Characterisation of the Genomic Region Upstream of PSG-11.

A Thesis Presented in Partial Fulfilment
of the Requirements for the degree of
Master of Science in Genetics

at

Massey University, Palmerston North
New Zealand

Terence Wayne Joe

1994

TABLE OF CONTENTS

B: METHODS AND MATERIALS

D: RESULTS AND DISCUSSION FOR COSMID CLONES #1 AND #4.

ABSTRACT

The genomic clone hC3.11, isolated in 1989 in our laboratory encompasses the majority of the PSG-11 gene and contains 8.5 kb of upstream intergenic sequence. Nucleotide sequence from the initial 1.5 kb of the hC3.11 clone revealed the presence of a number of unique PSG-like C-domains upstream of the PSG-11 gene. In this investigation, the sequencing of this region was completed resulting in 3.8 kb of contiguous sequence representing the area of interest. When compared to known PSG sequences the combined hC3.11 sequence was found to be similar to other PSG genes, but also contained several unique and previously unreported C-domain-like regions.

Chromosome walking techniques were used to investigate the area upstream of the PSG-11 gene. Two cosmid clones, #1 and #4, were isolated from a human genomic DNA library as potential candidates representing the a full length gene upstream of PSG-11. These were characterised by restriction enzyme mapping, cos-mapping and hybridisation analysis. Analysis of the data of these cosmid clones indicate that one of the clones represents an allelic variant of the hC3.11 region, whereas the other clone appears to contain a genomic fragment from another PSG locus.

Hybridisation analysis of the region stretching 9 kb upstream of the C-domain region of the hC3.11 clone failed to identify other PSG-related sequence. The absence of a PSG gene associated with the C-terminal domains, suggested that the hC3.11 C-domain region may be a remnant of evolutionary activity. It is proposed that the hC3.11 C-domain cluster represents a free-standing C-domain `cassette', which may be ubiquitous amongst PSG gene family members.

ACKNOWLEDGEMENTS

I would sincerely like to acknowledge and thank the many people who have supported and assisted me during the course of this degree. Firstly, many thanks to my supervisor, Dr Brian Mansfield for his much appreciated guidance, optimism and encouragement. A special thanks to Trish McLenachan, for all the enthusiasm, advice and invaluable help during this trial. Thanks also to the staff and populace of the Micro and Genetics faculty for providing an interesting setting for this drama.

To each of the members of Mansfield park who have passed through during my sentence, including Kyle, Ruth, Kay, Lester, Felix, Mike, Neville, and Sheralee, my thanks for all the help and many welcome hours of distraction.

A special mention must be made of Merie, Paul, Shalome, Joseph, Geoff, Morgan, Sheree, Delwyn and David for the laughs and the company as we endured our thesis years.

To my family for their unquestioning provision and support of my education, in particular to the memory of Amy Joe to whom this thesis is dedicated.

Finally, but foremost I thank and credit the α & θ for the Faith and Grace that began and finished this work.

ABBREVIATIONS

bp	-	Base pairs
C-terminus	-	Carboxyl terminus
CEA	-	Carcinoembryonic antigen
DNA	-	Deoxyribonucleic acid
EDTA	-	Ethylenediaminetetraacetic acid
IPTG	-	Iso-propyl β-D thioglylactosidase
kb	-	Kilobase
N-terminus	-	Amino terminus
ON-R	-	Right cosmid vector arm
ON-L	-	Left cosmid vector arm
PSG	-	Pregnancy-Specific β1-Glycoprotein
RNA	-	Ribonucleic acid
SDS	-	Sodium Dodecyl Sulphate
SSC	-	Standard Saline Citrate
U	-	Units
UV	-	Ultra-violet
X-Gal	-	5-bromo-4-chloro-3-indolyl-β-D-galactopyranosidase

LIST OF FIGURES, TABLES, AND EQUATIONS

Page

Page

INTRODUCTION

A.1: The Placenta

One of the greatest immunological challenges occurring naturally in humans is presented by development of the feto-placental unit during pregnancy. The successful allogeneic grafting of the fetal to maternal tissue takes place, despite the existence of a complex maternal immune system, which would be expected to defeat fetal implantation. A major factor in the survival of the placental allograft is the specific lack of rejection by the mother toward the developing foetus {124}. Although the basis of this immunological tolerance is not completely understood, it is likely that the relationship between the mother and the developing foetus involves basic humoral responses.

Experiments have shown that serum immunoglobulin concentrations increase during pregnancy, whereas cell-mediated responses decrease, suggesting that humoral response plays a predominant role in this tolerance. However, the multiplicity of molecules involved, some appearing at different times and in varying amounts, complicate the identification of the specific immunoglobulin subclasses involved {85}.

The human placenta is an organ comprised of tissues of two different genotypes, providing an `interface' for exchange of gases and nutrients between the mother and the developing foetus, while still preserving the individuality of both systems {85,124}.

Not only does the placenta temporarily serve as a fetal lung, kidney, liver and intestine, but it also acts as an active exocrine and endocrine gland. An array of complex endocrine functions are initiated and completed by the placenta, completely taking over functions of the maternal ovary and pituitary gland. A wide variety of hormones, enzymes, growth factors and other molecules essential for the survival and development of the human foetus are produced by the placenta {85,119,120,121}.

A.2: Pregnancy Specific Glycoprotein

Many pregnancy-specific proteins have been reported from gel electrophoretic and immunological studies of placental extracts and maternal sera. These are summarised in Table 1.

The first in this group to be identified was the Pregnancy-specific β_1-glycoprotein (PSG), originally known as pregnancy associated plasma protein C (PAPP-C).

In 1970, Tatarinov and Masyukevich isolated PAPP-C as a new protein from the serum of pregnant women {1}. This was soon discovered to be immunologically identical to the placental protein SP-1 (Schwangershafts protein-1) isolated by Bohn in 1971 {2}.

Initial studies focused on the development of sensitive assays for the protein. It was found that during human pregnancy, the protein SP-1 was present at the highest levels of any placenta-specific protein in maternal serum (350 µg/ml) {3}.

Human PSG are primarily synthesized in the placental syncytiotrophoblast cells during pregnancy, and are subsequently secreted into the maternal serum {1,2,4,5}. The secretory nature of the PSG has been demonstrated in vitro both in primary culture of trophoblasts {6}, and by transfecting cloned PSG cDNA into cultured mammalian cells {7,8,9,10}.

The PSG protein is first detectable in human serum 7 days post conception {11}. The levels increasing exponentially during pregnancy with a doubling time of 2-3 days and a half-life of 30h, to reach term concentrations of 200-400 µg/ml maternal serum {1,12}.

As with the majority of placental products, expression of the PSG protein is not limited exclusively to the placenta {99}. Low levels of PSG protein have been detected in non-placental tissue as well as various cell lines {2,5}. Studies have demonstrated PSG gene expression in non-placental tissues {17}, human fibroblasts {35,36,37,38} and malignant tumor cells {39,40}. Isolation of PSG clones from cDNA libraries created from testis {24}, fetal liver {8,26}, salivary gland {27}, intestine {27,41} tissues, HeLa {16} and myeloid cell lines {22,30} provided further evidence showing the expression of PSG in non-placental tissue.

Biochemical studies demonstrated that the human PSG gene products are actually a heterogeneous group of proteins consisting of at least 3 distinct placental protein species with molecular weights of 72, 64, and 54 kDa {11,14,15,16}. These immunologically similar proteins were found to have carbohydrate contents ranging from 28%-32% {14}.

To study their function(s) and to develop specific reagents for the individual PSG, it was considered important to identify all the PSG genes.

Despite the previous biochemical analysis of several PSG species, the complexity of the family was not fully appreciated until the PSG genes were cloned {17,18}. Several PSG cDNA clones were independently isolated and characterised by Watanabe and Chou {14}, Streydio et al. {15} , Rooney et al. {11}, Chan et al. {16} in 1988. Additional PSG cDNA's {7,9,10,19-34} and genes {20,22,25,29,34} were subsequently reported. Examination of these genes revealed the conserved nature of the PSG family, capable of producing an array of highly related, yet unique gene products.

PROTEIN	ABBREVIATION	ANALOGUE IN NON-PREGNANT ADULT	MOL.Wt. (kDa)
Human chorionic Gonadotrophin	hCG	Luteinising hormone	45-50
Human placental Lactogen	hCS	Prolactin, growth hormone	21-23
Human chorionic Thyrotrophin	hCT	Thyroid stimulating hormone	45
Human chorionic corticotrophin	hCCT	Adenocorticotrophic steroid	5
Human chorionic gonadotrophin releasing hormone	hC-LRH	Gonadotrophin releasing hormone	1
Scwangerschafts-spezifisches β1 glycoprotein	SP-1	Unknown	90-110
Pregnancy specific β1 glycoprotein	PSβG		
Trophoblast specific β1 globulin	TBG		
Pregnancy associated plasma protein C	PAPP-C		
Pregnancy associated plasma protein A	PAPP-A	Unknown	750
Pregnancy associated plasma protein B	PAPP-B	Unknown	1000
Heat Stable Alkaline Phosphotase	HSAP	Alkaline Phosphotase	
Cysteine aminopeptidase (oxytocinase)	CAP	Aminopeptidases	
Diamine oxidase (Histaminase)	DO	Histaminase	190
Placental protein 5	PP 5	Unknown	42

Table 1: A list of some of the many protein products produced by the human placenta. Adapted from Klopper *et al.* (128).

A.3: CLINICAL APPLICATIONS

Soon after the fundamental studies of Bohn {46,47} several clinical groups explored the possibility that the PSG could be of prime importance in the evaluation of pregnancies. This interest centred on both early and late pregnancy, and also on the production of PSG by tumours.

Bohn et al. demonstrated that PSG were essential for the maintainence of human pregnancy by showing that antibodies to PSG induced abortion in primates {48}. Another study using non-pregnant monkeys actively immunised with PSG, resulted in a loss of fertility with subsequent pregnancies often ending in abortion {49}. Moreover, a correlation between low maternal serum PSG levels during pregnancy and threatened abortion, was observed emphasising the importance of PSG in the maintainence of healthy primate pregnancy {49}.

The development of sensitive assays for PSG in maternal serum {50,51,52}, allowed several pregnancy related complications to be predicted, when used in conjunction with other tests and indicators.

For example, low levels of PSG in maternal serum during pregnancy can be indicative of ectopic pregnancy {57,58}, and when used in conjunction with ultrasound and/or a human placental lactogen test, threatened abortion can be predicted with approximately 97% accuracy {59,60,61}.

Such conditions as foetal intrauterine growth retardation and intrauterine foetal death are associated with low levels of PSG and are diagnosed in conjunction with ultrasound scans.{62}. The routine method of diagnosing Meckel's syndrome is by measuring the high concentrations of PSG in the amniotic fluid associated with this condition {63}.

One of the placental gene products often used in diagnosis of pregnancy related conditions, in conjunction with PSG, is the well characterised hormone, human chorionic gonadotrophin (hCG). This hormone, hCG, first detected in the serum and urine of pregnant women by Aschheim and Zondek in 1927 {53}, was demonstrated to be produced by the syncytiotrophoblast {54}, and was subsequently established as the most reliable marker of a viable trophoblast.

The pregnancy test of choice involves the measurement of human chorionic gonadotropin (hCG). PSG levels are used as an adjunct to the hCG test, and are also used to detect Gonadotrophin induced pregnancies {55}.

Babies afflicted with Down syndrome can be predicted with 72%-78% confidence when the concentration of PSG is measured in conjunction with human chorionic gonadotrophin (hCG), and α-fetoprotein and assessed along with maternal age {56}.

Human tumour cells have been found to produce immunoreactive PSG {64}, therefore, PSG have been used as a marker in the diagnosis of certain cancers and tumours. Since high levels of PSG have been associated with choriocarcinoma, hydatidiform mole and gestational trophoblastic disease, the concentration of PSG has been used both as an indicator, and as a prognosis index in the treatment of these conditions {5}. Searle et al. (1978) {66}, suggested that serum PSG concentrations alone are not of great value in the detection and monitoring of carcinoma in the breast, large bowel and ovary, since the increase in the concentration of PSG associated with these conditions, does not correlate with the extent of the disease. Therefore, as with pregnancy related conditions, the clinical measurement of PSG is usually performed in conjunction with other tests in the diagnosis of these tumours.

Measurement of human chorionic gonadotrophin (hCG) in plasma or urine is widely used in the diagnosis and management of trophoblastic tumours. Since the concentration of hCG is related to the total cell mass of the tumour cells, the rate of cell growth and regression of the tumour cell population can be predicted {65}.

The ratio of the two placental proteins hCG and PSG forms a prognosis index for hydatidiform mole and gestational trophoblastic disease. A value less than 5 indicates a 73% chance of persistent disease, whereas a value greater than or equal to five, indicates a 74% chance of remission {67}. In the treatment of breast cancer patients, the absence of PSG suggests an improved prognosis, while presence of PSG in breast cancer patients estimates a 40%-85% chance of mortality in less than 4 years {68,69,70}.

In the management of trophoblastic tumours, the measurement of hCG concentration is the most useful measurement in these patients. Only in isolated cases in which PSG persists after hCG has become undetectable, does the measurement of PSG become valuable {100}.

The determination of PSG concentration in the serum of patients with non-trophoblastic tumours such as carcinoma of the breast, intestine or ovary, does not give practical information on the extent and progression of the disease. However, detection of PSG in carcinoma tissue itself may have prognostic significance. Since PSG is said to have immunosuppressive properties {71,72}, it

seems likely that the production of this protein by malignant tumours might be a means by which the tumour escapes immunological recognition and continues to grow.

Therefore, there are possible practical implications in investigating the involvement of the PSG in circumventing the human immune system, to allow the design of more effective treatments for tumourous conditions.

Investigation into the nature of phosphorylation in PSG proteins could provide insight into the role of PSG in pregnancy and diseases. Phosphorylation of tyrosine residues in proteins, have been shown to play a major role in the control of cell growth and differentiation {101}. Therefore, similar events in the PSG could conceivably trigger a cascade of events involved in implantation and trophoblastic invasion.

A.4: THE STRUCTURE AND EVOLUTION OF CEA AND PSG

A.4.1: Structure of CEA

The human PSG are encoded by multiple, linked genes located on chromosome 19, q13.1-13.3 overlapping the region containing the closely related CEA gene subgroup {74,75,77}. Fluorescence in situ hybridisation to metaphase chromosomes localised the PSG subgroup telomeric to the CEA subgroup. Finer mapping suggests that most of the genes are contained within 800 kb of sequence flanked by SacII restriction sites {77}. In total, the CEA/PSG gene family region is estimated to span 1.1 to 1.2 Mb {76}.

Based on sequence comparisons, the PSG have been classified as a subgroup of the CEA family, for which carcinoembryonic antigen is the prototype.

Carcinoembryonic antigen (CEA) was found to be present in colonic tumors and foetal gut tissue by Gold and Freedman in 1965 {104}. It was initially thought to be absent in normal adult intestine, however, later studies revealed the presence of CEA in several normal tissues including human colon {105-108}. Despite the lack of tumor specificity, CEA is one of the most widely used human tumor markers for assessing the recurrence of colorectal, breast and lung cancers. The serum concentration of CEA represents an important parameter in the post-operative surveillance of cancer patients {109,110}

The CEA protein is a highly glycosylated molecule with a molecular weight of 180,000 daltons. Glycosylation inhibition studies show the protein to consist of a single polypeptide chain with an apparent molecular weight of approximately 80,000 daltons {111}. Amino acid sequence deduced from the nucleotide sequence of the CEA cDNA, shows that CEA is synthesised as a precursor of 702 amino acids. The leader peptide (34 amino acids) is followed by the mature CEA peptide (668 amino acids).

Due to the presence of three internal repeats, the peptide can be divided into a number of structural domains. A schematic diagram of the CEA domain arrangement is shown in Figure 1. The three repeat domains of 178 amino acids each reveal an exceptionally high degree of sequence similarity, having between 67% and 73% of their amino acids identical. Allowing for conserved changes/substitutions the degree of similarity is even higher. Each domain contains four cysteine residues at precisely the same positions. These CEA repeat domains are relatively long, and each can be further sub-divided into two subdomains (A,B) of approximately equal size. The amino acid sequence from these subdomains display similarity to each other, as well as to the N-terminal domain {123}. The degree of conservation at the nucleotide level is also very high (80%-83% identity). Other proteins with internal repeats have been reported in the literature, but the internal degree of similarity of the repeating domains of CEA is the highest reported so far {112}.

Analyses at the genomic level for members of the CEA family indicated a precise correlation between the exons and the A and B sub-domains {116,117,122}. A domain model was subsequently proposed for CEA, which assumes that the conserved neighbouring cysteine residues present in the repeat domains form disulphide bonds, creating a looped secondary structure characteristic of the immunoglobulin family {113,116,123}. This is shown in Figure 2, demonstrating the strong similarity in secondary structure amongst members of the immunoglobulin superfamily {84}.

The C-terminal domain of CEA consists of 27 amino-acids and is strongly hydrophobic. This provides a potential insertion region for the CEA protein into the plasma membrane and is of an appropriate length to span the lipid bilayer. Therefore, this provides a possible means to anchor CEA to the cell surface membrane {124}.

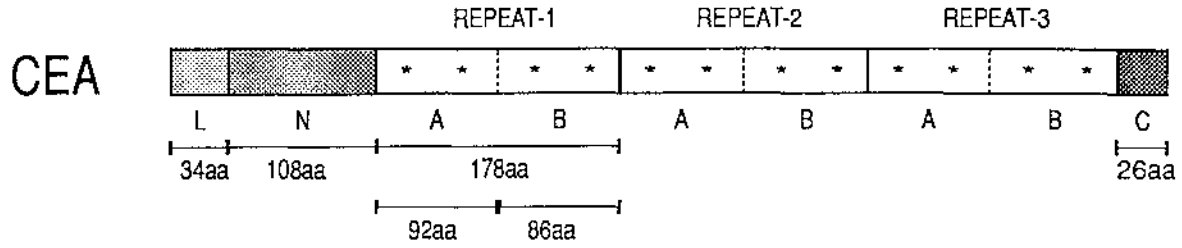


FIGURE 1: CEA THE PROTOTYPE DOMAIN MODEL.

The CEA protein domain arrangement based on deduced amino acid sequence. Domain sizes are indicated below the domain blocks. Invariant cysteine residues are marked (*). Adapted from Thompson et al. {25}.

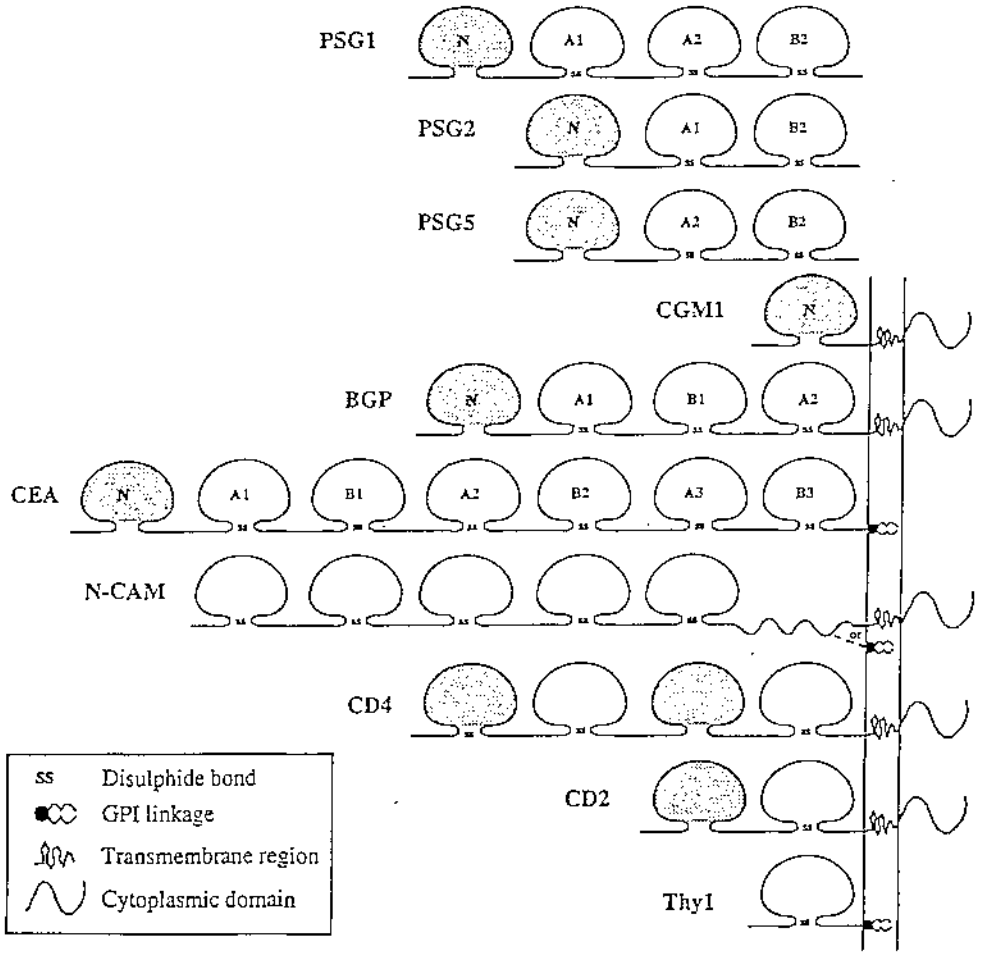


FIGURE 2: MEMBERS OF THE IMMUNOGLOBULIN SUPERFAMILY

A schematic representation of some members of the immunoglobulin family. Filled circles represent variable region (V)-like domains, other circles closed with (ss) represent constant domains. Adapted from Khan et al. {81}.

A.4.2: Structure of PSG

As with CEA, the PSG cDNA predict proteins with a high degree of sequence homology, characteristic domain structure, conserved disulphide bonds within domains, and β-sheet structure, which classify them as members of the immunoglobulin supergene family {84}.

Each of the PSG genes contain at least six exons. The first exon contains most of the putative protein leader sequence (L), the second exon encodes the remaining portion of the leader sequence, and the N-terminal domain (L-N). The subsequent exons code for the AI, BI, AII, BII, and C protein domains respectively. An example of typical PSG domain arrangement is shown in Figure 3.

The last exon encoding the BII domain, also includes the region encoding one of the putative C-terminal domains. There are usually several potential C-terminal domains present in this region, which are selected by alternate splicing for inclusion into each mature transcript. Each PSG gene contains a unique combination of these alternative C-terminal domains. It is unclear whether all of the putative C-domains are expressed or not in each PSG gene. Notwithstanding this, there have been a number of splice variants reported for many of the PSG genes.

With the exception of PSG-6 which contains 108 amino acids, the IgV-like N-terminal domain consists of 109 amino acids {8,29}. The IgC-like AI and AII repeat domains each consist of 93 amino acids, whereas the IgC-like BII domain comprises 86 amino acids.

In each of the IgC-like `A' and `B' domains, there are two conserved cysteine residues which are assumed to stabilise the IgG-like fold by forming a disulphide bridge. The distances between the cysteine residues are 47 amino acids in the A-domains, and 39 amino acids in the B-domain. These disulphide linkages are predicted to form the looped (β-sheet) secondary structure characteristic of the immunoglobulin superfamily. It has been postulated that a similar fold exists in the N-domain by interactions between hydrophobic amino acids replacing the cysteines at the conserved sites, stabilised by a salt bridge {20,73}.

The N, A, and B domains are highly conserved and display extensive sequence homology with approximately 90% identity within the PSG/CEA subfamilies, dropping to approximately 60-70% identity across subfamilies {20}.

In contrast to the highly conserved internal repeat domains, the C-termini are of variable length, typically 2-26 amino acids, and exhibit remarkable diversity for such a short region of sequence (showing only approximately 46% amino acid identity). Unlike the membrane bound CEA subfamily members, the majority of the PSG have short C-domains and are predicted to be secreted {15,21}. The only reported exception is a splice variant of the PSG-11 gene, PSG-11w, which has an 81 amino acid hydrophobic C-terminus which could potentially anchor the protein to the cell membrane {31}.

Comparisons of PSG cDNA to the known PSG genomic sequences, have shown that in certain cDNA species, some central domain sequences are not expressed. While most PSG cDNA transcripts predict a protein L-N-AI-AII-BII-C (type I), (eg: PSG-1,-3,-4,-6,-7,-11) other arrangements exist due to the presence of internal stop codons, alternate splicing, and mutations in splice acceptor sites {22,34}.
For example, in the PSG-5 gene alternate splicing excludes both a pseudo-exon BI (which contains an internal stop codon), as well as an apparently good exon, AI, resulting in the PSG-5-Im transcript with a cDNA arrangement of L-N-AII-BII-C {22}. In the PSG-8 gene (subgroup-1), the exon BI is excluded due to a mutation in the splice acceptor sequence {34}. In fact, the BI domain is never included in mature PSG transcripts. Some genes, such as PSG-5 contain stop codons, while in other cases the splice acceptor sites are absent or non-functional. The mechanisms governing splice selection are not clear.

The PSG are demonstrated to have a tissue-specific pattern of expression {9,25}. Coordinated expression of the PSG in placental {25,103}, foetal liver {9,7,26}, and submandibular salivary gland tissue {27}, suggest the presence of common regulatory elements for their transcription. This notion is supported by the tightly linked organisation of these genes {25,29}.

The preferential expression of PSG-1, PSG-2 and PSG-3 in placenta {103}, and likewise PSG-6 in hydatidiform mole {29}, suggest that additional gene-specific control elements may be present in particular tissue types to up- or down-regulate the individual levels of PSG gene expression.

Based on the pattern of gene splicing the human PSG proteins identified to date can be classified into: Type I PSG which consist of the organisation L, N, AI, AII, BII, C; or a Type II protein further divided into two subgroups, Type IIA: L, N, AI, BII, C; or Type IIB: L, N, AII, BII, C.

These are represented by the domain arrangements shown in Figure 4.

The PSG cDNA can also be grouped according to the similarity of their C-terminal domains, predicted from the deduced amino acid sequence as shown in Table 2.

The subgroup-1 PSG include the genes PSG-1, -7, and -8. A schematic representation of the subgroup-1 C-domain arrangement is shown in Figure 5. Transcripts from these genes are currently divided into four categories, C_a, C_b, C_c, and C_d, which arise from splicing of alternative exons encoding the C-domain amino acids and 3'-untranslated region (C/3' exon).

Subgroup-2 PSG include the genes PSG-2, -3, and -5. A schematic representation of the subgroup-2 C-domain arrangement is shown also in Figure 5. These appear to have a single C-domain amino acid sequence $C_{m/n}$, but their C-terminal amino acid and 3'-untranslated regions lie on at least two separate exons.

The subgroup-3 PSG, which include the PSG-6 and PSG-12 genes, have not been well characterised, but three different C-termini C_w, C_r, and C_s have been reported, although no single gene is reported to express all three. A schematic representation of the subgroup-3 C-domain arrangement is included in Figure 5.

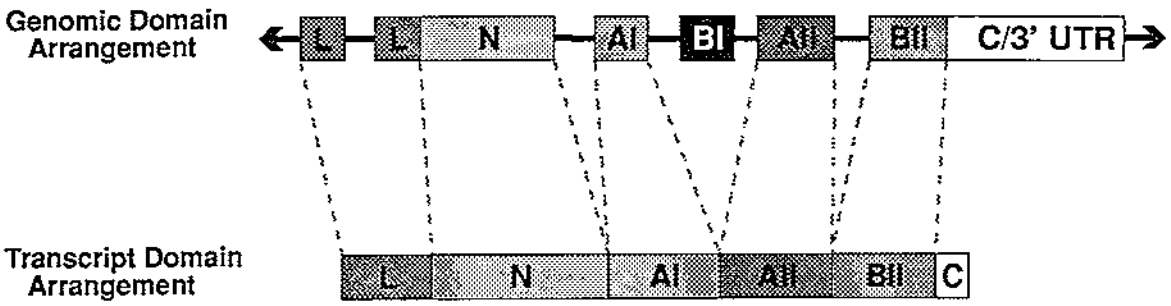


Figure 3: Domain Arrangement in Pregnancy Specific Glycoprotein
Physical map of a typical PSG gene. The upper line shows Exon-Intron organisation, while the lower line shows the most common splice form (Type I). Domain names are shown in the boxes. Not to scale. Adapted from Khan et al. {81}.

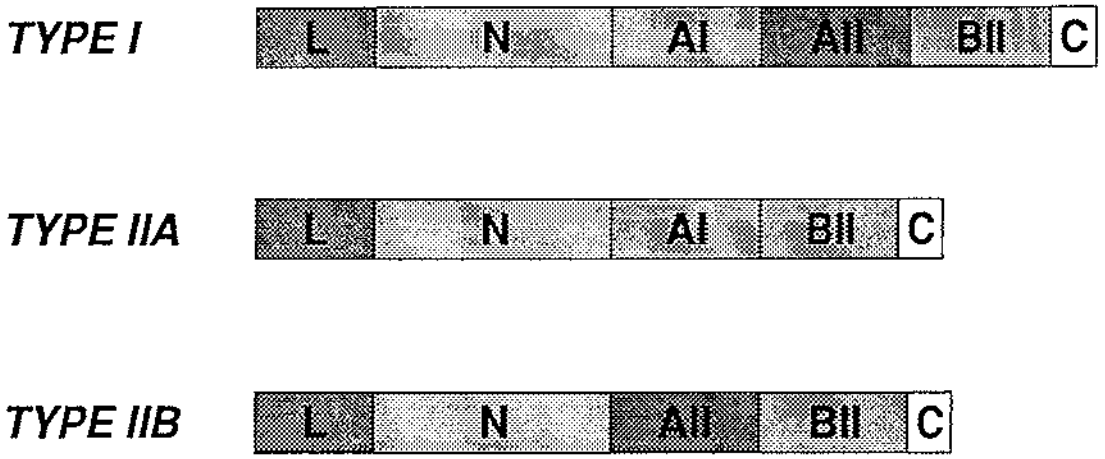


FIGURE 4: PSG Protein Domain Arrangements Predicted from Transcripts.
The structures of protein organisation produced by alternative splicing within the central domains. The PSG transcripts contain a leader (L) sequence, a varying number of IgC-like A/B domains, and a C-domain of varying length. Not to scale. Adapted from Lei et al. {127}.

Table 2: The amino acid sequences of the C-terminal domains of the PSG

SUBGROUP-1

Ca		Cb		Cc		Cd	
PSG1a	DWTVP	PSG1b	EAL	PSG1c	AYSSSINYTSGNRN	PSG1d	GKWIPASLAVGF
PSG4a	..IL.	PSG4b	---	PSG4c	--------------	PSG4d	------------
PSG7a	..SL.	PSG7b	.S.	PSG7c	...G.....D.	PSG7d	.....
PSG8a	...L.	PSG8b	...	PSG8c	.....AVY	PSG8d	..R..V.....I

SUBGROUP-2

Cm		Cn	
PSG3m	APSGTGHLPGLNPL	PSG2n	ASTRIGLLPLLNPT
PSG5m	....I.R..L...I		

SUBGROUP-3

Cr		Cs		Cw	
PSG6r	ETASPQVTYAGPNTWFQEILLL	PSG6s	GPCHGNQTESH		
PSG11r	.....S.....	PSG11s	.....DL...ES	PSG11w	GKWIPASLAVGFYVESIWLSEKSQENI
					FIPSLCPMGTSKSQILLLNPPNLSLQT
					LFSLFFCFLMADLVSGLKKVGRGLYQP

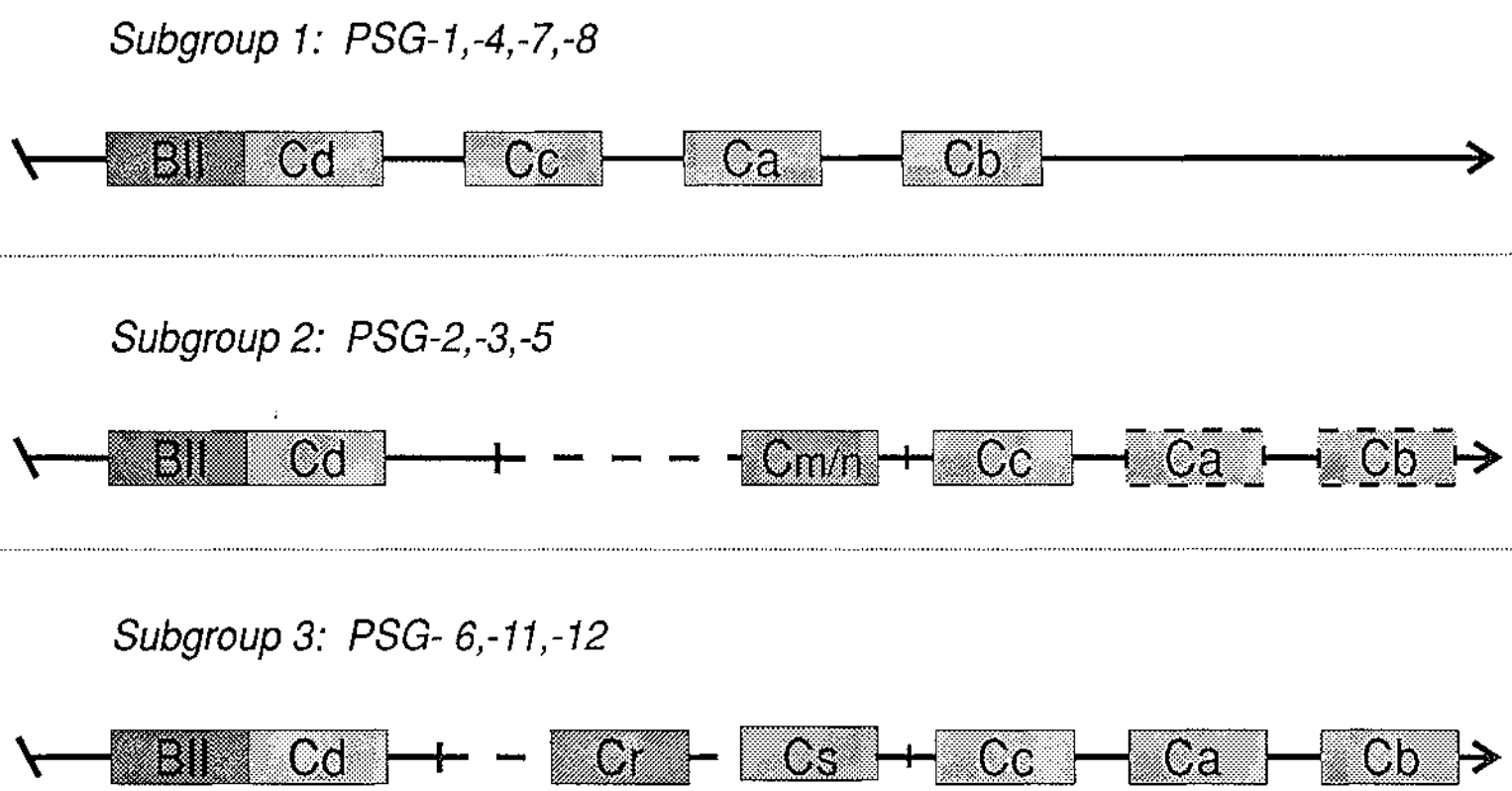


Figure 5: The Three PSG Subgroups.

The three PSG gene subgroups classified by C-domain organisation are shown above (not to scale). PSG domains are labelled respectively, dashed lines indicate areas of low homology, dashed boxes represent exons present but not yet demonstrated to be expressed. PSG C-domains are selected individually by alternative splicing for inclusion into cDNA transcripts.

A.4.3: Size of the CEA/PSG Gene Family

The size of the CEA/PSG family was estimated in 1992 by Khan and co-workers {81}, using PCR primers specific to the Ig-V like N-domain in PSG to amplify PSG clones from a human genomic library. From this study, it was found that the PSG subfamily contains at least 11 different genes: PSG-1,-2,-3,-4,-5,-6,-7,-8,-11,-12,-13.

Complete PSG genomic clones have been reported for PSG-4 and PSG-5 {77,22}, and clones containing parts of PSG-1,-6,-7,-8,-11 have also been published {34,77,78,79,80}.

However, it is possible that the PSG genes may occur in more than one copy in the genome, and that these copies may give rise to different, alternatively spliced transcripts.

The CEA subfamily currently consists of nine genes, including CEA, Non-specific Cross-reacting Antigen (NCA), Billiary Glycoprotein (BGP), and six genes referred to as CEA-Gene family Members (CGM-1,-2,-6,-7,-8, and -9) {125}. The members of this family are shown in Table 3. Since numerous CEA and PSG genes were isolated independently by various research groups, individual genes had each been assigned many different names. To standardise the nomenclature used for the PSG/CEA family, a workshop was held in Freiberg, Germany in 1989 to coordinate an internationally accepted standard classification {44}. Table 4 provides a summary of the standard nomenclature in use to date.

A.4.4: Evolution

Molecular cloning studies have revealed that human PSG consist of a large family of closely related glycoproteins that share extensive sequence identity with the carcinoembryonic antigen (CEA), the prototype of another extensive family of proteins.

Immunobiochemical studies have revealed a number of molecules closely related to CEA, such as non-specific cross reacting antigen (NCA) and billiary glycoprotein (BGP), substantiating the existence of a gene family {25}.

The existence of such a cluster of highly related genes, together with their strong sequence similarities suggests that the genes have evolved relatively recently by unequal crossing over, which has led to gene multiplication.

On the basis of sequence data, evolutionary trees have been constructed from which it has been proposed that all CEA/PSG family genes are derived from a common primordial gene unit. According to this hypothesis, duplication of this gene unit resulted in the formation of the PSG and CEA subfamilies. Recent amplification and recombination events have led to the subsequent divergence of these subfamilies. The variation resulting from these rearrangements, has led to the rapid expansion of both the CEA and PSG subfamilies {83,28}. This hypothesis is represented diagramatically in Figure 6.

This hypothesis was supported by parsimony analysis, using 24 PSG and CEA sequences, and 12 members of the immunoglobulin gene superfamily to root the family tree, showing that the CEA genes formed a monophyletic sister group to the PSG genes {81}. These results also verified a finding by Rudert et al. {83}, who suggested that the CEA subfamily and the PSG subfamily had diverged to form two major branches in the family tree.

Antigen	MW	CHO	Source
βE-Protein	?		colon tumour
Biliary glycoprotein I (BGP-I)	90K	40	bile
Biliary glycoprotein II (BGP-II)	?		infected bile
Biliary glycoprotein III (BGP-III)	?		infected bile
Breast carcinoma glycoprotein (BCGP)	?		breast tumour
Carcinoembryonic antigen 200 (CEA-200)	200K		colon tumour
Carcinoembryonic antigen 180 (CEA-180)	180K	60	colon tumour
Carcinoembryonic antigen 160 (CEA-160)	160K		colon tumour
CEA-associated protein (CEX)	?		colon/plasma
Colonic carcinoembryonic antigen 2 (CCEA-2)	?		colon/plasma
Colonic carcinoma antigen III (CCA-III)	60K		serum
Fetal sulphoglycoprotein antigen (FSA)	?		stomach
Gastric CEA-like antigen (Celia)	?		stomach juices
Meconium antigen (MA)	185K		meconium
Meconium antigen-100 (MA-100)	105K	50	meconium
Melanoma/carcinoma cross-reacting oncofetal Ag	95–150K		melanoma/carcinoma
Non-specific cross-reacting antigen 2 (NCA-2)	160K	50	meconium
Non-specific cross-reacting antigen 160 (NCA-160)	160K		granulocytes
Non-specific cross-reacting antigen 110 (NCA-110)	110K		spleen
Non-specific cross-reacting antigen 97 (NCA-97)	97K		colon tumour
Non-specific cross-reacting antigen 95 (NCA-95)	95K		granulocytes
Non-specific cross-reacting antigen 90 (NCA-90)	90K		granulocytes
Non-specific cross-reacting antigen 75 (NCA-75)	75K		colon tumour
Non-specific cross-reacting antigen 50 (NCA-50)	50K	30	granulocytes
Normal colon washings antigen (NCW)	?		colon
Normal faecal antigen 1 (NFA-1)	20–30K	13	faeces
Normal faecal antigen 2 (NFA-2)	170K		faeces
Normal faecal cross-reacting antigen (NFCA)	80–90K		faeces
Normal glycoprotein (NGP)	60K		lung
Tumour-extracted antigen (TEX)	85K	60	colon tumour
165K antigen	165K		meconium
160K antigen	160K		colon tumour
128K antigen	128K	50	colon tumour
90K antigen	90K		colon tumour
50K antigen	50K	25	colon tumour
40K antigen	40K		colon tumour

TABLE 3: THE CEA GENE SUBFAMILY

Members of the CEA gene subfamily. Adapted from Thompson et al. [25].

Old gene or clone name	New gene or mRNA name
CEA SUBFAMILY	
CEA	CEA
NCA	NCA
hsCGM6, GN-1, M6	CGM6
BGPI, TM-1 CEA	BGPa
TM-2 CEA	BGPb
TM-3 CEA	BGPc
TM-4 CEA	BGPd
hsCGM1	CGM1
hsCGM2	CGM2
PSG SUBFAMILY	
PSβG	PSG1
PSG93, PSβGD, hPSP11, FL-NCA-2, hPS3, PSG-1a, PSβG81	PSG1a
PSG16	PSG1b
PSβG-C	PSG1c
FL-NCA-1, PSG1d, SG9	PSG1d
PSBG-Ci, PSG95	PSG1e
PSβGD'	PSG1-IIa
PSG1-I	PSG1-I
PSβG-E, SG8, hSP184	PSG2n
pSP1-i, hC17, PS35, hTS-16, PSβG-A, SG5, hPS173	PSG3m
PSG4, hsCGM4, hHSP2, FL17	PSG4
hPS133, PSG9 (PSkα)	PSG4a
PSG5	PSG5
FL-NCA-3, hPS176	PSG5-In
PSβG HL (clone 22)	PSG5-Im
hsCGM3, FL26, PSGGB	PSG6
PSG6	PSG6r
hPS12, PSG10, hPS89	PSG6s
PSG7, PSGGA	PSG7
CGM35, PSG8	PSG8
hTS1	PSG8a
PS34, PSβG-G, PSβG-B, PSG7	PSG11s
hPS2, hPS91	PSG11-Iw
PSG14	PSG14
PSG15	PSG15
PSG9	PSG16a

Table 4: The Standardised Nomenclature for the CEA/PSG Gene Family.

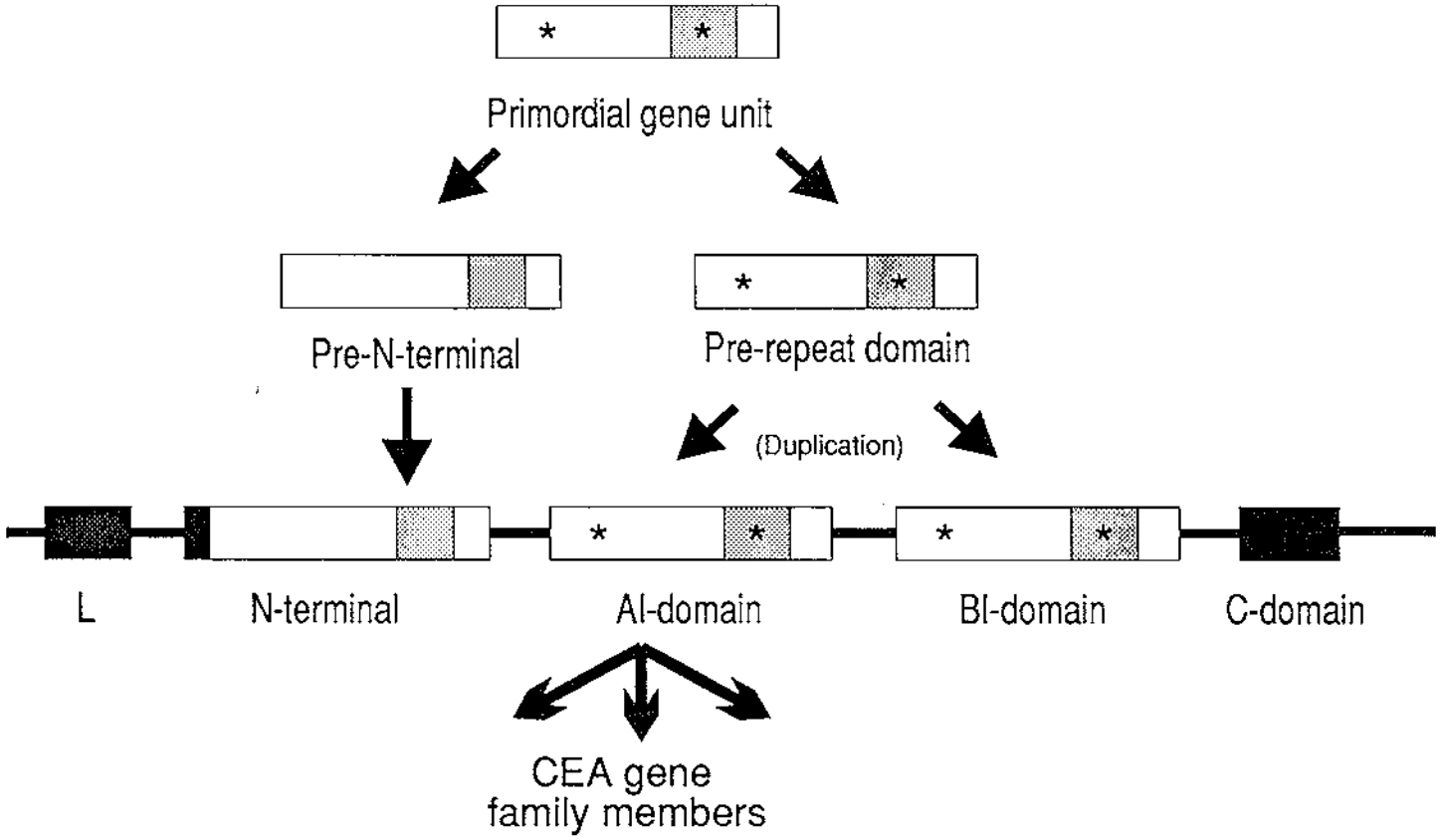


Figure 6: Evolution of the CEA/PSG Gene Family.

Hypothetical model for the evolution of the CEA gene family (not to scale).
Light grey boxes represent highly conserved areas within domains.
Invariant cysteine residues are shown as (*).
Adapted from Thompson and Zimmerman {25}.

A.5: Possible Biological Role(s) of the PSG

Both the CEA and PSG gene products are useful as clinical markers in the diagnosis of pregnancy-related disorders and malignancies, however, the exact biological role of individual members are still unknown.

Many roles have been proposed for the PSG including: growth factor activity {31}, prevention of immune rejection of the foetus {24,90}, assistance with trophoblastic cell invasion of the uterus {15}, and involvement in cellular interactions with the extracellular matrix {11,91}.

Despite the extensive protein similarity between the CEA and PSG sub-families, the two groups have evolved completely separate functions. The proteins of the CEA-subgroup appear to be membrane bound, while most of the PSG proteins appear to be secreted {42,43}. Various members of the CEA subfamily members such as CEA {94,95}, non-specific cross reacting antigens {95}, and mouse CEA-related members {97}, have been demonstrated to exhibit calcium independent cell adhesion properties, in vivo.

In contrast to this, the PSG do not seem to mediate direct cell-cell interactions {8}. Since PSG subfamily gene products are predominantly secreted proteins, they could only mediate cell adhesion through cellular receptors, similar to the fibronectin-integrin family {92}.

The tripeptide motif arginine-glycine-aspartic acid (RGD) is part of the receptor binding region present in proteins recognised to be important in the cellular recognition process, including the integrin ligands fibronectin and tenascin {92,93}. The majority of the PSG contain the RGD motif in their N-domain. The location of this tripeptide motif is conserved amongst these PSG subfamily members, between amino acid residues 127-129.

It is possible that covalent interactions between the PSG N-terminus and lymphocyte cells, perhaps involving interactions with other integral membrane proteins, could facilitate cell-cell signalling {87,88}. Since the cell-cell signalling process is thought to be responsible for controlling both cell adhesion, and migration on substrates {92,86}, this could implicate the PSG in the invasion and implantation processes of pregnancy.

The existence of other PSG which lack the RGD tripeptide motif, suggests involvement in biological processes other than cell-cell/cell-matrix interactions in trophoblastic invasion. Structural

differences amongst the PSG family members, for example between soluble and membrane-bound member(s), may indicate different functions for different members.

Results from experiments using mixed PSG purified in bulk from serum, suggest that the PSG may have an immunosupressive role antagonising normal T-cell interactions. It has been demonstrated that the PSG inhibit in a concentration dependent manner, E-rossette formation, suggesting an interference with the T-cell CD-2:erythrocyte LFA-3 interaction {129}. Moreover, the PSG affect the proliferative activity of phytohaemoglobin-stimulated lymphocytes {49}, and are also found to inhibit stimulated lymphocytes in a mixed lymphocyte assay {31}.

The high concentration of PSG in the maternal serum during pregnancy, and their localisation in sites requiring immunosupression (eg: uterus, testes, placenta) are consistent with their suspected role as immunosupressive agents {31}.

The specific reagents responsible for the supressive interactions could not be identified, since a number of different PSG proteins were present in mixtures used in the experiments. Moreover, since different receptors on the T-cell surface were involved in each of the above assays, it is unknown whether the PSG disrupt the interactions directly by binding several receptors, or indirectly, by binding another separate receptor.

The variable PSG C-terminal domain is thought to play a role in localising the PSG:lymphocyte complex (with different C-termini interacting with different lymphocytes), but this is as yet unconfirmed.

The various hypotheses concerning the biological role of the PSG may be investigated, as different forms of PSG are characterised. Cloning of cDNA from different PSG species should allow testing of the properties of each individual protein, following expression of the specific cDNA with appropriate systems.

Our research group is contributing to the understanding of the PSG on two fronts. In one programme, the biological role of PSG-11 is being examined. Initial binding studies show that monocyte cells from two different cell lines bind to a PSG-11 RGD motif in a concentration dependent manner. Introducing an RAD peptide as a competitor into the binding experiment did not reduce RGD binding, whereas a competitor peptide containing the RGD motif sucessfully competed for binding. This demonstrated the specificity of the receptor for the RGD motif, implying that monocyte cells have a receptor for PSG-11. This is consistent with the biological roles

currently proposed for this group of proteins, and isolation of the receptor protein is presently being pursued for further study.

The genomic organisation, regulation, and evolution of the PSG-11 gene is also being investigated in another programme of research. The PSG-11 cDNA and gene was isolated by our group in 1990, and regions of this gene were examined by various members of our group. The C-terminal regions of PSG-11 are presently being investigated to complete the characterisation of this gene.

A.6: OBJECTIVES OF THIS PROJECT

This project is centred around the second aspect, the investigation of the gene organisation around the PSG-11 locus.

The PSG-3 cDNA was isolated at Massey University in 1989 by P.A. McLenachan and B.C. Mansfield {21}. A partial cDNA containing the AII-BII-C-3'-untranslated regions from this gene (PSG-3), was used to probe a human genomic cosmid library under high stringency conditions {98}.

Several clones which cross-hybridised to the PSG-3 cDNA probe were identified and isolated. One of these clones containing a 30 kb insert, named cosmid hC3.11, which was partly characterised in 1990 by K.Beggs {80}.

Hybridisation analysis using PSG domain-specific probes showed that the cosmid hC3.11 contained a large portion of PSG-11 genomic sequence.

Subsequently, an 11 kb *Bam*HI fragment containing all but the C-domain and 3'-untranslated regions of PSG-11 was subcloned, and designated p3B.11. Further work is currently being undertaken to isolate the C-terminal regions of PSG-11, and to complete the characterisation of this gene. {45}

Interestingly, an 8.6 kb *Bam*HI fragment lying upstream contiguous to the p3B.11 fragment, hybridised to the PSG-3 3'-untranslated region probe. This fragment, G3B8.5, was also subcloned into pGEM2 vector, and examined in greater detail by P.A. McLenachan using restriction mapping and sequencing techniques. A region containing potential C-domains was identified, and part of this, a 1.5 kb fragment representing pBEE.5 in Figure 7 was sequenced. Several putative PSG C-terminal splice acceptor sites were identified. Reported human genomic PSG sequences usually

contained 2-4 C-termini per gene, whereas at least seven potential C-terminal splice sites were identified in the hC3.11 C-terminal cluster.

The presence of multiple C-termini upstream of the PSG-11 gene inferred that the C-terminal regions identified, may be a part of a unique PSG gene lying upstream of PSG-11.

It was proposed that this investigation should:

1) Undertake the complete sequencing and analysis of the C-terminal region contained in hC3.11.

2) Try to identify the upstream gene. This involved:

   (i) Identification of unique sequence suitable for use as a `gene-specific' probe.

   (ii) Use the sequence identified in (i) to screen a human genomic library to isolate overlapping or full-length clones containing the `gene' upstream of PSG-11.

   (iii) Characterise clones isolated in (ii) by hybridisation analysis using PSG `domain-specific' probes.

B: MATERIALS AND SOLUTIONS

B.1: MEDIA AND SOLUTIONS

Luria Broth	0.5% (w:v) yeast extract, 1% (w:v) tryptone, 0.5% (w:v) NaCl adjusted to pH 7.5 with NaOH.
Luria Agar	1.5% (w:v) agar in Luria broth.
Luria Agarose	1.5% (w:v) agarose in Luria broth.
Luria Top Agar	0.7% (w:v) agar in Luria Broth.
Luria Top Agarose	0.7% (w:v) agarose in Luria Broth.
HQ STET buffer	8% (w:v) sucrose, 5% (v:v) Triton X-100, 50 mM EDTA pH 8.0, 10 mM Tris-HCl pH 8.0
SM buffer	0.1 M NaCl, 0.05 M Tris-HCl pH 7.5, 0.01% (w:v) gelatin, 8 mM $MgSO_4$.H2O
TE buffer pH 7.6	10 mM Tris-HCl pH 7.6, 1 mM EDTA
TE buffer pH 8.0	10 mM Tris-HCl pH 8.0, 1 mM EDTA
TES buffer	10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.1 M NaCl
TNES buffer	10 mM Tris-HCl pH 8.0, 10 mM NaCl, 2 mM EDTA, 0.1% SDS
TAE buffer	40 mM Tris-acetate, 2 mM Na2EDTA
TBB buffer	89 mM Tris-HCl pH 8.0, 89 mM Boric acid, 5 mM EDTA
PEG solution	20% (w:v) polyethylene glycol 6000, 2 M NaCl
Sequencing Buffer	135 mM Tris, 45 mM Boric acid, 2.5 mM EDTA
20 x SSC	3 M NaCl, 0.3 M sodium citrate
10 x Denhardts Solution	0.2% (w:v) Ficoll 400, 0.2% (w:v) polyvinylpyrrolidone, 0.2% (w:v) bovine serum albumin, in 6 x SSC

Oligonucleotide Pre-hybridisation Solution	6 x SSC, 0.5% (w:v) SDS, 2 mM EDTA, 10 x Denhardts solution, 0.05% (w:v) Sodium Pyrophophate
Oligonucleotide Hybridisation Buffer	6 x SSC, 0.5% (w:v) SDS, 20 x Denhardts solution
Nick-translation Pre-hybridisation Solution	6 x SSC, 10 x Denhardt's solution
Nick-translation Hybridisation Solution	1 M NaCl, 50 mM Sodium Pyrophosphate pH 6.5, 2 mM EDTA, 0.5% SDS, 10 x Denhardt's solution

B.2: BACTERIAL CELL CULTURES

B.2.1: Cell Genotypes

E. coli DH-1 : F^-, end A1, hsdR17 (r^{K-},m^{K-}), sup E44, thi-1, recA1, gyrA96, relA1, lambda$^-$.

E. coli XL-1 Blue : end A1, hsdR17 (r^{K-},m^{K+}), supE44, thi, recA1, gyrA96, relA1, lambda$^-$, (F', proAB, $lacI^q$, lacZm15, Tn10(tet^R)).

E. coli C600 : F^-, supE44, thi-1, thr-1, leuB6, lacY1, tonA21, lambda$^-$.

B.2.2: Storage of Bacterial Cell Cultures

Cell stocks were stored in both 50% glycerol at -20°C, and 30% glycerol at -70°C. To revive cells for use, an innoculum was streaked for single colonies on a Luria agar plate using aseptic technique, inverted and incubated overnight at 37°C. Media for plating *E. coli* XL-1 blue contained 10 µg/ml tetracycline. After sealing with parafilm, plates were stored at 4°C.

B.2.3: Preparation of E.coli Competent Cells

DH-1

A single colony from an E. *coli* DH-1 culture plate was inoculated into a volume (5 ml) of Luria broth and grown overnight on a shaker platform at 220 rpm, at 37°C. The overnight culture was diluted 50 fold, grown to an A°550 of 0.45-0.5 then harvested by centrifugation for 10 minutes, speed 10 in an MSE (Minor) bench top centrifuge at 4°C. The cell pellet was resuspended in a volume of cold 50 mM $CaCl_2$, stood on ice for 30 minutes, centrifuged again, then resuspended in cold 50 mM $CaCl_2$ to 1/25 of the original volume. The competent cells were stored at 4°C, and were used within 2 days after preparation.

XL-1

Preparation of *E. coli* XL-1 competent cells differs from the protocol above only in the following points: the overnight XL-1 culture was diluted 100 fold and harvested when the A°660 reached 0.6-0.7.

B.2.4: Preparation of C-600 Plating Cells

An overnight culture was prepared by innoculating a single *E. coli* C-600 colony into one volume (10 ml) of Luria broth, and shaking 220 rpm overnight at 37°C. Cells were pelleted by centrifugation for 10 minutes, speed 10 in an MSE (Minor) benchtop centrifuge at 4°C, then resuspended in a half volume of cold 10 mM $MgSO_4$.

B.2.5: Infection of E.coli C-600 Cells

Phage lysates were prepared by eluting a plaque into 500 µl SM buffer overnight at 4°C. The eluate was cleared of debris by centrifugation in an Eppendorf Microcentrifuge for 3 minutes, and the supernatant was then stored under chloroform at 4°C. The lysate was then titred, typically, 1-10 µl of lysate was combined with 300 µl of C-600 plating cells, incubated for 30 minutes at 37°C, then plated onto an L-agarose plate using L-top agarose containing 10 mM $MgSO_4$. The plates were then incubated to confluence (10^5 pfu), while others were inverted at 37°C overnight.

METHODS

B.3: DNA PREPARATION

B.3.1: Rapid Isolation of Double Stranded DNA

This procedure, a modification of the Holmes and Quigley method, 1981, was performed essentially as described by Sambrook and Maniatis {13}

Transformants were typically inoculated into 5 ml of Luria broth and incubated overnight at 37°C on a shaker platform at 225 rpm. The cells were then pelleted by centrifugation for 45 seconds in an Eppendorf Microcentrifuge, at 10,700 g (12,500 rpm), and the supernatant was discarded. However, when cells infected with M13 phage were used, the cell-free supernatant was retained for the preparation of single stranded DNA.

The cell pellet was resuspended in 350 µl HQ STET buffer, made 0.7 mg/ml with respect to lysozyme, vortexed gently for 3 seconds, then boiled for 40 seconds. Following centrifugation for 10 minutes in an Eppendorf Microcentrifuge, the gelatinous pellet was discarded and the supernatant was combined with one volume of isopropanol, stood at room temperature for 10 minutes, then centrifuged in an Eppendorf Microcentrifuge for 20 minutes. The pelleted DNA was overlayed with cold 70% ethanol, centrifuged for 15 minutes, dried under vacuum in a Savant Speedvac, and resuspended in 50 µl sterile water.

B.3.2: Medium Scale Lambda DNA Preparation

Before preparing DNA, the titre of the phage lysates of interest were determined by infecting E. coli C-600 cells with serial dilutions of the lysates, and determining the number of plaque forming units per µl of lysate.

Following titration, the phage were plated at 10^5 PFU per 88 mm LB-agarose plate, grown to confluence at 37°C, overlaid with 10 ml SM buffer at 4°C overnight to elute the phage. Lysates were collected from the plates, centrifuged at 4500 gmax (5000 rpm) for 10 minutes, made 1 µg/ml with respect to DNAseI and RNAseA and incubated for 30 minutes at 37°C. One volume of PEG solution (20% PEG 8000, 2 M NaCl in SM buffer) was added and the mixture, stood on ice for 1 hour, before being centrifuged at 8800 gmax (7000 rpm in a Hereaus Christ bench centrifuge) for

30 minutes, at 4°C. The pelleted phage were resuspended in 500 µl SM buffer, vortexed, made 0.1% with respect to SDS, and 5 mM with respect to Na_2EDTA, then incubated for 15 minutes at 65°C. After vortexing briefly, the DNA was extracted with phenol:chloroform (1:1), precipitated with an equal volume of isopropanol at -70°C for 20 minutes, thawed, then centrifuged for 5 minutes in an Eppendorf Microcentrifuge. The supernatant was discarded and the DNA pellet was overlaid with cold 70% ethanol, centrifuged again, vacuum dried, then resuspended in 50 µl of TE buffer containing 0.01 µg/ml RNAse A.

B.3.3: Large Scale Plasmid Preparation

The transformed cells containing the construct of interest were typically inoculated into 5 ml L-broth and grown overnight at 37°C on a shaking platform at 225 rpm. This culture was used to make a 1 in 100 innoculation of 300ml of Luria broth, supplemented with either Ampicillin (100 µg/ml) or Tetracycline (6 µg/ml) depending on the selection required, and grown on a shaking platform at 225 rpm, at 37°C. After 5 hours growth, chloramphenicol was added to a final concentration of 150 µg/ml and the incubation continued overnight to amplify the plasmid copy number.

The culture was harvested by centrifugation at 10,000 g (9000 rpm in a GSA rotor) for 5 minutes at 4°C. The resulting cell pellet was washed by resuspension in 0.5 volume of ice cold TE buffer (pH 8.0), pelleted as above, and finally drained completely.

The DNA was extracted from the cell pellet by the alkaline lysis method essentially as described by Birnboim and Doly (1979) {89}, made 50 µg/ml with respect to ethidium bromide and purified by isopycnic centrifugation through a 5.9 M CsCl cushion at 285,000 g (55000 rpm in a TV865 rotor), for 5 hours at 15°C. The plasmid band, intercalated with ethidium bromide, was visualised using long wavelength UV-light and collected using a syringe. Ethidium bromide was removed by repeated extraction with water saturated isobutanol, and the plasmid finally dialysed against several changes of TE buffer (pH 7.6), before being stored in 1 ml aliquots at 4°C. The concentration and yield of plasmid was calculated by spectroscopy, assuming that an absorbance of 1.0 at 280 nm represented 50 µg/ml of DNA. The purity was judged by the 260 nm /280 nm ratio. Typically, yields ranged from 200-500 µg/ml.

B.3.4: Isolation of Human Genomic DNA

This method was performed essentially as described by Gustafson *et al.*, 1987 {82}.

A volume of human blood (typically 50 ml) was collected into vacuum sealed tubes containing acid citrate dextrose anticoagulant. The cells were pelleted by centrifugation for 15 minutes at 1300 g (2830 rpm, Heraeus Christ Labofuge GL). The buffy coat was removed, resuspended in 0.6 volumes of TE, then brought to 500 µg/ml Pronase, 50 µg/ml RNAse, and 0.5% SDS. The mixture was inverted several times then incubated for 2 hours at 37°C. Following the digestion an equal volume of phenol was added and the solution mixed gently by hand for 9 minutes. The aqueous phase was recovered by centrifugation for 10 minutes at 3400 g at 23°C. The phenol extraction was repeated, followed by a chloroform-isopropanol extraction. DNA was precipitated from the aqueous phase by addition of 1/50 volume of 5 M NaCl and two volumes of cold 95% ethanol. The solution was mixed, incubated at -70°C for 15 minutes, then centrifuged at 3400 g (4600 rpm Heraeus Christ Labofuge GL) for 10 minutes, at 4°C. The DNA pellet was rinsed with cold 70% ethanol, dried, then resuspended in TE buffer pH 7.6 to 1/10 of the original blood volume. Yields were estimated to average 29 µg DNA per ml of blood.

B.4: STANDARD PROCEDURES

B.4.1: Restriction Enzyme Digests of DNA

A typical digest would contain 2 µg of DNA, 10 units of the required restriction enzyme(s), in a total volume which was at least ten times the combined volume of the enzyme and DNA, in the reaction buffer supplied by the manufacturer. The digestions were performed at the recommended temperature for sufficient time to give a 20 fold overdigestion.

B.4.2: Ethanol Precipitation

To concentrate DNA, the sample was made 0.3 M with respect to NaOAc pH 5.5, and combined with 3 volumes of cold 95% ethanol. After standing at -70°C for 20 minutes, or at -20°C overnight, the DNA was pelleted by centrifugation at 10,700 gmax (12500 rpm) in an Eppendorf Microcentrifuge for 30 minutes at 4°C. The supernatant was discarded and the pellet was overlayed with cold 70% ethanol, centrifuged again for 15 minutes at 10,700 gmax (12,500 rpm), dried under vacuum in a Savant Speedvac, and resuspended in a small volume of TE buffer or sterile water.

B.4.3: Agarose Gel Electrophoresis

DNA samples were size fractionated by electrophoresis through agarose gels in TAE buffer. When fragments were to be gel purified, low melting point agarose gels (Seaplaque, FMC Bioproducts) were used. Small analytical gels were typically electrophoresed in TAE buffer at 100 V, for 1 hour. Larger preparative gels were electrophoresed overnight at 20 V. Gels were stained in ethidium bromide for 10 minutes, destained in Milli-Q water for 15 minutes, illuminated with shortwave UV light and photographed with Polaroid type 665 film. Fragment sizes were estimated by comparing mobilities against a 1 kb DNA ladder (BRL).

B.4.4: Phenol-Chloroform Extraction of DNA

To deproteinise DNA, one volume of DNA was combined with one volume of Tris-equilibrated phenol and vortexed vigorously for 30 seconds. After standing for 5 minutes, the aqueous phase was separated from the organic phase by centrifugation at room temperature in an Eppendorf Microcentrifuge for 5 minutes. The aqueous phase was extracted with one volume of Tris-equilibrated phenol:chloroform (1:1), centrifuged for 5 minutes, then extracted with one volume of

choloroform. DNA was recovered from the resulting aqueous phase by precipitation with ethanol in the presence of sodium acetate as described above.

B.5: DNA CLONING PROCEDURES

B.5.1: Ligation of DNA fragments

Ligation reactions containing 15 pmol of insert and 15 pmol of vector were performed in a reaction containing 60 mM Tris pH 8.0, 50 mM NaCl, 5 mM DTT, 1 mM ATP, and 1 unit of T4 DNA ligase in a total volume of 10-20 µl. Reactions were incubated at room temperature for 2 hours, followed by 12 hours at 4°C. To analyse the ligation reaction, an aliquot of the ligation reaction was electrophoresed through an agarose gel and assessed for ligation products.

B.5.2: Gel Purification

DNA fragments of interest were purified from cloned DNA by restriction enzyme digestion and subsequent gel purification.
Typically, 5 µg of DNA was digested using appropriate restriction enzyme(s). Following ethanol precipitation, the DNA was resuspended in sterile water, then size fractionated by electrophoresis through a 1.0% low melting point agarose gel (Seaplaque, FMC Bioproducts) in TAE buffer. Following staining with ethidium bromide, the gel was destained in Milli-Q water, and the band(s) of interest, visualised under long wavelength ultraviolet light, were excised using a sterile scalpel blade. The DNA was extracted from the gel slice by a centrifugation procedure. The bottom of a sterile 0.5 ml Eppendorf tube was pierced with a hot needle, packed 1/3 full with siliconised glass wool, overlaid with 200 µl of sterile water, then placed within a 1.5 ml Eppendorf tube. The assembly was then centrifuged in an Eppendorf microcentrifuge for 3 minutes, to elute the sterile water and rinse the filter unit. Using a fresh 1.5 ml collection tube, the gel slice was placed on the glass wool, and the complete asssembly was centrifuged at 2438 gmax (6500 rpm) in a variable-speed benchtop centrifuge (MSE Minor) for 10 minutes. The eluate containing DNA was either used directly or concentrated by ethanol precipitation.

B.5.3: Transfection and Transformation

Competent *E. coli* XL-1 and DH-1 cells were prepared by the $CaCl_2$ method {102}.

The plating and selection methods for sucessful transformants was dependent on the type of vector used:

For transfecting constructs into *E. coli* XL-1 cells, an aliquot (1-10 µl) of ligation reaction was combined with 300 µl of competent cells, stood on ice for 30 minutes, heat shocked at 42°C for 90 seconds, then mixed with 3.5 ml of molten top Luria agar and plated onto appropriate agar plates. After allowing the surface to set, the plates were inverted and incubated overnight at 37°C. To allow for colour selection via α-complementation, XL-1 cells were plated in the presence of 0.03% X-Gal and 5 mM IPTG.

E. coli DH-1 cells to be transformed with pGEM-2 constructs were treated similarly, except they were made up to 1 ml with Luria broth after the heat shock, shaken (225 rpm) for 30 minutes at 37° C. Up to 1/10 volume (100 µl) was plated onto Luria agar in the presence of ampicillin (100 µg/ml).

B.5.4: Electroporation and Electro-transformation

Electrocompetent cells were prepared by innoculating an overnight culture of *E. coli* DH-1 cells 1:100 into 1 litre of LB-broth and growing them on a shaking platform, at 37°C, to an optical density at 600nm of 0.64.

To harvest the cells, the culture was stood on ice for 15-30 minutes, then centrifuged at 2877 gmax (4000 rpm in a Heraeus Christ Labofuge GL) for 15 minutes at 4°C. The supernatant was discarded, the cell pellets resuspended in 1 litre of sterile cold water, centrifuged at 2877 gmax (4000 rpm, Heraeus Christ) for 15 minutes at 4°C, resuspended in 0.5 litres of cold sterile water then centrifuged again at 2877 gmax (4000 rpm, Heraeus Christ) for 15 minutes at 4°C. The pellet was resuspended in 20 ml of cold 10% glycerol, centrifuged at 2877 gmax (4000 rpm, Hereaus Christ) for 15 minutes at 4°C, then resuspended in a final volume of 2 ml of cold 10% glycerol. The concentration of cells in the suspension was between 1×10^{10} and 3×10^{10} cells/ml. These electrocompetent cells were stored in 40 µl aliquots at -70°C, and remained viable for about 6 months under these conditions.

To electroporate the cells, a vial of cells was thawed slowly to room temperature. A small volume (typically 1-10 µl) of the ligation reaction was then combined with the electrocompetent cells and the mixture stood on ice for 1 minute. The cell mixture was transferred into an ice cold electroporation cuvette, and positioned on a pre-chilled safety chamber slide in a Biorad Gene Pulser.

Electroporations were performed at 25 μF, 2.5 KeV, and 200 Ω producing an electric field of 12.5 KeV/cm, with a time constant of 4-5 milliseconds. Immediately after the shock, the cells were transferred into a 13 x 100 mm glass (Kimax) tube containing 1 ml of Luria broth and incubated for 30 minutes on a shaking platform at 225 rpm at 37°C to allow for recovery. Transformants were then selected by plating on L-agar containing either ampicillin (100 μg/ml), or tetracycline (15 μg/ml), depending on the selection required.

B.6: NUCLEOTIDE SEQUENCING

B.6.1: Small Scale DNA Preparation for Sequencing of M13

Recombinant M13 plaques were inoculated into Luria broth and grown on a shaker platform at 225 rpm at 37°C overnight. The cells were then pelleted for 4 minutes at 10,700 gmax (12500 rpm) in an Eppendorf Microcentrifuge, and used for the preparation of double stranded DNA . The cell-free supernatant was used for the preparation of single-stranded DNA.

To precipitate the M13 bacteriophage, a volume (1ml) of the supernatant was made 500 mM in NaCl and 4% in PEG 6000, stood at room temperature for 30 minutes, then centrifuged in an Eppendorf Microcentrifuge for 5 minutes at 10,700 gmax (12,500 rpm). The supernatant was discarded, and remaining traces of the PEG solution were removed carefully using tissue paper. The pelleted phage were resuspended in TE buffer, then phenol/chloroform extracted. Following ethanol precipitation, the resulting single-stranded DNA templates were dried under vacuum and resuspended in 30 μl TE buffer for sequencing.

B.6.2: Polyacrylamide Gel Electrophoresis

The 6% polyacrylamide sequencing gel containing 8 M urea was warmed by electrophoresis at 1400 V for 20 minutes in TBB buffer electrolyte before loading templates. For each template, typically 3 μl of the sequencing reactions was loaded into each of the 4 wells (G, A, T, C) defined by sharkstooth combs, in the order G,A,T,C. The 0.4 mm thick gel was electrophoresed at a constant power of 65 W, for four hours, before a duplicate set of lanes were loaded and electrophoresed for an additional 2 hours. Following electrophoresis, the gel electrophoresis assembly was dismantled and the gel was transferred onto 3MM filter paper and dried under vacuum at 80°C for 55 minutes. The gel was autoradiographed with Fuji RX X-ray film overnight and typically gave 300 bases of clear sequence per template.

B.7: SOUTHERN BLOT HYBRIDISATION ANALYSIS

B.7.1: Vacuum Blotting

The DNA of interest was digested with appropriate restriction enzymes, electrophoresed through a 0.7% (w:v) preparative agarose gel in TAE buffer for 16 hours at a constant voltage of 20 V, stained in ethidium bromide, destained in milli-Q water, and examined under short wavelength UV light. Photographs were taken with a metric ruler adjacent to the lane containing the 1kb DNA marker ladder (BRL) for accurate calibration of fragment sizes.

The vacuum blot was assembled on a Biorad Vacuum Blotter manifold and a vacuum of 50 mbar was applied using an electrically operated vacuum pump (Pharmacia). Both nylon and nitrocellulose type membranes were used on different occasions, in different experiments.

The gel was overlaid with 0.2 M HCl for 10 minutes to depurinate the DNA. This solution was replaced with denaturating solution (1 M NaOH, 0.5 M NaCl) for 10 minutes, and this solution was finally replaced with 3 M HCl, 1 M Tris for 10 minutes to neutralise the pH. Following removal of the neutralising solution, the gel was covered with 20 x SSC buffer and the transfer proceeded for 40 minutes under vacuum. Marks recording position of lanes on the gel, were made on the membrane, which was then baked at 80°C, under vacuum (-80 kpa) for 3 hours, then stored between sheets of filter paper at room temperature.

B.7.2: Labelling Probes

Probes were labelled by nick translation essentially as described by Rigby *et al* {114}. The nick translation reaction contained 0.2 µg DNA, 0.01 mM dATP, 0.01 mM dTTP, 0.01 mM dGTP, 10 U DNA Polymerase I, 50 µCi α-[^{32}P]dCTP (10 mCi/ml, 3.3x10^{-3} µmol/ml), 1 x DNA Polymerase I buffer (50 mM NaCl; 10 mM Tris-HCl, pH 7.5; 10 mM DTT; 10 mM $MgCl_2$), and 100 pg DNAseI, in a total volume of 25 µl, was incubated at 15°C for 15 minutes.

The reaction was diluted to 200 µl with TES to quench the reaction. Following the reaction, unincorporated nucleotides were removed by centrifugation through a 1 ml column containing Sephadex G-50 (Sigma). To make the column, a sterile 1 ml syringe (Monoject 501S-TB) was plugged with sterile glass wool, then filled with Sephadex G-50 resin (Sigma) pre-swollen in TES buffer. The bottom of the syringe was placed through the top of a pierced eppendorf tube, and the

entire assembly placed into a 50 ml Falcon tube. Centrifugation at 1550 gmax (speed 3, MSE Minor centrifuge) for 3 minutes packed the column, which was then rinsed by centrifugation with 200 µl of TES buffer. The eppendorf collection tube was then replaced and the nick-translation reaction applied to the spin column. The diluted probe (200 µl), was eluted through the column by centrifugation at 1550 gmax (speed 3, MSE Minor centrifuge) for 3 minutes. Specific activity of the probe was determined by thin layer chromatography using polyethylene imine (PEI) paper and a 2 N HCl solvent. Radioactive nucleotide incorporated into the DNA remained stationary at the point of origin on the paper, whereas the unincorporated nucleotide was mobile, and migrated with the solvent front. The PEI paper was cut in half, and the radioactivity of each half quantified by scintillation counting (Beckman LS7000). The following equations were used to calculate the efficiency of isotope incorporation into the probe and the specific activity:

Equation 1

$$\% \text{ Incorporation} = \frac{\text{cpm incorporated into DNA}}{\text{cpm incorporated} + \text{cpm unincorporated}} \times \frac{100}{1}$$

Equation 2

$$\text{Specific activity (cpm/µg DNA)} = \text{Total incorporated cpm (cpm/Reaction)} \times \text{Amount of DNA in reaction (µg)}$$

Typically, values of 60-70% incorporation and specific activities of 2×10^7 cpm/µg were obtained.

B.7.3: Hybridisation

Filters to be probed were prehybridised at 65°C in sealed pyrex hybridisation tubes containing 100 ml prehybridising solution (6 x SSC, 10 x Denhardt's solution) for 2-3 hours, after which the prehybridisation solution was discarded and replaced with 10ml of hybridisation solution (1 M NaCl, 50 mM Sodium Pyrophosphate, 2 mM EDTA, 0.5% SDS, 10 x Denhardt's solution). Immediately prior to use, the probe was diluted 4-fold in TNES buffer and Herring sperm DNA (1 mg/ml), and denatured by boiling for 10 minutes. The probe was then applied to the filters and the hybridisation was performed in glass cylinders for 20 hours at 65°C, rotating at 6 rpm in a hybridisation oven (Bachofer 400HY).

Wash solutions were typically heated to 68°C prior to use. The hybridisation solution was discarded, and the filters washed with 2 x SSC, 0.1% SDS twice for 15 minutes, followed by two 15 minute washes in 1 x SSC, at 65°C. Filters were then removed, air-dried on filter paper and autoradiographed against Fuji RX X-ray film at -70°C.

B.7.4: Oligonucleotide Labelling and Hybridisation

Oligonucleotides were 5'-end labelled. A typical end-labelling reaction contained approximately 7 pmol of oligonucleotide, 3 U of T4 polynucleotide kinase, 30 μCi of γ-[^{32}P]ATP (10 mCi/ml, 3.3 pmol/ml), 1 x kinase buffer (50 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 10 mM DTT), in a total volume of 20 μl. This was mixed gently, incubated at 37°C for 1 hour, combined with 10 volumes of TNES buffer, heated to 65°C for 3 minutes then used immediately. The activity of the probe was estimated by chromatography using a polyethyleneimine paper developed in 1.2 M KH_2PO_4.
The filters to be probed were prehybridised in 50 ml of buffer consisting of 6 x SSC, 0.5% SDS, 0.05% Sodium Pyrophosphate, 10 x Denhardt's solution for 2 hours at 65°C. This was discarded and replaced with 10 ml hybridisation buffer (6 x SSC, 0.05% (w:v) sodium pyrophosphate, 20 x Denhardt's solution), containing 9.9×10^{-3} pmol of the end-labelled probe. The filters were hybridised to the oligonucleotide probe for 1 hour at 37°C, then transferred to a plastic box for washing. Three ten minute washes were carried out at room temperature in 6 x SSC and 0.05% sodium pyrophosphate, with a final wash, for two minutes, at an appropriate temperature to set the stringency. This final temperature was determined experimentally for each oligonucleotide, guided initially by the theoretical T_m for the hybrid which was determined by Equation 3:

Equation 3

$T_m = 4(G+C) + 2(A+T)$ in 6 x SSC

Typically a stringency in the range of T_m -5°C to T_m -15°C was found appropriate.

The wet filters were autoradiographed with Fuji Rx X-ray film overnight at -20°C.

B.8: SCREENING OF A cDNA LIBRARY

A human genomic DNA library had been constructed by B.C Mansfield. DNA extracted from his peripheral blood lymphocytes, had been partially digested with S*au*3AI, cloned into the λ-vector EMBL3 and transfected into *E. coli* CES 200 cells and then amplified. The library was screened with a PSG central-domain probe, and clones which cross-hybridised were picked and patched to create a library enriched for PSG-like clones.

This enriched human genomic library (kindly supplied by P.McLenachan) was screened in this investigation with both the PSG-domain probe and a gene-specific probe (pBE.5), to identify cosmid clones of interest.

The phage titre of this enriched library on *E. coli* C-600 plating cells was calculated and the library then plated to give 100-200 plaques/ 88 mm plate on L-agarose plates.

Assymetric orientation marks were made on the outside of the plates with a permanent marker pen, and duplicated on the nitrocellulose filters during lifting for orientation. Lifts were typically made for 60 seconds.
Filters were air-dried on 3MM filter paper briefly, placed DNA-side up onto a sheet of 3MM filter paper saturated with denaturating solution (3 M NaCl, 1 M NaOH) for 5 minutes, then transferred onto 3MM filter paper saturated with neutralising solution (3 M NaCl, 1 M Tris-HCl) for the same period of time. Following a brief rinse in 2 x SSC the filters were dried under vacuum (-80 kpa) at 80°C for 3 hours, then stored for Southern hybridisation.

B.9: Cos-mapping

To partially digest DNA for use in cos-mapping, conditions were optimised for each particular DNA preparation and restriction enzyme. To do this, the ratio of enzyme to DNA and the duration of digestion were varied to obtain sucessful partial digests, as judged by gel electrophoresis profiles in comparison to complete digests. Typically, digests contained 3 µg of DNA prepared by the medium scale method, approximately 3 µl of a medium scale phage preparation and 5 U of the appropriate restriction enzyme, in the reaction buffer supplied by the manufacturer, and were incubated for 1 hour at 37°C. Partial digests were subsequently divided into half for hybridisation to the ON-R and ON-L that are specific to the right and left arms of the cosmid respectively.

B.9.1 End Labelling the Oligonucleotides

The oligonucleotides ON-L and ON-R (Amersham) were end-labelled with γ-[^{32}P]ATP in a reaction that contained 4.5 pmoles of oligonucleotide, 30 µCi of γ-[^{32}P]ATP (10 mCi/ml, 3.3 pmol/ml), 3 U polynucleotide kinase, and kinase buffer (50 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 10 mM DTT) in a total volume of 15 µl. The components were mixed thoroughly, incubated first at 37°C for 1 hour, then at 42°C for 3 minutes. Incorporation of the isotope was checked by thin layer chromatography using polyethylene imine (PEI) paper and a 1.2 M KH_2PO_4 solvent. The labelled oligonucleotides were diluted 125-fold in 0.5 M TE buffer pH 8.0, 0.2 M NaOH, 170 mM Bromophenol Blue, 39 mM Tris-HCl pH 7.6, 36 mM sucrose, 30 mM EDTA. The diluted probe was typically used immediately for hybridisation. If stored at -20°C, the probes retained sufficient activity for use in Cos-mapping for up to 2 weeks.

B.9.2: Oligonucleotide Hybridisation

For hybridisations, each half of the partial digest was combined with 12 fmoles (5 µl) of the appropriate diluted probe and mixed gently. The mixture was then denatured by incubation at 70°C for 3 minutes, allowed to anneal at 42°C for 30 minutes, and finally loaded directly onto a 0.4% (w:v) preparative agarose gel.

High molecular weight markers (BRL catalogue #5618SA) were used to calibrate the size of fragments resulting from these experiments. These markers were also hybridised to both the ON-R and ON-L labelled oligonucleotides in separate reactions, which were then pooled before electrophoresis.

The gel was electrophoresed for 24-30 hours at 40 V, dried onto DE81 cellulose acetate paper at 55°C, for 30 minutes using a Biorad gel dryer (model 583), then autoradiographed with Fuji RX X-ray film, overnight, at -70°C.

C: RESULTS AND DISCUSSION FOR THE hC3.11 SEQUENCE

C.1: Characterisation of hC3.11

This report details the investigation of the sequence located near the 5' end of the cosmid hC3.11. A schematic diagram representing the clone hC3.11, and subclones of this cosmid is shown in Figure 7. This cosmid was initially isolated by Sharon Sims in 1989, and characterised by Kyle Beggs {80} and Patricia McLenachan {45}. Their analysis of hybridisation data combined with DNA sequencing, revealed that the cosmid hC3.11 encoded the L, N, A, and B domain exons for PSG-11, as shown in Figure 7.

Furthermore, an 8.5 kb *Bam*HI fragment upstream of PSG-11 hybridised to a probe representing the 3'-untranslated region of PSG-3. Consequently, the 8.5 kb *Bam*HI fragment was subcloned into the vector pGEM2 and named pG3B8.5. The initial 1.5 kb of pG3B8.5 was examined in greater detail by P.A. McLenachan {45}, and was found to contain sequence resembling parts of several PSG C-terminal domain exons. This suggested that just upstream lay another PSG gene, in a tail to head configuration to PSG-11.

C.2: The Subclone pE2.2

To determine whether there was a complete C-terminal coding region, the downstream 2.2 kb *Eco*RI fragment adjacent to the previously sequenced 1.5 kb region, was subcloned and sequenced in both directions.

To do this, the plasmid pG3B8.5 was digested with *Eco*RI and the 2.2 kb fragment, isolated by gel electrophoresis, was ligated into the pGEM2 vector and transformed into *E. coli* DH-1 cells. Ten transformants were selected and analysed for inserts by restriction digests. One transformant, named pE2.2 was selected for further analysis.

Fine mapping of the insert with restriction enzymes did not identify any sites which would allow easy subcloning and sequencing of the insert. Therefore, a Bal 31 exonuclease strategy was employed. This strategy created a series of nested deletions over the entire length of the insert, allowing the region to be sequenced in both directions.

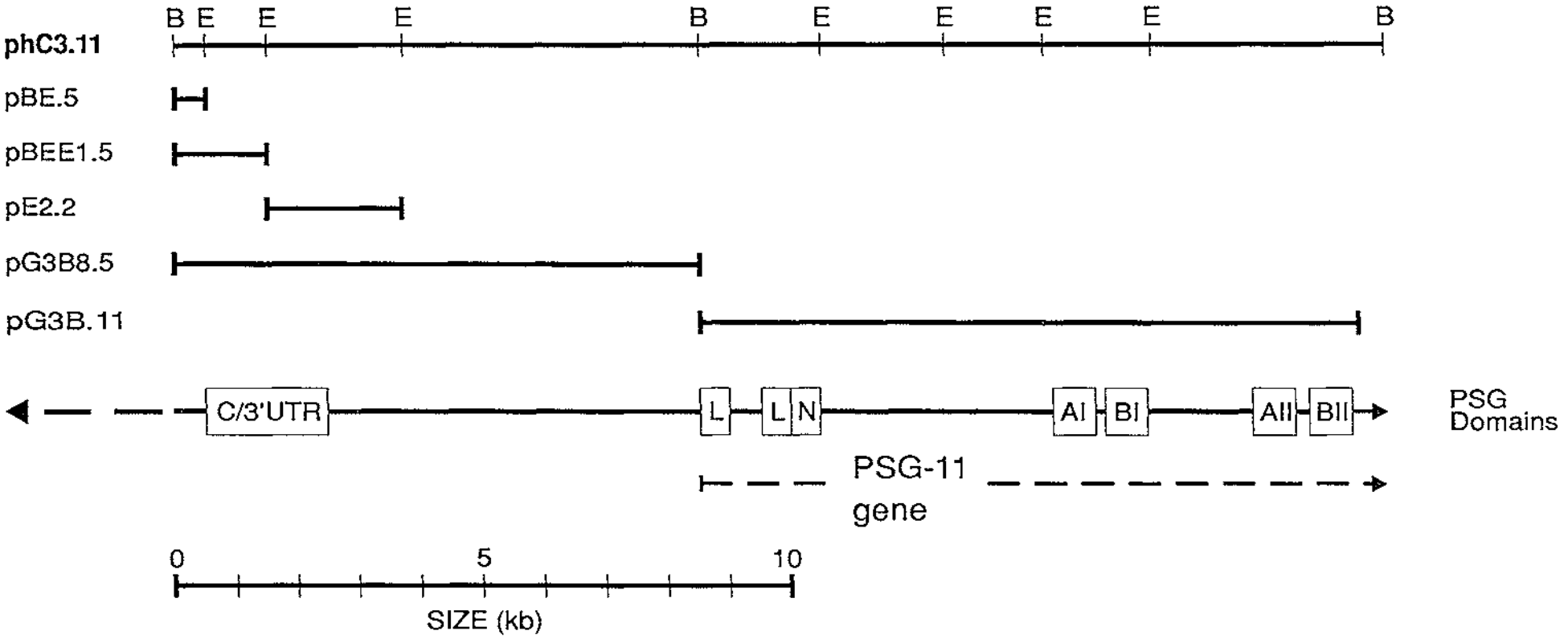


Figure 7: Subclones of Cosmid hC3.11

Fragments subcloned from the cosmid hC3.11 are shown above corresponding to their designated titles. Alignment to schematic representations of PSG domains below approximate their positions.

C.2.1: Titration of the Bal 31 Enzyme

To determine the rate of Bal 31 exonuclease activity, a series of titrations were performed with varying concentrations of enzyme. A number of different procedures were tried before the most consistent and convenient protocol was selected. Titrations using 0.5 U, 1 U, and 2 U of Bal 31 enzyme were performed using 8 µg of linearised insert DNA. Results from these experiments are shown in Figure 8. The rate of digestion using 0.5 U of Bal 31 was too slow (25 bp/min/end), whereas the rate using 2 U of enzyme was too rapid (175 bp/min/end). The optimum rate of digestion was achieved using 1 U of Bal 31 per 8 µg of DNA; this method is detailed below:

The circular plasmid construct pE2.2 (2.2 kb insert in pGEM2), was digested with the restriction enzyme HindIII to yield a 5.1 kb linear molecule, suitable for Bal 31 digestion. The optimal exonuclease conditions were with 1 unit of Bal 31 exonuclease, 8 µg of linearised plasmid DNA, in a buffer containing 0.6 M NaCl, 20 mM Tris-HCl (pH 8.0), 12.5 mM $MgCl_2$, 12.5 mM $CaCl_2$, and 1 mM EDTA, in a total reaction volume of 40 µl. The reaction was incubated at 30°C for time intervals of 0, 2, 5, 10, 15, 20, 30, and 40 minutes. At the end of each time interval, 5 µl aliquots were transferred into loading dye containing 40 mM EDTA to terminate Bal 31 activity. Samples from the individual time points were separately electrophoresed through a 1% agarose gel, stained with ethidium bromide, destained in milli-Q water, examined under short wavelength UV light and photographed.

The rate of digestion was calculated using Equation 4:

Equation 4

$$\text{Rate of exonuclease activity} = \frac{\text{Change in length of DNA (bp)}}{\text{Duration of digestion (min)}} \times 0.5 \text{ (bp/min/end)}$$

[The factor of 0.5 is included to account for simultaneous digestion at each end of the molecule]

The results of this titration are shown in Figure 8. Using the conditions specified above, the activity of the Bal 31 enzyme was 81.5 bp/min/end.

Figure 8: Titration of the Bal 31 Enzyme.

Reaction termination times are labelled in minutes.
Units of enzyme used are noted above the figures.
The BRL 1 kb DNA size ladder is labelled (L).
(See section C.2.1 for text).

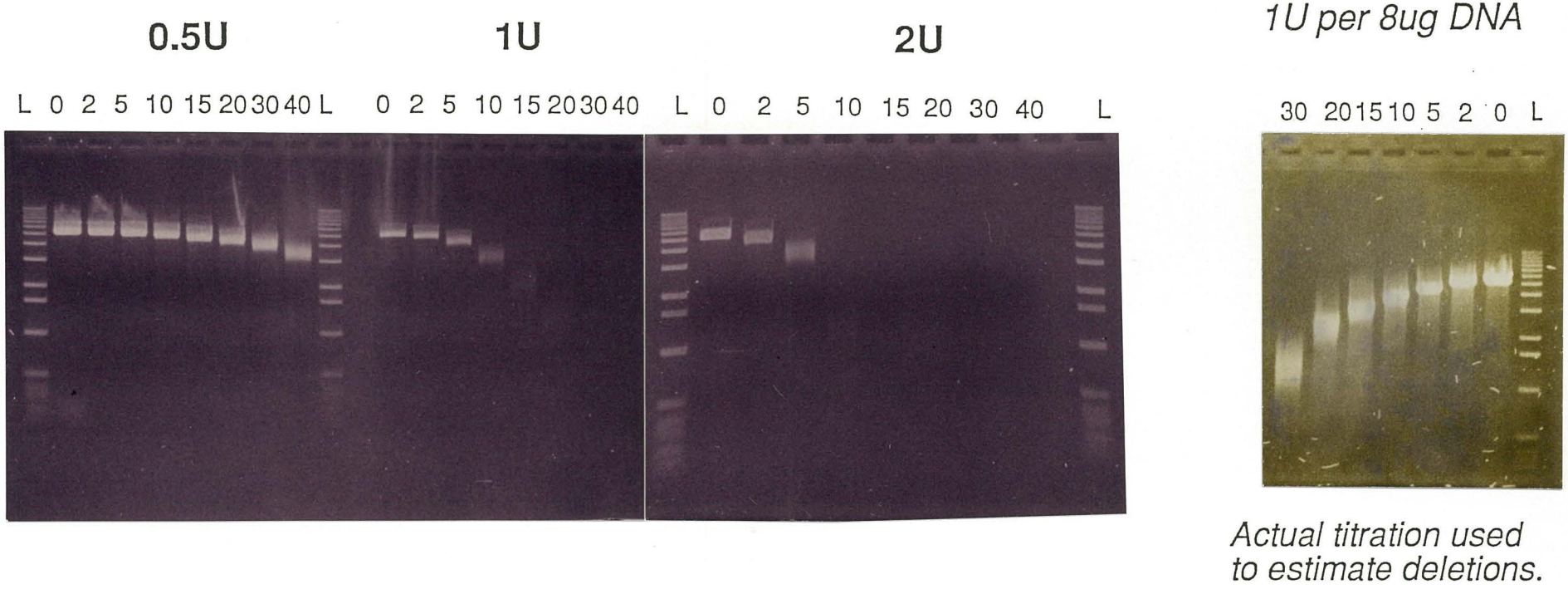


Figure 8: *The Bal31 enzyme titration. Reaction termination times are labelled in minutes. Units of enzyme used are noted above the figures. The BRL 1 kb DNA size ladder is labelled (L).*

C.2.2: The Bal 31 Strategy

After the rate of Bal 31 activity was determined, the analytical reaction was scaled up 3-fold for a preparative scale reaction. The only difference was that the time point fractions were pooled, and terminated in Tris-equilibrated phenol.

To obtain sequence from both ends of the clone, the clone pE2.2 was linearised with *Sph*I for one direction and with *Hind*III for the other direction. The linear DNA were then digested with Bal 31 to create a set of nested deletions which were spaced approximately 283 bp apart. These fragments were digested with *Eco*RI to release the insert fragments from vector fragments, resulting in each insert having one blunt end, and one *Eco*RI end.

To prevent the pGEM2 vector fragment from ligating into the sequencing vector, an SphI, or *Bam*HI digest (depending on the direction sequenced) was then performed to cut the vector sequences, leaving only the insert with blunt *Eco*RI ends.

To fractionate the Bal 31 deletion fragments, digests were pooled, then size fractionated by electrophoresis through a 1% low melting point agarose gel. Following staining with ethidium bromide and examination under long wavelength UV light, the gel was sliced to create sections containing inserts differing from each other, on average by 283 bp, and DNA from the individual sections gel purified.

The sequencing vector M13mp18 was digested with *Eco*RI and *Sma*I, the Bal 31 deletion derivatives ligated into the sequencing vector, and then transformed into *E. coli* XL-1 cells. Transformants were selected by the blue/white selection of α-complementation, and the double stranded and single stranded DNA prepared from these clones, were sequenced using the Sequenase version 2.0 (USB) modification of the dideoxy-termination method.

A schematic representation of the overlapping clones which were sequenced, spanning the 2.2 kb *Eco*RI fragment is shown in Figure 9.

The resulting sequence of hC3.11 is shown in Figure 10. The initial 1523 bp represents sequence previously obtained by P.McLenachan {45}, while the area from (1524 bp to 3796 bp) is from this investigation.

Figure 9: The Bal 31 Deletion Clones.

Schematic diagram showing the overlapping clones used in determining the sequence of the p2.2E clone in two directions. Fragment lengths are measured in kb.
(See section C.2.2 for text).

Figure 9: The Bal 31 Deletion Clones.

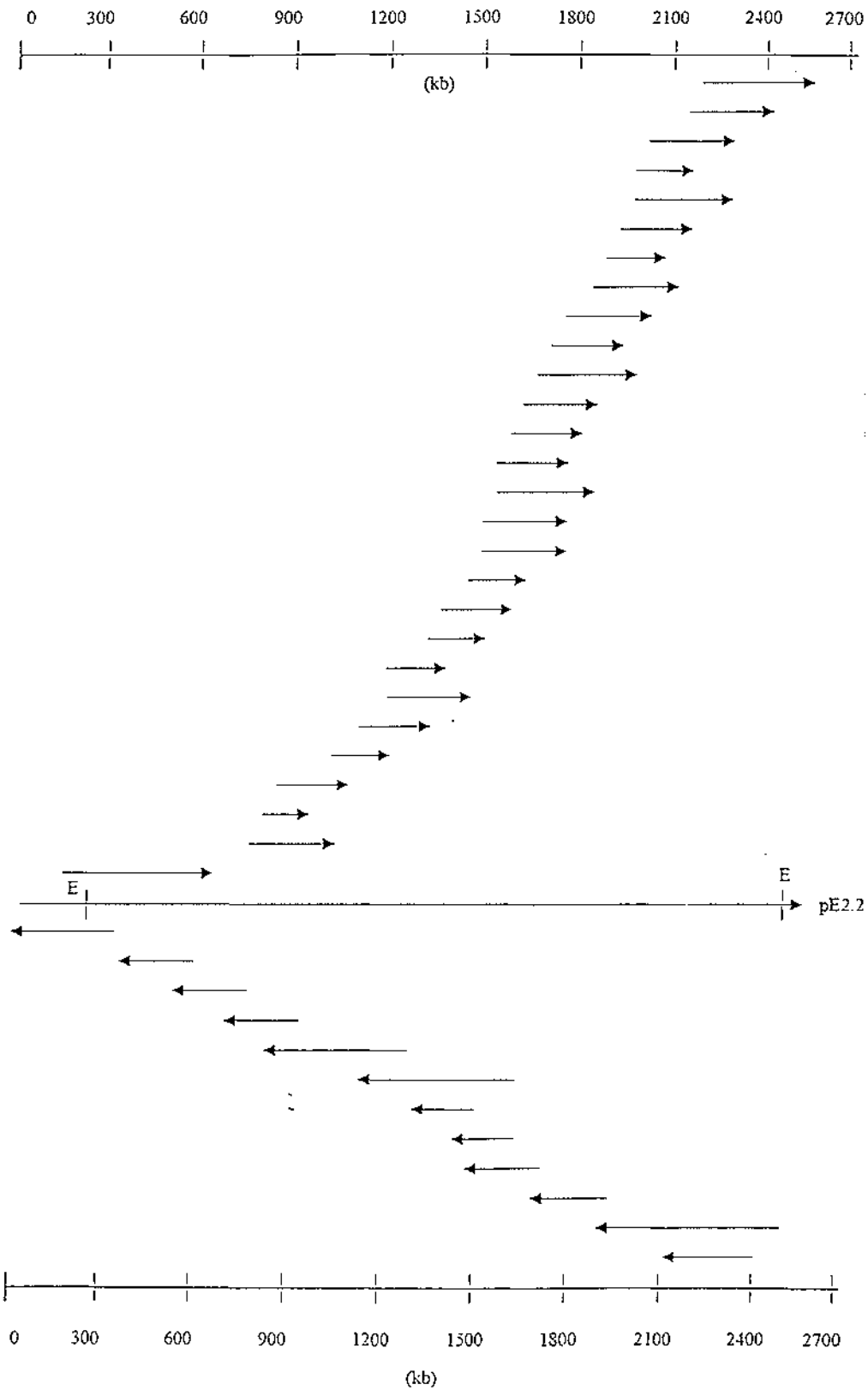
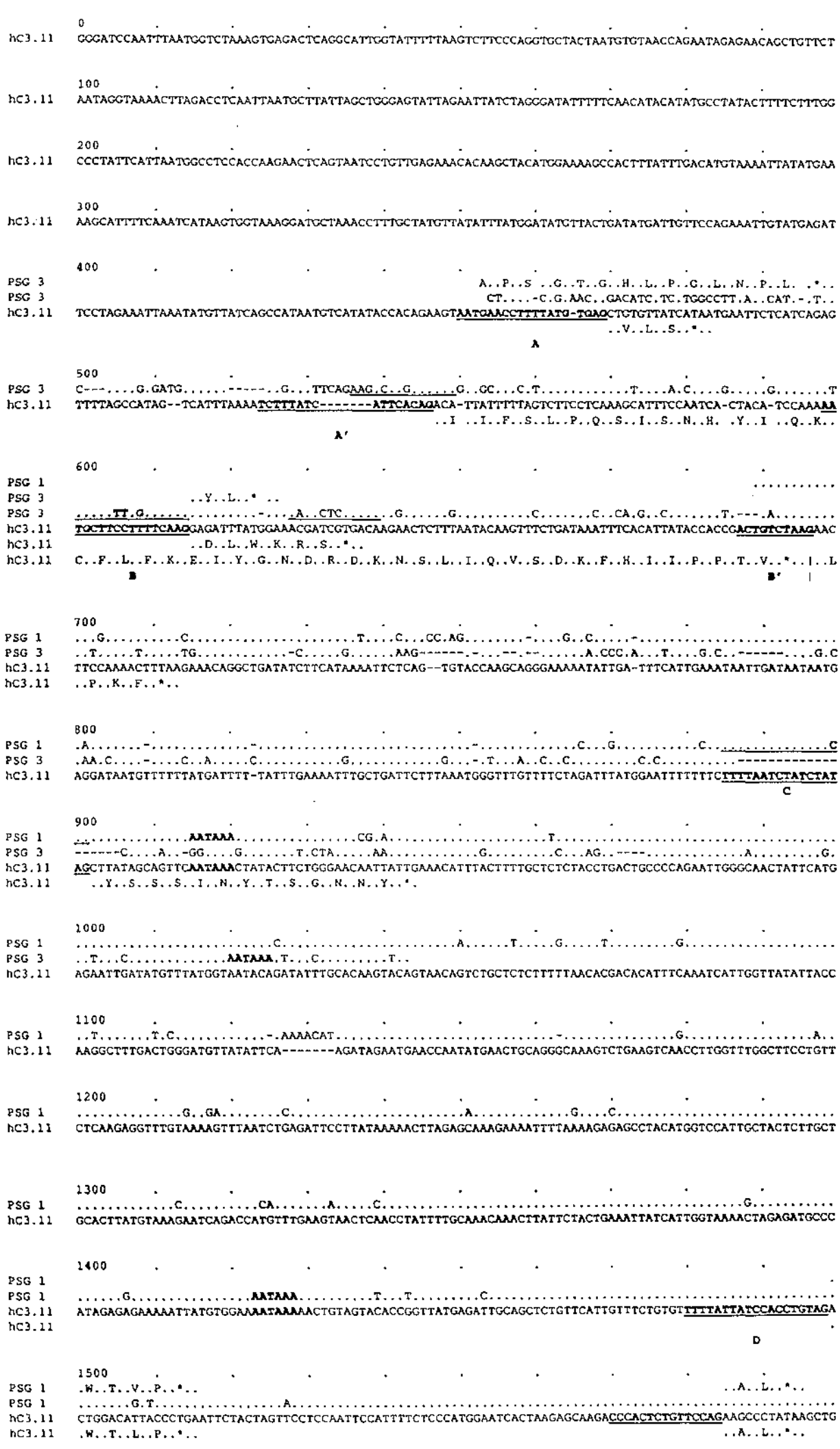


Figure 10: The Combined hC3.11 Sequence

The sequence of the C-terminal domains of hC3.11 aligned with other PSG sequences. The areas of identity to known PSG sequences to hC3.11 are represented by `. ', differences are designated by their respective nucleotide codes. Putative splice sites, which match the consensus sequence proposed by Shapiro & Senapathy {118} are underlined in bold type and are designated `**A** ' to `**H** ', while predicted open reading frames for C-termini are shown alongside the sequence in single letter amino acid code. Polyadenylation signals (AATAAA) are displayed in bold type. Spaces introduced for optimum sequence alignment are represented by `- '.


0
hC3.11 GGGATCCAATTTAATGGTCTAAAGTGAGACTCAGGCATTGGTATTTTTAAGTCTTCCCAGGTGCTACTAATGTGTAACCAGAATAGAGAACAGCTGTTCT
100
hC3.11 AATAGGTAAAACTTAGACCTCAATTAATGCTTATTAGCTGGGAGTATTAGAATTATCTAGGGATATTTTTCAACATACATATGCCTATACTTTTCTTTGG
200
hC3.11 CCCTATTCATTAATGGCCTCCACCAAGAACTCAGTAATCCTGTTGAGAAACACAAGCTACATGGAAAAGCCACTTTATTTGACATGTAAAATTATATGAA
300
hC3.11 AAGCATTTTCAAATCATAAGTGGTAAAGGATGCTAAACCTTTGCTATGTTATATTTATGGATATGTTACTGATATGATTGTTCCAGAAATTGTATGAGAT
400
PSG 3 A..P..S ..G..T..G..H..L..P..G..L..N..P..L. .*..
PSG 3 CT....-C.G.AAC..GACATC.TC.TGGCCTT.A..CAT.-.T..
hC3.11 TCCTAGAAATTAAATATGTTATCAGCCATAATGTCATATACCACAGAAGTAATGAACCTTTTATG-TGAGCTGTGTTATCATAATGAATTCTCATCAGAG
..V..L..S..*..
A
500
PSG 3 C---....G.GATG.....------..G...TTCAGAAG.C..G.....G..GC...C.T.....T....A.C....G.....G.....T
hC3.11 TTTTAGCCATAG--TCATTTAAAATCTTTATC-------ATTCAGAGACA-TTATTTTTAGTCTTCCTCAAAGCATTTCCAATCA-CTACA-TCCAAAAA
..I ..I..F..S..L..P..Q..S..I..S..N..H. .Y..I ..Q..K..
A'
600
PSG 1 .....
PSG 3 ..Y..L..* ..
PSG 3 .....TT.G.....-...A..CTC.....G.....G.....C.....C..CA.G..C.....T.---.A.....
hC3.11 TGCTTCCTTTTCAAGGAGATTTATGGAAACGATCGTGACAAGAACTCTTTAATACAAGTTTCTGATAAATTTCACATTATACCACCGACTGTCTAAGAAC
hC3.11 ..D..L..W..K..R..S..*..
hC3.11 C..F..L..F..K..E..I..Y..G..N..D..R..D..K..N..S..L..I..Q..V..S..D..K..F..H..I..I..P..P..T..V..*..|..L
B
B'
700
PSG 1 ...G.....C.....T....C...CC.AG.....-....G..C.....
PSG 3 ..T.....T.....TG.....-C.....G.....AAG------.-...--.--.....A.CCC.A...T....G.C..------....G.C
hC3.11 TTCCAAAACTTTAAGAAACAGGCTGATATCTTCATAAAATTCTCAG--TGTACCAAGCAGGGAAAAATATTGA-TTTCATTGAAATAATTGATAATAATG
hC3.11 ..P..K..F..*..
800
PSG 1 .A.....-.....-.....C
PSG 3 .AA.C....-....C..A.....C.....G.....G...-.T...A..C..C.....C.C.....----------
hC3.11 AGGATAATGTTTTTTATGATTTT-TATTTGAAAATTTGCTGATTCTTTAAATGGGTTTGTTTTCTAGATTTATGGAATTTTTTTCTTTTAATCTATCTAT
C
900
PSG 1 ..AATAAA.....CG.A.....T.....
PSG 3 ------C....A..-GG....G.....T.CTA.....AA.....G.....C...AG..----.....A.....G.
hC3.11 AGCTTATAGCAGTTCAATAAACTATACTTCTGGGAACAATTATTGAAACATTTACTTTTGCTCTCTACCTGACTGCCCCAGAATTGGGCAACTATTCATG
hC3.11 ..Y..S..S..S..I..N..Y..T..S..G..N..N..Y..*.
1000
PSG 1 .....C.....A.....T.....G.....T.....G.....
PSG 3 ..T...C.....AATAAA.T...C.....T..
hC3.11 AGAATTGATATGTTTATGGTAATACAGATATTTGCACAAGTACAGTAACAGTCTGCTCTCTTTTTAACACGACACATTTCAAATCATTGGTTATATTACC
1100
PSG 1 ..T.....T.C.....-.AAAACAT.....-.....G.....A..
hC3.11 AAGGCTTTGACTGGGATGTTATATTCA-------AGATAGAATGAACCAATATGAACTGCAGGGCAAAGTCTGAAGTCAACCTTGGTTTGGCTTCCTGTT
1200
PSG 1 .....G..GA.....C.....A.....G....C.....
hC3.11 CTCAAGAGGTTTGTAAAAGTTTAATCTGAGATTCCTTATAAAAACTTAGAGCAAAGAAAATTTTAAAAGAGAGCCTACATGGTCCATTGCTACTCTTGCT
1300
PSG 1 .....C.....CA.....A.....C.....G.....
hC3.11 GCACTTATGTAAAGAATCAGACCATGTTTGAAGTAACTCAACCTATTTTGCAAACAAACTTATTCTACTGAAATTATCATTGGTAAAACTAGAGATGCCC
1400
PSG 1 .
PSG 1 .....G.....AATAAA.....T...T.....C.....
hC3.11 ATAGAGAGAAAAATTATGTGGAAAATAAAAACTGTAGTACACCGGTTATGAGATTGCAGCTCTGTTCATTGTTTCTGTGTTTTTATTATCCACCTGTAGA
hC3.11 .
D
1500
PSG 1 .W..T..V..P..*..
PSG 1 .....G.T.....A.....
hC3.11 CTGGACATTACCCTGAATTCTACTAGTTCCTCCAATTCCATTTTCTCCCATGGAATCACTAAGAGCAAGACCCACTCTGTTCCAGAAGCCCTATAAGCTG
hC3.11 .W..T..L..P..*..
..A..L..*..

```
                1600
PSG 1           ...G.....................................G...A.CA.....................A......A.............ATG
hC3.11          GAGTTGGACAACTCAATGTAAATTTCATGGGAAAACCCTTGTACCTGACGTGTGAGCCACTCAGAACTCACCAAAATGTTCGACGCCATAACAACAGCCA

                1700
PSG 1           ..................ACA......G.....................................................................C.
hC3.11          CTCAAACTGTAAACCAGG---ACAAGTTGATGACTTCACACTGTGGACAGTTTTTCCCAAGATGTCAGAACAAGACTCCCCATCATGATGAGGCTCTCAC

                1800
PSG 1           ..................C......G...C.....................................C.................A....C.......
hC3.11          CCCTCTTAACTGTCCTTGCTGATGCCTACCTTTTTCACTTGGCAGGATAATGCAGTCATTAGAATTTCATATGTAGTAGCTTCTGAGGGTAATAACAGAG

                1900
PSG 1           ................................CA.....................................................................C....
hC3.11          TGTCAGATATGTCATCTCAACCTCAAACTTTTATGTAACATCTCAGGGGGAAATGTGGCTCTCTCCACCTTGCATACAGGGCTCCCAATAGAAATGAACA

                2000
PSG 1           ....................................G...G....CG..................A.......C.........C.............
hC3.11          CAGAGATATTGCCTGTGTGTTTGCAGAGAAGATGGTTTCTATAAAGAG-TAGGAAAGCTGAAATTATAGTAGAGTCTCCTTTAAATGCACATTGTGTGGA

                2100
PSG 1           .......TG..G.........................A......G..--.........................C.....................
hC3.11          TGGCTCTCACCATTTCCTAAGAGATACATTGTAAAACGTGGCAGTAATACTGATTCTAGCAGAATAAAACATGTACCACATTTGCTAATACTGTTCTCTT

                2200
PSG 1           ........................G.....................-.....................
hC3.11          AAAATAATTTTAAAAGAATGGGGTGAGCCCTCCCATGTGTCCAGGCCAGGTCTCTGAACAGAATCCTCCATCTGCAGTAACAATGCCTAAGAAGATGACA

                2300
hC3.11          TGGACTTGGTCCCGATATGCAGCCATTCCTGTATACCCTTCCCCTGCTGCAGGGCCGTACCATCCCAGGGCCCAAATCTTCAGCTGCAGAGCTGACAGAG

                2400
hC3.11          AACATGGGACACCCAGCATCCCTTACCTTCTTCCAATCCACTGCAGTGGCTACCGGCATGGCCCATTTATCCCTGAGGACACCCATCTGCTGACCCACAG

                2500
hC3.11          TTTCTAAGAGTCAGACTTTCCCGGCTTCTCTGAGCCCCAGCTACTTTCACTCTGCTGAACCCTTCTTCTCCCCACAGGTGTCATTGACTTAGCAGACACC
hC3.11                                                                                    ..F..I..Y..L..A..D..T..
                                                                                          F

                2600
hC3.11          TCTTTCAGCTGCAGCTAACAGGTAAGCCAAGACCCAGACCCCAGAGGATAAACAAGGATTTCAAAACCTACTGTGTGCAATGGAGATGCCCACTTGTGGG
hC3.11          S..F..S..C..S..*..

                2700
hC3.11          CGGCAGGACACCCAGTTGCCGAGGCAAGAGACTGAGTGCACGAGCCTTTCCAGTACAATAAAATATAAAAGAAGAATAGTTATACCAGATATAGATCTTA


                2800
hC3.11          GATATGATTATATATGAATATCATTAATCATTAATTGGTAGCAATTACTCTTTATTCCAATATTATAATAATCTTTGGTCTATAATCATAACCTAGGAAA

                2900
hC3.11          AGCCAGGCCATACAGAGATAGGAGCTGAGGAGACATAGTGAGAAGTGACCAGAAGACAAGAGTGCGAGCCTTCTGTTATGCCCAGATAGGGCCACTAGAG

                3000
hC3.11          GTCTCCTTGGTCTAGTGGTAACGCCAGGGTCTGGGAAGATGCCGGTTTCCAGGCGAACCATGATGTAGTGGTAGCCTCAGTGTCAAGGAAAAACACCCAC

                3100
hC3.11          TACTTAGCAGACTGGGAAAGGGAGTCTCCCTTTCCCCGGGGGAGTTTAGAGAAGACTCTGCTCCTCCACCTCTTGTGGAGGGCTTGACATTAGTCAGGTC

                3200
hC3.11          GACCCGCAGTTATTCAGATGCCTAACTGTCTCCCTGTGATGCTGTGCTTCAGTGGTCACACTCCTAGTCCGCCTTCATGTTCCATCCTGTACACGTGGCT

                3300
hC3.11          CTGCGTTTAGTTAGCAGTAGGAAATTAGCGAATGTACTAAAAGTCTGTAATAAGCAGAAATAATGGTGTAAGCTCTCTCTCTCTTTCTCCTCTCTCTCTC
                                                                                                                G

                3400
hC3.11          TGCCTCAGCTTCCAGGCAGGAAAGGGCCCTCTGTACAGTGGACATGTGACACATGTGGCCTTACCTATCAATTGGAGATGGCTCACACTCCTTATCCTGC
hC3.11                  ..S..R..Q..E..R..A..L..C..T..V..D..M..*..

                3500
hC3.11          CCCTTTGTCTAGTATCCAATAAATATCAGCACAGCCTGGCATTCGGGGCCACCACTGGTCTCCACATCTTAGTGGTAGTGGTCCCCCAACCCAGTTGTCT

                3600
hC3.11          TTTCTTTTATCTCTTTGTCTTGTGTCTTTGTTTCTACAATCTCTCATCTCTGCACACGGGGAGAAAAGCCCACCGACTCTGTGGGGCTGGTCCCTACACC

                3700
hC3.11          CACTGTACAGAGACATAAAGAAGTTGAGATGTATACAACTCCCCCAACAACACTGTATCACAAAACAATGTCCTCTCTGCCCCTCATCGTGGAATTC
```

C.3: RESULTS FOR THE COMBINED hC3.11 SEQUENCE

The combined hC3.11 sequence shown in Figure 10 is aligned against C-domain and 3'-untranslated region sequences from representative subgroup-1 (PSG-1), and subgroup-2 (PSG-3) PSG genes. As can be seen from the comparisons, there is extensive homology between the respective sequences. Since the PSG are known to select alternately spliced exons from their C/3'-untranslated regions, the hC3.11 sequence was examined for features such as splice acceptor sequences, polyadenylation addition sites, and C-terminal open reading frames.

When compared to other sequences on the GenBank and EMBL databases using a FASTA search, the combined hC3.11 sequence showed similarity to many PSG sequences, although the best matches were with those PSG genes belonging to subgroup-1 (ie: PSG-1,-4,-7,-8).

In an attempt to classify the PSG sequence within hC3.11, the coding potential of the 1.5 kb *Bam*HI/*Eco*RI sequence was examined for open reading frames, potential splice donor and acceptor sites, poly(A) addition consensus signals and similarities to other PSG cDNA and gene sequences.

The hC3.11 C-region sequence was analysed against a number of criteria, and consequently regions of sequence were categorised first into PSG C-domain subgroups (ie: subgroup-1, -2, or -3), and then into groups of alternate transcripts from each of these subgroups (ie: C_a, C_b, C_c, C_d etc.).

Splice acceptor sequences for the C-termini were predicted based on the PSG splice consensus shown in Figure 11.

Figure 11

```
                        TT TTTTTTT
---(PYRIMIDINE RICH)-              TT    NCAG G
                        CC CCCCCC
```

The generic splice acceptor consensus proposed by Shapiro and Senapathy {118}.

Overall, nine putative splice sites were predicted from the hC3.11 sequence. These sites were positioned at: **A**:450-469 bp, **A'**:524-546 bp, **B**:598-614 bp, **B'**:687-696 bp, **C**:885-902 bp, **D**:1480-1498 bp, **E**:1570-1584 bp, **F**:2559-2576 bp, and **G**:3371-3407 bp.

As would be expected for coding regions, each of these putative splice sites were followed by open reading frames coding for potential PSG C-termini. Assuming a +2 reading frame, as found in other PSG genes, four of the nine predicted PSG C-termini (following splice sites **A**, **C**, **D**, and **E**) were homologous to functional domains previously detected in other PSG transcripts.

The remaining putative exons do not resemble parts of cDNA transcripts reported to date.

Within the 3.8 kb hC3.11 sequence there were five polyadenylation addition consensus signals (AATAAA), located at 915-920 bp, 1433-1437 bp, 2162-2167 bp, 2756-2761 bp, and 3517-3522 bp. Two of these sites, beginning at 915 bp and 1433 bp, correspond to those expected for domains in subgroup-1 genes. The remaining polyadenylation addition sites, although not directly associated with subgroup-1 domains, may indicate exons upstream of these sites have potential to be expressed.

C.3.1: The 500 bp Intron Region

The initial 450 bp of the combined hC3.11 sequence did not contain any significant open reading frames, splice acceptor nor donor sites, suggesting that this region may be part of an intron. Comparison of this region to sequences on the GenBank and EMBL databases using a FASTA search did not yield any good matches to reported sequences.

C.3.2: Previously Unreported Splice Sites

The region of sequence immediately following the initial 450 bp intron region of hC3.11, showed a degree of similarity to C-terminal and 3'-untranslated sequence of the subgroup-2 gene PSG-3. The sequences are aligned beginning at position 455 bp, continuing for 590 bp, to the point at which the PSG-3 cDNA sequence ends at position 1043 bp.

The first potential splice acceptor site, **A**, occurs at position 450-469 bp. The putative C-terminal following this predicted splice site is only 3 amino acids long, typical of the subgroup-1 C_b domains, but shares little sequence similarity to those previously characterised.

Up to the splice site at **A'**, there is only 39% nucleotide similarity to other reported PSG sequences. However, there is a sudden appearance of nucleotide identity to the 3'-untranslated region of the subgroup-2 PSG-3 transcript, from position 540 bp onwards. This sequence identity starts, suprisingly, within the 3'-untranslated region of PSG-3, just 40 bp past the end of the PSG-3 C-domain coding region.

This remarkably sharp transition in sequence similarity would suggest the presence of an intron-exon boundary. The C-terminal regions of subgroup-2 PSG, such as PSG-3, are comprised of at least two exons containing a C-domain and a 3'-untranslated region {11}. Considering this, hC3.11 could be expected to have a splice acceptor site for the 3'-untranslated region exon (3' exon) of a PSG subgroup-2 transcript at this point. Indeed, a potential splice point which resembles the splice consensus sequence of Shapiro and Senapathy {118} is present at the point of sequence convergence.

This second splice acceptor site, **A'**, is predicted at position 524-546 bp. Not suprisingly, the sequence following the splice site in hC3.11 has the potential to encode part of an open reading frame. If the splice occurs at the +2 position of the open reading frame, as with other PSG C-domain exons, a 47 residue C-domain is predicted. Of the three possible reading frames, this codes for the largest protein. This putative protein sequence comprising mostly neutral and hydrophobic residues, could represent the 3' exon of a subgroup-2 gene, a C/3' exon of a potentially membrane spanning PSG, or perhaps both depending on the splicing pattern used. Sequence similarity with PSG-3 extends to position 1043 bp. A potential polyadenylation addition sequence exists at 915-920 bp, although PSG-3 uses one further downstream at 1020-1025 bp.

Typically, the PSG C-termini are only 3-14 amino acids in length {8}, therefore a 47 amino acid long C-domain is unusual. Unlike the hydrophobic M-domain of CEA (27 amino acids in length), PSG C-termini are not usually membrane spanning, although it is possible that some may be linked to cell membranes via phosphotidylinositol linkages {8}. One possible exception is the PSG-11w C-domain, consisting of 81 amino acids it is of a length which could potentially span a cell membrane {31}. Therefore, considering the existence of both large and small PSG C-termini, it is possible the putative 47 amino acid hC3.11 C-domain, if expressed, could represent an intermediate class of PSG C-domain.

The third potential splice acceptor site, **B**, is located at position 598-614 bp. The core acceptor sequence at this putative splice site agrees well with the consensus sequence proposed by Shapiro and Senapathy {118}.

This would suggest that splice site **B** could be potentially active. The seven residue putative C-domain following **B**, does not share sequence homology to any other PSG C-domain previously reported. The sequence predicted is predominantly comprised of hydrophilic amino acids and is of a length typical for PSG C-termini. This could represent another category of alternative PSG C-terminal exons.

Interestingly, an analogous splice site corresponding to **B** exists in the 3'-untranslated region of the PSG-3 cDNA, predicting that the PSG-3 transcript may have more than one splice pattern. In this case, the alternate PSG-3 C-domain (PSG 3') is predicted to consist of only two amino acids.

The fourth splice site predicted from the hC3.11 sequence, **B'**, differs significantly to the core consensus sequence, but shares a degree of similarity to the active subgroup-1 C_a and C_c sites suggesting it may be functional. A putative C-terminus of 5 amino acids is encoded in the region immediately following splice acceptor **B'**. Although the amino acid sequence does not resemble any previously reported C-termini from the subgroup-1 PSG genes, it is of an appropriate length for such an exon, since the subgroup-1 C_a domains are also 5 residues long.

Homology to PSG subgroup-1 (PSG-1) sequence commences abruptly at position 689 bp, which corresponds to a position in the middle of the predicted **B'** splice consensus site.

The eighth splice acceptor site, **F**, located at position 2559-2576 bp, conforms well to the splice consensus sequence described by Shapiro and Senapathy {118}. The sequence available for the PSG-8 and PSG-1 genes did not extend into this region, both finishing at approximately 2200 bp. Hence it was not possible to compare this region to cDNA sequence from other genes. The open reading frame immediately following splice acceptor site F, encodes a putative C-domain comprising 13 residues. A polyadenylation signal is located 180 bp downstream at position 2756-2761 bp, which could potentially be included in a cDNA transcript containing this C-domain.

The nineth, and final splice acceptor site, **G**, is located at position 3371-3407 bp. This particular splice site is preceded by a 29 bp sequence comprising alternating purines and pyrimidines (CT), which overlap to form part of the splice consensus signal. Following splice site **G**, a putative 13 amino acid C-domain, hC3.11g, was predicted from the nucleotide sequence. Approximately 70 bp downstream of the putative C-domain sequence, a polyadenylation site is present which could be utilised in the expression of this protein.

Both the nucleotide sequence, and the deduced amino acid sequences of the putative C-termini hC3.11f and hC3.11g, were compared with sequences on the EMBL and GenBank databases using a FASTA search. Neither of the sequences from the two putative exons matched well with other genes or proteins reported on these databases. Therefore, these putative proteins could represent another subgroup of PSG C-termini, but expression of these domains remains to be established experimentally.

The putative C-domain sequence from hC3.11g and hC3.11h are both 13 residues in length, similar to other PSG C-termini known to be expressed. Both of the open reading frames are predicted to splice at the +2 codon position, as with other PSG C-termini in this region. The amino acids predicted for these C-termini are predominantly neutral or hydrophilic, consistent with the majority of PSG C-termini and their suspected biological role(s).

C.4: ANALYSIS OF THE hC3.11 SEQUENCE

C.4.1: Defining the `Gene-Specific' Sequence

The first 500 bp of the sequence presented in Figure 10 appears unique. It shows no homology to PSG gene sequences or to other sequences reported on the GenBank nucleotide database.

C.4.2: PSG Subgroup-1-like Putative C-termini

Figure 5 shows a schematic diagram of the typical C-domain organisation found in subgroup-1 PSG genes, similar to the hC3.11 sequence.

Identity to subgroup-1 C_d domain 3'-untranslated sequence, was detected following the initial 500 bp intron region in the combined hC3.11 sequence. This sequence similarity commences abruptly at position 689 bp, which corresponds to a position (148 bp) past the boundary of the BII domain in the PSG-1d gene.

Homology to C_d /3'-untranslated sequence continues through to position 941 bp in the hC3.11 sequence. A polyadenylation signal located at 915-920 bp in hC3.11, corresponds to an analogous signal in PSG-1d. Therefore this looks like a functional subgroup-1 C_d exon, however the absence of a BII exon abutting to the sequence would suggest otherwise.

Coincidentally, a putative splice consensus sequence is also predicted at this location, however, there is a 3 bp deletion present in the PSG-3 sequence (Figure 10, nucleotides 687-689) which renders this putative splice acceptor site inoperable.

Consistent with exon organisation found in the PSG subgroup-1 genes, the region of sequence downstream from the C_d 3'-untranslated region, encodes putative C-termini which resemble the subgroup-1 C_c, C_a, and C_b domains.

The fifth putative splice acceptor site, **C**, at position 885-901 bp matched the consensus sequence extremely well, and was almost identical to the functional PSG-8_c C-domain splice consensus sequence.

Following this splice site, the amino acid sequence predicted for the putative C-domain, hC3.11c, clearly classifies it as a subgroup-1 domain. The putative hC3.11c C-domain and the C_c domain of

PSG-1 are both 14 residues long. They share 86% identity at the amino acid level, differing only in the two final residues (assuming the splice donor does not change the first residue). By comparing the nucleotide sequences from hC3.11 to PSG-1_c, the two amino acid differences are found to result from three point mutations.

Sequence comparisons against another subgroup-1 gene, PSG-8, revealed that the PSG-8 gene encodes a 12 residue C_c domain. This is two residues shorter than the PSG-1_c and hC3.11c C-termini, the result of a 5 bp insertion, shifting the reading frame in the PSG-8 gene. (Figure 10, nucleotides 903-944)

Interestingly, a deletion of 19 bp in the corresponding position of the PSG-3 3'-untranslated region disrupts the near identity between the hC3.11 and PSG-3 sequences, precisely removing this potential splice site from PSG-3, preventing alternate splicing of the transcript (Figure 10, nucleotides 886-906).

The hC3.11 sequence contains a polydenylation signal at position 915-920 bp, corresponding to analogous sites found in the PSG-1 gene. This particular polyadenylation signal is located in a favourable position for potential inclusion into an hC3.11_c cDNA transcript.

The sixth putative splice acceptor site, **D**, is located approximately 560 bp further downstream at position 1480-1498 bp. This splice sequence shows a high degree of similarity to the consensus sequence of Shapiro and Senapathy {118} which suggests it may be functional, although this remains to be established experimentally. The open reading frame following this splice site encodes a putative PSG C-domain comprising 5 amino acids. This putative domain is identical to the C_a domain of PSG-8, and strongly resembles the C_a domain of PSG-1, differing only in the penultimate amino acid (See Table 2).

The *Eco*RI site at position 1514 bp demarcates the beginning of the sequence from the 2.2 kb *Eco*RI fragment, from the previously obtained sequence.

The seventh putative splice acceptor sequence, **E**, is located 60 bp further downstream between nucleotides 1570-1584 bp. This splice site matches the splice consensus sequence well, and is almost identical to functional PSG C_b domain splice acceptor sequence. The open reading frame immediately following E encodes a three residue putative C-domain amino acid sequence, identical to subgroup-1 (PSG-8_b and PSG-1_b) C_b domains.

A polyadenylation signal located 570 bp downstream from the end of this open reading frame, at position 2162-2167 bp. In both PSG-8 and PSG-1 genes, analogous polyadenylation signals are present at corresponding positions. Therefore, this signal could potentially be included in an hC3.11e transcript, in the event of expression.

In summary, this region of hC3.11 sequence bears strong resemblance to subgroup-1 type sequence. The high degree of similarity of these putative C-termini to types known to be expressed suggests they could be functional. It will be necessary in future studies to demonstrate the detection of these transcripts before further conclusions can be made.

C.5: DISCUSSION OF THE hC3.11 SEQUENCE

Recent studies have suggested a subgroup-1 gene, PSG-4, is located some distance upstream of PSG-11 {115}. To date, only one C-domain has been confirmed for the PSG-4 gene {77}. However, the hC3.11 sequence does not show identity to the PSG-4 C-terminus sequence.

A significant feature of the hC3.11 sequence, is that it contains what appears to be a subgroup-1 PSG C-domain `cassette', containing the $C_c$, $C_a$, and $C_b$ exons. Based on the organisation of C-domains present in other subgroup-1 PSG genes, a BII-$C_d$ domain exon would be expected immediately upstream of the C-terminal `c-a-b' cassette (See Figure 12).

Interestingly, this is not the case. Although a short region of subgroup-1 C_d 3'-untranslated sequence is present preceding the `c-a-b' cassette, neither a Cd-domain nor a BII domain was detected by hybridisation within the 9 kb lying upstream of the C-domain cluster.

Considering that the distance separating the PSG genes within each gene cluster has been estimated be 6-9 kb, an initial interpretation of the data would suggest that it is unlikely that the C-domain cluster in hC3.11 is linked to a functional PSG gene unit.

Although the hC3.11 C-terminal cluster is the first to be encountered upstream of the PSG-11 gene, it does not appear to represent PSG-4. This is consistent with mapping data, which places PSG-4 well upstream of PSG-11 {115}.

Comparisons were made to other PSG genes including the subgroup-2 gene, PSG-5. When `m'-type C-domain sequence from PSG-5 was compared to the combined hC3.11 sequence no significant homology was detected between hC3.11 and the PSG-5 sequence. This is consistent

with the sequence data which shows that the hC3.11 C-domain cluster resembles subgroup-1 PSG sequence much more strongly than subgroup-2 or subgroup-3 sequence.

C.5.1: Is the Upstream Gene a Subgroup-3 Gene?

All members of the PSG gene family reported to date, conform to the patterns of organisation for their L, N, A and B domains previously shown in Figure 4.

In addition to this, the PSG can be further categorised into three subgroups according to their genomic C-domain organisations. These three organisational structures are shown in Figures 12-14. Initially, these subgroups appear quite distinct.

Concurrent studies in our laboratory of the C-terminal regions of the PSG-11 gene have found that there is an interesting pattern of sequence similarity between the C-domain region of subgroup-1 and subgroup-3 genes. This is particularly interesting since the C-terminal region of the PSG are noted for their remarkable diversity.

When further sequence comparisons were made between members from all three subgroups, a common C-terminal sequence element was detected in all three subgroups.

C.5.1.1: Subgroup-1

The PSG-1, -4, -7, and -8 genes comprise the first category of PSG genes, subgroup-1. A schematic diagram representing the typical subgroup-1 gene arrangement is shown in Figure 12.

In the subgroup-1 genomic arrangement, the BII domain sequences abut the C_d domain sequence. The C_c, C_a, C_b domains are located on separate exons further downstream. A splice site at the BII/C_d junction can be spliced in an alternative manner, to include one of the other downstream C-domains into the transcript, in place of the Cd domain.

Our group has recently proposed, based on studies of the PSG-11 gene that the C_c, C_a, C_b domains from subgroup-1 form a cassette, which is present in the other two subgroup genes although apparently not expressed. The hC3.11 sequence that I report may be a free standing version of this cassette.

C.5.1.2: Subgroup-2

The subgroup-2 genes include PSG-2, -3, and -5. A schematic diagram representing the PSG-5 gene arrangement is shown in Figure 13. As with subgroup-1, a C_d domain is present immediately following the BII domain.

In PSG-5, there is presently approximately 540 bp of sequence available following the end of the BII domain. Approximately 148 bp into this sequence, there is a marked decrease in the degree of homology to subgroup-1 sequence, followed by an area of sequence unique to subgroup-2 genes.

A further 4.3 kb downstream, there is a C_m domain, followed by a 3'-untranslated region. In this untranslated region, homology to the 3'-untranslated sequence from a subgroup-1 C_d domain is again apparent. The homology to subgroup-1 sequence, continues further downstream to an area encoding a C_c domain. This C-domain may be a remnant of a subgroup-1-like `c-a-b' cassette. Since the sequence presently available for this area does not extend beyond the C_c'-like domain, we can only speculate as to the presence of C_a and C_b domains downstream.

C.5.1.3: Subgroup-3

The subgroup-3 genes include PSG-11, -6, and -12. A schematic diagram representing the typical subgroup-3 gene arrangement is shown in Figure 14. As with the first two subgroups, the BII domain is immediately followed by a Cd domain. In PSG-12 a stop codon is present at position 11337-11339 bp terminating the open reading frame, producing a typical C_d domain comprising 12 amino acids.

The PSG-11w cDNA is significantly different to the majority of the PSG with respect to it's C-terminus. A nucleotide substitution occurs at position 849 bp, altering the stop codon normally present at this point to a tyrosine. This allows the open reading frame to be extended to 243 bp in length, resulting in the production of an uncharacteristically long hydrophobic C-domain of 81 residues {31}.

As was the case with subgroup-2 genes, subgroup-3 sequence is also similar to subgroup-1 sequence. This sequence similarity continues past the internal domains into the C-domain region until approximately 148 bp downstream of the BII domain, where similarity decreases markedly. Sequence unique to the subgroup-3 genes continues a further 4 kb downstream. Toward the end of this unique sequence, there are two C-domains present, C_r and C_s, unique to these genes.

Following the C_r and C_s domain sequence, there is an increase in the degree of similarity to subgroup-1 `c-a-b' cassette sequence. The point at which the homology increases, corresponds to the point at which the hC3.11 sequence abruptly resumes similarity to `c-a-b' cassette sequence. However, neither C_r nor C_s domain sequences were detected in the hC3.11 sequence, implying hC3.11 is unlikely to `carry' a subgroup-3 gene.

Considering the absence of an upstream BII domain, the precision of the sequence break points, and the arrangement of the C-domains, it is probable that the hC3.11 sequence represents a free standing version of the `$C_c$, $C_a$, $C_b$ cassette'. The existence of such a cassette could perhaps be explained by genomic rearrangement occurring in the evolutionary history of the PSG family. It is possible that this C-terminal region represents a fragment of the `gene unit' from which the subgroup-1, and perhaps, other PSG subgroup C-termini arose. Since the evolutionary history of the PSG indicates that the genes evolved by duplication and amplification of a common primordial gene unit, this explanation would be consistent with previous evolutionary events.

Alternatively, an event such as double-crossover, resulting in the translocation of this cluster from a functional PSG gene, provides another possible explanation for the existence of this region. In addition, if there were multiple copies of the PSG genes, a corresponding reduction in the structural/functional constraints on the region would presumably allow for increased variability.

Considering the capricious nature of evolutionary events, it is possible that a combination of these factors and events generated this C-domain cluster.

However, these hypotheses must be regarded as speculative until they can be substantiated by evolutionary analysis. Therefore, further sequencing and analysis of genomic PSG C-domain regions are required to clarify the history concerning the evolution of these areas in PSG.

Figure 12: The Subgroup-1 Arrangement.

Schematic diagram of the typical subgroup-1 C-domain arrangement. C-domains are labelled in the corresponding boxes, and are selected individually for inclusion into cDNA transcripts. Not to scale. (See section C.5.1.1 for text).

Figure 13: The Subgroup-2 Arrangement.

Schematic diagram of the arrangement of the C-domain region in the subgroup-2 gene PSG-5. C-domains are selected individually for inclusion into cDNA transcripts. Dashed boxes indicate the C-domain is present, but expression has not been detected. Dashed lines indicate sequence unique to subgroup-2 genes. Not to scale. (See section C.5.1.2 for text).

Figure 14: The Subgroup-3 Arrangement.

Schematic diagram of the typical subgroup-3 C-domain arrangement. C-domains are selected individually for inclusion into cDNA transcripts. Dashed lines indicate sequence unique to subgroup-3genes. Not to scale. (See section C.5.1.3 for text).

FIGURE 12: THE SUBGROUP-1 ARRANGEMENT

The hC3.11 sequence.

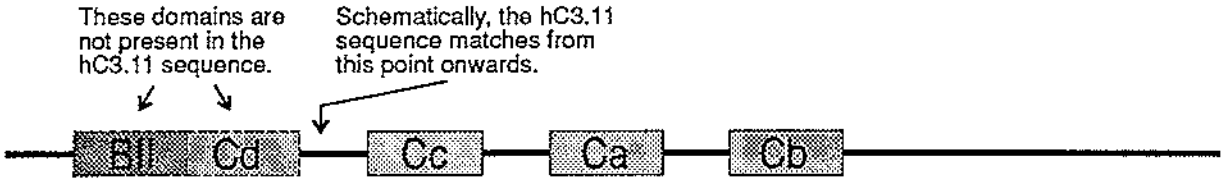


FIGURE 13: THE SUBGROUP-2 ARRANGEMENT

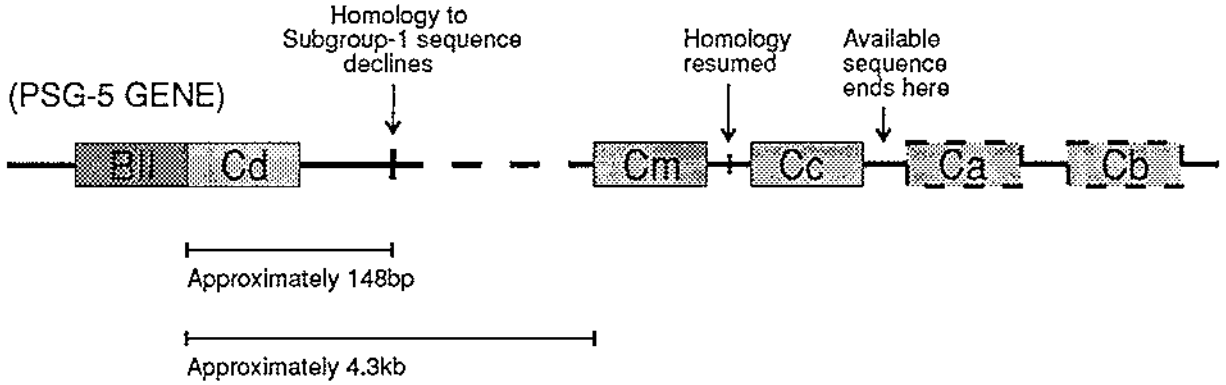
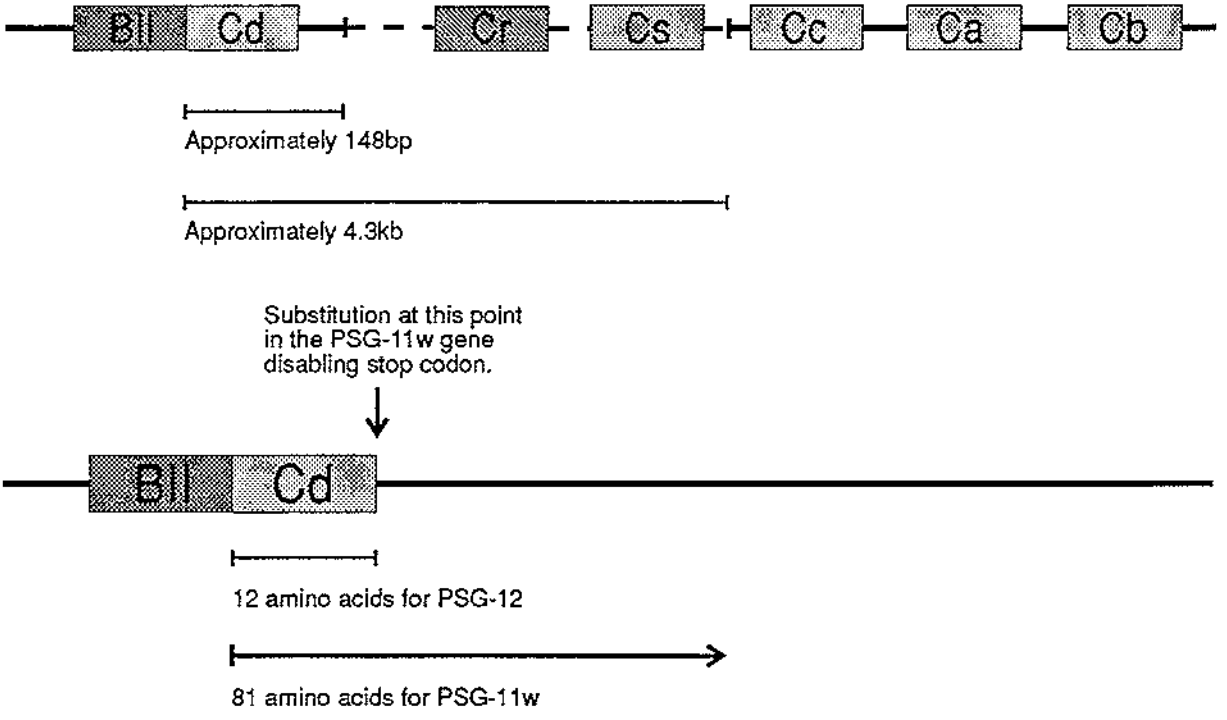


FIGURE 14: THE SUBGROUP-3 ARRANGEMENT



D: RESULTS AND DISCUSSION FOR THE COSMID CLONES

D.1: Defining the Clones

D.1.1: The Clone hC3.11

The PSG-like C-domains in hC3.11 identified in section C.4.2, suggested there may be associated PSG domains further upstream of this region.

The PSG genes are usually organised in sequential domain arrangements as shown in Figure 3. Since PSG C-domain clusters are located at the 3' end in typical genomic arrangements, it was suspected a novel PSG gene was present upstream of the hC3.11 C-domain cluster.

To investigate the existence of such a gene, the region upstream of PSG-11 was examined, using chromosome walking techniques, to isolate overlapping genomic clones.

D.2: THE PROBES

D.2.1: The `Gene-Specific' Probe

A 500 bp *Bam*HI/*Eco*RI fragment (Figure 7, pBE.5) lying immediately upstream of the putative C-domains, had previously been sequenced and shown to lack homology to any reported sequence. Therefore, it was assumed this sequence contained a PSG intron, and as such it was selected for use as a gene-specific probe.

A `shotgun' cloning strategy was chosen to subclone the 500 bp *Bam*HI/*Eco*RI fragment from pG3B8.5 because the 500 bp fragment represented such a small portion of the 8.5 kb clone.

The pG3B8.5 subclone (see Figure 7) contains three internal *Eco*RI restriction sites bordered by the two *Bam*HI restriction sites. To clone the 500 base pair fragment, the pG3B8.5 construct was digested to completion using restriction enzymes *Bam*HI and *Eco*RI. This digest resulted in two *Eco*RI fragments (1 kb, 2.2 kb), two *Bam*HI/*Eco*RI fragments (500 bp, 5.5 kb), and the pGEM2 vector with one *Bam*HI and one *Eco*RI end.

Following recovery of the digested DNA by ethanol precipitation, the fragments were ligated back together. The rationale was that the pGEM2 vector fragment would ligate to the two B*am*HI/E*co*RI fragments, and thus recover the 500 bp fragment of interest in the pGEM2 vector. Competent *E. coli* DH-1 cells were transformed with the ligation mix using the $CaCl_2$ method, and twelve transformants were selected.

To determine which of these clones contained the 500 bp insert, DNA was prepared from the transformants by the rapid boil method and digested with H*ind*III and S*ph*I.

Of the twelve clones selected, four contained the 500 bp B*am*HI/E*co*RI fragment. The other clones were recyclised vector, some of which contained the remaining fragments of pG3B8.5. One of the clones containing the 500 bp B*am*HI/E*co*RI fragment, designated pBE.5, was selected for further use and a large scale plasmid preparation made. This was the `gene-specific' probe, pBE.5.

D.2.2: The PSG-Domain Probe

To ensure that clones hybridising to the gene specific probe were also PSG-like, a PSG-11s cDNA probe (pBB5) was chosen for use in differential hybridisation. The PSG-11s cDNA, is a full length copy of the PSG-11s clone containing the L, N, AI, AII, BII and Cs domains. This probe will hybridise via the L, N, A and B domains, to all members of the PSG family, given the high degree of sequence conservation in these regions.

D.3: Isolation of Clones #1 and #4

When the PSG-enriched genomic library (Section B.8) was screened with both the PSG-cDNA probe and the gene-specific probe, 88% of the plaques cross-hybridised to the PSG-11 probe as expected. Of these, eleven clones also hybridised to the gene-specific probe and were selected for further study. To purify these eleven clones, each plaque was eluted into SM buffer, titred, and subsequently plated at low density on L-agarose plates. Plaque lifts were made with nitrocellulose filters, which were hybridised to the gene-specific probe to identify clones of interest.

From this second round of hybridisation, four phage clones were selected for further study, and DNA was prepared from them using the medium scale DNA preparation method. These clones were designated phage #1, #2, #3, and #4.

Restriction digests of the four cosmid clones revealed that clones #1 and #4 were distinctly different, whereas clones #2 and #3 appeared by their restriction patterns to be mixtures of clones of #1 and #4. Subsequently, only phage clones #1 and #4 were pursued further, with no further attempts to purify clones #2 and #3.

To confirm the results of plaque hybridisation, digests of clones #1 and #4 were transferred onto nitrocellulose filters by vacuum transfer, then probed with both the PSG cDNA probe and the gene-specific probe.

While clone #4 hybridised to both of the probes, clone #1 failed to hybridise to either of the probes. Therefore, in the case of clone #1, it was thought that the incorrect plaque had been picked. In order to isolate another cosmid clone for analysis, the first round lysate of phage #1 was plated again at low density for plaque purification, and the hybridisation experiments were repeated. From this plate four clones were selected and these were designated clones #1a, #1b, #1c, and #1d.

Restriction digests of these four new clones revealed they were identical as would be expected, but that they were different to clone #4. Clone #1d was designated clone #1, and was used in subsequent experiments

D.3.1: The Lorist Vector Clones

While this work was in progress, Dr A. Olsen from Los Alamos National Laboratories, CA, USA, provided our group with two Lorist cosmid clones, designated F11193 and F20478, which were believed to contain the PSG-11 locus. These cosmids were subsequently shown by P. McLenachan {45} to contain a complete PSG-11 gene, and a substantial region of upstream sequence.

Restriction maps for clones #1, #4, hC3.11, and the Lorist vector clones F11193 and F20478 are shown in Figure 15.

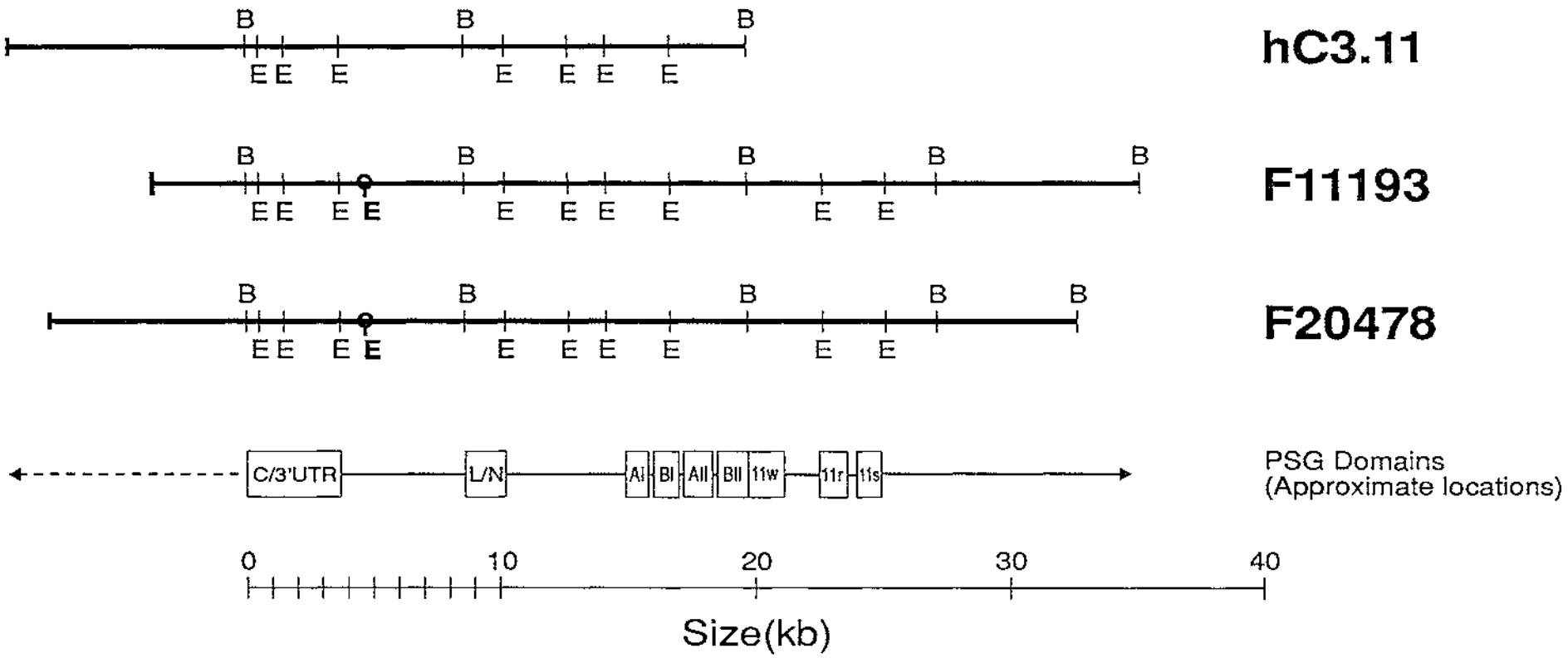


FIGURE 15: Comparison of the Cosmid Clones used in this Study.

Cosmid hC3.11 is aligned with Lorist Vector Clones F11193 and F20478.
Restriction enzyme sites are labelled BamHI (B) and EcoRI (E).
An allelic EcoRI site is marked (⚲).

D.4: RESULTS FROM SOUTHERN HYBRIDISATION AND RESTRICTION MAPPING

D.4.1: Hybridisation Analysis of Clones #1 and #4

To map both of the cosmid clones #1 and #4, restriction digests were performed using the enzymes *Eco*RI, *Bam*HI, and *Hind*III. These complete digests were divided in half and loaded onto two preparative agarose gels, giving duplicate sets of lanes for each clone. Following electrophoresis, the DNA was transferred onto nylon or nitrocellulose membranes by Southern blotting to give a set of identical filters for each clone. One filter was then hybridised with the PSG cDNA probe, and the other with the gene-specific probe. From this experiment, hybridising fragments were identified and correlated with the restriction maps. The results are shown in Figures 16 and 17.

D.4.1.1: Hybridisation to Cosmid Vector Arms

Both the clones #1 and #4, were isolated from a library constructed using the vector EMBL3. Digestion of the EMBL3 vector with the restriction enzyme *Sal*I, results in cleavage at the *Sal*I sites flanking the multiple cloning site.

The expected sizes for the left and right vector arms of EMBL3 are 19.9 kb and 8.8 kb respectively.

To ensure that the EMBL3 vector arms did not hybridise to either of the probes used in the hybridisation experiments, *Sal*I digests were performed on each of the clones, releasing both vector arms from the insert fragment. Hybridisation experiments were subsequently performed on these digests using both the `gene-specific' and PSG-probes. The results are shown in Figure 16b & 16c, lane 3; and Figure 17b & 17c, lane 3.

In clone #1, two fragments from the *Sal*I digest hybridised to the PSG-domain probe (Figure 16b: 12.2 kb, and 8 kb). A single 6.5 kb fragment hybrised to the `gene-specific' probe (Figure 16c).

In clone #4, three *Sal*I fragments hybridised to the PSG-domain probe (Figure 17b: 6.6 kb, 5.8 kb, and 4.3 kb), while a single 5.3 kb fragment hybridised to the `gene-specific' probe (Figure 17c). Separate *Bam*HI/*Eco*RI digests of pBE.5 and pG3B8.5 were included in the Southern hybridisation experiments as positive controls for both the `gene-specific' probe and the PSG-domain probe.

Figure 16: Patterns of Hybridisation Exhibited by Clone #1.

a) Photograph of digests of Clone #1 analysed by gel electrophoresis.

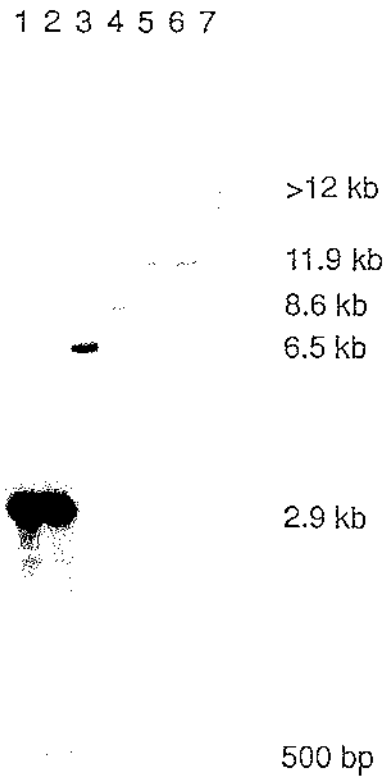
Lane

1: Positive control (pG3B8.5 digested with B*am*HI/E*co*RI).
2: Positive control (pBE.5 digested with B*am*HI/E*co*RI).
3: Clone #1 digested with S*al*I.
4: Clone #1 digested with H*ind*III.
5: Clone #1 digested with E*co*RI.
6: Clone #1 digested with B*am*HI/E*co*RI).
7: Clone #1 digested with B*am*HI.
L: BRL 1 kb DNA ladder.

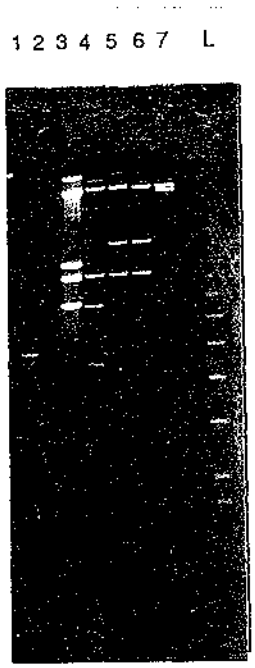
b) Digests of Clone #1 hybridised to the PSG-domain probe. Lanes are identical to those listed in 16a.

c) Digests of Clone #1 hybridised to the `gene-specific' probe. Lanes are identical to those listed in 16a.

(b) Clone #1 hybridised to the PSG-domain probe.



(c) Clone #1 hybridised to the 'gene-specific' probe.



(a) Photograph of digests of

Lane:

1: Positive control (pG3B8.5 digested with BamHI/EcoRI)

2: Positive control (pBE.5 digested with BamHI/EcoRI)

3: Clone #1 digested with Sal1.

4: Clone #1 digested with HindIII.

5: Clone #1 digested with EcoRI.

6: Clone #1 digested with BamHI/EcoRI.

7: Clone #1 digested with BamHI.

L: 1kb DNA ladder (BRL).

Figure 16 : Hybridisation Patterns Exhibited by Clone #1.

Figure 17: Patterns of Hybridisation Exhibited by Clone #4.

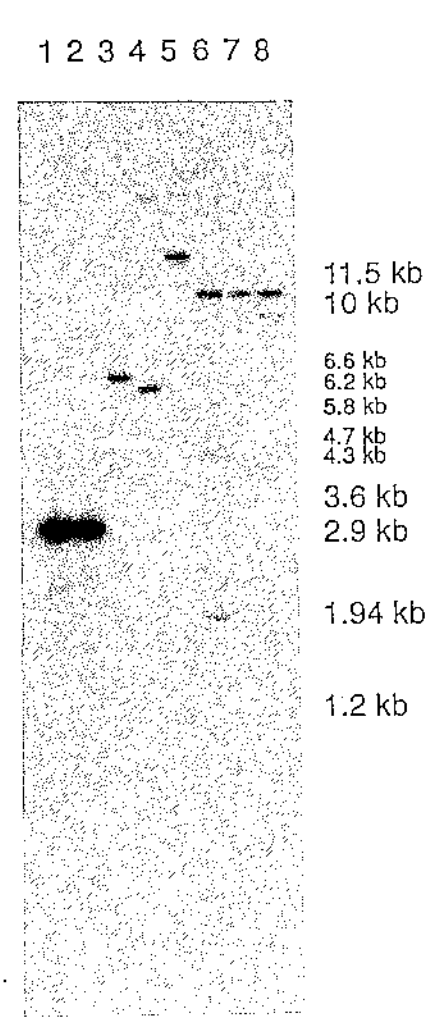
a) Photograph of digests of Clone #4 analysed by gel electrophoresis.

Lane

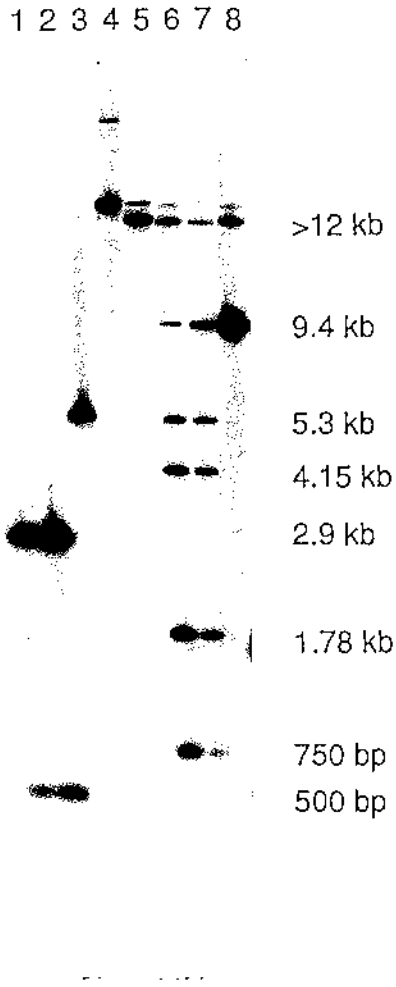
1: Positive control (pG3B8.5 digested with B*am*HI/E*co*RI).
2: Positive control (pBE.5 digested with B*am*HI/E*co*RI).
3: Clone #4 digested with S*al*I.
4: Clone #4 digested with H*ind*III.
5: Clone #4 digested with E*co*RI.
6: Clone #4 digested with B*am*HI/E*co*RI.
7: Duplicate of digest #6, performed independently.
8: Clone #4 digested with B*am*HI.
L: BRL 1 kb DNA ladder.

b) Restriction enzyme digests of Clone #4 hybridised to the PSG-domain probe. Lanes are identical to those listed in 16a.

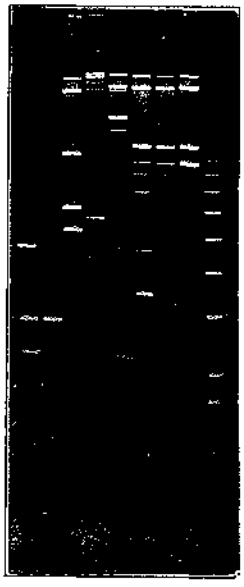
c) Restriction enzyme digests of Clone #4 hybridised to the `gene-specific' probe. Lanes are identical to those listed in 16a.



(b) Clone #4 hybridised to the PSG-domain probe.



(c) Clone #4 hybridised to the 'gene-specific' probe.



(a) Photograph of digests of Clone #4

Lane:

1: Positive control (pG3B8.5 digested with BamHI/EcoRI)
2: Positive control (pBE.5 digested with BamHI/EcoRi)
3: Clone #4 digested with Sal1.
4: Clone #4 digested with HindIII.
5: Clone #4 digested with EcoRI.
6: Clone #4 digested with BamHI/EcoRI.
7: Duplicate of digest 6, performed independently.
8: Clone #4 digested with BamHI.
L: 1 kb DNA ladder (BRL).

Figure 17: Patterns of Hybridisation Exhibited by Clone #4.

D.4.1.2: HYBRIDISATION OF AN N-DOMAIN PROBE TO CLONES #1 AND #4

To determine if a near complete gene unit was present on either of the clones, an N-domain specific oligonucleotide probe was also used. The Lorist vector based clones F11193 and F20478, and the clone phC3.11 containing the PSG-11 N-domain were included as positive controls.

While the `PSG-domain' cDNA probe only defined hybridising areas to be similar to N, A, B, or C type PSG-domains, an N-terminal specific probe was used to further define the nature of the domains in clones #1 and #4.

Restriction digests of cosmid clones #1, #4, F11193, F20478, and hC.311 were transferred to nitrocellulose filters and hybridised to an N-domain specific probe, which had been end labelled with γ-[^{32}P]ATP (see section B.7.4).

In a preliminary experiment, hybridisation of the probe and the washing of the filters was performed at 50°C. This resulted in many extra bands present on the autoradiograph. The temperature was subsequently raised in increments to 55°C, at which temperature consistent, specific hybridisation was observed. At higher temperatures, all of the hybridised probe was washed from the filters. Results from this experiment are shown in Figure 18.

In hC3.11, three fragments hybridised to the N-domain probe. These were: a 1.85 kb *Bam*HI/*Eco*RI fragment, a 6.4 kb *Eco*RI fragment, and a 10.9 kb *Bam*HI fragment.

Digests of the clone F20478 show that a 1.82 kb *Bam*HI/*Eco*RI fragment, a 5.3 kb *Eco*RI fragment, and a 10.9 kb *Bam*HI fragment hybridised to the N-domain probe. The clone F11193 displayed an identical hybridisation pattern as F20478, but with an additional 8.8 kb *Bam*HI fragment also hybridising to the N-terminal probe.

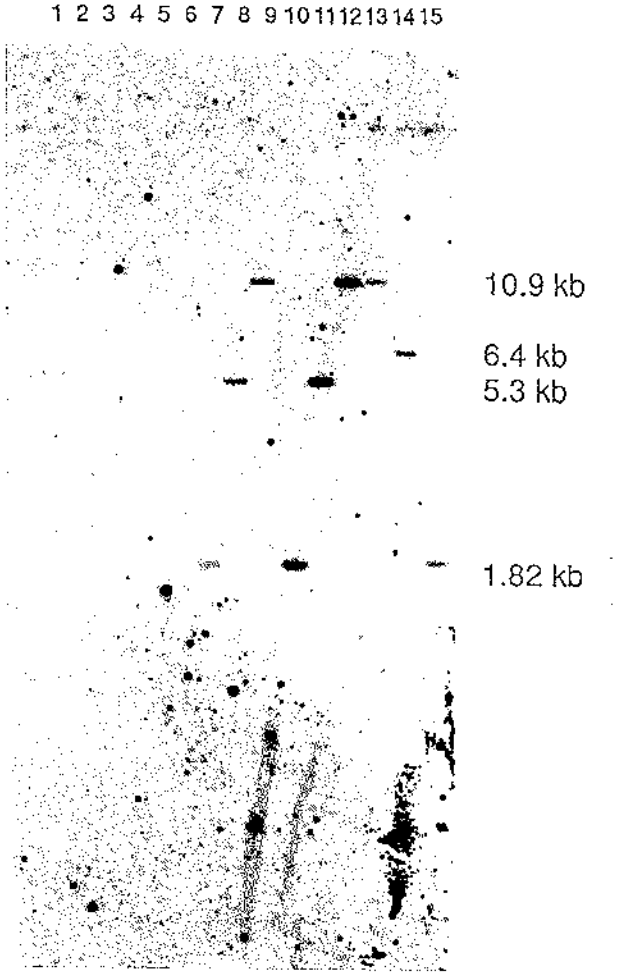
Figure 18: N-terminal Oligonucleotide Hybridisation.

Results of the N-terminal oligonucleotide hybridisation. A photograph of the restriction enzyme digests, analysed by gel electrophoresis, is positioned to the extreme right of the figure. The sequence of the N-terminal oligonucleotide is:

$^{5'}$ GTA CCA GAT GTA GCC AGC AAG $^{3'}$

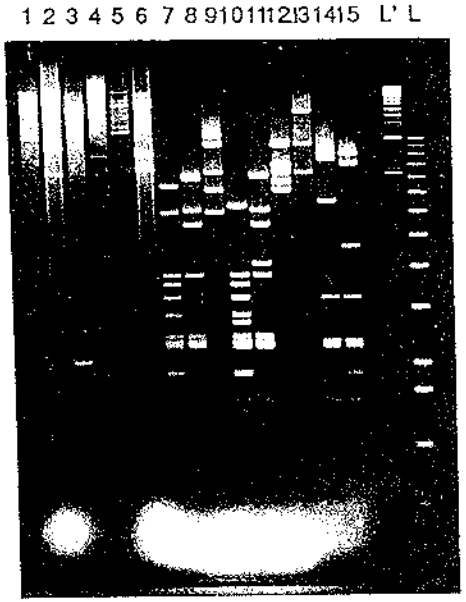
The filter was hybridised at 50°C, with the final wash temperature at 55°C. The lanes are as marked in the key beside the autoradiograph.
(See section D.4.1.2 for text).

Figure 18: *N-terminal Oligonucleotide Hybridisation.*



Lane

1: Clone #1 digested with HindIII.
2: Clone #1 digested with EcoRI.
3: Clone #1 digested with BamHI.
4: Clone #4 digested with HindIII.
5: Clone #4 digested with EcoRI.
6: Clone #4 digested with BamHI.
7: Clone 20478 digested with BamHI/EcoRI.
8: Clone 20478 digested with EcoRI.
9: Clone 20478 digested with BamHI.
10: Clone 11193 digested with BamHI/EcoRI.
11: Clone 11193 digested with EcoRI.
12: Clone 11193 digested with BamHI.
13: Clone hC3.11 digested with BamHI.
14: Clone hC3.11 digested with EcoRi.
15: Clone hC3.11 digested with BamHI/EcoRi.
L': High Mr DNA ladder (BRL).
L: 1kb DNA ladder (BRL).



D.4.1.3: Hybridisation to Genomic DNA

Human genomic DNA was prepared from fresh peripheral blood lymphocytes. A *Bam*HI/*Eco*RI double digest was performed using 20 µg of the genomic DNA, which was subsequently electrophoresed through an agarose gel then Southern blotted onto a nylon filter. The filter was hybridised to a nick-translated pBE.5 probe, to detect genomic *Bam*HI/*Eco*RI fragments which cross-hybridise to this 'gene-specific' probe.

Results from this experiment are shown in Figure 19. *Bam*HI/*Eco*RI genomic DNA fragments of the following sizes hybridised to the 'gene-specific' probe: 8.9 kb, 7.8 kb, 6.4 kb, 5.3 kb, 4.35 kb, 3.2 kb, 2.95 kb, 2.25 kb, 2 kb, 1.62 kb, and 800 bp.

The number of hybridising fragments suggested that the probe was not as unique as was first anticipated. In addition, the expected 500 bp *Bam*HI/*Eco*RI fragment was not apparent. In retrospect, this hybridisation should have been done earlier, although initially the sequence appeared unique when compared to sequence databases (see section C.4.1). Difficulties with cross-hybridisation have been reported by other groups, attempting to use chromosome-walking techniques to isolate PSG members, due to the high degree of sequence conservation within this gene family.

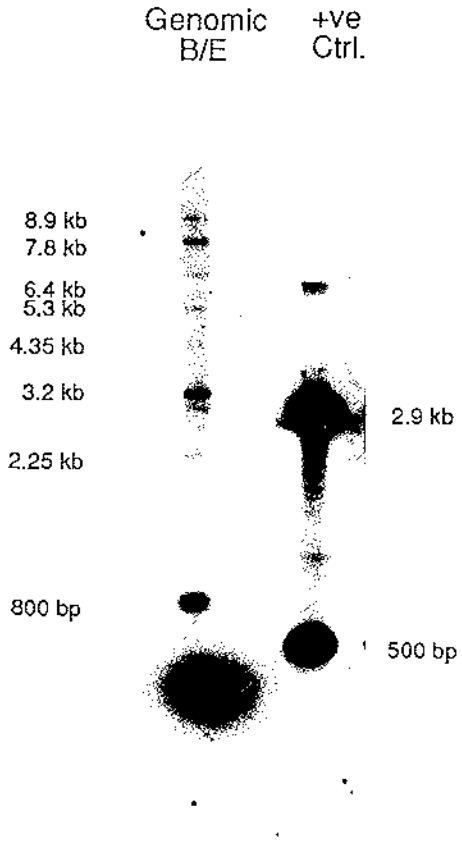


Figure 19: Hybridisation of Genomic DNA to the 'Gene-specific' probe. *Human genomic DNA digested with BamHI/EcoRI, probed with the 'Gene-specific' probe (pBE.5). The positive control lane (+ve ctrl.) is clone pBE.5 digested with BamHI/EcoRI.*

D.4.2: Cos-Mapping Clones #1 and #4

While the experiments above identified hybridising fragments, to utilise the data it was necessary to prepare restriction enzyme maps for the clone(s) of interest. Consequently, a cosmid mapping system was used to construct enzyme maps for clone #1 and #4. Results from a typical Cos-mapping experiment are shown in Figure 20 which are summarised in Figure 21.

The cosmid mapping system relied on the accurate measurement of DNA fragment sizes following gel electrophoresis. Since it is difficult to size large DNA fragments accurately by gel electrophoresis, the accumulation of these inaccuracies can lead to discrepencies in the construction of restriction enzyme maps. Such inaccuracies were incurred in the calculation of restriction maps for clones #1 and #4.

However, despite the differences estimating total insert size, the combined restriction maps for each of the clones were found to be internally consistent by both multiple enzyme digests and hybridisation analysis.

To allow meaningful comparison of the three maps for each clone, the length of each cosmid map was standardised by empirical means. Of the three restriction maps for each clone, the *Hind*III map and the *Eco*RI map were assumed to be the most accurate for clone #1, and #4 respectively. The rationale for this is that these particular enzymes generate the smallest, and therefore most accurately sized fragments in each of the clones.

The insert fragment sizes in each of the cosmid maps were converted to standardised lengths by the following equation:

Equation 5

$$\text{Standardised fragment size (kb)} = \frac{\text{Fragment size (kb)}}{\text{Total length of cosmid (kb)}} \times \text{Total length of 'model' map (kb)}$$

Where the model 'maps' were *Hind*III in the case of clone #1, and *Eco*RI in clone #4.

The combined hybridisation and Cos-mapping data on the standardised maps are shown in Figures 22 and 23.

Figure 20: A Typical Cos-mapping Gel

The molecular weight markers are labelled in kb, and lanes are as follows:

Lane

1: Clone #4 digested with *Hind*III, hybridised with ON-R.
2: Clone #1 digested with *Hind*III, hybridised with ON-R.
3: Clone #4 digested with B*am*HI, hybridised with ON-R.
4: Clone #1 digested with B*am*HI, hybridised with ON-R.
5: Clone #4 digested with E*co*RI, hybridised with ON-R.
6: Clone #1 digested with E*co*RI, hybridised with ON-R.
7: Clone #4 digested with *Hind*III, hybridised with ON-L.
8: Clone #1 digested with *Hind*III, hybridised with ON-L.
9: Clone #4 digested with B*am*HI, hybridised with ON-L.
10: Clone #1 digested with B*am*HI, hybridised with ON-L.
11: Clone #4 digested with E*co*RI, hybridised with ON-L.
12: Clone #1 digested with E*co*RI, hybridised with ON-L.
L: BRL High Molecular Weight DNA Markers (5618SA) hybridised to both ON-R and ON-L.

(See section D.4.2 for text).

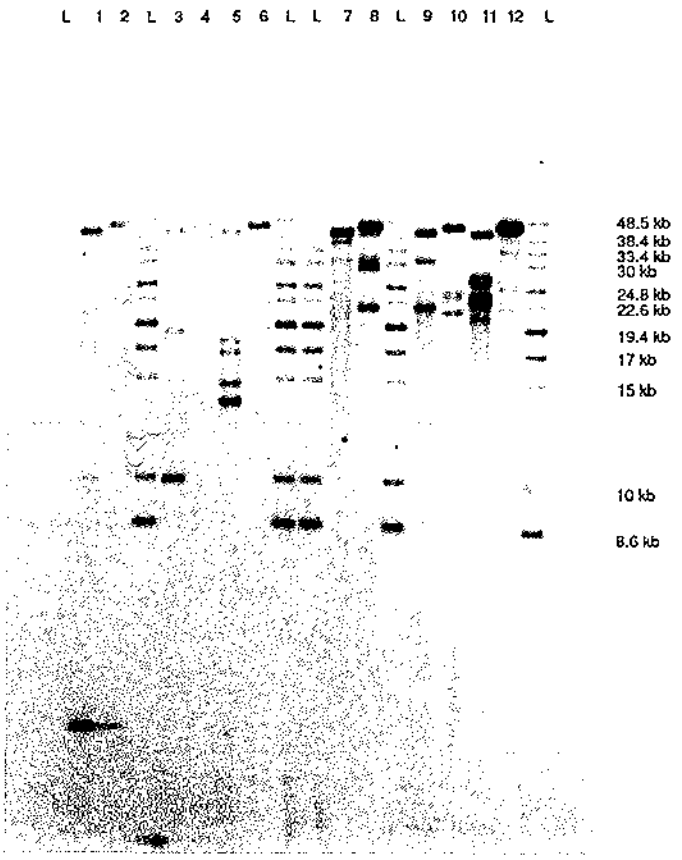


FIGURE 20: A TYPICAL COS-MAPPING GEL.

Figure 21: Summary of Mapping Data For Clone #1 and Clone #4.

Schematic diagrams summarising the Cos-mapping data from Figure 20. Restriction sites are labelled B*am*HI (B), E*co*RI (E), and HindIII (H). (See section D.4.2 for text).

Figure 21a: Restriction maps of Clone #1

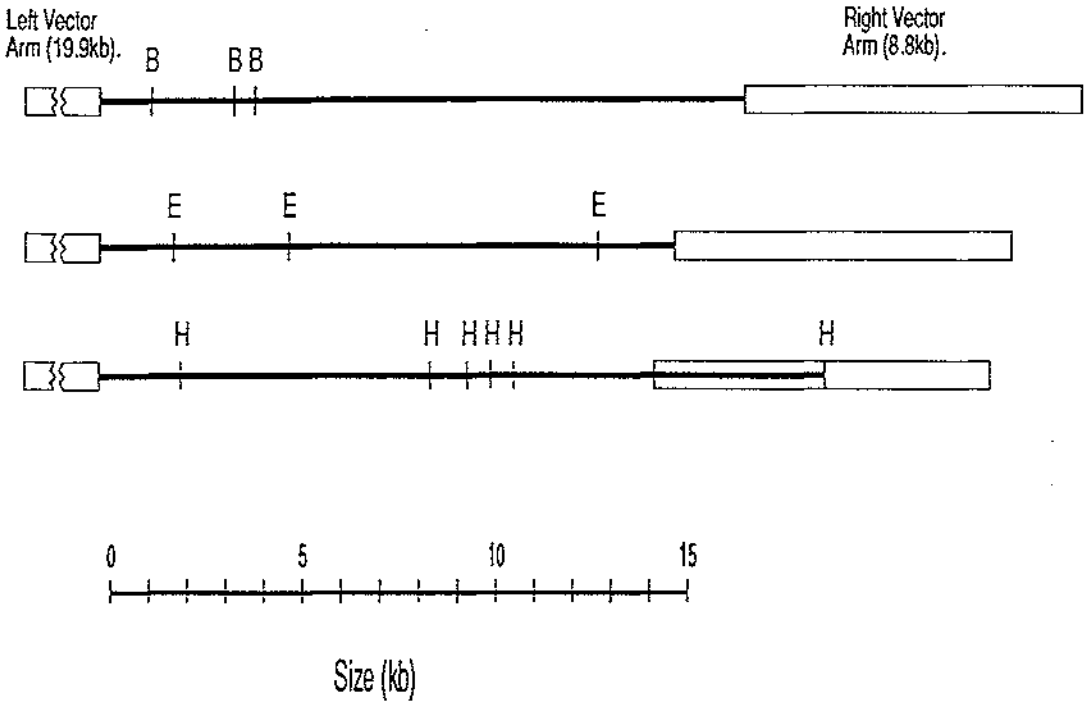


Figure 21b: Restriction maps of Clone #4

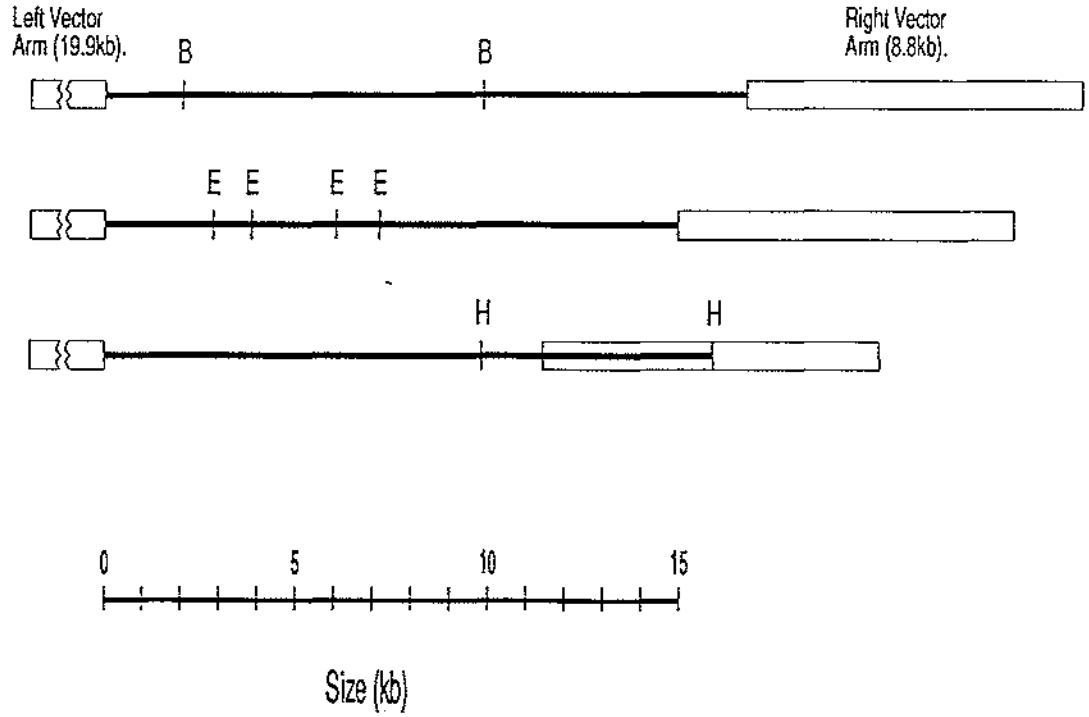


Figure 22: Standardised Maps for Clone #1.

Schematic diagrams summarising the combined mapping and hybridisation data for Clone #1. Restriction enzyme sites are labelled B*am*HI(B), E*co*RI (E), and H*ind*III (H). The fragment lengths have been standardised using Equation #5. (See section D.4.2 for text).

Figure 22: Standardised Maps for Clone #1

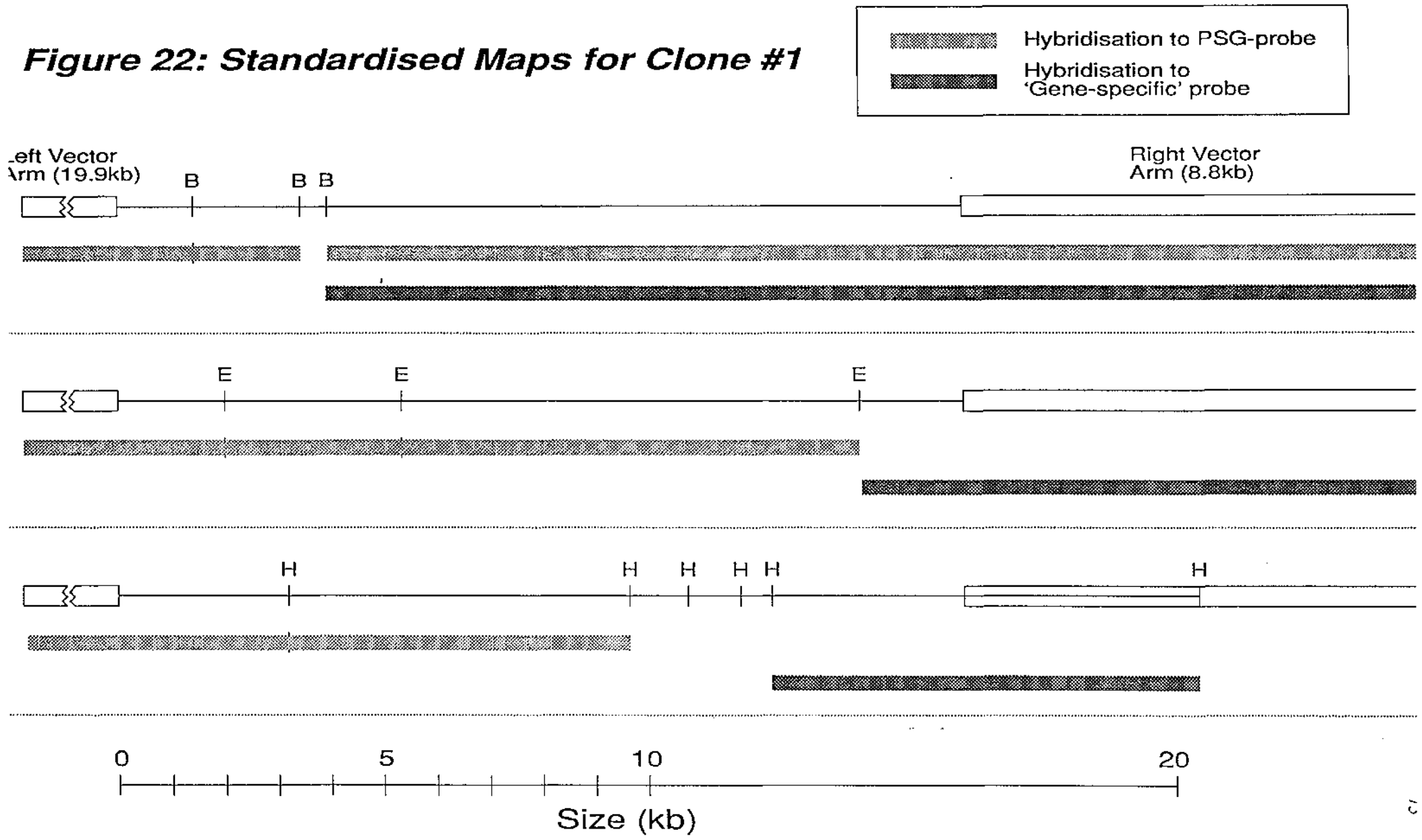
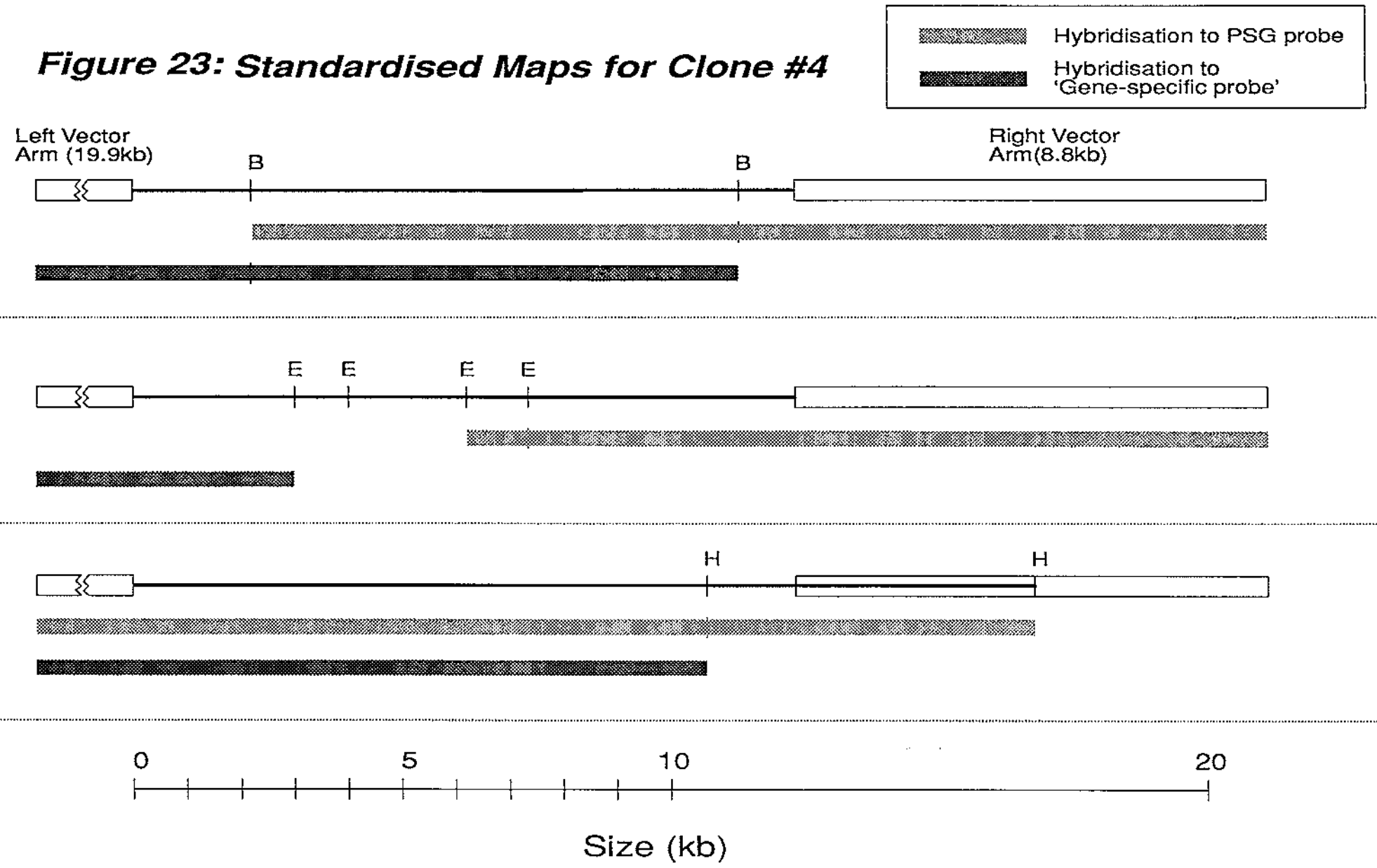


Figure 23: Standardised Maps for Clone #4.

Schematic diagrams summarising the combined mapping and hybridisation data for Clone #4. Restriction enzyme sites are labelled B*am*HI(B), E*co*RI (E), and H*ind*III (H). The fragment lengths have been standardised using Equation #5. (See section D.4.2 for text).

Figure 23: Standardised Maps for Clone #4



D.5: ANALYSIS OF CLONES #1 AND #4

D.5.1: *Cosmid* Clone #1

The combined restriction mapping and hybridisation data for clone #1 are summarised in Figure 22.

The hybridisation patterns displayed by the *Eco*RI and *Bam*HI maps of clone #1, suggest that homology to the PSG-domain probe covers the length of the clone. However, the large 8.8 kb *Eco*RI and 21.5 kb *Bam*HI fragments, are overlapped by the smaller fragments of the *Hind*III map which resolve the region hybridising to the PSG-domain probe to within the first 9 kb of the insert in clone #1.

Of the three maps, the *Eco*RI map provided the finest resolution of the region hybridising to the 'gene-specific' probe. A 10.8 kb *Eco*RI fragment containing the 3' cosmid arm and 2 kb of insert hybridised to the 'gene-specific' probe, whereas the adjacent 8.8 kb *Eco*RI fragment did not, localising the hybridising region to the final 2 kb of the cosmid insert. The hybridisation of an 11.2 kb *Bam*HI/*Eco*RI fragment containing the 8.8 kb 3' cosmid arm and 2 kb of insert, provided further evidence to support this assignment.

D.5.2: Cosmid Clone #4

The assignment of hybridising fragments in clone #4 proved more difficult than for clone #1, because of difficulties in obtaining complete digests of this clone. However, by careful analysis of the partial fragments, when they occurred, three internally consistent restriction maps were obtained. The combined hybridisation and mapping data for clone #4 is summarised in Figure 23.

There were four *Eco*RI sites present in clone #4, evenly distributed through the cosmid insert, that gave good resolution for hybridisation.

Two adjacent *Eco*RI fragments 1.15 kb and 13.8 kb (comprised of 5 kb insert and the 8.8 kb 3' cosmid vector arm), located at the 3' end of the cosmid clone hybridised to the PSG-domain probe. This indicated that PSG-domain-like sequence was present within 6 kb of the 3' end of the insert in clone #4. This conclusion is consistent with the hybridisation patterns observed in both the *Bam*HI and *Hind*III maps.

Figure 17c (lane 5) shows clone #4 digested with *Eco*RI. Only fragments larger than 12 kb cross-hybridised to the `gene-specific' probe. Considering the problems encountered in obtaining consistent digestion of clone #4 with EcoRI, the experiment was repeated to confirm the results. Three separate experiments consistently demonstrated there were no fragments smaller than 12 kb cross-hybridising to the `gene-specific' probe.

It was deduced that one of the larger fragments contained the 19.9 kb 5' cosmid vector arm and 2.2 kb of insert. The other large hybridising fragments probably represent either uncut DNA or undenatured DNA, probably as multimers or concatemers.

Lanes 6 and 7 in Figure 17c show duplicate *Bam*HI/*Eco*RI double digests of clone #4. Seven fragments from these digests cross-hybridised to the `gene-specific' probe. On examination, a number of the fragment sizes were found to be inconsistent with the existing restriction maps for the cosmid clone. However, when these anomalous bands were assessed as products of partial digestion, the data was helpful in locating the sequence cross-hybridising to the `gene-specific' probe in clone #4.

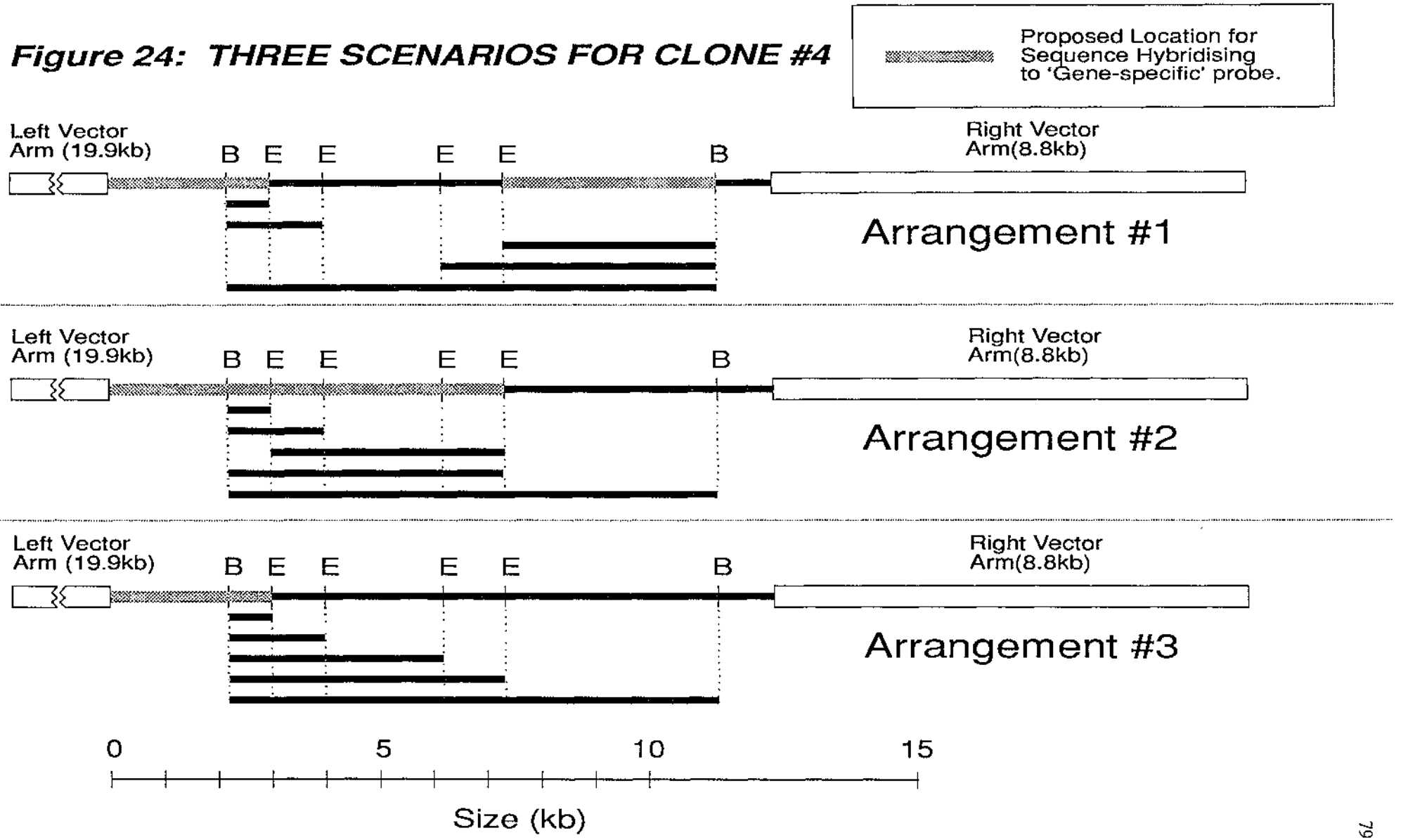
The *Bam*HI/*Eco*RI fragments of clone #4 hybridising to the `gene-specific' probe were the following sizes: 750 bp, 1.78 kb, 4.15 kb
5.5 kb, 9.4 kb, and >12 kb. The fragments larger than 12 kb were assumed to be partially digested, containing the large vector arms.

Three arrangements of these *Bam*HI/*Eco*RI fragments hybridising to the `gene-specific' probe are presented in Figure 24.

Figure 24: Three Scenarios for Clone #4.

Schematic diagrams representing three arrangements of partially digested DNA fragments from Clone #4. The partial fragments are presented as thick bold lines beneath the cosmid clone. Proposed locations of the region(s) cross hybridising to the `gene-specific' probe are presented as grey lines. Restriction enzyme sites are labelled B*am*HI(B), and E*co*RI (E).
(See section D.5.2 for text).

Figure 24: THREE SCENARIOS FOR CLONE #4



The first scenario presumes there are two discrete regions within the clone which hybridise to the `gene-specific' probe. It is predicted that one hybridising area is associated with the 800 bp *Bam*HI/*Eco*RI fragment, and the other with the 4.15 kb *Eco*RI/*Bam*HI fragment. If this arrangement is correct, then the 13.8 kb *Eco*RI fragment containing the 3' cosmid arm and the last 5 kb of the insert (ie: the `second' hybridising region) should hybridise to the `gene-specific' probe. However, this did not occur, despite the 13.8 kb *Eco*RI band being clearly present in the photograph the gel (Figure 17a), therefore the data is inconsistent with this first model.

In the second possible arrangement, also shown in Figure 24, the 4.15 kb and 5.5 kb bands would be located further 5' in the cosmid near the hybridising 800 bp *Bam*HI/*Eco*RI fragment. However, assuming this were the correct organisation, an *Eco*RI fragment of 1 kb, 2.2 kb, or 1.15 kb should hybridise to the `gene-specific' probe, since the 4.15 kb partial fragment which would contain at least one of these fragments. The photograph of the gel (Figure 17a) clearly demonstrates that small fragments of the expected size were present, yet *Eco*RI fragments less than 12 kb did not cross-hybridise to the `gene-specific' probe. Therefore this arrangement is also inconsistant with the data.

The third model presents the most likely scenario, being the most consistent with the combined data. In this model it is assumed that the 1.78 kb, 4.15 kb, 5.5 kb, 9.4 kb fragments (and perhaps those fragments >12 kb) result from partial digestion, and that the cross-hybridising sequence is associated with the 800 bp *Bam*HI/*Eco*RI fragment. In this arrangement, all of the hybridising bands overlap the 800 bp *Bam*HI/*Eco*RI fragment.

Assuming this is the correct arrangement, the experimental data can be interpreted in the following way:

1) The 1.78 kb, 4.15 kb, and 5.3 kb fragments are the result of partial digestion, with cleavage occurring at the first *Bam*HI site, then subsequently at each of the downstream *Eco*RI sites in turn.

2) The 9.4 kb fragment could be a *Bam*HI internal fragment resulting from partial digestion of the clone also overlapping the 800 bp *Bam*HI/*Eco*RI cross-hybridising region.

3) If the first *Bam*HI site was not cleaved during the *Bam*HI/*Eco*RI double digest, the estimated 23 kb *Eco*RI fragment would contain the 5' cosmid arm and the 3 kb of insert to the first *Eco*RI site.

4) The slightly larger 23.3 kb fragment could be a *Bam*HI fragment containing the 5' cosmid arm and 2.2 kb of insert; depending on whether the cross-hybridising region extended upstream of the first *Bam*HI site.

5) The largest fragments may represent uncut DNA, or `conformational artifacts' (eg: concatemers or multimers).

These explanations are consistent with the fragment sizes and hybridisation patterns resulting from the *Bam*HI/*Eco*RI digest.

The combined data from the restriction maps for clone #4 suggest the hybridising region includes the 800 bp *Bam*HI/*Eco*RI fragment, and perhaps extends into the area 5' of this fragment. In conclusion, the data consistently demonstrates that the sequence cross-hybridising to the `gene-specific' probe is located within the initial 3 kb of insert within clone #4.

D.6: DISCUSSION FOR CLONES #1 AND #4

D.6.1: The Cos-Mapping

The lambda terminase system used to map the cosmid clones required the DNA to be partially digested, whereas Southern hybridisation experiments required the sample DNA to be digested to completion. Since difficulties were encountered in obtaining consistently predictable digestion with some of the restriction enzymes, this often complicated analysis of the results and hindered the progress of this investigation.

Analysis of the Cos-mapping results was often complicated by the partial bands being very faint, (see Figure 25); whereas the Southern hybridisation data was sometimes difficult to interpret due to the presence of additional hybridising fragments.

In addition to being the most frequent cutting of the three enzymes used to map clone #4, *Eco*RI was also the most difficult restriction enzyme to attain satisfactory digestion with. Attempts to digest clone #4 to completion often resulted in multiple partial bands. These unexpected bands may have resulted from star activity due to regions of sequence closely resembling the *Eco*RI recognition sequence, or perhaps overdigestion of the cosmid DNA.

Although a consistent standard protocol was used, attempts to partially digest the cosmid DNA for Cos-mapping often resulted in near complete digests. When this occurred, the `partial' fragments were poorly represented in the fragment population, making them very faint or even undetectable. Hence, this initially led to the construction of erroneous maps, since some restriction sites were not detected by the mapping system.
Despite the difficulties encountered in the analysis of the data, the three final restriction maps for each of the two clones, were found to be internally consistent. Moreover, analysis of the partially digested products revealed additional useful information.

A good example of this is the construction of the *Bam*HI/*Eco*RI map for clone #4. Ambiguity in positioning the `gene-specific' cross-hybridising region in clone #4 with respect to the *Bam*HI and *Eco*RI restriction fragments was resolved by simple addition and subtraction of completely digested fragment sizes to match the sizes of corresponding partial fragments. In this manner, the partial fragments revealed which completely digested fragments lay adjacent to each other, thereby determining the order and location of the hybridising fragments.

Due to lengthy difficulties encountered with the Cos-mapping approach, practicality dictated the use of only three enzymes to map each clone. Consequently, further mapping of the two cosmid clones using different restriction enzyme combinations, would enhance the resolution of the restriction maps, and thus increase the accuracy in locating area(s) of interest within each clone.

Restriction maps for clones hC3.11, and the Lorist vector clones F11193 and F20478 are shown in Figure 15. Maps for clones #1 and #4 are shown in Figures 22 and 23.

The clones #1 and #4, hC3.11, and the Lorist vector clones (F11193 and F20478) were isolated from three separate libraries, created from DNA extracted from three different individuals. Considering the heterogeneity of the sources of the human genomic DNA, allelic differences are not unexpected. Nonetheless, a cautious approach is required when attributing anomalous differences to allelic variation, since the possibility of multiple loci existing for particular genes cannot yet be eliminated.

The combined data from restriction mapping, hybridisation experiments and fragments of sequence obtained from the Lorist vector clones, all support the `authenticity' of hC3.11, being representative of the genomic arrangement at this particular locus.

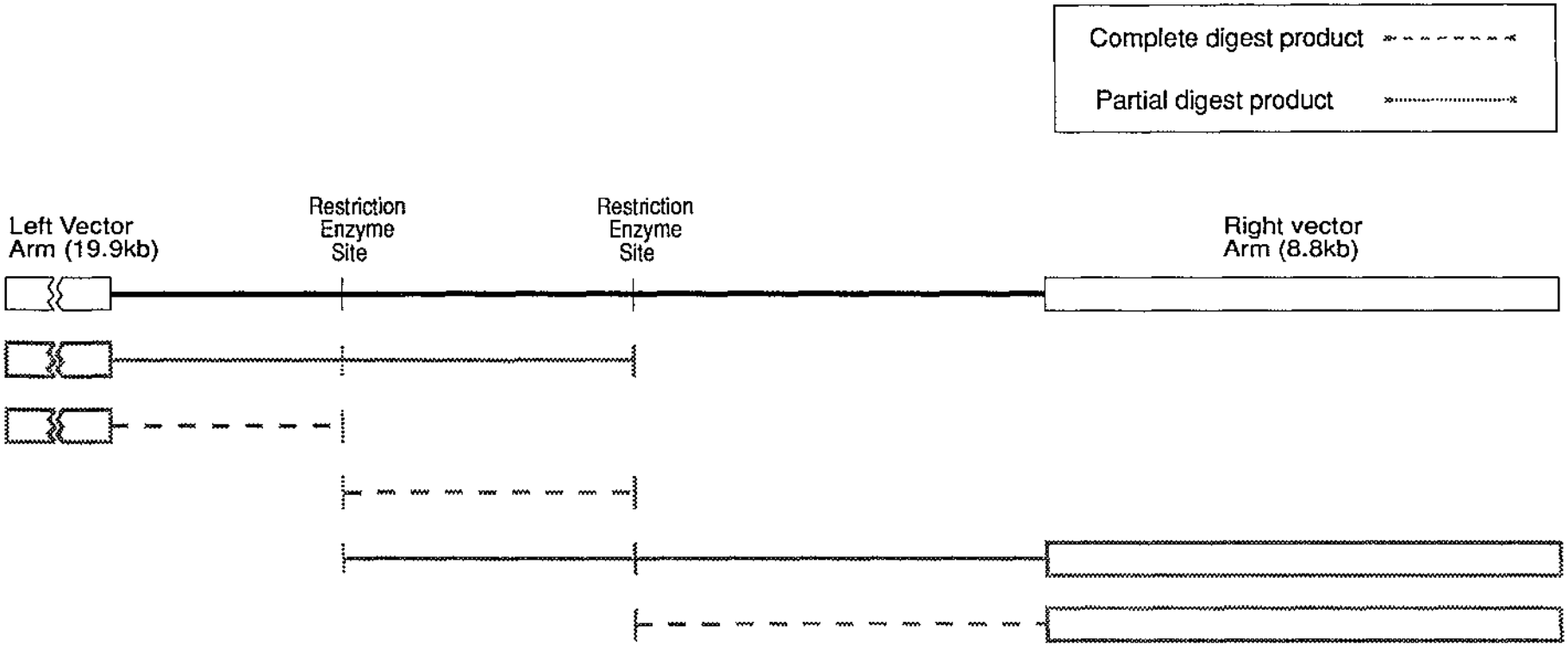


Figure 25: PRODUCTS FROM COS-MAPPING DIGESTS.

In this hypothetical arrangement, only fragments attached to the vector arms would be detected by the ON-R and ON-L probes. If partial fragments are not well represented, internal restriction sites may not be detected.

D.6.2: Clones #1 and #4

Restriction mapping established the distinct identities of clones #1 and #4, subsequently confirmed by hybridisation experiments using both the `gene-specific' and PSG-probes.

The size and order of the restriction fragments in clone #4 resemble those in hC3.11 to a significant degree. For example, a 2.2 kb *Eco*RI fragment flanked by two 1 kb *Eco*RI fragments, is also found in hC3.11 as shown in Figure 7.

However, subtle but significant size differences were apparent between hC3.11 and clone #4. For example, the 800 bp *Bam*HI/*Eco*RI hybridising fragment found in clone #4 is significantly larger than the 500 bp *Bam*HI/*Eco*RI hybridising fragment found in hC3.11. In addition to this, one of the 1 kb fragments flanking the 2.2 kb *Bam*HI fragment was estimated to be 1.15 kb. The latter difference may simply be due to the small inaccuracy in estimating fragment size, but the former difference is clear and marked.

Interestingly, an additional *Eco*RI restriction site was present in clone #4 corresponding to a site found in both the Lorist clones F11193 and F20478. This particular site is not present in the hC3.11 clone, which may indicate an allelic difference at this particular locus.

The restriction patterns in clone #1 did not resemble those found in F11193, F20478, or hC3.11. Moreover, the patterns of hybridisation to both the PSG-probe and the `gene-specific' probe differed from those displayed by either hC3.11 or the Lorist clones. This strongly suggests that clone #1 overlaps neither hC3.11, F20478, nor F11193. Since it appears that the cosmid clone #1 does not overlap hC3.11, the insert of clone #1 probably represents a fragment from another PSG-like gene.

The PSG-11s cDNA probe, BB5, was initially used to screen the genomic library to enrich for PSG-like members. Because this PSG-domain probe contains highly conserved L, N, A and B domains, cross-hybridising CEA related family members were not excluded from the enriched library. Indeed, extensive cross-hybridisation of the `gene-specific' probe is displayed in the genomic hybridisation experiment (see Figure 19).

Other research groups investigating the PSG have also reported difficulties in chromosome walking due to the extensive cross-hybridisation amongst the PSG family members, and even across the related gene families (ie: CEA, NCA, BGP etc).

A study by Rudert et al. {83} has observed that intron and 3'-untranslated sequences in CEA, are more highly conserved than the structural domains. It is suggested that the selective conservation of certain intron regions may reflect unknown functional constraints on these regions.

Other studies have demonstrated the conservation of functional elements in both intron {126} and 3'-untranslated regions {127,128}. The clustering of the CEA family genes on chromosome 19, and the occurrence of simple sequences that could serve as potential recombination sites between non-homologous genes {129,130} in the introns of several CEA related genes {20} support this assumption.

Such putative recombination sites could also facilitate exon shuffling {131}, whereby genes varying in the number of repeated domains may have arisen, accounting for the domain arrangements in NCA, PSG and CEA. The selective conservation of certain intron regions may reflect unknown functional constraints, contributing to the degree of sequence similarity observed in these regions. Conservation of functional elements in intron {126} as well as 3'-untranslated regions {127,128} have been reported.

Since the `gene-specific' probe was comprised of intron sequence, this provides a possible explanation for the extensive cross-hybridisation observed in these experiments.

The crux of the problem involves designing probes which have a sufficient degree of specificity to distinguish the gene of interest from other closely related genes. This first requires that all PSG C-termini be characterised and sequenced, before specific oligonucleotides can be designed and manufactured to target specific genes.

Therefore, the possibility that clone #1 contains a non-PSG DNA fragment, cannot be eliminated until further characterisation of the clone is undertaken. Hybridisation experiments involving specific oligonucleotide probes and/or sequencing regions of the clone to identify the gene would resolve this matter.

D.6.3: The Cosmid Vector Arms

To demonstrate that the results obtained from the hybridisation experiments were not due to the cosmid vector arms hybridising to the probes, both clone #1 and #4 were digested to completion

with *Sal*I. These digests were electrophoresed, Southern blotted onto nitrocellulose filters, then hybridised to both the PSG-domain probe and the `gene-specific' probe.

The fragment sizes expected for the left and right cosmid arms are 19.9 kb and 8.8 kb respectively. Fragments corresponding to these sizes were clearly visible in the gel photo, indicating that digestion was satisfactory. The presence of multiple hybridising fragments indicated there were internal *Sal*I sites in both clones #1 and #4, although these sites were not mapped.

As expected, fragments hybridising to both the PSG-domain probe and the `gene-specific' probe, were not of sizes corresponding to either cosmid arm. This demonstrated that the cosmid vector arms did not cross-hybridise to either of the probes (See lane 3, Figures 16 and 17).

D.6.4: The N-Terminal Probe

Since the `gene-specific' probe used to isolate clones #1 and #4 was derived from a region of 3' untranslated sequence, clones isolated using this probe would be expected to contain PSG-like C-domain, 3' untranslated region sequence, and perhaps also portions of the internal A/B domains. Indeed, the hybridisation experiments performed on both clones #1 and #4 as shown in Figures 16 and 17 provide experimental evidence to substantiate this.

To investigate the possible presence of an N-domain within the clones #1 and #4, digests of both cosmids were probed with an oligonucleotide specific for the PSG N-domain (see Figure 18). Neither clone #1 nor clone #4 displayed hybridisation to the N-domain probe when washed at the correct stringency. This result may not be entirely unexpected considering that the insert in each clone is only approximately 10 kb in length. To date, entire lengths of reported PSG genes have been 11-17 kb. Both clones (#1 and #4) were originally isolated using a probe derived from 3'-untranslated sequence, therefore it is reasonable to expect that both 5' and 3' regions of the gene may not have be included in their entirety, due to the limited length of the cosmids.

The difference in size of the hybridising *Eco*RI fragment between the Lorist vector clones and hC3.11, is probably due to an allelic difference. An additional *Eco*RI site which is present in the both Lorist vector clones, is not present in the hC3.11 clone, resulting in the 1.1 kb difference in size of the hybridising *Eco*RI fragment. Incidentally, an analogous *Eco*RI site is also present in clone #4, which may suggest that the hC3.11 arrangement is the exception.

Since the human DNA used to construct the EMBL and Lorist libraries, from which the Lorist based clones and hC3.11 were isolated, was obtained from two separate individuals, allelic differences amongst the isolated clones could be expected.

In conclusion, the hybridisation experiment using the N-terminal oligonucleotide probe demonstrated that neither clone #1 nor clone #4 contained PSG N-domain sequence.

D.6.5: The Validity of the `Gene-specific' probe

The genomic hybridisation results in Figure 19, show a large number of genomic fragments cross-hybridise to the `gene-specific' probe, suggesting this probe is not as unique as initially expected. The 500 bp clone pBE.5 selected as the `gene-specific' probe would be expected to hybridise in a genomic hybridisation. However, the genomic autoradiograph failed to show a fragment of a corresponding size.

Given the intensity of the hybridisation pattern, it is unlikely that the bands represent partial digests. Even if partial fragments were present, the band of interest at 800 bp is unlikely to be an artifact of partial digestion, since this is the smallest hybridising fragment.

The genomic DNA used to construct the libraries from which hC3.11 and the Lorist clones were isolated were not available, therefore it was not possible to perform genomic hybridisation experiments to provide confirmation of allelic variance. Despite this, there is indirect evidence to suggest the existence of allelic difference amongst these clones. The restriction patterns of clone #4 display a significant degree of resemblance to the hC3.11 clone. An allelic difference in the form of an additional *Eco*RI site, is evident between the Lorist vector clones (F20478, F11193) and hC3.11, as shown in the N-terminal domain hybridisation experiment (see Figure 18).

Moreover, if the 800 bp B*am*HI/E*co*RI hybridising fragment represented a genomic locus other than that of the 500 bp B*am*HI/E*co*RI region of hC3.11, a discrete hybridisation signal would be expected for each locus. The results of this genomic hybridisation experiment indicate a single predominant hybridising fragment present at 800 bp, suggesting that the size differences of the hybridising fragments are due to allelic differences between the sources of the clones, rather than the detection of separate loci.

Assessing the combined data from the hybridisation experiments performed on clone #1, it is likely that this clone contains a fragment representing a genomic locus other than that in hC3.11. Neither

the hybridisation nor the restriction patterns bear any resemblance to either hC3.11 or the Lorist vector clones.

Clone #1 contains an insert fragment which was produced by a partial *Sau*3AI genomic digest. The recognition sequence for *Sau*3AI is the 4 bp `_GATC_' sequence, whereas *Bam*HI is more specific recognising the 6 bp sequence `_GGATCC_'. Therefore, while the *Sau*3AI ends of the insert are compatible to *Bam*HI cohesive ends, they do not necessarily generate *Bam*HI cleavage sites. This presents a number of possibilities, depending on the exact nature of the ligation in clone #1. If a *Bam*HI site was regenerated at the 3' end of the insert, then the final 1.98 kb *Eco*RI fragment in clone #1 may represent a band of the same size on the genomic autoradiograph. If the region of cross-hybridising sequence extended back to the 8.69 kb *Eco*RI fragment, this may correspond to another of the larger hybridising fragments. If a *Bam*HI site was not generated following ligation into the vector, then it is still possible that one of the other bands larger than 2 kb represent the hybridising region in clone #1, since would be *Bam*HI sites downstream in the genomic arrangement.

As expected, the results of hybridisation experiments performed on digests of clones #1 are consistent with this genomic hybridisation experiment, since the DNA originated from a common source. However, considering the extensive cross-hybridisation of the `gene-specific' probe, hybridising fragments which are of similar sizes predicted sizes for clones #1, may not necessarily correspond to the loci in question

The conclusions which may be drawn from this experiment, are firstly that the 500 bp *Bam*HI /*Eco*RI fragment in hC3.11 is not as unique as first thought, and secondly that the 800 bp *Bam*HI /*Eco*RI fragment probably represents an allelic variant to the 500 bp fragment in hC3.11.

A more unique piece of sequence needs to be identified as a gene-specific probe, if further clones representing this genomic region are to be isolated. Any candidates should be tested by genomic hybridisation before being used to screen phage libraries.

E: CONCLUSIONS

A nucleotide sequence of 3796 bp has been sequenced in both directions and shown to closely resemble subgroup-1 PSG gene sequence, containing what appears to be a free-standing 'C_c-C_a-C_b' -domain cassette. This cassette is thought to be ubiquitous amongst the members of the PSG gene family. In addition to the C-termini comprising the 'C_c-C_a-C_b' cassette, a number of previously unreported C-termini are also predicted. Whether the putative exons reported in this hC3.11 C-domain cluster are expressed or not, remains to be established experimentally. A paper reporting sequence from this investigation, to be submitted for publication, is included in the appendix {130}.

Overlapping clones for chromosome walking were sought, subsequently clones #1 and #4 were identified and mapped. Restriction profiles and hybridisation patterns indicate that cosmid clone #1 does not overlap the genomic region represented by hC3.11. Since it is possible that the DNA fragment contained in this clone may represent either a CEA or PSG subfamily gene, the exact identity of this clone remains inconclusive.

In contrast to this, the mapping and hybridisation data indicate that clone #4 closely resembles the patterns displayed by the hC3.11 cosmid. However, since the possibility of multiple loci for the PSG genes has not been eliminated, clone #4 may not necessarily represent the same genomic locus as the hC3.11 clone.

Characterisation of the recently acquired Lorist vector clones F11193 and F20478 {45}, revealed they overlap the hC3.11 genomic locus, as well as a considerable upstream region. In light of this, further characterisation of the clones #1 and #4 would not be a priority, since Southern hybridisation information from the Lorist vector clones has clarified the PSG domain organisation in this area.

E.1: FUTURE STUDIES

Other groups have also reported similar difficulties when attempting to isolate complete PSG transcriptional units from phage libraries {125}. A high degree of specificity is required to eliminate the extensive cross-hybridisation encountered amongst the CEA/PSG family members during chromosome walking experiments.

Nucleotide sequencing of genomic PSG fragments identified in genomic mapping experiments would enable unique sequence to be used in the design and synthesis of oligonucleotide probes.

These oligonucleotide probes would potentially distinguish individual PSG species from other members of the PSG/CEA family under high stringency conditions.

Further studies to definitively map the genomic location of the individual PSG genes would also resolve whether there were multiple loci for the PSG genes or not. Mapping techniques employing yeast chromosome vectors have been used to sucessfully map large sections of the human genome, and data amassed from the worldwide genomic sequencing project could reveal further information about the PSG/CEA family.

Another avenue of study which could be pursued, is the investigation of evolutionary mechanisms operating in the CEA/PSG gene family. Examination of existing PSG sequences indicate the 'c-a-b' cassette is present in all PSG genes characterised to date. Evolutionary trees may now be constructed using this information, perhaps revealing further insights into the evolutionary history of this gene family.

The isolation and expression of individual PSG species is a crucial step in assigning biological role(s) to specific proteins. Recently in our laboratory, a putative receptor protein for the PSG-11 protein. While exhibiting several binding characteristics in common with integrin family members, the molecular weight of the putative PSG-11 receptor is significantly smaller than would be expected for an integrin {126}. The specific receptors for the different PSG proteins may now be identified, and the biological role(s) of the individual PSG proteins can be investigated in greater detail.

Future investigations into the biological role of the PSG and the factors controlling their expression, may also provide a valuable model from which aspects of alternate splicing and the mechanisms governing splice selection could be studied.

F: BIBLIOGRAPHY

{1} Tatarinov, Y.S., Masyukevich V.N.: Immunological identification of a new β_1-globulin in the blood serum of pregnant women. *Byull.Eksp.Biol.Med. USSR* 69:66-68, (1970).

{2} Bohn, H.: Nachweis und charakterisierung von Schwangerschafts protein in der menschlichen Plazenta, sowie ihre quantitative immunologische Bestimmung im serum schwangerer Frauen. *Arch.Gynaek.* 210:440-457, (1971).

{3} Bohn, H.: Isolierung und Charakterisierung des shwangerschafts-spezifischen β_1-Glykoproteins. *Blut* 24:292-302, (1972).

{4} Lin, T.M., Halbert, S.P., Spellacy, W.N.: Measurement of pregnancy-associated plasma proteins during human gestation. *J.Clin.Invest.* 54:576-582, (1974).

{5} Tatarinov, Y.S.: Trophoblast-specific beta-1-glycoprotein as a marker for pregnancy and malignancies. *Gynecol.Obstet.Invest.* 9:65-97, (1978).

{6} Chou, J.Y., Zilberstein, M.: Expression of the pregnancy-specific beta-1-glycoprotein gene in cultured human trophoblasts. *Endocrinology* 127:2127-2135, (1990).

{7} Khan, W.N., Hammarstrom, S.: Carcinoembryonic antigen gene family: Molecular cloning of cDNA for a PSBG/FL-NCA glycoprotein with a novel domain arrangement. *Biochem.Biophys.Res.Commun.* 161:525-535, (1989).

{8} Zimmermann, W., Weiss, M.,Thompson, J.A.: cDNA cloning demonstrates the expression of pregnancy-specific glycoprotein genes, a subgroup of the carcinoembryonic antigen gene family, in fetal liver. *Biochem.Biophys.Res.Commun.* 163:1197-1209, (1989).

{9} Khan, W.N., Osterman, A., Hammarstrom, S.: Molecular cloning and expression of a cDNA for a carcinoembryonic antigen-related fetal liver glycoprotein. *Proc.Natl.Acad.Sci. USA* 86:3332-3336, (1989).

{10} Plouzek, C.A., Watanabe, S., Chou, J.Y.: Cloning and expression of a new pregnancy-specific β_1-glycoprotein member. *Biochem.Biophys.Res.Commun.* 176:1532-1538, (1991).

{11} Rooney, B.C., Horne, C.H.W., Hardman, N.: Molecular cloning of a cDNA for human pregnancy-specific β_1-glycoprotein: homology with human carcinoembryonic antigen and related proteins. *Gene* 71:439-449, (1988).

{12} Lin, T.M., Halbert, S.P., Kiefer, D.: Quantitative analysis of pregnancy-associated plasma proteins in human placenta. *J.Clin.Invest.* 57:466-472, (1976).

{13} Sambrook, J., Fritsch, E.F., Maniatis, T.: *Molecular cloning: a laboratory manual.* second edition, Section 1.29, (1989).

{14} Watanabe, S., Chou, J.Y.: Isolation and characterisation of complementary cDNAs encoding human pregnancy-specific β_1-glycoprotein. *J.Biol.Chem* 263:2049-2054, (1988).

{15} Streydio, C., Lacka, K., Swillens, S., Vassart, G.: The human pregnancy-specific β_1-glycoprotein (PSáG) and the carcinoembryonic antigen (CEA)-related proteins are members of the same multigene family. *Biochem.Biophys.Res.Commun.* 154:130-137, (1988).

{16} Chan, W-Y., Borjigin, J., Zheng, Q-X., Shupert, W.L.: Characterisation of cDNA encoding human pregnancy-specific β_1-glycoprotein from placenta and extraplacental tissues and their comparison with carcinoembryonic antigen. *DNA* 7:545-555, (1988).

{17} Mueller, U.W., Jones, W.R.: Identification of an SP-1-like protein in non-pregnancy serum: isolation using monoclonal antibody. *J.Reprod.Immunol.* 8:111-120, (1985).

{18} Osborne, Jr. J.C., Rosen, S.W., Nilsson, B., Calvert, I., Bohn, H.: Physiochemical studies of pregnancy-specific β_1-glycoprotein: unusual ultracentrifugal and circular dichoric properties. *Biochemistry* 21:5523-5528, (1982).

{19} Chan, W-Y., Tease, L.A., Borjigin, J., Chan, P-K., Rennert, O.M., Srinivasan, B., Shupert, W.L., Cook, R.G.: Pregnancy-specific β_1-glycoprotein mRNA is present in placenta as well as non-placental tissues. *Hum.Reprod.* 3:667-685, (1988).

{20} Thompson, J., Mauch, E-M., Chen, F-S., Hinoda, Y., Shrewe, H., Berling, B., Barnert, S., Von Kleist, S., Shively , J.E., Zimmerman, W.: Analysis of the size of the carcinoembryonic antigen (CEA) family: isolation and sequencing of N-terminal domain exons. *Biochem.Biophys.Res.Commun.* 158:996-1004, (1989).

{21} McLenachan, T., Mansfield, B.C.: Expression of CEA-related genes in the first trimester human placenta. *Biochem.Biophys.Res.Commun.* 162:1486-1493, (1989).

{22} Oikawa, S., Inuzuka, C., Kuroki, M., Matsuoka, Y., Kosaki, G., Nakazato, H.: A pregnancy-specific β_1-glycoprotein, CEA gene family member expressed in a human promyelocytic leukemia cell line, HL-60.: structures of protein, mRNA and gene. *Biochem.Biophys.Res.Commun.* 163:1021-1031, (1989).

{23} Niemann, S.G., Flake, A., Bohn, H., Bartels, I.: Pregnancy-specific β_1-glycoprotein: cDNA cloning, tissue expression, and species specificity of one member of PSβG family. *Hum.Genet.* 82:239-243, (1989).

{24} Borjigin, J., Tease, L.A., Barnes, W., Chan, W-Y.: Expression of the pregnancy-specific β_1-glycoprotein genes in human testis. *Biochem.Biophys.Res.Commun.* 166:622-629, (1990).

{25} Thompson, J., Zimmermann, W.: The carcinoembryonic antigen gene family: Structure, Expression and Evolution. *Tumor Biol.* 9:63-83, (1988).

{26} Khan, W.N., and Hammarstrom, S.: Identification of a new carcinoembryonic antigen (CEA) family member in human fetal liver- cloning and sequence determination of pregnancy-specific glycoprotein-7. *Biochem.Biophys.Res.Commun.* 168:214-225, (1990).

{27} Zoubir, F., Khan, W.N., Hammarstrom, S.: Carcinoembryonic antigen family members in submandibular salivary gland: demonstration of pregnancy-specific glycoproteins by cDNA cloning. *Biochem.Biophy.Res.Commun.* 169:203-216, (1990).

{28} Streydio, C., Swillens, S., Georges, M., Szpirer, C., Vassart, G.: Structure, evolution, and chromosomal localisation of the human pregnancy-specific β_1-glycoprotein family. *Genomics* 6:579-592, (1990).

{29} Leslie, K.K., Watanabe, S., Lei, K-J., Chou, D.Y., Plouzek, C.A., Deng, H-C., Torres, J., Chou , J.Y.: Linkage of two pregnancy-specific β_1-glycoprotein genes: one is associated with hydatidiform mole. *Proc.Natl.Acad.Sci. USA* 87:5822-5826, (1990).

{30} Barnett, T.R., Pickle, II ,W., Elting, J.J: Characterisation of two new members of the pregnancy-specific β_1-glycoprotein family from the myeloid cell line KG-1 and suggestion of two distinct classes of transcription unit. *Biochemistry* 29:10213-10218, (1990).

{31} Zheng, Q-X., Tease, L.A., Shupert, W.L., Chan, W-Y.: Characterisation of cDNA's of the human pregnancy-specific β_1-glycoprotein family, a new subfamily of the immunoglobulin gene superfamily. *Biochemistry* 29:2845-2852, (1990).

{32} Arkawa, F., Kuroki, M., Misumi, Y., Matsuo, Y., Matsuoka, Y.: The nucleotide and deduced amino acid sequences of a cDNA encoding a new species of pregnancy-specific β_1-glycoprotein (PSβG). *Biochim.Biophys.Acta.* 1048:303-305, (1990).

{33} Chan, W-Y., Zheng, Q-X., McMahon, J., Tease, L.A.: Characterisation of new members of the pregnancy-specific β_1-glycoprotein family. *Mol.Cell.Biochem.* 106:161-170, (1991).

{34} Oikawa, S., Inuzuka, C., Kosaki, G., Nakazato, H.: Exon-intron organisation of a gene for pregnancy-specific β_1-glycoprotein, a subfamily member of CEA family: implications for it's characteristic repetitive domains and C-terminal sequences. *Biochem.Biophys.Res.Commun.* 156:68-77, (1988).

{35} Chou, J.Y., Sartwell, A.D., Wan, J-Y, Watanabe, S.: Characterisation of a pregnancy-specific β_1-glycoprotein synthesised by human placental fibroblasts. *Mol.Endocrinology* 3:89-96, (1989).

{36} Rosen, S.W., Kaminska, J., Calvert, I.S., Aaronson, S.A.: Human fibroblasts produce `pregnancy-specific' β_1-glycoprotein *in vitro*. *Am.J.Obstet.Gynecol.* 134:734-738, (1979).

{37} Engvall, E., Miyashita, M., Ruoslahti, E.: Monoclonal antibodies in analysis of oncodevelopmental proteins SP-1 *in vivo* and *in vitro*. *Cancer.Res.* 42:2028-2033, (1982).

{38} Chou, J.Y.: Production of pregnancy-specific β_1-glycoprotein by placental cells and human fibroblasts. *Oncodev.Biol.Med.* 4:319-326, (1983).

{39} Nozawa, S., Engvall, E., Kano, S., Kurihara, S., Fishman, W.H.: Sodium butyrate produces concordant expression of "early placental" alkaline phosphotase, pregnancy-specific β_1-glycoprotein and hCG beta-subunit in a newly established cell line (SKG-IIIa). *Int.J.Cancer.* 32:267-272, (1983).

{40} Azer, P.C., Braunstein, G.D., Van de Velde, R.L., Van de Velde, S., Kogan, R., Engvall, E.: Ectopic production of pregnancy-specific β_1-glycoprotein by a neotrophoblastic tumour *in vitro*. *J.Clin.Endocrinol.Metab.* 50:234-239, (1980).

{41} Shupert, W.L.: Pregnancy-specific β_1-glycoprotein in human intestine. PhD. Dissertation, University of Oklahoma, Oklahoma city, (1990).

{42} Takami, N., Misumi, Y., Kuroki, M., Matsuoka, Y., Ikehara, Y.: Evidence for carboxyl-terminal processing and glycolipid anchoring of human carcinoembryonic antigen. *J.Biol.Chem.* 263:12716-12720, (1988).

{43} Barnett, T.R., Kretchmer, A., Austen, D.A., Goebel, S.J., Hart, J.T., Elting, J.J., Kamarck, M.E.: Carcinoembryonic antigens: alternative splicing accounts for the multiple mRNAs that code for novel members of the carcinoembryonic antigen family. *J.Cell.Biol.* 108:267-276, (1989).

{44} Barnett, T.R, Zimmermann, W.: Workshop reports: proposed nomenclature for the carcinoembryonic antigen (CEA) gene family. *Tumor.Biol.* 11:59-63, 1990.

{45} McLenachan, P.A.: Dissertation, MSc thesis, in preparation, (1994).

{46} Bohn, H.: Detection and characterisation of pregnancy proteins in the human placenta and their quantitative immunochemical determination in sera from pregnant women. *Arch.Gynäkol.* 210:440-457, (1971).

{47} Bohn, H.: Studies on the pregnancy-specific β_1-glycoprotein (SP-1). *Arch.Gynäkol.* 217:209-218, (1974).

{48} Bohn, H., Schmidtberger, R., Zilg, H.: Isoleirung des Schwangerschaftsspezifischen Beta-1-Glykoproteins (SP-1) und antigenverwandter Proteine durche Immunoadsorption. *Blut* 32:103-113, 1976).

{49} Bischof, P.: Placental proteins *Contrib.Gynecol.Obstet.* 12:1-96, (1984)

{50} Grudzinskas, J.G., Gordon, Y.B., Jeffrey, D., Chard, T.: Specific and Sensitive determination of Pregnancy-Specific β_1-Glycoprotein by Radioimmunoassay. A new pregnancy test. *Lancet* I, pg: 333-340, (1977).

{51} Tatarinov, Y.S., Sokolov, A.V.: Development of a immunoassay for pregnancy-specific β_1-glycoprotein and it's measurement in serum of patients with trophoblastic and non-trophoblastic tumors. *Int.J.Cancer.* 19:161-165, (1977).

{52} Sørenson, S., Borggaard, B., Rolf, L.: A radioimmunoassay of the pregnancy-specific β_1-glycoprotein (SP-1). *Scand.J.Clin.Invest.* 37:537-543, (1977).

{53} Aschheim, S., Zondek, B.: Hypophysenvorderlappenhormon und ovarialhormon in Harn von Schwangeren. Klin Wochenschr, 6:1322, (1927).

{54} Brody, S.: Protein hormones and hormonal peptides from the placenta. In Klopper, A. and Diczfalusy, E. (Eds.): *Foetus and placenta.* Surrey, Adlard and Sons, pg: 299, (1969).

{55} Rosen, S.: New placental proteins : Chemistry, physiology, and clinical uses. *Placenta* 7:575-594, (1986).

{56} Petrocik, E., Wassman, E.R., Lee, J.J., Kelly, J.C.: Second trimester maternal serum pregnancy-specific beta-1-glycoprotein (SP-1) levels in normal and Down syndrome pregnancies. *Am.J.Med.Genet.* 37:114-118, (1990).

{57} Lavy, G., DeCherney, A.H.: The hormonal basis of ectopic pregnancy. *Clin.Obstet.Gynecol.* 30:217-224, (1987).

{58} Sterzik, K., Rosenbusch, B., Benz, R.: Serum specific protein-1 and beta-human chorionic gonadotrophin concentration in patients with suspected ectopic pregnancies. *Int.J.Gynecol.Obstet.* 28:253-256, (1989) .

{59} Hertz, J.B., Schultz-Larsen, P.: Placental proteins in threatened abortion. In Hau, J. (ed): *Pregnancy Proteins in Animals.* New York: Walter de Gruyter, pg:31-40, (1986).

{60} Masson, G.M., Anthony, F., Wilson, M.S.: Value of scwangershaftsprotein (SP-1) and pregnancy associated protein-A (PAPP-A) in the clinical management of threatened abortion. *Br.J.Obstet.Gynecol.* 90:146-149, (1983).

{61} Wurz, H., Geiger, W., Kunzig, H.J., Jabs-Lehman, A., Bohn, H., Luben, G.: Radioimmunoassay of SP-1 in maternal blood and in amniotic fluid in normal and pathological pregnancies. *J.Perinat.Med.* 9:67-78, (1981).

{62} Tamsen, L., Johansson, S.G.O., Axelsson, O.: Pregnancy-specific β_1-glycoprotein (SP-1) in serum of pregnant women with pregnancies complicated by intrauterine growth retardation. *J.Perinat.Med.* 11:19-25, (1983).

{63} Heikinheimo, M., Aula, P., Rapola, J., Wahlström, T., Jalanko, H., Sepälä, M.: Amniotic fluid pregnancy-specific β_1-glycoprotein (SP-1) in Meckels' syndrome: a new test for prenatal diagnosis? *Prenat.Diagn.* 2:103-108, (1982).

{64} Chou, J.Y.: Unpublished results.

{65} Bagshawe, K.D.: *Choriocarcinoma- The clinical Biology of the tropohoblast and it's tumors.* London: Arnold, (1969).

{66} Searle, F., Leake, B.A., Bagshawe, K.D., Dent, J.: Serum-SP-1-pregnancy-specific β_1-glycoprotein in choriocarcinoma and other neoplastic diseases. *Lancet* I 8064:579-580, (1978).

{67} Kim, S.J., Jung, J.K., Kang, E.C.,Bae, S.N.: Diagnostic value and limit of placental proteins in the management of gestational trophoblastic disease. In Mochizuki, M., Hussa, R. (eds.): *Placental protein hormones.* New York: Elsevier Science Publishers BV, pg: 221-228, (1988).

{68} Horne, C.H.W., Reid, I.N., Milne, G.D.: Prognostic significance of innappropriate production of pregnancy proteins by breast cancers. *Lancet* 2:279-282, (1976).

{69} Fagnart, O.C., Cambiaso, C.L., Lejeune, M.D., Noel, G., Maisin, H., Masson, P.L.: Prognostic value of concentration of pregnancy-specific β_1-glycoprotein (SP-1) in serum of patients with breast cancer. *Int.J.Cancer.* 36:541-544, (1985).

{70} Wright, C., Angus, B., Napier, J., Wetherall, M., Udagawa, Y., Sainsbury, J.R.C., Johnston, S., Carpenter, F., Horne, C.H.W.: Prognostic factors in breast cancer: immunohistochemical staining for (SP-1) and NCRC 11 related to survival, tumor epidermal growth factor and oestrogen receptor status. *J.Pathol.* 153:325-331, (1987).

{71} Johannssen, R., Haupt, H., Bohn, H., Heide, K., Seiler, F.R., Schwick, H.G.: *Immunitaetsforsch* 152:280-286, (1976).

{72} Cerni, C., Tatra, G., Bohn, H.: Immunosupression by human lactogen (HPL) and the pregnancy-specific beta-1 gycoprotein (SP-1). Inhibition of mitogen induced lymphocyte transformation. *Arch.Gynaekol.* 223:1-7, (1977).

{73} Williams A.F, Barclay A.N,: The immunoglobulin superfamily-domains for cell surface recognition. *Annu.Rev.Immunol.* 6:381-405, (1988).

{74} Barnett, T.R., Pickle, II W., Rae, P.M.M., Hart, J., Kamarck, M., Elting, J.: Human pregnancy-specific β_1-glycoproteins are coded within chromosome 19. *Am.J.Hum.Genet.* 44:890-893, (1989).

{75} Neimann, S.C., Schonk, D., van Dijk, P., Wieringa, B., Graeschik, K.H., Bartels, I.: Regional localisation of the gene encoding pregnancy-specific β_1-glycoprotein-1(PSβG) to human chromosome 19q13.1. *Cytogenet.Cell.Genet.* 52:95-97, (1989).

{76} Brandriff, B.F., Gordon, L.A., Tynan, K.T., Olsen, A.S., Mohrenweiser, H.W., Fertitta, A., Carrano A.V., Trask, B.J.: Order and Genomic distances among members of the Carcinoembryonic antigen (CEA) gene family determined by fluorescence *in situ* hybridisation. *Genomics* 12:773-779, (1992).

{77} Thompson, J., Koumari, R., Wagner, K., Barnert, S., Schleussner, C., Schrewe, H., Zimmermann, W., Müller, G., Schempp, W., Zaninetta, D., Ammaturo, D., Hardman, N.: The pregnancy-specific glycoprotein genes are tightly linked on the long arm of chromosome 19 and are coordinately expressed. *Biochem.Biophys.Res.Comm.* #2 167:848-857, (1990).

{78} Chou, J.Y., Leslie K.Y., Lei, K-J.: *Proceedings of the International Workshop on CEA*, Montreal, pg: 5, (1990).

{79} Brophy, B.K., MacDonald, R.E., McLenachan, P.A., Mansfield, B.C.: cDNA sequence of the pregnancy-specific β_1-glycoprotein-11s (PSG-11s). *Biochim.Biophys.Acta.* 1131:119-121, (1992).

{80} Beggs K.T.: Dissertation, BSc.(Hon) Thesis, Massey University, (1990).

{81} Khan, W.N., Teglund, S., Bremer, K., Hammarstrom, S.: The Pregnancy-Specific Glycoprotein family of the Immunoglobulin Family: Identification of New Members and Estimation of Family Size. *Genomics* 12:780-787, (1992).

{82} Gustafson,S., Proper, J.A., Bowie, E.J.W., Somer, S.S.: Parameters affecting the yield of DNA from human blood. *Analytical Biochemistry* 165:294-299, (1987).

{83} Rudert, F., Zimnmermann, W., Thompson, J.A.: Intra- and Interspecies analysis of the Carcinoembryonic Antigen (CEA) family reveal independant evolution in primates and rodents. *J.Mol.Evol.* 29:126-134, (1989).

{84} Williams, A.F.: A year in the life of the immunoglobulin superfamily. *Immunol.Today* 8:298-303, (1987).

{85} Landers, D.V., Bronso, R.A., Pavia, C.S., Stites, D.P.: Reproductive Immunology, in *Basic and Clinical Immunology*, Seventh Edition edited by Stites, D.P. and Terr, A.I., published by Appleton and Lange, Chapter 17 pg: 200-215, (1991).

{86} Ruoslahti, E.,Pierschbacher, M.D.: Arg-Glu-Asp.: a versitile cell recognition signal. *Cell* 44:517-518, (1986).

{87} Grand, R.J.A.: Acylation of viral and eukareote protein. *Biochem.J.* 258:625-638, (1989).

{88} Fergusson, M.A.J., Williams, A.F.: Cell surface anchoring of proteins via glycosyl-phosphotidyl-inositol structures. *Annu.Rev.Biochem.* 57:285-321, (1988).

{89} Birnboim, H.C., and Doly, J.: A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Research* 7:1513, (1979).

{90} Watanabe, S., Chou J.Y.: Human Pregnancy-specific β_1-Glycoprotein: A New Member of the Carcinoembryonic Antigen Family. *Biochem.Biophys.Res.Commun.* 152:762-768, (1988).

{91} Zoubir,F., Khan, W.N., Hammarstrom, S.: Carcinoembryonic Antigen Gene Family Members in Sumandibular Salivary Gland: Demonstration of Pregnancy-Specific Glycoproteins by cDNA Cloning. *Biochem.Biophys.Res.Commun.* 169:203-216, (1990).

{92} Ruoslahti, E., Pierschbacher, M.D.: New perspectives in cell adhesion: RGD and integrins. *Science* 238:491-497, (1987).

{93} Cunningham, B.A., Hemperly, J.J.,Murray, B.A., Prediger, E.A., Brackenbury, R., Edelman, G.M.: Neural cell adhesion molecule: structure, immunoglobulin-like domains, cell-surface modulation, and alternative RNA splicing. *Science* 236:799-806, (1987).

{94} Benchimol, S., Fuks, A., Jothy, S., Beauchemin, N., Shirota,K., Stanners, C.P.: Carcinoembryonic antigen, a human tumor marker, functions as an intercellular adhesion molecule. *Cell* 57:327-334, (1989).

{95} Oikawa, S., Inuzuka, C., Kuroki, M., Matsuoka, Y., Kosake, G., Nakazata, H.: Cell adhesion activity of non-specific cross-reacting antigen (NCA) and carcinoembryonic antigen (CEA) expressed on CHO cell surface: homophilic and heterophilic adhesion. *Biochem.Biophys.Res.Commun.* 164:39-45, (1989)

{96} Rojas, M., Fuks, A., Stanners, C.P.: Billary glycoprotein, a member of the immunoglobulin supergene family, functions *in vitro* as a Ca^{+2}-dependant intercellular adhesion molecule. *Cell.Growth.Diff.* 1:527-533, (1990).

{97} Turbide, C., Rojas,M., Stanners, C.P., Beauchemin, N.: A mouse carcinoembryonic antigen gene family is a calcium-dependant cell adhesion molecule. *J.Biol.Chem.* 266:309-315, (1991).

{98} Mansfield, B.C.: Personal Communication.

{99} Bohn, H.: Isolation and characterisation of placental proteins with special reference to Pregnancy-Specific β_1-Glycoprotein and other proteins specific to the placenta. In Klopper, A., Chard, T. (eds): *Placental Proteins.* Springer-Verlag, Berlin, Heidelberg, New York pg: 71-86, (1979).

{100} Tatra, G.: Clinical Aspects of Pregnancy-Specific β_1-Glycoprotein (PSβG). In Klopper, A., Chard, T. (eds): *Placental Proteins.* Springer-Verlag, Berlin, Heidelberg, New York pg: 135-140, (1979).

{101} Yarden, Y.: Growth factor receptor tyrosine kinases. *Annu Rev.Biochem.* 57:443-478, (1981).

{102} Sambrook, J., Fritsch, E.F., Maniatis, T.: *Molecular cloning: a laboratory manual.* Second edition, Section 1.82, (1989).

{103} Streydio, C., Vassart, G.: Expression of human pregnancy-specific β_1-glycoprotein (PSG) genes during placental development. *Biochem.Biophys.Res.Commun.* 166:1265-1273, (1990).

{104} Gold, P., Freedman, S.O.: Demonstration of tumor specific antigens in human colonic carcinomata by immunological tolerance and absorption techniques. *J.Exp.Med.* 121:439-462, (1965).

{105} Chu, T.M., Reynoso, G., Hansen, H.J.: Demonstration of carcinoembryonic antigen in normal human plasma. *Nature* 238:152-153, (1972)

{106} Kupchik, H.Z., Zamcheck, N.: Carcinoembryonic antigen in liver disease. Isolation from human cirrhotic liver and serum from normal liver. *Gastroenterology* 36:75-101, (1972).

{107} Fritsche, R., Mach, J.P.: Isolation and characterisation of carcinoembryonic antigen (CEA) extracted from normal colon mucosa. *Immunochemistry* 14:119-127, (1972).

{108} Egan, M.L., Pritchard, D.G., Todd, C.W.: Isolation and characterisation of carcinoembryonic antigen-like substances in colon lavages of healthy individuals. *Cancer.Res.* 37:2638-2643, (1977).

{109} Thompson, D.M.P., Krupey, J., Freedman, S.O., Gold, P.: The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. *Proc.Natl.Acad.Sci. USA* 64:161-167, (1969).

{110} Go, V.L.W.: Carcinoembryonic antigen, clinical applications. *Cancer* 37:562-566, (1976).

{111} Kuroki, M., Kuroki, M., Ichiki, S., Matsuoka, Y.: Identification and partial characterisation of the unglycosylated peptide of the carcinoembryonic antigen synthesized by human tumor cells in the presence of tunicamycin. *Mol.Immunol.* 21:743-746, (1984).

{112} Zimmermann, W., Ortlieb, B., Friedrich, R., von Kleist, S.: Isolation and characterisation of cDNA clones encoding the human carcinoembryonic antigen reveal a high conserved repeating structure. *Proc.Natl.Acad.Sci. USA* 84:2960-2964, (1987).

{113} Beauchemin, N., Benchimol, S., Cournoyer, D., Fuks, A., Stanners, C.P.: Isolation and characterisation of full length functional cDNA clones for human carcinoembryonic antigen (CEA). *Mol.Cell.Biol.* 7:3221-3230, (1987).

{114} Rigby, P.W.J., Diekmann, M., Rhodes, C., Berg, P.: Labelling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. *J.Mol.Biol.* 113:237, (1977).

{115} Teglund, S.: *CEA/PSG Workshop Oral Sessions*, Umea, Sweden, (1992).

{116} Thompson, J.A., Pande, H., Paxton, R.J., Shively, L., Padma, A., Simmer, P.L., Todd, C.W., Riggs, A.D., Shively, J.E.: Molecular cloning of a gene belonging to the carcinoembryonic antigen family and discussion of a domain model. *Proc.Natl.Acad.Sci. USA* 84:2965-2969, (1987).

{117} Oikawa, S., Kosaki, G., Nakazato, H.: Molecular cloning of a gene for a member of carcinoembryonic antigen gene family; signal peptide and N-terminal domain sequences of non-specific cross reacting antigen (NCA). *Biochem.Biophys.Res.Commun.* 146:464-469, (1987).

{118} Shapiro, M.B., Senapathy, P.: RNA splice junctions of different classes in eukareotes : sequence statistics and functional implications in gene expression. *Nucleic Acids Research* Vol 15. 7:7155-7173, (1987).

{119} *The Placenta, Biological and clinical aspects.* Edited by Karman, S., Moghissi, M.D. Charles C. Thomas Publisher, (1974).

{120} Beer, A.E. and Billingham, R.E.: *The immunobiology of human reproduction.* Prentice-Hall, (1976).

{121} Klopper, A., Chard,T.: *Placental Proteins.* Springer-Verlag, Berlin, Heidelberg, New York, (1979).

{122} Thompson, J., Rudert, F., Mauch, E-M.: Intra- and inter species evolution of the CEA gene family. *Int.Soc.Oncodev.Biol.Med.: XVth Annu Meet*, Quebec, Abstr.133, (1987).

{123} Oikawa, S., Imajo, S., Noguchi, T., Kosaki, G., Nakazata, H.: The carcinoembryonic antigen (CEA) contains multiple immunoglobulin-like domains. *Biochem.Biophys.Res.Commun.* 144:634-642, (1987).

{124} Mendenhall, H.W.: The immunology of the fetal-maternal relationship. In *Immunology of Human Reproduction,* edited by Scott, J.S. and Jones, W.R. Academic press, London, Grune and Statton, New York, (1976).

{125} Chou, J.Y., Plouzek, C.A.: *Pregnancy-Specific* β_1*-Glycoprotein.* Thieme Medical Publishers, New York, (1992).

{126} Rutherfurd, K.J., Chou, J.Y., Mansfield, B.M.: Pregnancy-specific β_1-glycoprotein 11s (PSG11s) Binds to Receptors on Promonocyte cells. In press, (1994).

{137} Lei, K-J., Sartwell, A.D., Pan, C-J., Chou, J.Y.: Cloning and expression of genes encoding human Pregnancy-Specific Glycoproteins. *J.Biol.Chem.* Vol.263 23:16371-16378, (1992).

{128} Gordon, Y.B., Chard, T.: The Specific Proteins of the Human Placenta: Some New Hypotheses. *In Placental Proteins.* Edited by Klopper, A., Chard, T. (Eds.) Springer-Verlag, Berlin, Heidelberg, New York, pg: 1-21, (1979).

{129} Kan, M., Tatarinov, Y.: *Proceedings of the International Workshop on CEA*, Montreal, pg: 11, (1990).

{130} Joe, T.W., McLenachan, P.A., Mansfield, B.C.,: Sequence of a novel prenancy-specific β_1-glycoprotein C-terminal domain. Submitted to *Biochim.Biophys.Acta.* for publication, (1994).

G: APPENDIX **Short Sequence Paper**

Sequence of a novel pregnancy-specific β1-glycoprotein C-terminal domain.

Terence W. Joe, Patricia A. McLenachan, Brian C. Mansfield[1]

Department of Microbiology and Genetics, Massey University,

Palmerston North, New Zealand.

Key words: pregnancy-specific β_1-glycoprotein, gene sequence, C-terminal domains.

The sequence given in Figure 1 has been submitted to GenBank, accession number L17043

[1]Corresponding author: Phone [redacted]

FAX [redacted]

E-mail B.MANSFIELD@MASSEY.AC.NZ

Summary

A sequence related to the C-terminal coding regions of a subgroup 1 pregnancy-specific β_1-glycoprotein (PSG) has been characterised upstream of the PSG11 gene. The sequence can encode the C_c, C_a, and C_b domains but there appear to be no other gene exons within at least 10kb. Based on the organisation of other PSG genes, the organisation of these sequences is novel.

The human pregnancy-specific β_1-glycoproteins (PSG) form a family of glycoproteins expressed in the placenta throughout pregnancy [reviewed 1]. Cloning studies have shown that there are at least 11 genes encoding the PSG, clustered on chromosome 19 region q13.1-13.2 [2,3,4]. Each gene is composed of at least six exons. The first exons encode the leader sequence (L) and the N-terminal domain (N) of the proteins. Subsequent exons encode the central protein domains AI, BI, AII, BII. Over these exons there is greater than 90% nucleotide identity between the genes. The C-terminal domains of the proteins are created through alternative splicing of C-terminal exons. There are three subgroups of the PSG, defined by the organisation of the C-terminal exons. Subgroup 1 PSG genes express four alternative C-terminal domains, C_d, C_c, C_a, C_b, which are encoded by separate exons [5,6]. The subgroup 2 genes express a $C_{m/n}$ domain. The organisation of these genes is unique. Unlike the subgroup 1 and 3 genes where the C-terminal coding sequence and 3'-untranslated region are within one exon, the $C_{m/n}$ coding sequence and first 44 bases of 3'-untranslated region are encoded within a separate exon from the rest of the 3'-untranslated region [4,7]. The subgroup 3 genes can express up to three C-terminal domains, C_w, C_r, C_s, encoded by separate exons [1,8,9,10,11]. Recently we reported the complete sequence of the subgroup 3 gene PSG11 and demonstrated, through sequence comparison, that both the subgroup 2 and 3 genes have the subgroup 1 C_c, C_a, and C_b coding sequences present [12], although they do not seem to be expressed. It was demonstrated that in evolution, the subgroup 1 C_d sequence had been disrupted by the sequences encoding either $C_{m/n}$ or C_r and C_s of the subgroup 2 and 3 genes. The C_c, C_a, and C_b sequences present in subgroups 2 and 3 gene were referred to as a "cab" cassette, since the point at which the subgroup 2 and 3 genes resumed identity to the subgroup 1 C_c, C_a, and C_b sequences was identical, as if the sequences were an evolutionary unit. We have now sequenced a PSG-related region upstream of the PSG11 gene and show that it contains a "cab" cassette,

apparently unlinked to a PSG transcription unit and propose that it is an evolutionary relic of the evolution of the PSG gene subgroups.

Three cosmid clones, hC3.11, F11193 and F20478, which encompassed the PSG11 gene have previously been characterised. Approximately 9 kb upstream of the PSG11 gene we identified, by hybridisation, a PSG-like sequence (Figure 1). The sequence of this region is given in Figure 2. The first 500 bases of sequence have little relationship to sequences reported on the GenBank nucleotide database. From about 500 bp onwards there is increasing similarity to the cDNA sequence of the subgroup 2 gene PSG3. Between nucleotides 522 and 537 is a potential splice acceptor site. This occurs immediately prior to the sequence corresponding to the 3'-untranslated region exon of the subgroup 2 genes, suggesting that this could be the C-terminal coding sequence for a subgroup 2 gene. Another potential splice site, also present in the PSG3 transcript, between nucleotides 588 and 602, is also present. The similarity to the PSG3 transcript continues for 490 bp to the end of the PSG3 transcript, with 83% nucleotide identity. From nucleotide 682 onwards, a "cab" cassette is present, consistent with the organisation of the PSG genes.

No genomic sequence has been reported previously for this region of a subgroup 2 gene. However, the sequence has several features which suggest that it is not part of a subgroup 2 gene. Firstly, the subgroup 2 genes lack the C_C exon splice acceptor site present in the subgroup 1 genes (Figure 2, nucleotides 872 to 885), which prevents expression of the $C_{C\text{-like}}$ domain sequences in the subgroup 2 genes. In the sequence reported, however, this splice site is intact and a C_C coding sequence almost identical to that of PSG1c is predicted. Secondly, the associated poly(A) addition consensus sequence (AATAAA) present in the subgroup 2 transcript (Figure 2, nucleotides 1004 to1009) is absent. Thirdly, over the 350bp of sequence common to the subgroup 1 and 2 genes (nucleotides 677 to 1027), there is 94.3% identity to the subgroup 1 PSG1 gene and 84% identity to the subgroup 2

gene PSG3. If spaces required for optimal alignment are considered, the identities drop to 93.7% subgroup 1 identity and 72% subgroup 2 identity. This implies that the "cab" cassette is more closely related to the functional sequences of a subgroup 1 gene than the non-expressed subgroup 2 sequences.

Beyond the "cab" cassette region (nucleotides 2244 to 3769), the sequence shows little sequence similarity to previously reported sequences on the GenBank nucleotide database. There are two potential splice acceptor sequences at nucleotides 2532 to 2549 and 3363 to 3380, each followed by poly(A) addition consensus sequences at nucleotides 2729 to 2734 and 3490 to 3495, but the domains encoded are unlike any reported for the PSG and their significance is unclear.

To investigate if the sequence reported is part of a PSG gene, the cosmids were hybridised with cDNA probes representing the central A and B domains of the PSG genes. No hybridisng regions were identified within the 10kb lying upstream of the "cab" cassette. Since the length of a PSG transcription unit is 14-17kb in length [1,12] and the L, N, A and B exons are spread over 10-12kb, it would seem unlikely that a "cab" cassette separated from a BII domain by more than 10kb is part of a functional PSG gene. This sequence is unlikely to be a cloning artifact, since it is present in three independent clones isolated from two libraries created from independent tissue sources, using different cloning vectors and hosts. This may suggest that this cassette is a relic of an evolutionary event that created the C-terminal diversity of the PSG gene family [12].

This work was supported in part by a Health Research Council of New Zealand project grant to BCM. KJR was supported by a University Grants Committee post-doctoral fellowship.

References

1. Chou, J.Y. and Plouzek, C.A. (1992) Seminars Reprod. Endocrin. 10, 116-126.
2. Khan, W.N., Teglund, S., Bremer, K. and Hammarstrom, S. (1992) Genomics 12, 780-787.
3. Streydio, C., Swillens, S., Georges, M., Szpirer, C. and Vassart, G. (1990) Genomics 6, 579-592.
4. Thompson, J., Koumari, R., Wagner, K., Barnert, S., Schleussner, C., Schrewe, H., Zimmerman, W., Muller, G., Schempp, W., Zaninetta, D., Ammaturo, D. and Hardman, N. (1990) Biochem. Biophys. Res. Commun. 167, 848-859.
5. Oikawa, S., Inuzuka, C., Kosaki, G. and Nakazato, H. (1988) Biochem. Biophys. Res. Commun. 156, 68-77.
6. Lei, K-J., Sartwell, A.D., Pan, C-J. and Chou, J.Y. (1992) J. Biol. Chem. 267, 16371-16378.
7. Oikawa, S., Inuzuka, C., Kuroki, M., Matsuoka, Y., Kosaki, G. and Nakazato, H. (1989) Biochem. Biopys. Res. Commun. 163, 1021-1031.
8. Chan, W-Y., Zheng, Q-X., McMahon, J. and Tease, L.A. (1991) Mol. Cell. Biochem. 106, 161-170.
9. Zimmermann, W., Weiss, M. and Thompson, J.A. (1989) Biochem. Biophys. Res. Commun. 163, 1197-1209.
10. Arkawa, F., Kuroki, M., Misumi, Y. and Matsuoka, Y. (1990) Biochim. Biophys. Acta 1048, 303-305.
11. Brophy, B.K., MacDonald, R.E., McLenachan, P.A. and Mansfield, B.C. (1992) Biochim. Biophys. Acta 1131, 119-121.
12. McLenachan, P.A., Rutherfurd, K.J., Beggs, K.T., Sims, S.E. and Mansfield, B.C. (1994) Genomics (submitted).

Figure 1

Maps of the cosmid clones hC3.11, F11193 and F20478 showing the relation of the sequence reported to the PSG11 gene. The L, L/N, AI, pseudo-BI, AII, BII-C_W, C_T, and C_S exons are boxed and lie, in order, 5'-3'. The region containing the "cab" cassette is underlined. Restriction sites marked are *Eco* RI (E) and *Bam* HI (B). An *Eco* RI polymorphism is marked (*).

Figure 2

Nucleotide sequence of the PSG C-terminal sequences lying upstream of the PSG11 gene aligned with analogous regions of a PSG subgroup 1 gene (PSG1) and a subgroup 2 (PSG3) cDNA. Potential splice acceptor sites are marked in bold and underlined. Potential poly(A) addition consensus sequences are marked in bold. Potential translations are given in one letter code beneath the sequence. Translations of the PSG1 gene and PSG3 cDNA sequences are given above the sequences. Sequence identity is represented by ".". Spaces introduced into the sequences to optimise alignments are represented by "-". The sequence has the GenBank accession number L17043.

Figure 1

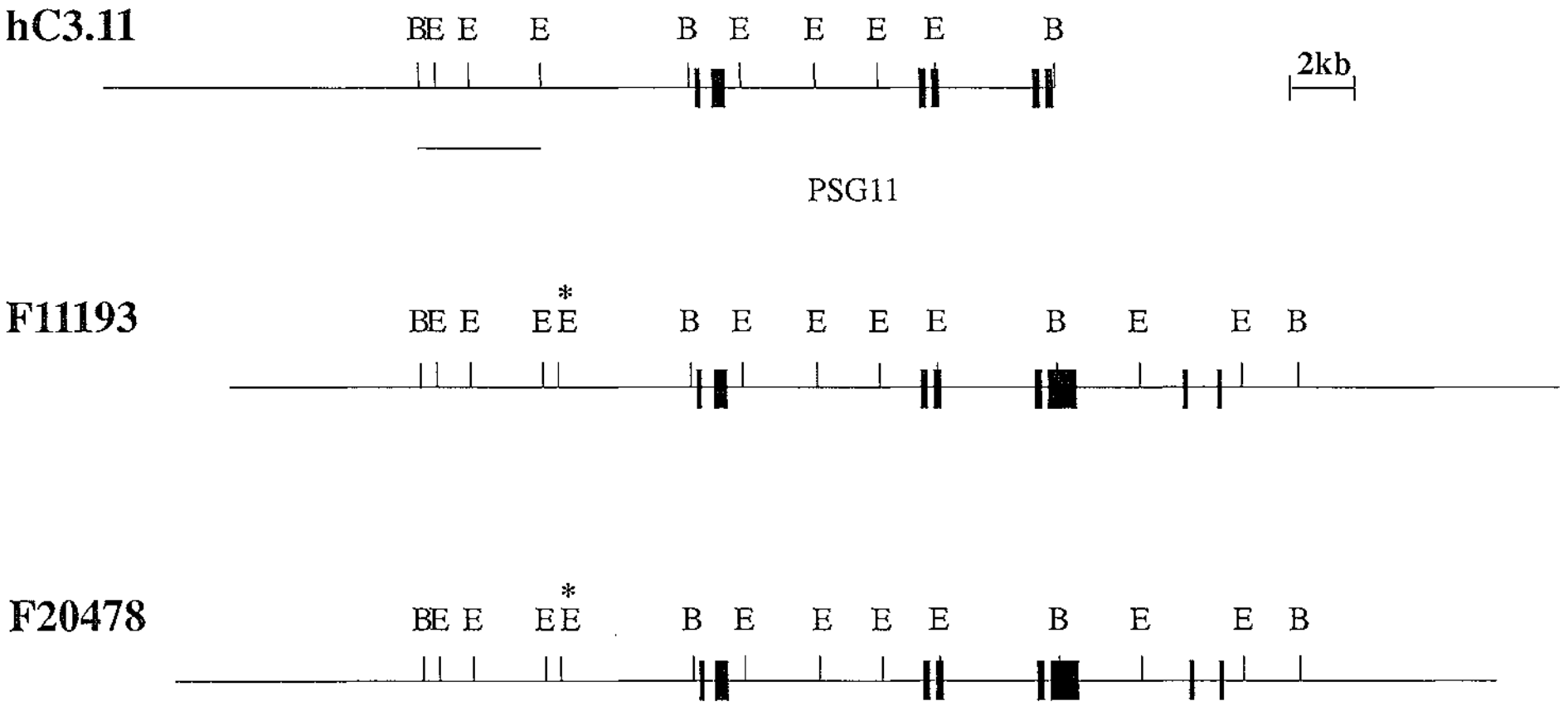
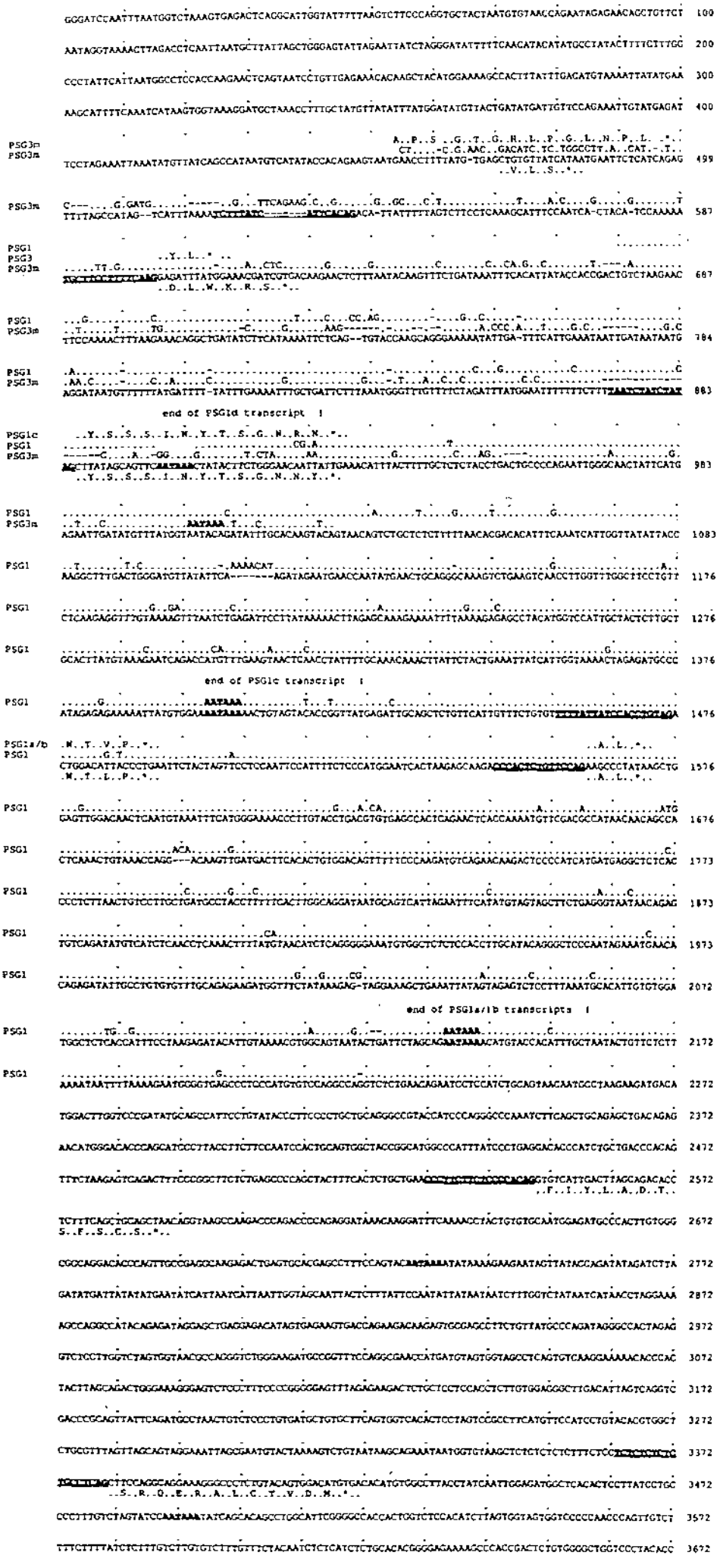

hC3.11
BE E E B E E E E B
2kb
PSG11
F11193
BE E E*E B E E E E B E E B
F20478
BE E E*E B E E E E B E E B

Figure 2


PSG3m
PSG3m
PSG3m
PSG1
PSG3
PSG3m
PSG1
PSG3m
PSG1
PSG3m
end of PSG1d transcript
PSG1c
PSG1
PSG3m
PSG1
PSG3m
PSG1
PSG1
end of PSG1c transcript
PSG1
PSG1a/b
PSG1
PSG1
PSG1
PSG1
PSG1
PSG1
end of PSG1a/1b transcripts
PSG1
PSG1