γ and A_γ globin gene nucleotide sequences provide further evidence of gene conversions in hominine evolution. *Mol. Biol. Evol.* **2:** 370–389.
27. SPANDIDOS, D. A. (1984). Transfer of human globin genes to human erythroleukemia cells. *Mol. Biol. Med.* **2:** 167–175.
28. TASHIAN, R. E. (1989). The carbonic anhydrases: Widening perspectives on their evolution, expression and function. *BioEssays* **10:** 186–192.
30. VENTA, P. J., MONTGOMERY, J. C., AND TAHSIAN, R. E.

(1987). Molecular genetics of carbonic anhydrase isozymes. *In* "Isozymes: Current Topics in Biological and Medical Research" (M. C. Razatti, J. G. Scandalios, and G. S. Whitt, Eds.), pp. 58–72. A. R. Liss, New York.

29. VENTA, P. J., MONTGOMERY, J. C., WIEBAUER, K., HEWETT-EMMETT, D., AND TASHIAN, R. E. (1984). Organisation of the mouse and human carbonic anhydrase II genes. *Ann. N. Y. Acad. Sci.* **429:** 309–323.
31. WEIBAUER, K., GUMUCIO, D. L., JONES, J. M., CALDWELL, R. M., HARTLE, H. T., AND MEISLER, M. H. (1985). A 78-kilobase region of mouse chromosome 3 contains salovary and pancreatic amylase genes and a pseudogene. *Proc. Natl. Acad. Sci. USA* **82:** 5446–5449.
32. WEISS, E., GOLDEN, L., ZAKUT, R., MELLOR, A., FAHRNER, K., *et al.* (1983). The DNA sequences of the H2-K^b gene: Evidence for gene conversion as a mechanism for the generation of polymorphism in histocompatability antigens. *EMBO J.* **2:** 453–462.

Human carbonic anhydrase I cDNA

Jonathan H.Barlow, Nicholas Lowe, Yvonne H.Edwards and Peter H.W.Butterworth

Department of Biochemistry and MRC Human Biochemical Genetics Unit, University College, London WC1E 6BT, UK

Submitted February 9, 1987

```
                          1                                                  20
                    Met Ala Ser Pro Asp Trp Gly Tyr Asp Asp Lys Asn Gly Pro Glu Gln Trp Ser Lys Leu Tyr Pro Ile Ala
AAAAGAAAACTCAGTAGAAGATA ATG GCA AGT CCA GAC TGG GGA TAT GAT GAC AAA AAT GGT CCT GAA CAA TGG AGC AAG CTG TAT CCC ATT GCC

                                                  40
Asn Gly Asn Asn Gln Ser Pro Val Asp Ile Lys Thr Ser Glu Thr Lys His Asp Thr Ser Leu Lys Pro Ile Ser Val Ser Tyr Asn Pro
AAT GGA AAT AAC CAA TCC CCT GTT GAT ATT AAA ACC AGT GAA ACC AAA CAT GAC ACC TCT CTG AAA CCT ATT AGT GTC TCC TAC AAC CCA

                      60                                                  80
Ala Thr Ala Lys Glu Ile Ile Asn Val Gly His Ser Phe His Val Asn Phe Glu Asp Asn Asp Asn Arg Ser Val Leu Lys Gly Gly Pro
GCC ACA GCC AAA GAA ATT ATC AAT GTG GGG CAT TCT TTC CAT GTA AAT TTT GAG GAC AAC GAT AAC CGA TCA GTG CTG AAA GGT GGT CCT

                                                 100
Phe Ser Asp Ser Tyr Arg Leu Phe Gln Phe His Phe His Trp Gly Ser Thr Asn Glu His Gly Ser Glu His Thr Val Asp Gly Val Lys
TTC TCT GAC AGC TAC AGG CTC TTT CAG TTT CAT TTT CAC TGG GGC AGT ACA AAT GAG CAT GGT TCA GAA CAT ACA GTG GAT GGA GTC AAA

                      120                                                 140
Tyr Ser Ala Gly Leu His Val Ala His Trp Asn Ser Ala Lys Tyr Ser Ser Leu Ala Glu Ala Ala Ser Lys Ala Asp Gly Leu Ala Val
TAT TCT GCC GAG CTT CAC GTA GCT CAC TGG AAT TCT GCA AAG TAC TCC AGC CTT GCT GAA GCT GCC TCA AAG GCT GAT GGT TTG GCA GTT

                                                 160
Ile Gly Val Leu Met Lys Val Gly Glu Ala Asn Pro Lys Leu Gln Lys Val Leu Asp Ala Leu Gln Ala Ile Lys Thr Lys Gly Lys Arg
ATT GGT GTT TTG ATG AAG GTT GGT GAG GCC AAC CCA AAG CTG CAG AAA GTA CTT GAT GCC CTC CAA GCA ATT AAA ACC AAG GGC AAA CGA

                      180                                                 200
Ala Pro Phe Thr Asn Phe Asp Pro Ser Thr Leu Leu Pro Ser Ser Leu Asp Phe Trp Thr Tyr Pro Gly Ser Leu Thr His Pro Pro Leu
GCC CCA TTC ACA AAT TTT GAC CCC TCT ACT CTC CTT CCT TCA TCC CTG GAT TTC TGG ACC TAC CCT GGC TCT CTG ACT CAT CCT CCT CTT

                                                 220
Tyr Glu Ser Val Thr Trp Ile Ile Cys Lys Glu Ser Ile Ser Val Ser Ser Glu Gln Leu Ala Gln Phe Arg Ser Leu Leu Ser Asn Val
TAT GAG AGT GTA ACT TGG ATC ATC TGT AAG GAG AGC ATC AGT GTC AGC TCA GAG CAG CTG GCA CAA TTC CGC AGC CTT CTA TCA AAT GTT

                      240                                                 260
Glu Gly Asp Asn Ala Val Pro Met Gln His Asn Asn Arg Pro Thr Gln Pro Leu Lys Gly Arg Thr Val Arg Ala Ser Phe CTS
GAA GGT GAT AAC GCT GTC CCC ATG CAG CAC AAC AAC CGC CCA ACC CAA CCT CTG AAG GGC AGA ACA GTG AGA GCT TCA TTT TGATGATTCTG

AGAAGAAACTTGTCCTTCCTCAAGAACACAGCCCTGCTTCTGACATAATCCAGTTAAAATAATAATTTTTAAGAAATAAATTTATTTCAATATTAGCAAAAAAAAAAAAAA(GAATTC)
```

Human carbonic anhydrase I (CAI) is an erythrocyte-specific zinc metalloenzyme which catalyses the reversible hydration of CO_2. It is a member of a multigene family occurring on the long arm of chromosome 8 (Edwards, Y.H. et al. (1986) Ann. Hum. Genet. 50, 123-129; Davis, M.B. et al (1987) Som. Cell Mol. Genet., in press). Human CAI cDNA-containing clones were isolated from a λ gt11 expression library prepared from human reticulocyte poly-A^+ RNA. Analysis of a number of clones indicates the use of multiple polyadenylation sites: the example shown above has the shortest 3'-untranslated sequence.

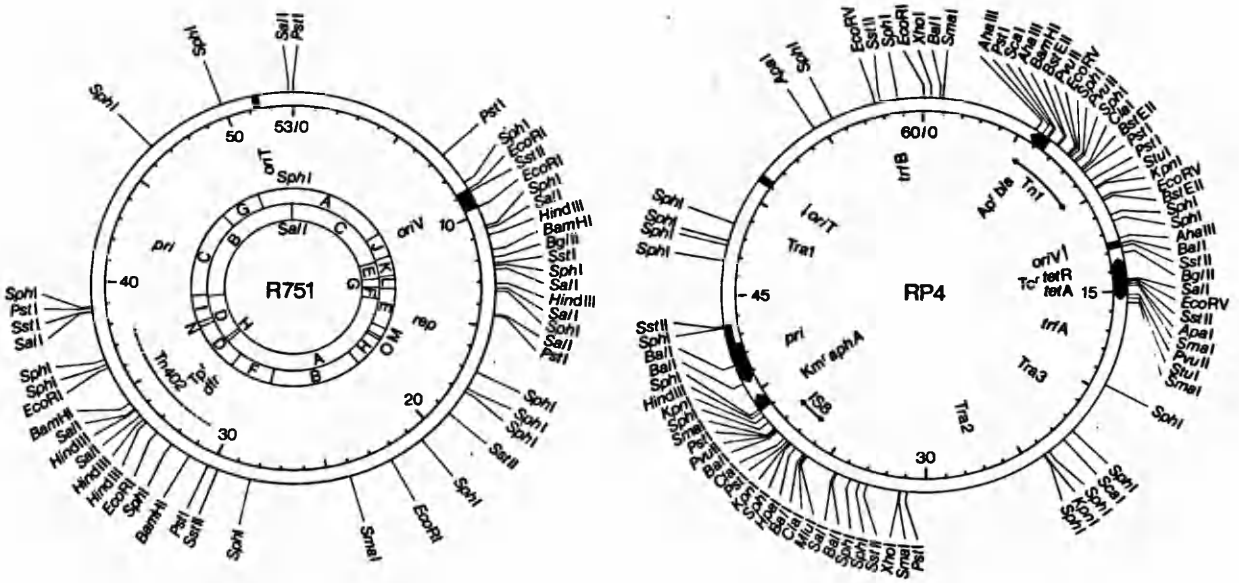
Conservation of a common 'backbone' in the genetic organization of the IncP plasmids RP4 and R751

Werner Pansegrau and Erich Lanka

Max-Planck-Institut für Molekulare Genetik, Abt. Schuster, D-1000 Berlin 33, Ihnestrasse 73, FRG

Submitted February 13, 1987

RP4 and R751 are related selftransmissible broad host range plasmids belonging to the IncP subgroups α and β respectively (1). A composite restriction map of plasmid R751 for cleavage sites of *Sph*I and *Sal*I is presented. All sites including those characterized by others (2,3) were adjusted to each other. The maps are calibrated in kilobase coordinates. Although recognition sites for restriction endonucleases appear to be clustered mainly in two regions of both plasmids, a relationship is not obvious because sites do not match. However, heteroduplex studies (4) as well as DNA hybridization experiments (3), reveal a certain degree of homology. The genetic loci for *ori*T (origin of transfer), *pri* (DNA primase), and *dfr*, Tp^r (dihydrofolate reductase) have been determined by molecular cloning of *Sph*I fragments into the vector plasmid pBR329. Including data on *ori*V (5,6) the R751 physical and genetic map is presented in a version which readily demonstrates striking similarities in the genetic organization as compared to RP4 (7). The relative order of *ori*V, *rep* (*trf*A), *pri* and *ori*T on both plasmids indicates the evolutionary conservation of the arrangement of essential replication and conjugation functions.



(1) YAKOBSON,E. and GUINEY,D.G. (1983) Mol. Gen. Genet. **192**, 430-438.
(2) MEYER,R.J. and SHAPIRO,J.A. (1980) J. Bacteriol. **143**, 1362-1373
(3) WARD,J.M. and GRINDSTED,J. (1982) Plasmid **8**, 244-252.
(4) VILLARROEL,R., HEDGES,R.W., MAENHAUT,R., LEEMANS,J., ENGLER,G., VAN MONTAGU,M. and SCHELL,J. (1983) Mol. Gen. Genet. **189**, 390-399.
(5) SMITH,C.A. and THOMAS,C.M. (1985) Nucl. Acids Res. **13**, 557-571.
(6) SMITH,C.A. and THOMAS,C.M. (1987) Mol. Gen. Genet. (in press).
(7) LANKA,E., LURZ,R. and FÜRSTE,J.P. (1983) Plasmid **10**, 303-307.

Multiple GF-1 binding sites flank the erythroid specific transcription unit of the human carbonic anhydrase I gene

Hugh J.M. Brady, Jane C. Sowden, Mina Edwards, Nicholas Lowe and Peter H.W. Butterworth

Department of Biochemistry, University College London, Gower Street, London WC1E 6BT, England

Received 14 September 1989

Six potential GF-1 sites which bind an erythroid factor are present in the 5′ and 3′ regions flanking the erythroid-specific transcription unit of the human carbonic anhydrase I (HCAI) gene. When two of these sites are placed upstream of a minimal eukaryotic promoter they confer up-regulated expression in erythroid over non-erythroid cells. The presence of the erythroid factor in TPA-treated HEL cells in which the level of HCAI transcript has greatly decreased and in non-HCAI-expressing K562 cells suggests that in these cases the presence of the factor is not sufficient for HCAI expression.

GF-1; Erythroid specific transcription factor; Human carbonic anhydrase I; *Trans*-acting protein

1. INTRODUCTION

Certain conserved DNA sequence elements to which transcription factors bind are necessary for the expression of most eukaryotic genes by RNA polymerase II; other elements, binding specific *trans*-acting protein factors have been shown to determine cell-specific expression [1]. In erythroid cells, the promoters of globin genes contain the conserved 'TATA' or 'CATA', 'CAAT' and 'CACCC' sequence cassettes and recent work has identified a sequence motif 'GATAAG' (or closely related variants thereof) which binds an erythroid-specific protein [2–5]. This sequence element is conserved across species and is found in either orientation in the regulatory regions of erythroid-specific genes. A cDNA encoding this factor has recently been cloned and designated GF-1 [6]. This paper defines the erythroid-specific transcription unit of the HCAI gene, the expression of which is characteristic of erythroid cells of the adult phenotype [7], and examines the binding of the erythroid-specific factor to sequences flanking it.

2. MATERIALS AND METHODS

2.1. *Transcription unit mapping*

Total human reticulocyte RNA was prepared by the guanidinium hydrochloride/caesium chloride method [8]. Primer extension analysis [9] of the 5′-end of HCAI mRNA used a single-stranded DNA oligonucleotide primer (3′-CACCAGGACAGACCGTCGGA-5′) complementary to a sequence in the 5′-leader region of the HCAI gene (from +33 to +52). The primer was 5′-end labelled with T4 polynucleotide kinase and [γ-^{32}P]ATP and hybridised to 25 μg of RNA followed by extension with reverse transcriptase.

S_1-nuclease mapping used single-stranded DNA probes generated from M13 templates [10]. For 5′-end mapping, a *Pvu*II-*Hin*dIII fragment containing the first HCAI exon was subcloned into M13mp18 and the complementary strand synthesised using the above primer. For 3′-end mapping, a *Hin*dIII-*Mbo*I fragment from the 3′-untranslated region (221–572 bp downstream from the stop codon) was subcloned into M13mp19 and synthesis of the complementary strand initiated using the 17mer Amersham sequencing primer. 3×10^5 cpm of 5′-end probe was hybridised to 15 μg total human reticulocyte RNA at 62 or 70°C for 3 h. 1.4×10^5 cpm of 3′-end probe was hybridised to 9 μg of RNA for 1 h at 50 or 58°C. S_1-nuclease digestion was carried out for 2 h at 20°C.

2.2. *Cell lines and tissue culture*

The erythroid cell lines used were K562 [11], K562-SA1 [12], HEL (92.1.7) [13] and mouse erythroleukemic (MEL) cells F412B2 (TK^-) [14]. K562 have an embryonic/foetal phenotype, the others an adult phenotype. All were grown in Dulbecco's MEM (DMEM) with 10% foetal calf serum (Gibco) plus penicillin (100 U/ml), streptomycin (100 μg/ml) and amphotericin B (2.5 μg/ml). Another MEL cell line C88 ($APRT^-$) [15] was also used and grown in α-MEM supplemented with 10% foetal calf serum and 50 μg/ml diaminopurine. HeLa cells were grown in DMEM as above. HL-60 (myeloid) [16] and HUT-78 (lymphoid) [17] were grown in RPMI 1640 plus 10% foetal calf serum with antibiotics and amphotericin B as above. All media were supplemented with 2 mM glutamine.

HEL cells were induced to undergo a macrophage-like shift by treatment with 12-*O*-tetradecanoyl-phorbol-13-acetate (TPA) [18]. TPA was dissolved in dimethylsulphoxide (DMSO) and used at 10^{-6} M for 4–8 days. Control cultures contained equivalent amounts of DMSO (0.01%).

2.3. *Protein preparation*

Whole cell extracts used for gel retardation were prepared by modification of the method of Dale et al. [19]: frozen cell pellets were made of total volume 0.2 ml containing $2–3 \times 10^7$ cells. 1.0 ml ice-cold extraction Buffer A (10 mM Hepes, pH 7.9, 0.4 M NaCl, 1.5 mM $MgCl_2$, 0.1 mM EGTA, 0.5 mM DTT, 0.5 mM PMSF and 5% glycerol) was added to a single pellet for lysis. The lysate was cen-

Correspondence address: H.J.M. Brady, Department of Biochemistry, University College London, Gower Street, London WC1E 6BT, England

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/89/$3.50

trifuged at 100000 × g at 4°C for 15 min. The supernatant was desalted with Buffer B (same as Buffer A but with 50 mM NaCl) on a NAP-5 column (Pharmacia) and stored at −70°C.

Nuclear proteins for footprinting were prepared from 10^8 cells which were washed twice in phosphate-buffered saline and twice in Buffer I (0.05% Nonidet P-40, 10 mM Hepes pH 7.9, 10 mM NaCl, 3 mM $MgCl_2$). The lysate was resuspended in 10 ml Buffer II (10 mM Hepes, pH 7.9, 10 mM NaCl, 3 mM $MgCl_2$) and sedimented twice through 10 ml 30% sucrose in Buffer II at 1000 × g at 4°C for 5 min. Nuclei were resuspended in 3 ml Buffer III (20 mM Hepes pH 7.9, 0.42 M NaCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 0.5 mM DTT, 0.5 mM PMSF, 25% glycerol) and the protein extract prepared as described by Wildeman [20] apart from the addition of 0.5 mM PMSF to the buffers used. Typically 5–10 $mg \cdot ml^{-1}$ protein was obtained from 10^8 cells and stored in aliquots at −70°C.

2.4. *Oligonucleotides and probe preparation*

The oligonucleotide 'αg2' derived from mouse α_1-globin was a gift from Dr M. Plumb (Beatson Institute, Glasgow). Other oligonucleotides (antisense strand shown below) were made as complementary single-stranded sequences and annealed before use:

Oligo A: 5′-GTATTTTTATTGATTATTGTGCTG-3′
Oligo B: 5′-ACCACTTCCCTTATCAGGTTCTC-3′
Oligo C: 5′-CCCACTCTAATCACCACAGGGCCA-3′
Oligo E: 5′-TGATCAAATGATTATCTTTTATAT-3′
Oligo F: 5′-CTATTTTTATCTTTAATTGACACA-3′
Oligo αg2: 5′-GATCCGGGCAACTGATAAGGATTCCC-AGATC-3′
Oligo CACCC: 5′-CTGATTAAATCCACACCCCA-3′

The oligonucleotides were 5′-end labelled as above and 5′-overhangs were filled in using excess dNTPs and Klenow fragment and purified by electrophoresis on a 10% polyacrylamide gel.

Fragment 'D', a 57 bp *Nhe*I-*Eco*RI fragment (195–251 bp downstream from the 'stop' codon) which lies between the two polyadenylation sites, was dephosphorylated using calf intestinal phosphatase and ^{32}P end-labelled as above.

2.5. *Gel retardation assay*

Gel retardation assays using whole cell extracts were carried out essentially as described by Dale et al. [19]. 10 μl of extract was preincubated with 1 μl of 5 $mg \cdot ml^{-1}$ poly(dI-dC)·poly(dI-dC) for 15 min at 20°C. Additional components were added to a final concentration of 0.5 × Buffer B, 2% Ficoll (w/v), 0.25 $mg \cdot ml^{-1}$ BSA, 10–20000 cpm end-labelled DNA and 100 ng competitor DNA where indicated in a final volume of 40 μl. The mixture was incubated for a further 15 min at 20°C. Samples were electrophoresed on a 5% nondenaturing polyacrylamide gel in 0.5 × TBE (89 mM Tris, 89 mM boric acid and 5 mM EDTA) at 150 V for 2 h.

2.6. *Footprinting analysis*

The 255 bp *Pvu*II-*Ava*II fragment (−219 to +35) was subcloned into the *Sma*I site of Bluescript plasmid (KS+, Stratagene). Both strands were labelled for DNase I footprinting of the promoter region of HCAI: the coding and noncoding strands were 5′-end labelled at the polylinker *Hin*dIII site and the *Dde*I site at +14, respectively, and fragments were purified after secondary digestion with *Hae*II (−107) for the coding strand and with *Pst*I (in the polylinker) for the noncoding strand. Markers were prepared by Maxam-Gilbert sequencing of the 5′-end labelled fragments.

Nuclear protein (50–100 μg) was preincubated with 1 μg poly(dI-dC)·poly(dI-dC) in 40 μl binding buffer (50 mM KCl, 5 mM $MgCl_2$, 1 mM EDTA, 10 mM Tris-HCl pH 8.0, 1 mM DTT, 12.5% glycerol and 0.1% Triton X-100) at 4°C for 30 min. Labelled fragments (20000 cpm) were added and incubated at 4°C for 30 min. DNase I digestion at 0.5 $\mu g \cdot ml^{-1}$, in the presence of protein and 0.01 $\mu g \cdot ml^{-1}$ in the absence of protein, was at 20°C for 2 min, followed by the addition of 0.1 vol. of 'stop' solution (1 mM EDTA, 10% SDS, 1 $mg \cdot ml^{-1}$ tRNA). DNA was purified by organic extraction and resolved on an 8% denaturing polyacrylamide gel.

2.7. *Transfection*

F412B2 MEL cells and HeLa cells were transfected using calcium phosphate/DNA precipitation as described by Rosenthal [21]. 1.5 × 10^6 F412B2 cells were plated on 100 mm Corning tissue culture dishes (Bibby) 20–22 h before transfection whereas 3 × 10^6 HeLa cells were plated on 75 cm^2 Falcon tissue culture flasks (Becton and Dickinson) at the equivalent time. The precipitate was left in contact with F412B2 cells for 24 h and with HeLa cells for 6 h, followed by glycerol shock for 2 min. Both types of cell were left for 48 h after adding the precipitate before harvesting. Cell lysates were then made by 3 cycles of freeze-thawing.

2.8. *Chloramphenicolacetyl transferase (CAT) and β-galactosidase assays*

β-Galactosidase assays were performed exactly as described by Herbomel et al. [22]. Equivalent amounts of β-galactosidase activity for each transfected plate or flask were then assayed for CAT activity exactly as described by Gorman et al. [23].

2.9. *Northern analysis*

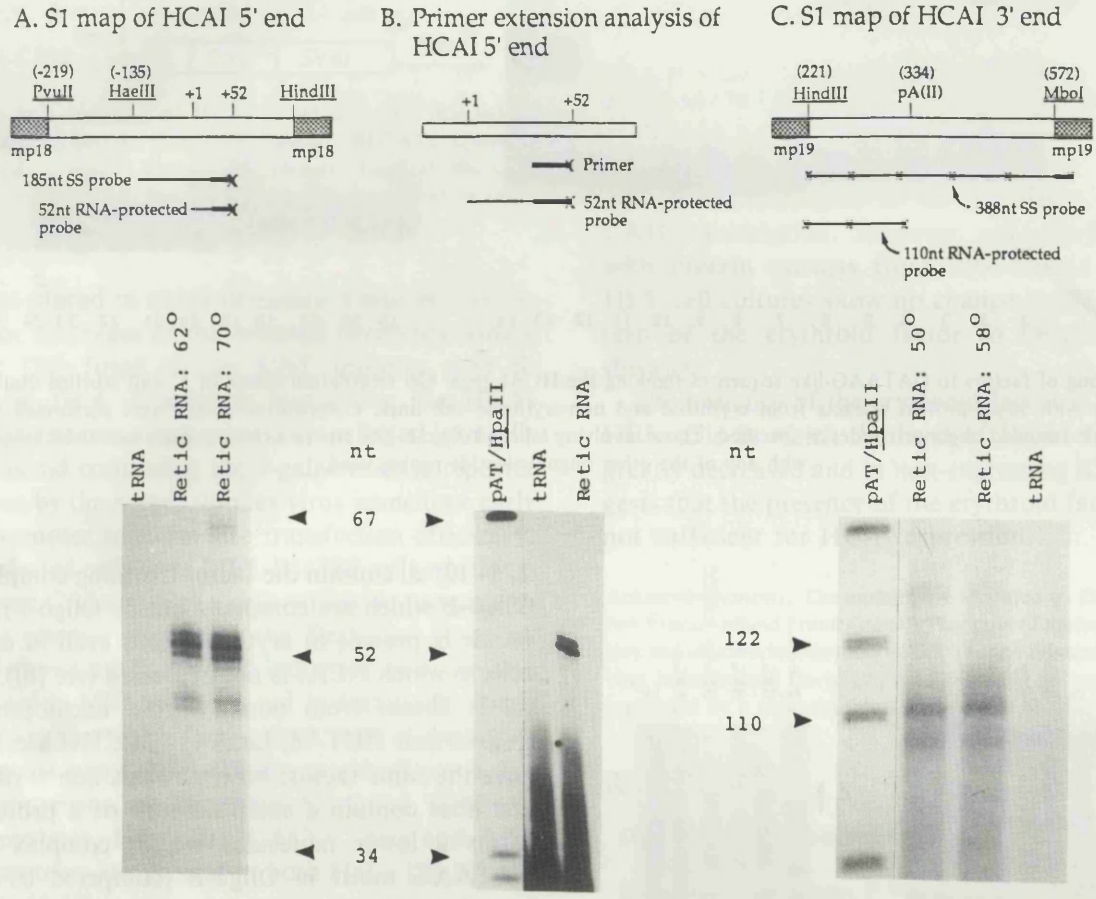
RNA was separated and transferred onto Gene Screen Plus [24]. HCAI mRNA was detected by hybridising the filter with 1 × 10^3 cpm HCAI cDNA [25] labelled with [γ-^{32}P]dCTP using random primers. Hybridisation and washing conditions were as recommended for Gene Screen Plus.

3. RESULTS AND DISCUSSION

S_1-mapping and primer extension studies have defined the position of the 5′-end of the transcription unit (fig.1A,B). S_1-mapping has also identified the polyadenylation site, pA(II) at the 3′-end of the most abundant HCAI mRNA species (fig.1C) which lies 225 bp downstream from an alternative (yet rarely used) site of 3′-end maturation, pA(I), previously described from an analysis of cDNA clones [25]. Consensus sequences for the binding of general transcription factors are apparent (fig.1D). At −28 there is a globin-like 'CATA' motif [26] and three potential 'CAAT' box sequences [26] between −60 and −90. The flanking sequences also contain consensus binding sites for characterised transcription factors: for the 'CACCC'-binding factor [27] at −209 and −47; for AP-1 [28] at −324 and 801 bp downstream from the end of the protein-coding sequence; for Sp1 [29] at −93 and Oct-1 [30] at −81. Based on previously reported consensus sequences [3–5] for the binding of an erythroid-specific transcription factor, GF-1, six potential sites are found flanking the HCAI gene: sites A, B and C at −290, −190 and −149, respectively, and sites D, E and F located 223 bp, 581 bp and 833 bp downstream from the 'stop' codon. Site D lies between the two polyadenylation sites and site E has the sequence motif in two orientations.

Gel retardation assays show that all six GATAAG-like sequences flanking the HCAI transcription unit bind the same erythroid-specific protein (fig.2). Double-stranded oligonucleotides (23- or 24-mers) containing sites A, B, C, E and F and a 57 bp fragment containing site D were used. Each gives rise to a banding pattern containing a more abundant upper band

and a much less abundant lower band when incubated with a protein extract from erythroid (MEL) cells. In each case, competition using Oligo-B or Oligo-αg2 (in which the only common sequence is a GATAAG-like motif) shows binding to be specific to the GATAAG-like motif (lanes 15–26). Protein extracts from erythroid cell lines regardless of developmental phenotype (K562, K562-SA1, MEL and HEL, lanes 1,



D. (i) 5'-region flanking the transcription start site

```
                                    AP-1                                                    GF1(A)
-360  TGAGTTCAAAATCGACATCAAGATATAAAGGAATCAGTGACTAAATATATTTCATATATGGTATTTTTATTGATTATTGTGCTGTCTTGACCTAGTATGG
            TaqI
                                                          'CACCC'                   GF1(B)
-260  AGGCCTTGGCTAGAGGCTGGTCAGTTTCCTCTCTTGAGCAGCTGATTAAATCCACACCCCAACCACTTCCCTTATCAGGTTCTCACACTCTGGGGCCACT
                                    PvuII
            GF1(C)                                                                  SpI          Oct-1
-160  ATGTACCCACTCTAATCACCACAGGGCCAGACATCAGACAATTAAGGACAGCGCCCATGCCCCAAAGCCCGCCAAAATTATGCAAATTATTCAAAATTAT
       RsaI                 HaeIII                 HaeII                          'CAAT'   'CAAT'   'CAAT'
                  'CACCC'          'TATA'                    +1
- 60  TCAACCTAGCTAACCCCACCCTTTTTGCTGTACATAAGCTGCCCATTCCCCCTCCAGCCTGTGGTACCCAGTCCTCAGGTGCAACCCCCTGCGTGGTCCT
                                                                                DdeI
```

(ii) 3' region flanking the translation stop signal

```
Arg Ala Ser Phe CTS CTS                                                                    pA(I)
AGA.GCT.TCA.TTT.TGA.TGA ---------------------------- TTTTTAAGAAATAAATTTATTTCAATATTAGCAAGACAGCATGCC -------
                    1                                77
                GF1(D)                                                  pA(II)
----- TCTGTAAGCATAAGCTTATCTTAAATTCAA ----- TATTTGAACCAATAAAATAATTTTATCTCTTTCTTTCTGTTGTGCATTCAGTTTCTA -----
      209                                  302
                GF1(E)                AP-1                                  GF1(F)
----- TGATCAAAATGATTATCTTTTATAT ------- CAATCTTTTAGTCAGATGTGCGCTTCACCTTATGCTATTTTTATCTTTAATTGACACA -------
      572                                794
```

Fig.1. The HCAI transcription unit. (A) S_1-mapping and (B) primer extension analysis defining the transcription start site. (C) S_1-mapping of the most 3'-polyadenylation site. (D) DNA sequences flanking the 5'- and 3'-end of the HCAI transcription unit showing relevant restriction endonuclease cleavage sites and consensus sequences for binding of ubiquitous and cell-type specific transcription factors and for 3'-end maturation.

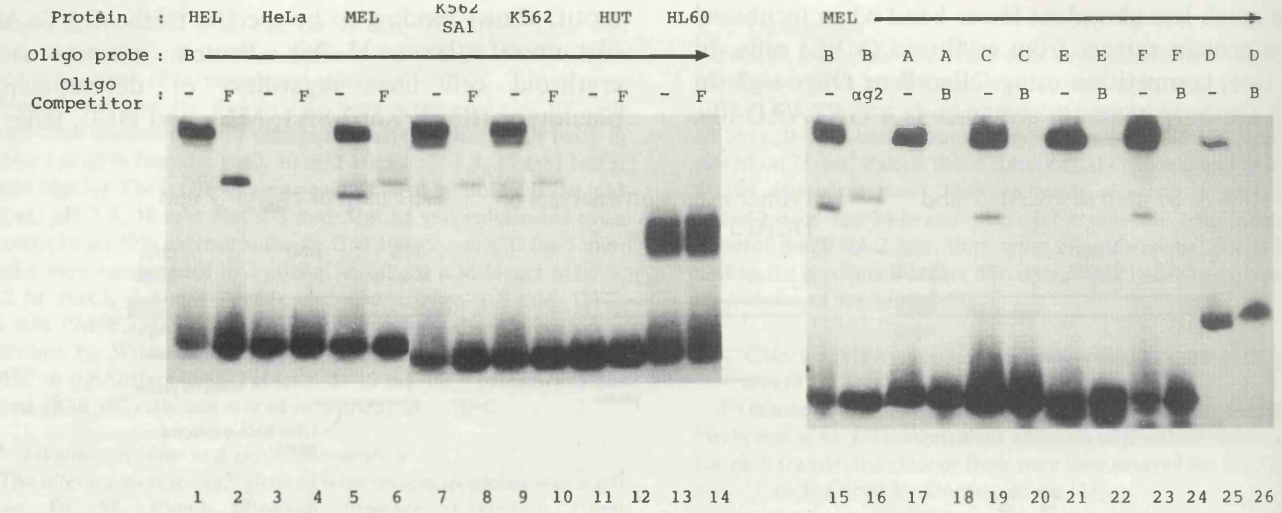


Fig.2. The binding of factors to GATAAG-like sequences flanking the HCAI gene. Gel retardation assays of 5′-end labelled double-stranded oligonucleotides with 50 μg protein extracts from erythroid and non-erythroid cell lines. Competition assays were performed with 150 ng unlabelled double-stranded oligonucleotides as specified. Those involving labelled oligo-B give rise to a characteristic band which does not occur with any of the other oligonucleotide probes used.

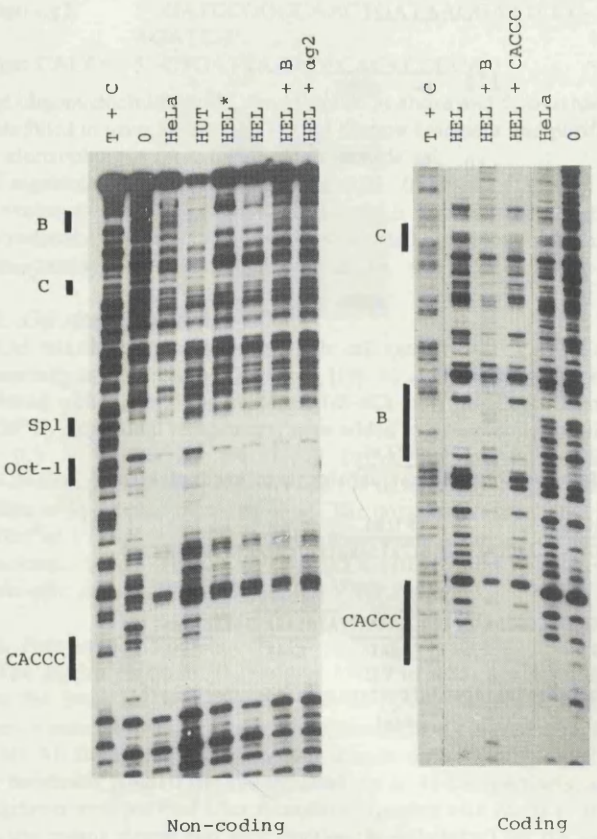


Fig.3. Footprint analysis of the −219 to +14 region of the HCAI gene. T+C indicates Maxam-Gilbert sequencing reactions; '0' indicates DNase I digestion without protein. The boxes to the left of the panels denote the footprint around each consensus: CACCC (−209 and −47), GF-1 binding sites B and C (−190 and −149, respectively), Sp1 (−93) and Oct-1 (−81). (A) Analysis of the non-coding strand from *Pvu*II (−219) to (+14) after binding with HEL* (50 μg) and HEL (100 μg), HeLa (100 μg) or HUT-78 (100 μg) nuclear extracts. Competition of the footprint over GF-I (site B) in HEL extracts was by the addition of 200 ng double-stranded oligonucleotide B or αg2. (B) Analysis of the coding strand from *Pvu*II (−219) to *Hae*II (−107) after binding with protein extracts from HEL (100 μg) and HeLa (100 μg); competitor for footprint at site B was 200 ng Oligo-B.

2, 5–10) all contain the factor (forming complexes with Oligo-B which are competed out by Oligo-F); thus the factor is present in erythroid cells even in embryonic cells in which HCAI is not expressed (see [6]). The factor is absent from non-erythroid haemopoietic cells (HL-60 and HUT-78, lanes 11–14). HeLa cells do not have the same factor; however, this non-erythroid cell line does contain a small amount of a protein which forms a lower molecular weight complex with the GATAAG motif in Oligo-B (competed by Oligo-F, lanes 3 and 4). Comparing the six binding sites with the other published consensus sequences [3–5] suggests a core recognition site of 3′-Py-$^{T}_{A}$-A-T-C-$^{A}_{T}$-5′.

DNase I footprinting of the HCAI promoter region by HEL, HeLa and HUT-78 nuclear proteins (fig.3) shows protection around the GATAAG motif at Site B, exclusively with proteins from erythroid cells. The region between −193 and −179 containing Site B is footprinted by HEL cell proteins with the induction of a hypersensitive site at −180. The footprint is specifically competed out by the addition of GATAAG motif-containing double-stranded oligonucleotides B and αg2 but not the 'CACCC' oligonucleotide. No footprint is evident on either DNA strand for Site C at −150 which suggests non-equivalence in function between the multiple GATAAG-like elements. Footprints over the Sp-1, Oct-1 and 'CACCC' consensus sequences are also observed which are not erythroid specific [26,28,29].

To show in vivo effects of erythroid specific factor binding, the 5′ *Taq*I-*Rsa*I fragment (−348 to −157) of

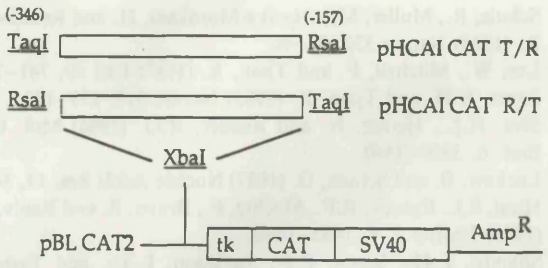


Fig.4. *Taq*I-*Rsa*I fragment of HCAI (−348 to −157) inserted (in either orientation) into the unique *Xba*I site of pBLCAT2, upstream of a minimal thymidine kinase (tk) promoter fused to the CAT reporter gene, the SV40 small t intron and polyadenylation site.

HCAI was placed in either orientation into an expression vector upstream of the minimal thymidine kinase promoter [31] fused to the CAT reporter gene illustrated in fig.4. Constructs containing the HCAI fragment, or vector alone, were cotransfected into cells with a plasmid containing the β-galactosidase reporter gene driven by the herpes simplex virus immediate early gene 4 promoter to normalise transfection efficiency. The transfected cells were MEL F412B2 cells which express mouse CAI and HeLa cells which do not express CAI. The plasmid containing the *Taq*I-*Rsa*I fragment of HCAI shows a 2.5–2.8-fold induction over the control plasmid in MEL cells but not in HeLa cells (table 1). This fragment which flanks the 5′-end of the HCAI gene contains two GATAAG-like motifs (Sites A and B), and consensus sequences for AP-1 and 'CACCC' binding proteins. However, 'CACCC'-box and AP-1 binding proteins are present in both HeLa [27,28] and MEL cells [4,32].

HEL cells constitutively express HCAI. When treated with the phorbol ester TPA, a shift takes place from erythroid to myeloid lineage as evidenced by the morphological, biochemical and functional changes they undergo [17]. Northern analysis (fig.5) shows that the treatment of HEL cells with TPA reduces the steady-state level of HCAI mRNA 7–8-fold compared with untreated HEL cells (from scanning densitometry). This is in contrast to the induction of CAII mRNA observed in TPA-treated HL60 cells [33] which

Table 1

Effect of an HCAI 5′-flanking region on minimal promoter function

	HeLa	F4
pBL CAT2	1.0	1.0
pHCAI CAT T/R	0.5	2.8
pHCAI CAT R/T	0.9	2.5

Each construct (illustrated in fig.4) was transfected separately into HeLa and MEL F412B2 (F4) cells. Normalised volumes of extracts from transfected cells (see section 2) were assayed for CAT activity and subsequently analysed by scanning densitometry. The data derived from each construct are given relative to pBL-CAT2 in each cell line

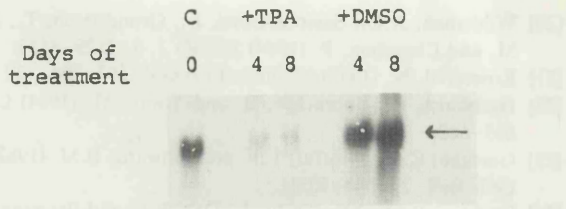


Fig.5. Northern analysis of equivalent amount of total RNA from control and TPA-treated HEL cells probed with ^{32}P-labelled HCAI cDNA.

indicates a difference in the regulation of CAI and CAII transcription. However, gel retardation assays with protein extracts from TPA-treated and control HEL cell cultures show no change in the binding pattern of the erythroid factor to Oligo-B (data not shown).

The presence of the erythroid factor in TPA-treated HEL cells in which the level of HCAI transcript has greatly decreased and in non-expressing K562 cells suggests that the presence of the erythroid factor (GF-1) is not sufficient for HCAI expression.

Acknowledgements: The authors are indebted to Drs Mark Plumb, Jon Frampton and Frank Grosveld for gifts of various cell lines, vectors and oligonucleotides and to Drs Yvonne Edwards, Ali Imam, Irving Johnston and David Linch for help and advice. This work was supported by a generous grant from the Wellcome Trust.

REFERENCES

[1] Maniatis, T., Goodbourn, S. and Fischer, J.A. (1987) Science 236, 1237–1245.
[2] Kemper, B., Jackson, P.V. and Felsenfeld, G. (1987) Mol. Cell. Biol. 7, 2059–2069.
[3] Evans, T., Reitman, M. and Felsenfeld, G. (1988) Proc. Natl. Acad. Sci. USA 85, 5976–5980.
[4] DeBoer, E., Antonio, M., Mignotte, V., Wall, L. and Grosveld, F. (1988) EMBO J. 7, 4203–4212.
[5] Plumb, M., Frampton, J., Wainwright, H., Walker, M., MacLoed, K., Goodwin, G. and Harrison, P. (1989) Nucleic Acids Res. 17, 73–92.
[6] Tsai, S.-F., Martin, D.I.K., Zon, L.I., D'Andrea, A.D., Wong, G.G. and Orkin, S.H. (1989) Nature 339, 446–451.
[7] Boyer, S.H., Siegel, S. and Noyes, A.N. (1983) Dev. Biol. 97, 250–253.
[8] Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979) Biochemistry 18, 5294–5299.
[9] Sutton, R.E. and Boothroyd, J.C. (1986) Cell 47, 527–535.
[10] Burke, J.F. (1984) Gene 30, 63–68.
[11] Lozzio, C.B. and Lozzio, B.B. (1975) Blood 45, 321–334.
[12] Spandidos, P.A. (1984) Mol. Biol. Med. 2, 167–175.
[13] Martin, P.J. and Papayannopoulou, T. (1982) Science 216, 1233–1235.
[14] Spandidos, D. and Paul, J. (1982) EMBO J. 1, 15–20.
[15] Deisseroth, A. and Hendrick, D. (1978) Cell 15, 55–63.
[16] Collins, S.J., Gallo, R.C. and Gallagher, R.E. (1977) Nature 270, 347–349.
[17] Gootenberg, S.E., Ruscetti, F.W., Mier, J.W., Gazdar, A. and Gallo, R.C. (1981) J. Exp. Med. 154, 1403–1418.
[18] Papayannopoulou, T., Nakamoto, B., Yokochi, T., Chait, A. and Kannagi, R. (1983) Blood 62, 832–845.
[19] Dale, T., Ali Imam, A.M., Kerr, I.M. and Stark, G.R. (1989) Proc. Natl. Acad. Sci. USA 86, 1203–1207.

[20] Wildeman, A.G., Sassone-Corsi, P., Grundstrom, T., Zenke, M. and Chambon, P. (1984) EMBO J. 3, 3129–3133.
[21] Rosenthal, N. (1987) Methods Enzymol. 152, 704–720.
[22] Herbomel, P., Bourachot, B. and Yaniv, M. (1984) Cell 39, 653–662.
[23] Gorman, C.M., Moffat, L.F. and Howard, B.M. (1982) Mol. Cell. Biol. 2, 1044–1051.
[24] Fourney, R.M., Miyokoshi, J., Day, R.S. and Paterson, M.C. (1988) BRL Focus 10, 5–7.
[25] Barlow, J.H., Lowe, N., Edwards, Y.E. and Butterworth, P.H.W. (1987) Nucleic Acids Res. 15, 2386.
[26] Proudfoot, N.J., Shander, M.H.M., Manley, J.L., Gefter, M.L. and Maniatis, T. (1980) Science 209, 1329–1336.
[27] Schule, R., Muller, M., Oksuka-Murakani, H. and Renkawitz, R. (1988) Nature 332, 87–90.
[28] Lee, W., Mitchell, P. and Tjian, R. (1987) Cell 49, 741–752.
[29] Jones, K.H. and Tjian, R. (1985) Nature 317, 179–182.
[30] Sive, H.L., Heintz, N. and Roeder, R.G. (1986) Mol. Cell. Biol. 6, 3329–3340.
[31] Luckow, B. and Schutz, G. (1987) Nucleic Acids Res. 15, 5490.
[32] Hirai, S.I., Ryseck, R.P., Mechta, F., Bravo, R. and Yaniv, M. (1989) EMBO J. 8, 1433–1439.
[33] Shapiro, L.H., Venta, P.J., Ya-Shion, L.Yu. and Tashian, R.E. (1989) FEBS Lett. 249, 307–310.