Characterisation of the 3' Region of the PSG11 Gene

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.

Patricia Ann McLenachan

1995

ABSTRACT

The pregnancy-specific beta-1 glycoproteins (PSG) are a family of abundant proteins essential to pregnancy that are encoded by 11 genes located on chromosome 19q 13.1-13.3. The genes can be divided into three subgroups based on the organisation of their 3' coding regions. In 1989, our group isolated cosmid hC3.11, which contained most of the PSG11 gene, but which did not include the 3' coding region. This thesis reports subsequent work to characterise two further cosmids which span the PSG11 locus and which do include the 3' coding region. These cosmids were mapped and partially sequenced.

Three exons encoding potential alternative C-terminal domains were identified: Cw, Cr and Cs. The Cs domain lies 4.6kb from the end of the B2 domain. This is the first report of genomic sequence for this particular domain and for a functional PSG subgroup 3 gene.

Downstream from this exon are sequences homologous to the C-termini of subgroup 1 PSG genes. This finding suggests that subgroup 1, 2 and 3 genes are related via insertion/deletion events.

Data from seven PSG genes from all three subgroups and from four different regions were used to construct evolutionary trees. Variability patterns in the data were examined and these showed that the mechanism of sequence evolution for the N-terminal domain, the A1 domain, and to a certain extent, the B2 domain was not neutral. Sequences from these regions were shown to be unsuitable for determining historical relationships between PSG genes. In contrast, the data from the C-terminal region showed a better fit with the assumptions of sequence evolution (e.g. all changes are independent and identically distributed) required by currently implemented analysis methods. Evolutionary tree reconstruction using this region gave a phylogeny that was consistent with one based on the genomic organisation of the genes.

ACKNOWLEDGEMENTS

This thesis has been a long time in the making, consequently, the number of people who have provided inspiration, encouragement and practical support has been large.

Many thanks must first go to my supervisor, Dr.Brian Mansfield, for whom I have worked for the past eight years. During this time I have picked up many tricks and techniques but have, above all, gained an increasing confidence in myself due to Brian's support and encouragement. To all the people who, over the last eight years, have passed through the now disbanded Mansfield Park, a special thanks for sharing the highs and the lows that come with research in the field of molecular biology. To Kay Rutherfurd, Neville Honey, Max Scott and the present-day members of the Cake Club - a special thanks and please let me know when you celebrate the next good result.

My Head of Department, Professor Tim Brown, has been steadfastly encouraging and has provided practical support. Also, I have been fortunate to have been able to discuss ideas along the way with Professor David Penny and Dr. Robert Hickson.

Special thanks go to the past and present members of Scott Base, especially Carolyn Young, for working and playing hard.

Eric and Betty Terzaghi have been a constant source of encouragement, inspiration and practical support over the years. Heart-felt thanks to you, we miss you.

The writing of this thesis has resulted in the complete disruption of my household for an extended period of time- to my children, Sam, Amy and our latest addition Toby - thanks for your help and foreberance. Special thanks too, to my mother Rita, for looking after all of us so well, during the final throes.

To my very special friends Rick, Kathy and Megan, thanks for looking after the boy and for Friday night take-outs and videos. Hope we can do it all again.

To Amy and Adalie for lunches at Bellas, baby therapy and your special friendship, many thanks.

To Maren, Brian ,Karen and David, many thanks for fun, friendship and lots of encouragement.

And last of all, to my husband Peter - for discussions and ideas, for your hard work, your constructive criticism and encouragment, for minding Toby and for taking me to Bath - what can I say- thank you, thank you, thank you, I owe you bigtime.

ABBREVIATIONS

aa	amino acid
ATP	adenosine tri-phosphate
B-ME	beta-mercaptoethanol
BSA	bovine serum albumin
CGM	**c**arcinoembryonic antigen **g**ene family **m**ember
CIAP	calf intestinal alkaline phosphatase
DMF	dimethylformamide
DNA	deoxyribonucleic acid
DNase	Deoxyribonuclease
DTT	Dithiothreitol
E.coli	*Escherischia coli*
EDTA	ethylene-diamine tetraacetic acid
EtBr	ethidium bromide
IPTG	Isopropyl B-D-thioglylactoside
KOAc	potassium acetate
NaOAc	sodium acetate
PCR	Polymerase Chain Reaction
pfu	plaque forming unit
PSG	pregnancy specific β-1 glycoprotein
PVP	polyvinylpyrididine
RNase	Ribonuclease
SDS	sodium dodecyl sulphate
SSC	Standard Saline Citrate
Tris-HCl	Tris hydroxy-methyl aminomethane, made to the appropriate pH with concentrated HCl
TE	Tris-EDTA Buffer
TBB	Tris Borate Buffer:.
TAE	Tris Acetate EDTA Buffer:
UTR	untranslated region
UV	ultra violet
vol	volume
X-gal	5-bromo-4-chloro-3-indolyl-B-D galactopyranoside

LIST OF FIGURES.

List of Tables.

CHAPTER 1. INTRODUCTION

1.1 The Placenta

The placenta is the organ primarily responsible for the interchange of substances between the developing foetus and the mother. As such, it performs a number of functions such as gaseous exchange, the supply of nutrients, the removal of waste products, the production and exchange of proteins and hormones as well as providing some immunological and physical protection as a barrier. The placenta is therefore a unique organ with several distinctive characteristics. It has a dual origin being derived from both maternal and foetal tissues. It has extensive contact between foetal chorionic villi and the maternal blood system. Because of its function it has both a limited life span and an extracoporeal position with respect to the foetus (Hamilton *et al.*, 1972).

Although much of the interchange of materials between the foetus and the mother is achieved by simple diffusion, the placenta cannot be considered to be an inert barrier. There is evidence that many substances, including nutrients are actively transported against a concentration gradient. Pinocytosis is thought to be involved in the transport of some immunologically important macromolecules from the maternal blood system to the foetus (Hamilton *et al.*, 1972). In a recent theoretical paper (Guilbert *et al.*, 1993), the placenta is likened to a completely syncitialised macrophage which "surrounds and protects the antigenically foreign conceptus in a manner analogous to the formation of cyst-like structures by giant cells (syncitialised macrophages) that envelop potentially harmful pathological agents". Indeed, it appears placental tissue and macrophages share many characteristics such as phagocytosis, syncitialisation, invasiveness and the expression of proteins such as CD4, CD14, IgG receptor (FcR), colony stimulating factor-1 (CSF-1), granulocyte macrophage-CSF (GM-CSF), interleukins 1 and 6 (IL-1, IL-6), tumour necrosis factor (TNF), transforming growth factor (TGF), platelet derived growth factor (PDGF) and receptors for these cytokines.

The immunological relationship between mother and foetus

The placenta and developing foetus are not antigenically inert either. Why this unit is maintained, without eliciting an immunological graft rejection response, is a question for which there is as yet no satisfactory explanation. The unique immunological relationship between the mother and the developing foeto-placental unit is extremely complex. Two hypotheses proposed for explaining the specific non-reactivity of the maternal immune system to the conceptus are immuno-tolerance and immunoenhancement.

When the immune system is exposed to either very high or very low doses of antigen, it may fail to respond to the antigen; this is known as immuno- tolerance. It is possible that during gestation, small numbers of foetal cells sloe off into the maternal blood system, providing a low dose of antigen. Although the two systems are quite separate, there is some evidence that foetal erythrocytes, lymphocytes and syncitial sprouts mix with the maternal blood. Erythrocytes do not possess transplantation antigens and the antigenic status of the sprouts is unknown. Lymphocytes however, probably do possess transplantation antigens and may contribute to "low zone" immuno-tolerance (Hamilton *et al.*, 1972).

Immunological enhancement is "the specific frustration of both the antigenic stimulus and the hosts cellular immune response, mediated by humoral isoantibody" (Beer and Billingham, 1976.). The mechanism of enhancement is not fully understood but it seems to involve the masking of antigen by antibody which inhibits attack by cytotoxic T cells and/or helper T cells and/or blocking of receptors on cytotoxic T cells by antigen shed from the target cell (Roitt, 1980). Some kind of central inhibitory action by antibodies or antibody/antigen complexes, to impair the development of the host cell-mediated immune response is also thought to be involved (Beer and Billingham, 1976).

Some aspects of the unique immunological relationship between the mother and the developing foetus are partially understood:

The maternal immune system appears to be "cognisant" of the antigenically foreign conceptus. Indeed, recognition of genetic disparity seems to be essential for successful implantation and

placental expansion (Guilbert *et al.*, 1993, Beer and Billingham, 1976). Also it appears that the uterus is able to express a normal immune response in other circumstances - that is - it does not appear to be an immunologically privileged site (Hamilton *et al.*, 1972).

There is some evidence from mouse and human studies, that pregnant individuals are biased in their immune response, toward humoral (antibody) responses and away from cell-mediated immunity (Guilbert *et al.*, 1993, Mosmann and Moore, 1991). T_H1 and T_H2 lymphocytes are at least partially responsible for cell-mediated and antibody responses respectively and produce a number of cytokines that cross-regulate an equilibrium between the two subsets of cells in non-pregnant individuals. Interleukin-10 (IL-10) is one such cytokine. It is produced by T_H2 cells and inhibits the synthesis of cytokines by T_H1 cells. IL-10 has been found to be expressed constitutively at the foeto-maternal interface during all three trimesters of pregnancy (Guilbert *et al.*, 1993). Other cytokines such as TGF-B, have been similarly detected. As well as functioning as growth stimulators, with a role in placental expansion, they may also help to establish the switch toward antibody responses during pregnancy. Interactions at the level of cytokine regulation are however, extremely complex and require further investigation to uncover all the pathways involved.

1.2 The PSG - A Major Class of Placental Proteins

One of the most abundant classes of proteins produced by the placenta are the pregnancy-specific β-1 glycoproteins (PSG). The PSG are a heterogeneous family of proteins whose levels increase in an exponential manner during pregnancy, reaching final concentrations of 200-400μg/ml in maternal serum (Oikawa *et al.*, 1989, Lin *et al.*, 1974). High levels of PSG during pregnancy correlate with stable pregnancy while decreased levels are associated with threatened miscarriage (Tamsen *et al.*, 1983).

The PSG are classified as a subgroup of the carcinoembryonic antigen (CEA) gene family on the basis of sequence identity. Both PSG and CEA proteins are predicted to contain the looped domains characteristic of members of the immunoglobulin superfamily (Paxton *et al.*1987, Zheng *et al.*, 1990 and FIG.1). The observation of high levels during stable pregnancies, suggests the PSG may have an essential role during pregnancy, though this has yet to be

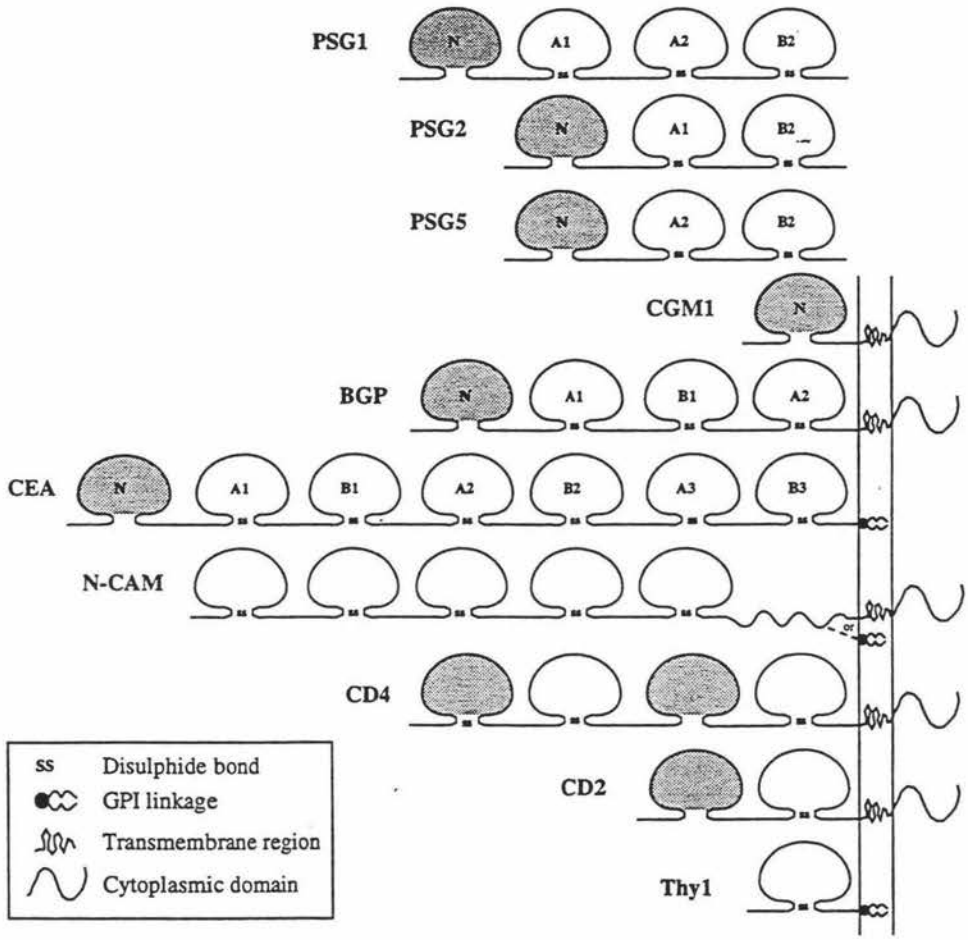


FIG. 1. Schematic representation of some members of the immunoglobulin superfamily (from Khan *et al.*, 1992).

Filled circles represent variable region (V) - like domains. Circles closed with the letters ss represent C2-type constant domains. Homologous domains are indicated with the letters N, A1, A2, B1, B2.

determined. Detailed characterisation of the PSG genes will provide a basis for future research into a biological role for PSG.

A brief historical perspective

In 1970, Tatarinov and Masyukevich isolated a new protein from human pregnant serum which they called PAPP-C (Tatarinov and Masyukevich, 1970). A year later, Bohn isolated a protein from human placenta named Schwangerschafts protein-1 (SP1), which proved to be immunologically identical to PAPP-C (Bohn, 1971). The protein was later found to be a collection of glycoproteins with molecular masses between 70kda and 32kda and a carbohydrate content of 28-32% (Watanabe and Chou, 1988a) and were subsequently called pregnancy-specific β-1 glycoproteins. Initially PSG were thought to be specific to the placenta but later studies detected the proteins in foetal liver cells (Zimmerman *et al.*, 1989), salivary gland and colon (Zoubir *et al.*, 1990), testes (Borjigin *et al.*, 1990), in various tumour cell lines, as well as in the sera of patients with hydatidiform mole, invasive mole and choriocarcinoma (Chou and Plouzek, 1992). However, the placenta remains the major site of production of PSG.

Early experimental work using antibodies to the protein showed that human PSG were a heterogeneous group of proteins, which were thought to be related via post-translational modifications. The extent of this heterogeneity was not appreciated however until molecular studies were undertaken and it became apparent that there were several PSG genes. By the late 1980s, a number of cDNA clones had been independently isolated (Watanabe and Chou, 1988a and b, Rooney *et al.*, 1988, Chan *et al.*, 1988a, Streydio *et al.*, 1988, Khan and Hammarstrom, 1989, McLenachan and Mansfield, 1989, Zimmerman *et al.*, 1989). Subsequently, genomic clones were isolated and the genomic organisation of some PSG established (Oikawa *et al.*,1988, Oikawa *et al.*, 1989a). The genes were mapped as a cluster on chromosome 19 (Streydio *et al.*, 1990, Thompson *et al.*, 1990) and estimates of the size of the family based on the number of 5' exons detectable in the genome, suggested that there were at least 9 different genes. Many of these had alternative splice and/or polyadenylation variants, giving rise to more than thirty different transcripts.

Table 1. CLASSIFICATION OF THE HUMAN CEA/PSG GENE FAMILY.

(from: Chou and Plouzek, 1992, Barnett and Zimmerman, 1990)

CEA subfamily.

current name of gene or mRNA	old names of genes or mRNA
CEA	CEA
CEAa	CEA
CEAb	CEA
NCA	NCA
BGPa	BGPI, TM-1
BGPb	TM-2 CEA
BGPc	TM-3 CEA
BGPd	TM-4 CEA
BGPe	4-22
BGPf	4-13
BGPg	W211
BGPh	W233
BGPi	W239
CGM1	hsCGM1
CGM1a	CGM1a, W264
CGM1b	W282
CGM1c	CGM1c
CGM2	hsCGM2
CGM6	hsCGM6, GN-1, M6, NCA-W272
CGM7	W236
CGM8	CGM8

PSG subfamily.

current name of gene or mRNA	old names of gene or mRNA
PSG1	PSβG
PSG1a	PSG93, PSβG-D, hPSP11, FL-NCA-2, hPS3, PSG1a, PSβG81
PSG1b	PSG16
PSG1c	PSβG-C
PSG1d	FL-NCA-1, PSG1d, SG9
PSG1e	PSβG-Ci, PSG95
PSG1-IIa	PSβGD
PSG1-I	PSG1-I
PSG2n	PSβG-E, SG8, hPS184
PSG3m	pSP-i, hC17, PS35, hTS16, PSβGA, SG5, hPS173
PSG4	PSG4, hsCGM4, hHSP2, FL17
PSG4a	PSG4, hPS133, PSG9
PSG5	PSG5
PSG5-In	FL-NCA-3, hPS176
PSG5-Im	PSβG-HL (clone 22)
PSG6	hsCGM3, FL26, PSGGB
PSG6r	PSG6
PSG6s	hPS12, PSG10, hPS89
PSG7	PSG7, PSGGA
PSG8	CGM35, PSG8
PSG8a	hTS1
PSG11s	PS34, PSG-G, PSβG B, PSG7
PSG11-Iw	hPS2, hPS91
PSG14	PSG14
PSG15	PSG15
PSG16a	PSG9

In 1989, following a combined CEA/PSG workshop in Freiberg, Germany, a standard nomenclature was proposed by Barnett and Zimmerman (1990) and adopted world-wide. The current gene and transcript designations are presented in Table 1.

The PSG genes are numbered 1 through 16 , e.g. PSG1. The C-terminal splice variants are indicated with a lower case letter, eg.PSG1a, while central domain splice variants are denoted by the roman numerals e.g. PSG1-IIa (Table 1 and FIG.2).

Most recent studies of the PSG include the fine mapping of the PSG and CEA genes to determine their relative positions and orientations on chromosome 19q13.1-3 (Brandriff *et al.*, 1992, Thompson *et al.*, 1992, Trask *et al.*, 1993, Tynan *et al.*, 1992, Olsen *et al.*, 1994, Teglund and Hammarstrom, 1994), promoter analysis (Lei *et al.*, 1992, Chou and Plouzek, 1992) and initial studies on differential expression of specific PSG during pregnancy (Streydio and Vassart, 1990, Chamberlin *et al.*, 1994). The evolutionary history of the PSG genes has been investigated and an evolutionary tree based on N-terminal nucleotide sequences proposed by Khan *et al.* (1992). A possible biological role for the PSG as haematopoetic growth factors has recently been investigated by Wu *et al.* (1993).

Primary structure of the PSG

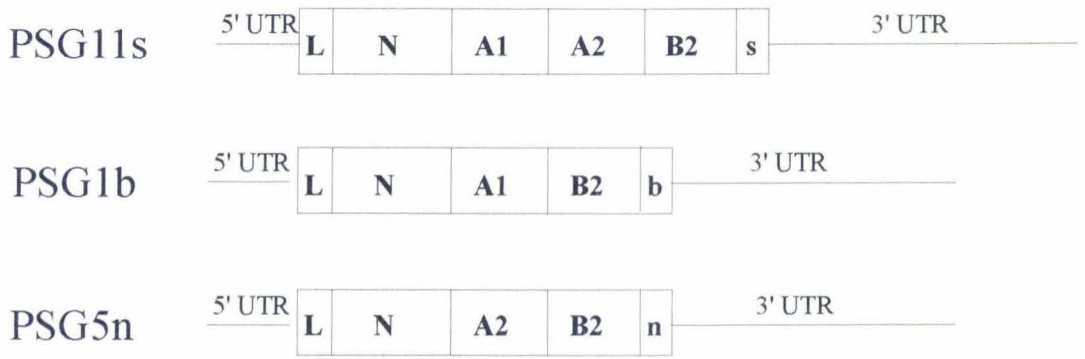
To date no direct protein sequences have been determined for the PSG. There are also no crystal structure determinations.

The primary structure of PSG, predicted from cDNA sequence, is as follows:
a short leader peptide, (L), typically 34 amino acids long, is followed by an N-terminal domain (N) of 109 amino acids (except for PSG6 which has 108), one or two similar central domains (A) of 93 amino acids, a single related central domain (B) of 86 amino acids and a C-terminus (C) of varying length (3-81 amino acids -see Table 2). The PSG proteins can be divided into two types depending on the number of central domains occurring in the protein. Type I proteins contain 3 central domains, always A1, A2 and B2. Type II proteins contain only 2 central domains, either A1, B2 (type IIa) or A2, B2 (type IIb) (FIG. 2, Chou and Plouzek, 1992).

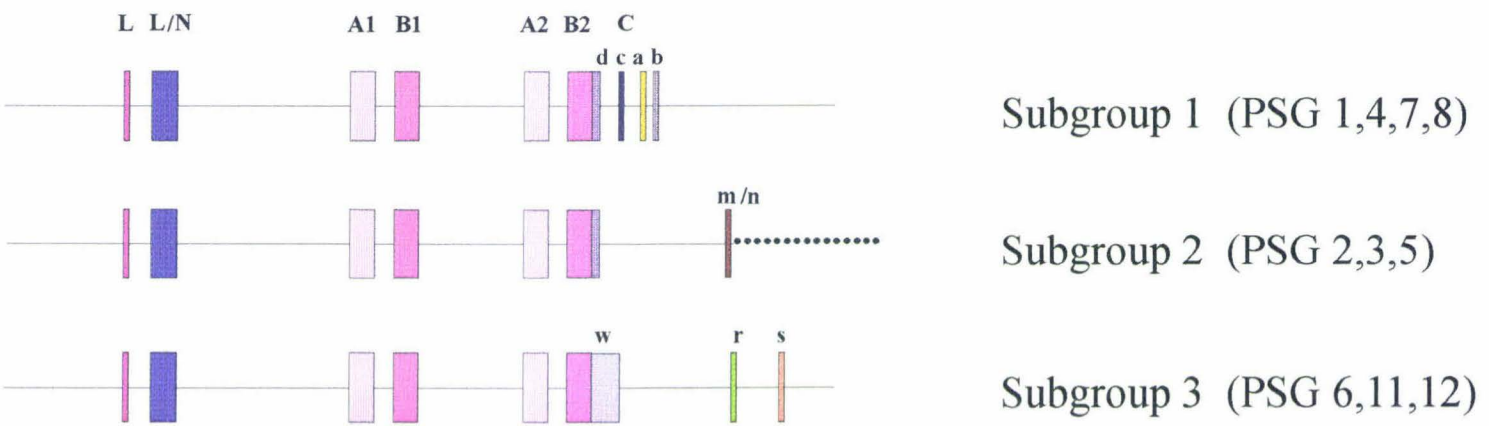
FIG. 2. Structure and organisation of the PSG cDNA and genes

The domain structure of the PSG cDNA is predicted from amino acid sequence. There is a direct correspondence between the domains and the exons. All PSG genes contain six exons. The first (L), encodes part of the leader peptide. The second (L/N), encodes the rest of the leader peptide and a complete Ig-V-like, N-terminal domain (N). The central exons (A1, B1, A2, B2) encode the IgC2-constant-type domains. The B1 exon is non-functional in all PSG characterised to date. The PSG transcripts can vary in the number of central domains they possess (Types I, IIa and IIb). Also, each PSG gene can produce multiple transcripts via alternative splicing of the C-terminal exons (Ca, b, c, d, m/n, w, r and s). The PSG genes can be divided into subgroups 1, 2 and 3 on the basis of the organisation of their 3' regions. The dotted line indicates unavailable sequence data.

SOME cDNA OF THE PSG SUBFAMILY



GENOMIC ORGANISATION OF PSG GENES



There is extensive similarity among members of the PSG family. The N-terminal domains of different PSG share around 90% and 96% similarity at the amino acid and nucleotide levels respectively, the A domains share 91% and 96% similarity respectively and the B domains share 86% and 94% similarity respectively. In addition, the N, A and B domains are similar to each other.

The C-terminal domains initially appeared to be the most variable part of the PSG . However as the genomic organisation of the different subgroups became known, it was found that comparable regions at the C terminus also share considerable nucleotide and amino acid similarity, varying between 63% (PSG 6 and 8), and 98% (PSG 2 and 5). A comparison of the C-terminal domains of reported PSG transcripts is presented in Table 2.

The N-terminal domains of all PSG, with the exception of PSG 1, 4 and 8, contain an RGD (Arg-Gly-Asp) tripeptide. This motif is the minimum functional unit for the binding of a variety of adhesive proteins of the extracellular matrix and the blood, such as fibronectin, laminin, tenascin and collagen, to their receptors, the integrins. While this motif can arise without a functional role, its presence in the N domain provides a basis for the hypothesis that it may contribute to the biological activity of the protein. RGD-mediated events include interactions with the immune system, cell recognition, cellular architecture type interactions, and tumour metastasis (reviewed in Springer, 1990 and Ruoslahti and Piersbacher, 1987). The absence of the RGD tripeptide in some PSG suggests that they may have different biological roles.

The A and B domains contain cysteine residues at conserved locations that have the potential to form disulphide bonds and make a looped secondary structure, characteristic of members of the immunoglobulin family (Paxton *et al.*, 1987, Zheng *et al.*, 1990). The N domain, in contrast, does not have these conserved cysteine residues but retains the potential to fold, through interactions between hydrophobic residues at the predicted sites, and by a conserved salt bridge (Bates *et al.*, 1992). The N domain is similar to the Ig-variable (IgV) type domains and the A and B domains to the Ig-C2 type constant domains (Williams,1987).

Table 2. PREDICTED AMINO ACID SEQUENCES FOR THE C-TERMINAL DOMAINS OF REPORTED PSG TRANSCRIPTS.

Subgroup 1 genes	Alternative C-terminal domains			
	Ca	**Cb**	**Cc**	**Cd**
PSG1	DWTVP	EAL	AYSSSINYTSGNRN	GKWIPASLAVGF
PSG4	..IL.	na	na	na
PSG7	..SL.	.S.	...G.....D.	.....
PSG8	...L.	...	.....AVY	..R..V.....I

Subgroup 2 genes	Alternative C-terminal domains
	Cm/n
PSG2	ASTRIGLLPLLNPT
PSG3	.PSGT.H..G...L
PSG5	.PSG..R.....I

Subgroup 3 genes	Alternative C-terminal domains		
	Cr	**Cs**	**Cw**
PSG6	ETASPQVTYAGP NTWFQEILLL	GPCHGNQTESH	na
PSG11	na	.....DL...ES	GKWIPASLAVGFYVESIWLSEK SQENIFIPSLCPMGTSKSQILL LNPPNLSLQTLFSLFFCFLMAD LVSGLKKVGRGLYQP

dots indicate sequence identity .
na : sequence not available.

The C-terminal domains vary from 3 to 81 amino acids. The majority are around 12 residues long. The exception is PSG11w which has an unusually long Cw domain of 81 residues. While most PSG are secreted, PSG11w is retained within the cell (Chen *et al.*, 1993). The C-terminal domains also vary in their hydrophobicity, e.g., PSG1d, 7d and 8d have hydrophobic Cd domains and PSG1c, 7c and 8c have hydrophilic Cc domains. The 22 amino acid Cr domains of PSG11 and PSG6 are hydrophobic while the 12 and 11 amino acid Cs domains of PSG11 and 6 respectively are hydrophilic.

The significance of the variability at the C terminus of PSG proteins is unknown. The cell adhesion molecule N-CAM has three splice variants with different C termini. One variant is a glycosyl phosphitidyl inositol (GPI)-anchored form (ssd N-CAM) while the other two are transmembrane forms (sd N-CAM and ld N-CAM). The GPI-anchored form is targeted to the apical surface of polarised epithelial cells, whereas the sd and ld forms are expressed on the basolateral surface (Powell *et al.*, 1991). The different isoforms then, determine the cellular destination of a particular N-CAM molecule. While a similar role for the variable C-termini of PSG molecules is hard to imagine, as the majority are secreted, it is possible that particular C-termini are involved in, or modify, particular interactions with particular cells in some way.

Genomic structure and organisation of the PSG

There have been many estimates of the number of genes in the PSG subfamily, based on a variety of methods including hybridisation cloning and sequencing of the N domains (Thompson *et al.*, 1989. Thompson *et al.*, 1990) as well as detection by PCR in blood and placental tissue (Khan *et al.*, 1992, Wu *et al.*, 1993). Most recently, a high resolution physical map has been constructed from overlapping cosmid clones, that spans the region of chromosome 19 which contains the CEA/PSG genes (Olsen *et al.*, 1994). This study has identified a total of twenty-nine genes belonging to the CEA/PSG family, including the PSG genes, genes for CEA, non-specific cross-reacting antigen (NCA), biliary glycoprotein-1 (BGP) and a number of CEA gene family members (CGM1-18). A centromeric cluster of six genes (cen- CGM10-CGM7-CGM2-CEA-NCA-CGM1), spanning 200kb, is separated by 560kb from cluster of twenty-three genes (BGP-CGM9-CGM6-CGM8-CGM12-PSG3-PSG8-

CGM13-PSG12-PSG1-PSG6-PSG7-CGM14-PSG13-CGM15-PSG2-CGM16-PSG5-PSG4-CGM17-PSG11-CGM18-CGM11-tel), which spans 860kb.

All PSG genes characterised to date contain at least six exons encoding a 5'UTR and the first 21aa of the leader peptide (L, exon 1), the rest of the leader peptide and the N domain (L/N, exon 2), an A domain (A1, exon 3), a B domain that is never translated (B1, pseudoexon 4), a second A domain (A2, exon 5) and a second B domain (B2, exon 6) (FIG.2). There is an exact correlation between the protein domains and the exons. Following the B2 domain exon are a variable, gene-dependent number of exons encoding alternative C-terminal domains and 3' UTR that are selected by alternative splicing.

It is not clear whether the difference in the number of central domains between type I (L-N-A1-A2-B2-C) and type II (L-N-A1-B2-C or L-N-A2-B2-C) proteins arises by alternative splicing or from genes with defective A1 or A2 exons (Chou and Plouzek, 1992).

It is clear, however, that the C-termini are alternatively spliced. PSG members can be divided into three different subgroups (1, 2 and 3) according to the organisation of their C-terminal domains (FIG.2).

The genomic organisation of the C-terminal domains of subgroup 1 genes has been determined independently for PSG1, 4, 7 and 8 (Lei *et al.*,1992, Thompson *et al.*, 1990, Leslie *et al.*, 1990, Oikawa *et al.*, 1988). The PSG1 locus appears to produce five PSG, each containing a unique C terminus (PSG 1a, 1b, 1c, 1d, 1e,). These transcripts are generated by alternative splicing and the use of alternative polyadenylation signals. The Cd domain exon lies adjacent to the B2 exon in all PSG genes. However only transcripts from subgroup 1 genes (PSG 1,4,7, and 8) appear to use this exon .

The genomic organisation for PSG5, a subgroup 2 gene, has been determined also (Thompson *et al.*, 1990, Oikawa *et al.*, 1989a) and the Cm/n region lies about 4.3 kb downstream from the end of the B2 domain. It appears the region lies on two exons with the Cm/n domain and the first 44bp of the 3'UTR on the first exon and the remainder more than 2kb further downstream (Thompson *et al.*, 1990). Only the first exon containing the Cm/n domain has been

sequenced and very little of the intervening intron sequence is available. The second exon is predicted from cDNA sequence. The Cm and Cn regions differ only in a deletion in the 3'UTR.

The PSG 6, 11 and 12 are subgroup 3 genes. The cDNA PSG6s, PSG6r (Zimmerman *et al.*, 1989, Barnett *et al.*, 1990, Zheng *et al.*, 1990), PSG11s and PSG11w (Arakawa *et al.*, 1991, Brophy *et al.*, 1992, Zheng *et al.*, 1990, Chan *et al.*, 1990) have been isolated and characterised.

The PSG 12 gene is unreported although a pseudogene, PSG12ψ, has been characterised (Lei *et al.*, 1993). A region that could encode a Cr domain was identified approximately 3.3kb downstream from the end of the B2 domain. A region that could encode a Cs domain was not located but was predicted to lie more than 5kb downstream from the end of the B2 domain.

Post translational modifications

The PSG are highly glycoslyated proteins with carbohydrate contents ranging from 28% to 32%. (Watanabe and Chou, 1988b). It is thought that the N-domain glycan moieties may be essential for the stability of the PSG protein (Chou and Plouzek, 1992).

Sequence analysis of the deduced PSG protein also indicates the PSG may be highly phosphorylated proteins . The proteins contain consensus sequences for phosphorylation by casein kinase II and protein kinase C. There is also a consensus sequence present in the B2 domains of all PSG, for tyrosine kinase. No experimental work has investigated the phosphorylation of PSG proteins.

PSG synthesis and expression

The primary site of synthesis of the PSG is in the syncytiotrophoblast cells of the placenta. These cells originate from the cytotrophoblast cells and form a ring around the developing blastocyst subsequent to implantation. By twelve days, the syncitium is full of large vacuoles or lacunae which form an interconnecting network. Syncitial cells penetrate into the endometrium and erode the endothelial lining of the surrounding maternal capillaries. As a consequence of

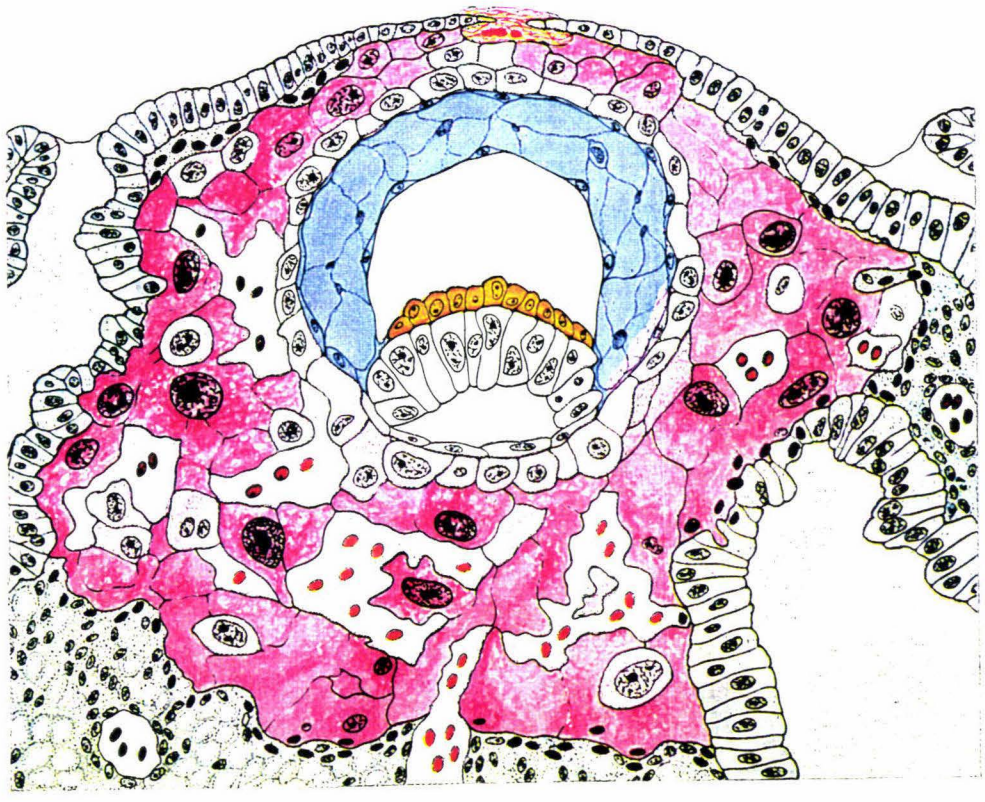
this, the lacunae fill up with maternal blood, setting up the uteroplacental circulation system by day thirteen (see FIG.3). By the end of the third week primary villi have developed. Lacunae are lined with syncitiotrophoblasts; these are the cells that have direct contact with the maternal blood and are intimately involved with the foetal/maternal exchange.(Ramsey and Donner, 1980)

Syncitiotrophoblasts are known to synthesise and secrete a wide variety of products which include oestrogenic hormones (mainly estriol), progesterone, human chorionic gonadotrophin and human chorionic somatomammotropin (Hamilton *et al.*, 1972). The PSG can be detected three to four days post fertilisation and are regulated independently of other products such as hCG.

The PSG are secreted into maternal serum throughout the duration of pregnancy in a pulsatile fashion with the pulse frequency increasing as the pregnancy progresses. Throughout gestation, PSG serum levels rise exponentially to reach a final concentration of 200-400ug/ml, primarily as a consequence of the increasing mass of the placenta. The decay of radioimmuno assayable PSG in post partum serum appears to be faster during the first 24 hours than later, with a first half-life (0-24 hours after delivery) of 20 hours and a subsequent one (after 24 hours) of 72 hours (Huttenmoser et al., 1987).

Three major PSG mRNA can been detected in placental tissue and in primary cultures of trophoblasts (Chou and Plouzek, 1992). They are 2.3, 2.2 and 1.7 kb in size and each represents a group of transcripts of similar size. Streydio and Vassart (1990) used specific oligonucleotides to detect PSG1a, 1c, 1e, 2, 3 and 11 by hybridisation in human placenta at different stages of gestation. They concluded that throughout the duration of pregnancy, PSG are expressed in a constant way, with no evidence for developmental regulation, at least for the PSG investigated. Some PSG are preferentially expressed however, PSG1a, 2 and 3 were present at higher levels than PSG1c, 1e and 11.

Further evidence for differential expression patterns of individual PSG genes in the placenta has arisen from a recent analysis of PSG gene promoters. Chamberlin *et al.* (1994) characterised the promoters of six PSG genes and have shown that they fall into two classes



Colour scheme:

Maternal tissue	Grey	Endoderm	Orange
Syncytiotrophoblast	Red	Extra-embryonic mesoderm	Purple
Cytotrophoblast	Pink	Maternal blood corpuscles	Red within dark circles
Ectoderm	White with black nuclei		

FIG. 3. A schematic representation of a human implantation site at an estimated age of 11-12 days (from Hamilton *et al.*, 1972).

which differ in their requirement for activator elements. Modulation of PSG gene expression is, at least in part, achieved by the abundance or availability of certain transcription factors and their interactions with both positive and negative DNA elements in the PSG promoters.

There is some experimental evidence for tissue specific expression of PSG. Work done by Leslie *et al.* (1990) indicates that the PSG6r transcript may be expressed only in hydatidiform mole. It may be possible to exploit this and develop mole-specific probes to enable early detection and diagnosis.

While the placenta is the major site of PSG synthesis and expression, PSG proteins are neither female or pregnancy specific. PSG cDNA clones have been isolated from libraries of testis (Borjigin *et al.*, 1990), foetal liver (Khan and Hammarstrom, 1989, Zimmerman *et al.*, 1989), salivary gland and intestine (Zoubir *et al.*, 1990), HeLa cells (Chan *et al.*, 1988a and b), myeloid cell lines (Barnett *et al.*, 1990, Oikawa *et al.*, 1989a), leukocytes, neutrophils and polymorphonuclear cells (Wu *et al.*, 1993). PSG protein synthesis however, has only been demonstrated in primary cultures of placental trophoblasts, fibroblasts, amnion cells and some human tumour cell lines (Chou and Plouzek, 1992).

All PSG except PSG11w are secreted. PSG11w is predicted to have a hydrophobic C terminus of 81 amino acids (Chan *et al.*, 1991). Preliminary studies on the *in vitro* expression of this transcript in eukaryotic cells suggest that PSG11w remains within the cells, and in fact, is not transported to the golgi bodies but remains bound to the endoplasmic reticulum until it is degraded (Chen *et al.*, 1993).

Clinical applications.

There are a number of clinical applications for the levels of PSG in pregnant serum but these are not used diagnostically in New Zealand.

PSG levels have been used to diagnose pregnancy but are generally considered to be less sensitive than the current hCG tests. Low levels of PSG can predict threatened abortion (Hertz and Schultz-Larsen, 1986, Masson *et al.*, 1983) or ectopic pregnancy (Sterzik *et al.*, 1989) and

other complications such as foetal growth retardation and intrauterine foetal death when taken in conjunction with ultrasound scans (Chou and Plouzek, 1992, Bischof, 1984, Tamsen *et al.*, 1983). High PSG levels in the amniotic fluid correlate with Meckels syndrome (Heikinheimo *et al.*, 1982). Downs syndrome can be predicted with around 78% accuracy when PSG levels are considered in conjunction with hCG, α-fetoprotein and maternal age (Petrocik *et al.*, 1990). PSG levels and hCG concentrations are useful for detecting gonadotrophin induced pregnancy, as arises in artificial fertility programmes (Rosen, 1986).

PSG levels are used as a prognosis index in breast cancer patients (Horne *et al.*, 1976, Fagnart *et al.*, 1985, Wright *et al.*, 1987) and as an indicator to monitor treatment of choriocarcinomas, hydatidiform mole and gestational trophoblast disease (Tatarinov, 1978). The uses of PSG6r as a probe for the early detection of hydatidiform mole and possibly choriocarcinomas is currently in progress (Leslie *et al.*, 1990).

The PSG in other animals.

Proteins similar to the PSG found in humans have been detected in other mammals such as rodents, sheep and cows. They are however, structurally distinct from human PSG (Thompson *et al.*, 1991, Chan *et al.*, 1988c, Turbide *et al.*, 1991).

1.3 The CEA Subfamily.

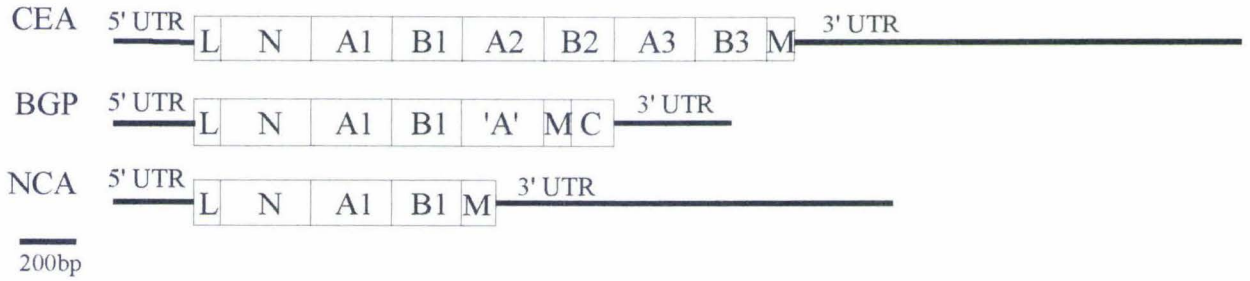
In 1965, Gold and Freedman identified carcinoembryonic antigen as a foetal and colonic cancer antigen present in colonic tumours and foetal gut (Gold and Freedman, 1965). Although CEA has since been detected at low concentrations in a range of normal tissues (Chu *et al.*, 1972), in some benign tumours (Kuroki *et al.*, 1984) and in other tumours e.g. breast tumours, it is still one of the most widely used human tumour markers for assessing the treatment of colorectal, breast and lung cancers.

Since the discovery of CEA, other closely related genes and proteins have been identified. As well as the PSG subfamily of proteins, two other CEA subgroup members are well characterised. These are non-specific cross-reacting antigen (NCA) and biliary glycoprotein-1

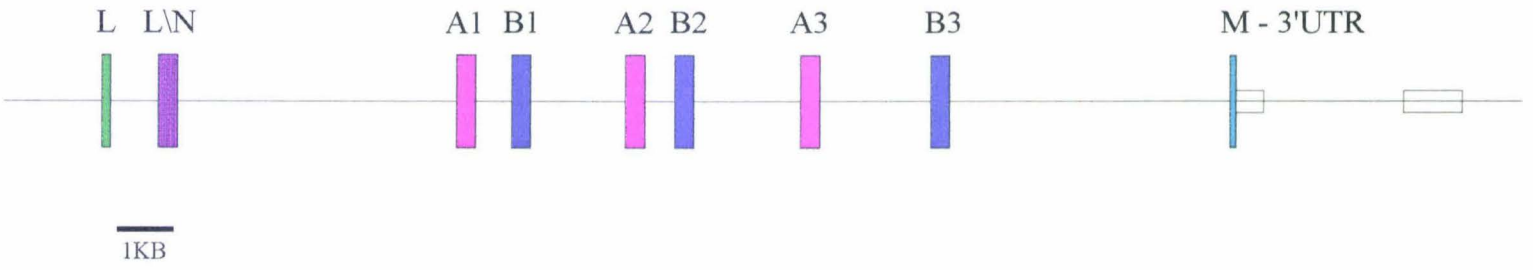
FIG. 4. Structure and organisation of the CEA subfamily cDNA and the CEA gene.

The domain structure of three CEA subfamily cDNA is based on deduced amino acid sequence (CEA from Oikawa et al., 1987, NCA from Neumaier et al., 1988 and BGP from Hinoda et al., 1988). The CEA gene (Schewe et al., 1990) contains exons that encode part of the leader peptide (L), the rest of the leader peptide and a complete N-terminal domain (N), central repeat domains (A1, B1, A2, B2, A3, B3) and a transmembrane C-terminal domain (M) and 3'UTR. The 3'UTR sequences are shown as clear boxes on the map. The genes for NCA and BGP are not fully characterised but it is known that the different numbers of central repeat domains occurs as a result of gene organisation rather than alternative splicing. The domain 'A' is a BGP-specific A-like domain and the domain 'C' is a cytoplasmic domain.

SOME cDNA OF THE CEA SUBFAMILY



GENOMIC ORGANISATION OF CEA



(BGP) which can be isolated from the colonic mucosa, serum, saliva, bile and faeces of normal individuals (Neumaier *et al.*, 1988, Barnett *et al.*, 1989, Hinoda *et al.*, 1988). Both these protein classes are immunologically crossreactive and share substantial sequence identity with CEA. Unlike the PSG , the CEA, NCA and BGP proteins can be membrane associated.

Other members of the CEA subgroup include a 128-kDa colon tumour associated antigen (TEX) and two meconium antigens of molecular weights 160-kDa and 100-kDa (Neumaier *et al.*, 1988). Whether these are products of uncharacterised genes or isoforms of known genes remains to be determined.

Recently other CEA gene-family member (CGM1-18) genes have been identified (Olsen et al., 1994) but as yet their transcriptional status is not known.

Primary structure and genomic organisation of the CEA subfamily

The primary structures of CEA, BGP and NCA proteins as predicted from nucleotide sequence data, are shown in FIG. 4. A short leader peptide of 30-34aa (L) is followed by an N-terminal domain of 108aa. The number of central domains (A-type, 93aa and B-type, 85aa) varies between subfamily members but in contrast to the PSG subfamily, this is as a result of gene organisation and not alternative splicing. As with the PSG, the N domain is Ig-V-like and the central domains are IgC2-like (Williams, 1987). The N domain and central A and B domains of PSG and CEA share around 60% homology (Streydio *et al.*, 1988).

In contrast to the PSG, members of the CEA subfamily are typically membrane bound and have a hydrophobic membrane spanning domain (M) of 26aa at the C-terminus. BGP proteins have a cytoplasmic domain (C) of 71aa in addition to the M domain.

The genomic organisation of CEA is shown in FIG.4. The genomic organisation of NCA and BGP has not yet been fully established.

The structure of CGM13-CGM18 has been characterised by the partial sequencing of these new genes (Teglund and Hammarstrom, 1994). They share identical gene organisation. A

single A domain is separated from a single B domain by 0.4kb. These domains share more than 95% identity at the nucleotide level. Approximately 6kb downstream from the end of the B domain is a region that shares 95% identity to PSG C termini and includes exons for the PSG C-terminal domains Ca, Cb and Cc. These CGM genes do not have exons encoding CEA/PSG -like N domains. It is not yet known whether these genes produce functional transcripts.

Biological roles of CEA subfamily

CEA, NCA and BGP proteins are highly glycosylated. CEA has 28 potential glycosylation sites and the carbohydrate moiety may comprise up to 60% of the total molecular weight (Thompson *et al.*, 1991).

A biological activity has been established in vitro for CEA, NCA and BGP. CEA and NCA have been shown function in vitro as intercellular Ca^{2+}-independent adhesion molecules (Benchimol *et al.*, 1989, Oikawa *et al.*, 1989b) which indicates a possible role in intestinal tissue organisation during development.

CEA and NCA both have a specific affinity for binding certain strains of *E.coli*, by way of a bacterial lectin / CEA carbohydrate interaction (Thompson *et al.*, 1991). Thus CEA may have a role in establishing intestinal flora while NCA on the surface of granulocytes, may facilitate phagocytosis.

Another possible role for CEA on the surface of migrating embryonic cells or metastasising tumour cells, may be in mediating interactions with basement membranes. CEA has been shown to facilitate binding of a colonic adenocarcinoma cell line to collagen type 1 *in vitro* (Thompson and Zimmerman, 1988). Also there is some evidence that CEA may direct metastases from colorectal cancers to the liver by binding a specific receptor on Kupffer cells (cells of the liver, responsible for removing CEA from the circulation) then interacting with CEA on tumour cell surfaces in a homotypic fashion, immobilising them and enabling the establishment of secondary tumours (Thomas and Toth, 1990).

BGP, unlike CEA and NCA, is reportedly Ca^{2+}- and temperature-dependent in its binding properties. Sequence homology to a rat ecto-ATPase may indicate a possible enzymatic role for these members of the CEA gene family (Turbide *et al.*, 1991, Rojas *et al.*, 1990).

1.4 The Biological Role(s) of PSG

Unlike the CEA subgroup, PSG are predominantly secreted proteins. It is likely therefore that they act through cellular receptors to mediate cellular interactions.

Four potential roles for the PSG have been suggested, they are:

- immunosuppression and the prevention of immune rejection of the foetus (Watanabe and Chou, 1988a, Borjigin *et al.*, 1990)

- an involvement in the invasion of the uterus by the trophoblast (Streydio *et al.*, 1988)

- a role in cellular interactions with the extracellular matrix (Rooney *et al.*, 1988)

- and growth factor activity (Wu *et al.*, 1993, Zheng *et al.*, 1990)

Experimental results, using bulk PSG (i.e. unfractionated PSG isolated from maternal serum) show that the PSG may interact directly with T cells, interfering with their normal interactions and thereby bringing about some form of immunosupression. It has been reported that PSG inhibit E-rosette formation, in a concentration dependent manner, which implies the (T-cell) CD2 : LFA-3 (erythrocyte) interaction is disrupted (Kan and Tatarinov, 1990). The PSG have also been reported to affect the proliferative activity of phytohaemagglutinin-stimulated lymphocytes (Bischof, 1984) and inhibit stimulated lymphocytes in a mixed lymphocyte assay (Zheng *et al.*, 1990).

Since bulk PSG was used for the above experiments, specific interactions could not be identified. The purity of bulk PSG is questionable, especially in the light of contamination by highly potent bioactives such as cytokines and components of the extracellular matrix. Furthermore, the above assays each involve different T-cell receptors so it is not clear whether

different PSG bind different receptors or whether they act indirectly, by binding initially uninvolved receptors. One hypothesis, currently under investigation by our group, is that PSG11 binds specific receptors that are involved in the switch to and/or the maintenance of antibody-mediated immunity and suppression of cell-mediated immunity, typical of pregnancy.

The presence of the RGD motif in some PSG suggests a role either in the invasion of the maternal tissue by the trophoblast or in the mediation or co-ordination of cell-cell interactions during embryogenesis (Rooney *et al.*, 1988, Streydio *et al.*, 1988). As discussed previously the RGD tripeptide is present in extracellular matrix proteins, has been shown to be a signal that is recognised by specific cellular receptors and has a role in controlling cell adhesion and cell migration on substrates (Streydio *et al.*, 1988, Ruoslahti and Piersbacher, 1987). Further, there is evidence of some amino acid sequence homology, albeit weak, between the PSG domains and N-CAM, fibronectin and vitronectin, three well characterised cell adhesion molecules (Rooney *et al.*, 1988). A role such as the two described above may well be consistent with the expression of PSG in tumours such as breast tumours, hydatidiform mole or choriocarcinomas.

Some viruses contain the RGD motif in their protein coats and are able to infect human cells through integrin receptors on the cell surface (Bergelson and Finberg, 1993). Perhaps one of the roles of the PSG is to block these receptors, in a non-specific manner during pregnancy.

Another role proposed for the PSG is as a growth enhancer for the cells of the haematopoeitic system. Using reverse transcriptase PCR, Wu *et al.* (1993) were able to isolate most known PSG transcripts from bone marrow and peripheral blood. The levels of expression of PSG transcripts from T lymphocytes were comparable to placental levels. The placenta is known to be a rich source of haematopoietic growth factors and PSG transcripts have also been isolated from foetal liver, a primary site of haematopoeisis in the developing foetus. This potential biological activity is currently under investigation.

1.5 Evolution of the PSG and CEA Multigene Family

Like most multigene families e.g. proteins of the blood clotting system, globin genes (Li and Graur, 1991) collagens (Buttice *et al.*, 1990), the PSG /CEA multi-gene family shows an exact correspondence between protein structural domains and exons (see FIGs 2 and 4). The different domains within each PSG locus also show significant similarity to each other. This indicates both that duplication of domains has occurred and that a single ancestral domain was a precursor in the evolution of this gene family (Streydio *et al.*, 1990).

A possible scenario for the evolution of the PSG family is as follows :

- The precursor (or primordial exon) was A or B domain-like and contained two conserved cysteine (Cys) residues with possibly an Ig-like fold. This molecule may have undergone a series of internal duplications to give a number of similar domains

- One of these domains could have then lost its Cys residues through reduced selective pressure, to become the N terminus

- Then followed a series of further duplications of the A-B unit, and/or rearrangement by unequal crossing over to give a variable number of central domains ,and/or duplication of complete gene units to give multiple genes (Streydio *et al.*, 1990)

Relevant to the question of origins of PSG is the process of concerted evolution. This is a term used to describe the observed homogeneity among members of a gene family, when processes are thought to act to prevent individual members of a family from evolving independently of other members. Thus, the family evolves as a unit, in a concerted fashion. Both the mechanism of unequal crossing over and that of gene conversion are thought to contribute to the process of homogenisation in concerted evolution (Li and Graur, 1991).

It would appear that the CEA/PSG genes have evolved by concerted evolution at least in part, by a mechanism of unequal crossing over . This is suggested by the observation that CEA family members (e.g. CEA, NCA, BGP) each have a different number of central domains (see FIG. 4). These may well be functional products of unequal cross over events. In contrast to the

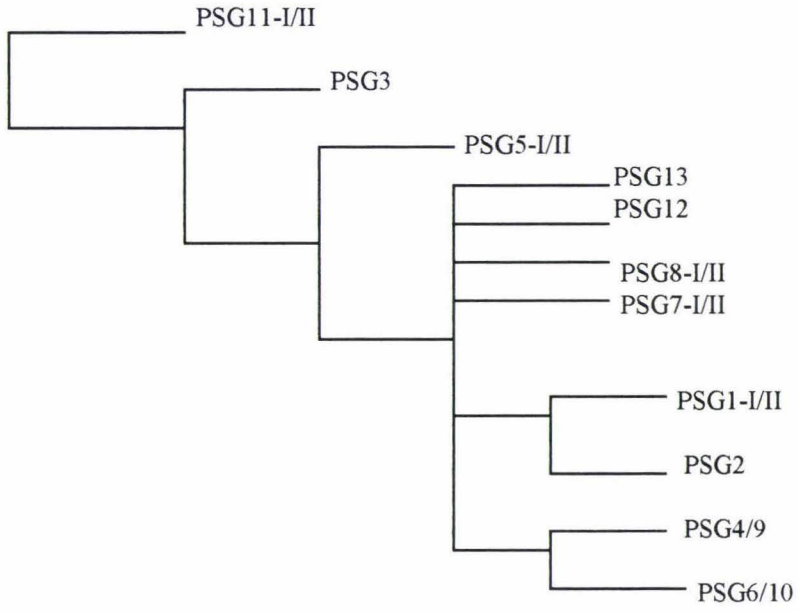


FIG. 5. Strict consensus of tied optimal trees found under parsimony using N domain sequences (from Khan *et al.*, 1992).

CEA subfamily, in the PSG subfamily all genes have the same four central domains (see FIG. 2) and different subfamily members appear to have arisen by the duplication of complete loci. This is further suggested from the intron and pseudogene sequences which show extensive similarity also (Rudert *et al.*, 1989).

A multi-gene family encoding proteins similar to the PSG/CEA human proteins has been found in rodents. Early analysis (Rudert *et al.*, 1989, Thompson *et al.*, 1989) indicated that the two families are non identical and that they have been evolving by parallel gene duplication and subsequent divergence since the mammalian radiation.

The reconstructed evolutionary tree shown in FIG.5 is that found by Khan *et al.* (1992) in a comparative analysis of N-terminal PSG sequences. It was suggested by these authors to represent the evolutionary relationships between PSG genes.

However, the PSG can be divided into three subgroups according to their types of C-termini (see Table 2 and FIG 2). In the tree presented by Khan *et al.* (1992) in FIG. 5, the different subgroups do not cluster together as would be expected if they shared a more recent common ancestor with other members of the same subgroup. In their reconstructed tree some subgroup 1 members join more closely with members of subgroups 2 or 3 (e.g.PSG1 and PSG2, PSG4 and PSG6).

1.6 Aims and Objectives.

No complete subgroup 3 PSG gene has been isolated or characterised.

A partial cDNA for PSG3 was isolated by myself and Dr. Mansfield in 1989, from a placental cDNA library (McLenachan and Mansfield, 1989). This cDNA was used by Ms.S.Sims to probe a human genomic cosmid library. Several cosmids were isolated and one of these, hC3.11, was partly characterised by myself and others of our group (Beggs, 1990, Joe, 1994, Joe *et al.*, 1994,) and was found to contain the L, N, A1, B1, A2 and B2 exons of the PSG11 gene. A map of the cosmid hC3.11 showing regions that have been sequenced is presented in FIG.6.

FIG. 6. A map of the cosmid hC3.11 which contains part of the PSG11 gene.
Sites for the restriction enzymes B*am*HI and E*co*RI are labelled. The exons for the Leader sequence (L), the N-terminal domain (N), the central domains (A1, A2, B2) and the B1 pseudoexon are shown as coloured boxes. Broken arrows above the map indicate regions of sequence obtained by our group (Beggs, 1990, Joe, 1994 , Joe *et al*, 1994, McLenachan *et al*, 1994). The upstream region of PSG11 that is largely uncharacterised is arrowed. The 0.5kb BamHI - EcoRI fragment used as a probe (see text) is marked *****.

The primary aim of this project was to complete the characterisation of the PSG 11gene. The isolation, mapping and sequence determination from this locus has now been published (McLenachan *et al.*, 1994). A secondary aim was to examine the evolution of the PSG genes using new sequences from the 3' region of a subgroup 3 gene. The hypothesis of relationship as proposed by Khan *et al.*, (1992) (FIG.5) was tested and this work is currently submitted for publication (McLenachan *et al.*, 1995). An important focus of the evolutionary analysis presented here, has been the attempt to separate historical signals from other (perhaps functional) patterns in the data.

Extensive characterisation of this extremely complex gene family will enable rational investigations into a biological role for PSG to be carried out and may provide further clues as to evolutionary history of the family.

CHAPTER 2. MATERIALS AND METHODS.

MATERIALS.

Agarose

Biorad Ultrapure DNA grade agarose.

Low melting point Sea plaque Agarose, FMC.

Cell Strains and relevant genotypes

E. coli C600 :

supE44,hsdR,thi-1,thr-1,leuB6,lacY1,tonA21

E. coli DH1 :

supE44,hsdR17,recA1,endA1,gyrA96,thi-1,relA1

E.coli XL1-Blue:

supE44,hsdR17,recA1,endA1,gyrA46,thi,relA1,lac$^-$

F'[proAB$^+$,lacI^q,lacz-M15 Tn10(tet)^r]

Enzymes

DNase 1: from BRL, 10mg/ml.

DNase-free RNase: RNase-A (Sigma) was dissolved in sterile distilled water to 2mg/ml and boiled for 15 minutes.

CIAP: calf intestinal alkaline phosphatase, from Boehringer, 1U/µl.

Proteinase K: from Sigma, made up to 25mg/ml in sterile distilled water.

Restriction Enzymes: All enzymes used were purchased from BRL unless otherwise noted and were at a concentration of 10U/µl.

T4 Kinase: from BRL, 10U/µl.

T4 Ligase: from New England BioLabs, 40U/µl.

Genomic Library

A human genomic library was constructed by Brian Mansfield.

DNA extracted from human blood lymphocytes, was partially digested with S*au*3A and ligated into the B*am*HI site of EMBL-3. The ligation was transfected into *E.coli* strain CES 200 and amplified. The total number of individual recombinants was 2.13×10^6 and the average size of the inserts was 18kb (B.C. Mansfield, personal communication.).

Herring testes DNA

Ty XIV from Sigma, made to 4mg/ml.

Mapping Kit

Lambda Terminase system and Lambda Mapping system from Amersham, catalogue nos RPN1720 and RPN1721 respectively.

Media

L Broth

1% (w/v) Tryptone, 0.5% (w/v) Yeast extract, 0.5% (w/v) NaCl, pH7.5 (Miller, 1972).

L Agar

L broth containing 1.5% (w/v) agar.

Top Agar

L broth containing 0.7% (w/v) agar and 100mM $MgSO_4$.

When required, media was supplemented with one or more of the following antibiotics:
Ampicillin- to a final concentration of 100μg/ml, from a stock solution of 100mg/ml in sterile distilled water.
Kanamycin A- to a final concentration of 30μg/ml, from a stock solution of 30mg/ml in sterile distilled water.
Tetracycline hydrochloride- to a final concentration of 10μg/ml, from a stock solution of 10mg/ml in methanol.
Chloramphenicol - to a final concentration of 15μg/ml, from a stock solution of 15mg/ml in absolute ethanol.
All antibiotics were purchased from Sigma.

Membranes

nitrocellulose (Satorius NC-extra 0.4μm cat. no 12806-41BL)
nylon (Satorius Sartorlon 0.2μm cat. no. 25007-41BL)

Phenol

All phenol used was Tris-equilibrated according to a standard method (Sambrook *et al.*, 1989). Phenol was liquified by heating to 68°C, prior to the addition of 8-hydroxyquinone to a final concentration of 0.1% (w/v). The phenol was then washed twice with an equal volume of 1M Tris-HCl pH8 and then twice with 0.1M Tris-HCl pH8, 0.2% (v/v) β-ME. It was stored under 0.1M Tris-HCl, pH8, at 4°C, in a dark bottle and discarded after 6 months.

Sequencing Kit

Sequenase Version 2.0 DNA Sequencing Kit from USB, product no. 70770.

Solutions

Denhardts (10x)
0.2% (w/v) Ficoll 400, 0.2% (w/v) PVP, 0.2% (w/v) BSA in 6xSSC.

Gel Loading Buffer (10x)

0.44% (w/v) bromophenol blue, 0.44% (w/v) xylene cyanol, 27.5% (w/v) Ficoll 400. (Sambrook *et al.*, 1989).

SSC: Standard Saline Citrate (1x)

150mM NaCl, 50mM tri-sodium citrate

TE buffer: Tris EDTA buffer

10mM Tris-HCl, pH8.0, 1mM EDTA

TBB: Tris Borate buffer

89mM Tris-HCl, 8.9mM Boric Acid, 5mM EDTA

TAE: Tris Acetate EDTA buffer

40mM Tris-acetate, 2mM EDTA

Vectors

The vector pGEM-2 (Promega Biotech) was routinely used for subcloning, restriction mapping and double stranded sequencing.

The vectors M13mp18 and M13mp19 were used for single stranded sequencing.

METHODS

2.1 Phenol-Chloroform Extraction of DNA

DNA was extracted with an equal volume of phenol. The phases were mixed for 1 minute and separated by centrifugation at room temperature either for 5 minutes in an Eppendorf microcentrifuge or, for larger samples, for 15 minutes at 3400g (4600rpm in a Heraeus Christ Labfuge GL centrifuge). The upper aqueous phase was then removed and extracted with an equal volume of chloroform:isoamyl alcohol (24:1 v/v).

2.2 Ethanol Precipitation of DNA

To precipitate DNA, 0.1vol of 3M NaOAc pH 5.5 and 3vol of cold (-20°C) 95% ethanol were added to the solution and the mixture stood for 15 minutes at -70°C, or for at least 2 hours at -20°C. The DNA was then pelleted by centrifugation for 15 minutes at 4°C in a microcentrifuge, or for 30 minutes at 4°C at 6000g (6000rpm in the Heraeus Christ centrifuge). The pellet was washed in 70% ethanol, centrifuged for a further 5 minutes, dried for 5-10 minutes under vacuum in a Savant Speedvac Concentrator and finally resuspended in sterile distilled water or TE buffer.

2.3 Restriction Enzyme Digests

Digests of Genomic DNA

Typically, 20 - 60μg of human genomic DNA was digested for 16 hours at 37°C with 100U of enzyme, in the buffer provided by the enzyme manufacturer. The volume of DNA to be digested was less than or equal to 10% of the final reaction volume. After checking the digest had gone to completion, by electrophoresing a small amount on a

mini-gel, the DNA was then precipitated with ethanol and resuspended in sterile water or TE buffer.

Digests of Cosmid and Plasmid DNA

Typically, 50 - 250ng of DNA was digested in a total volume of 25µl with 10U of enzyme in the buffer supplied by the enzyme manufacturer. Following an incubation for 1 hour at 37°C, DNase-free RNase was added to a concentration of 2µg/ml and the incubation continued for a further 2 minutes. The digest was then stopped by the addition of 0.1vol loading dye.

2.4 Agarose Gel Electrophoresis

Digested DNA was size fractionated by electrophoresis through a horizontal agarose gel matrix in TAE buffer. The concentration of the agarose was varied according to the size range of the fragments to be resolved. Typically 0.7% (w/v) agarose was used for fragments between 1kb and 15kb. Fragments less than 1kb were separated through a matrix of 1.2% (w/v) agarose. Mini gels were electrophoresed in TAE buffer at 100V for 1 - 1.5 hours while large gels were electrophoresed at 30V/cm for 16 hours. Following electrophoresis, the DNA was visualised by staining in 5µg/ml ethidium bromide for 10 minutes and then destaining in distilled water for 10 minutes prior to viewing on a short wave U.V transilluminator. Photographs were taken with Polaroid Type 665 film using an orange filter.

The BRL 1kb size markers were used to calibrate each gel. Large gels intended for hybridisation experiments were photographed alongside a ruler to allow the sizing of fragments after blotting and hybridisation.

2.5 Isolation of DNA

Cosmid DNA

Cosmid DNA was prepared according to a standard method from Sambrook *et al* (1989).

To prepare large quantities of cosmid DNA, 600ml of L broth, supplemented with Kanamycin at 30μg/ml, was inoculated 1 in 100 from an overnight culture and incubated on a shaking platform at 225rpm, at 37°C, for 16 hours. The cells were pelleted by centrifugation for 10 minutes at 3500g (4500rpm, GSA), at 4°C, resuspended in 10ml of ice cold Solution 1 (50mM glucose, 25mM Tris-HCl pH8, 10mM EDTA) then made 250μg/ml with lysozyme. Following a 5 minute incubation at room temperature, 2vol of ice-cold Solution 2 (0.2M NaOH, 1%(w/v) SDS) were added and the solution stood on ice for 10 minutes. Then 0.5vol of Solution 3 (5M KOAc, pH4.8) was added and the mixture stood on ice for 10 minutes before clarification by centrifugation for 5 minutes at 4°C at 12000g (10000rpm, SS34). The supernatant was transferred to two 30ml glass Corex tubes and extracted with an equal volume of phenol: chloroform: isoamyl alcohol (24:24:1 v/v/v). The DNA was then precipitated by the addition of an equal volume of isopropanol, stood at -20°C for 2 hours, then pelleted by centrifugation for 15 minutes at 4°C at 27000g (15000rpm, SS34). The pellet was washed with 70% ethanol, air dried briefly and resuspended in 400μl of TE buffer. Typical recoveries of cosmid DNA were 1mg per 600ml culture.

Lambda DNA

Lambda DNA was extracted from phage lysates prepared on *E.coli* C600, grown on L plates made with agarose.

Typically, phage lysates were prepared by plating 10^5 pfu of the appropriate phage per 88mm plate. The plaques were grown to confluence (about 6 hours) at 37°C then cooled

to 4°C and overlaid with 10ml of SM buffer (100mM NaCl, 10mM MgSO4, 50mM Tris-HCl pH7.5, 0.01% (w/v) gelatin). Following a 16 hour incubation at 4°C, the buffer overlay was collected and centrifuged for 10 minutes at 4°C at 4500g (5000rpm in the Heraeus Christ centrifuge) to remove cellular debris. An aliquot of the supernatant, typically 5ml, was used for DNA extraction; the remainder was stored at 4°C over chloroform to prevent bacterial growth.

For DNA extraction from the phage particles, contaminating bacterial DNA and RNA were first removed by making the lysate 1μg/ml with respect to each of RNase A and DNase I and incubating for 30 minutes at 37°C. The phage were precipitated with an equal volume of PEG solution (20% (w/v) PEG 6000, 2M NaCl in SM buffer), stood on ice for at least 2 hours then pelleted by centrifugation for 30 minutes at 4°C at 8800g (7000rpm in the Heraeus Christ centrifuge). The pellet was resuspended in 0.5ml SM buffer, brought to 0.1% (w/v) SDS, 5mM EDTA and incubated at 68°C for 15 minutes. Following a phenol-chloroform extraction, the DNA was precipitated by the addition of an equal volume of isopropanol, and incubated for 20 minutes at -70°C. The DNA was pelleted by centrifugation for 5 minutes at 4°C in a microcentrifuge, washed with 70% ethanol, dried under vacuum and resuspended in 50μl of TE buffer containing 0.01μg/ml DNase-free RNase. Typical recoveries were 250μg/ml per 5ml lysate.

Human Genomic DNA

High molecular weight genomic DNA was prepared from human term placental tissue obtained by informed consent according to protocols approved by the Massey University Ethics committee and the Palmerston North Hospital Ethics committee.

Freshly delivered placenta was washed in ice-cold phosphate buffered saline (1.7M KH_2PO_4, 5mM Na_2HPO_4, 150mM NaCl, pH 7.4), frozen in liquid air and stored at -70°C. For DNA extraction, about 300mg of tissue was crushed, resuspended in TE buffer containing 1% (w/v) SDS, 50μg/ml proteinase K, in a volume of 20 ml, and incubated overnight at 37°C. Following digestion, the suspension was extracted with an

equal volume of phenol and the phases separated by centrifugation for 5 minutes at 1700g (1000rpm in the Heraeus Christ centrifuge) at 4°C. This was repeated four or five times until the aqueous phase became clearer and the material trapped at the interface was minimal. The aqueous phase was then extracted three or four times with chloroform:isoamyl alcohol (24:1 v/v) until there was little or no material trapped at the interface. The DNA was precipitated out of the aqueous phase by the dropwise addition of 0.58 vol of isopropanol and collected by spooling on to a sterile glass rod. After air-drying, the DNA was resuspended in approximately 3ml TE buffer. Typically 600µg/ml DNA was obtained per 300mg of tissue.

Plasmid DNA

Rapid purification of Plasmid DNA

Rapid isolation of plasmid DNA was performed essentially according to the method of Holmes and Quigley (1981)

A colony of transformed cells was picked into 5ml L broth containing 100µg/ml ampicillin and grown for 16 hours at 37°C on a shaking platform at 225rpm. A 1.5ml aliquot was then removed, centrifuged in a microcentrifuge for 5 minutes at room temperature, and the resulting cell pellet resuspended in 350µl HQStet buffer (8% (w/v) sucrose, 5% (v/v) Triton X-100, 50mM EDTA pH8, 10mM Tris-HCl pH8). Lysozyme was then added to a concentration of 0.7mg/ml and the suspension placed in a boiling water bath for 40 seconds. Following a 10 minute centrifugation in a microcentrifuge at room temperature the resulting gelatinous pellet was removed with a sterile toothpick and an equal volume of isopropanol was added to the supernatant. After standing for 10 minutes at room temperature, the DNA was pelleted from the solution by centrifugation for 15 minutes in a microcentrifuge at 4°C, washed with 95% ethanol, dried under vacuum and resuspended in 50µl of either sterile water or TE buffer. Yields of 250mg/ml were routinely obtained from cells transformed with pGem-2 based plasmids.

Large Scale Extraction of Plasmid DNA

A large volume, typically 250ml, of L broth containing 100μg/ml ampicillin was inoculated 1 in 100 from an overnight culture of a clone of interest and grown on a shaking platform at 225 rpm for 3 - 5 hours at 37°C to medium cell density. Chloramphenicol was added to 150μg/ml and the incubation continued for a further 16 hours. The cells were then harvested by centrifugation for 5 minutes at 10000g (9000rpm,GSA), washed by resuspension in an equal volume of ice-cold sterile TE buffer and pelleted again. DNA was extracted according to the alkaline lysis method (Birnboim and Doly, 1979). Briefly, the cells were resuspended in 9 ml Solution 1 (50mM glucose, 25mM Tris-HCl pH8.0, 10mM EDTA). Lysozyme was then added to 5mg/ml and the mixture stood for 30 minutes on ice prior to the addition of 10 ml of Solution 2 (0.2N NaOH, 1% (w/v) SDS). Following a further 20 minute incubation on ice, 8.9 ml Solution 3 (5M KOAC pH 4.8) was added, gently mixed, and the solution clarified by centrifugation for 1 hour at 4°C at 27000g (15000rpm, SS34). The supernatant was decanted into 12 ml isopropanol, stood at 4°C for at least 1 hour and the DNA pelleted by centrifugation for 30 minutes at 4°C at 1900g (4000rpm, SS34). Following a 95% ethanol wash, the DNA pellet was drained well, resuspended in 9 ml TE buffer and centrifuged through a CsCl - ethidium bromide gradient for 5 hours at 15°C at 285000g (55000 rpm, TV865) in a Sorvall ultracentrifuge. The DNA was collected and the ethidium bromide extracted with an equal volume of water-saturated isobutanol. This step was repeated until the pink colour of the dye had been removed completely from the aqueous phase. The DNA was finally dialysed against 2.5 litres of cold sterile TE buffer. Typically 1.5mg of DNA with an A_{260}/A_{280} ratio of 1.8 was obtained per 250ml of culture.

Estimation of DNA Concentration

DNA concentration was estimated from the optical density measured at 260nm and 280nm, assuming that $1A_{260}$ is equivalent to 50μg/ml. The purity of the DNA sample was estimated by the ratio of A_{260}/A_{280}.

When the volume of DNA was small, or the concentration was less than 500ng/μl, a fluorimeter (Hoefer TKO 100) was used to estimate the concentration accurately. The apparatus was calibrated first, using Hoechst Dye (0.1μg/ml H33258 in 100mM NaCl, 10mM Tris-HCl, 1mM EDTA pH7.4) and DNA of a known concentration (calf thymus DNA, 100μg/ml). A small sample of DNA was then combined with Hoechst Dye and the fluorescence intensity measured in the fluorimeter. The concentration of DNA was determined from a standard curve.

2.6 Mapping Procedures

Mapping the cosmid clones

Attempts to apply the *cos*-mapping system to the cosmid clones were unsuccessful due to the inability of the lambda terminase preparation to linearise the cosmids. The cosmids had been constructed in the vector, Lawrist-5 (De Jong *et al.*, 1989) which contains two asymmetric S*fi*I sites. Three oligonucleotides were designed so that two linkers, specific for the left and right arms of an S*fi*I cut cosmid, were formed when the oligonucleotides were annealed, as seen in FIG.7 (OL-1, OL-2 and OL-3). This allowed a modified approach of the S*fi*I Linker Mapping technique (Promega Protocols and Applications Guide, 1991, see FIG. 7) to be used.

Briefly, 1μg of cosmid DNA was cut with 8U S*fi*I and dephosphorylated in a reaction containing 5ng BSA, 10mM Tris-HCl pH7.8, 50mM NaCl, 10mM $MgCl_2$, 10mM β-ME, and 1U CIAP, at 50°C for 1 hour. Following digestion, the DNA was extracted with phenol and chloroform, ethanol precipitated and resuspended to a concentration of 250μg/ml.

The complementary oligonucleotide OL-3, was radiolabelled in a reaction containing 20pmol oligonucleotide, 60μCi γ[^{32}P]ATP (3000Ci/mol, 10μCi/μl,) 50mM Tris-HCl pH7.8, 10mM $MgCl_2$, 5mM DTT, 0.1mM spermidine and 5U T4 Kinase, at 37°C for 10 min. Following the labelling reaction, the enzyme was inactivated by heating at 100°C for

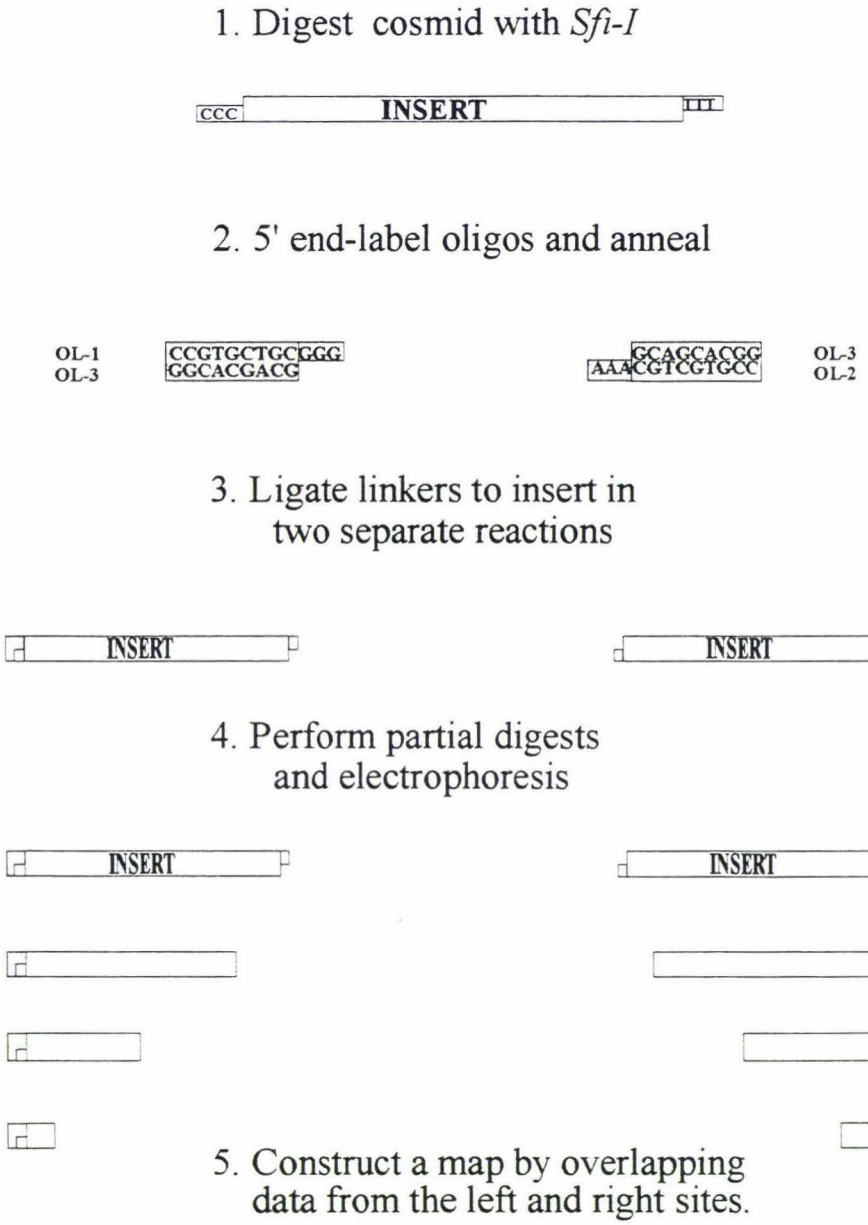


FIG. 7. S*fi*I Linker Mapping.

1 min, the reaction split in half and the complementary oligonucleotide simply mixed with 10pmol of unlabelled left or right oligonucleotide to form labelled linkers.

Next, the linkers were ligated to either the left or right S*fi*I site in a reaction containing 250ng S*fi*I-cut and dephosphorylated cosmid DNA, 5pmol labelled linker, 30mM Tris-HCl pH7.8, 10mM $MgCl_2$, 10mM DTT, 1mM ATP and 2.5U T4 ligase, at 14°C for 16 hours.

Conditions for partial digestion of the cosmids were established using 250ng of S*fi*I cut cosmid DNA , decreasing concentrations of enzyme (1U, 0.5U and 0.1U) and examining the time course of the 40 min reactions at 10 min intervals. Aliquots of the digests were removed into stop buffer (final concentration 5% (v/v) glycerol, 0.05% (w/v) SDS, 100mM EDTA, 0.01% (w/v) bromophenol blue), combined, and electrophoresed through a 0.7% (w/v) agarose gel in TAE buffer. For the enzyme B*am*HI, reaction conditions of 0.5U enzyme in buffer (REACT 3, BRL), at 37°C for 40 minutes (combining four 10 minute samples into stop buffer), were found to give a good spread of partial fragments. These conditions were then applied to the cut, labelled cosmid DNA.

The resulting cosmid fragments were separated by electrophoresis through a 0.4% (w/v) agarose gel in TBB buffer for 24 hours at 2V/cm.

High molecular weight markers (BRL) were hybridised with labelled *cos* oligonucleotides specific for the left or right lambda arms as described in the protocol provided with the Amersham Lambda Mapping System and electrophoresed as size standards.

Following electrophoresis and autoradiography, it was noted that there was a high background, presumably caused by unincorporated radionucleotide distributed throughout the gel and particularly at the buffer front. It was found that two or three changes of running buffer during the course of electrophoresis, with the first change 3

hours after beginning the run, greatly reduced this background. With this modification it was possible to dry the gel directly onto DE81 cellulose paper for autoradiography. This avoided the need to blot the gel onto nitrocellulose as recommended in the standard protocol.

After drying, the gels were exposed to X-ray film (Fuji RX of Kodak XAR) for at least 16 hours.

A standard migration curve was plotted from the size markers. Maps of the cosmids were constructed from both the left and the right *Sfi*I sites and consensus maps constructed from these data.

2.7 Cloning Procedures

Preparation of Insert DNA

DNA was digested with appropriate restriction enzymes to release the fragment of interest. The digest was then electrophoresed through a 1% (w/v) low melting point agarose gel in TAE buffer for 1-1.5h at 100V, stained with ethidium bromide and the DNA visualised under long wave UV (350-450nm). The fragment of interest was cut out of the gel and separated from the agarose by centrifugation through glass wool for 2 minutes in an microcentrifuge. (Heery *et al.*, 1990) The eluted fragment was used either directly in a ligation or concentrated by ethanol precipitation for further use.

Linearisation and Dephosphorylation of vector DNA

The vector was linearised with a restriction enzyme, then dephosphorylated by the addition of 1U of CIAP for 10 minutes at 37°C. This was followed by the addition of a further 1U of CIAP and a further 10 minute incubation. The reaction was terminated by the addition of proteinase K to a final concentration of 50μg/ml and SDS to a final concentration of 1% (w/v). Following a 1 hour incubation at 37°C the DNA was

extracted with phenol and chloroform, ethanol precipitated, washed with 70% ethanol, dried and resuspended in sterile distilled water to a concentration of 250ng/µl.

Ligation

Ligation reactions were performed typically in a final volume of 10µl with equimolar amounts of vector and insert DNA at a total DNA concentration of approximately 25ng/µl, using 40U of T4 DNA Ligase in 66mM Tris-HCl pH7.6, 6.6mM $MgCl_2$, 10mM DTT, 66µM ATP, for 16 hours at 4°C.

2.8 Transformations

Preparation of $CaCl_2$-Competent Cells

E. *coli* cells were made competent by treatment with $CaCl_2$ according to standard procedures (Cohen *et al.*,1972).

L broth (50ml) was inoculated 1:50 with an overnight culture of the appropriate *E.coli* strain and grown at 37°C on a shaking platform at 225rpm to mid log phase (A_{550} = 0.5, usually 2 - 4 hours). The cells were then pelleted by centrifugation for 10 minutes at 2400g (3800rpm, MSE bench centrifuge) at 4°C, resuspended in an equal volume of 50mM $CaCl_2$, stood on ice for 30 minutes, pelleted again and resuspended in 1/25 vol of 50mM $CaCl_2$. The cells were kept on ice and used within 48 hours.

Transformation of E.coli DH1 cells

Competent cells, typically 300µl, were mixed with 5µl of the ligation reaction in a glass Kimax tube and stood on ice for 20 minutes. Following a heat shock at 42°C for 90 seconds, 700µl of L broth was added and the cells incubated for 30 minutes at 37°C on a

shaking platform at 225rpm. An aliquot, typically 100µl, was spread on a fresh L plate supplemented with ampicillin (100µg/ml) and the plates incubated for 16 hours at 37°C. Ampicillin resistant colonies were picked and grown for DNA extraction.

Transformation frequencies using $CaCl_2$ competent cells were typically between 10^4 and 10^6 transformants/µg DNA.

Transfection of E.coli XL1-Blue

Competent cells, typically 300µl, were mixed with 5µl of the ligation reaction in a glass Kimax tube and stood on ice for 20 minutes. Following a heat shock at 42°C for 90 seconds, the solution was made 0.3% (v/v) with respect to X-gal (50µl of 2% (w/v) X-gal in DMF), 6mM with respect to IPTG (25µl of 100mM IPTG in H_2O), combined with 3.5ml of L top agar and plated onto dry L plates. The plates were incubated for 16 hours at 37°C.

2.9 Electroporation

Preparation of Electrocompetent Cells

One litre of L broth was inoculated 1:100 with a fresh overnight culture of *E.coli* DH1 and grown at 37°C on a shaking platform at 225rpm to a density of A_{600}=0.6. The cells were chilled on ice for 15 minutes then pelleted by centrifuging for 15 minutes at 4°C at 2600g (4000rpm,GSA). The supernatant was discarded and the cells were washed 3 times, first with an equal volume of ice-cold, sterile distilled water, second with a half volume of ice cold sterile distilled water and third with a quarter volume of ice cold sterile 10% (v/v) glycerol. Finally the cells were resuspended in 2ml of sterile 10% (v/v) glycerol, aliquoted (120µl) into microcentrifuge tubes, snap frozen in liquid air and stored at -70°C.

Electroporation

Electroporation was performed using a Bio-Rad Gene Pulser apparatus and Pulse Controller.

Prior to electroporation, electrocompetent cells were thawed on ice. Then, 1-5µl of the ligation reaction was mixed with 40µl of thawed cells in an ice cold microcentrifuge tube, transferred to a cooled 0.2mm electroporation cuvette and placed in the cooled gene pulser apparatus. A pulse was delivered to the cells using the settings 2.5V, 25µF, 200 Ω, then 700µl ice cold L broth was added immediately to the cuvette. The cells were transferred to a glass Kimax tube and incubated for 30 minutes, at 37°C, at 225rpm, on a shaking platform. Following incubation, an aliquot of cells (typically 10-100µl) was spread on to an L plate supplemented with ampicillin (100µg/ml). The plates were incubated for at least 16 hours at 37°C.

Transformation frequencies for electroporation were typically 10^8 to 10^9 transformants/µg DNA.

2.10 Sequencing

Preparation of double stranded template DNA

Large scale plasmid preparations were used directly for double stranded sequencing.

For DNA extracted by the rapid boil method of Holmes and Quigley (section 2.5), a quick clean-up was performed. Following isopropanol precipitation, the DNA was resuspended in 100µl of sterile distilled water. Into this, 100mg CsCl was dissolved and ethidium bromide added to 1% (w/v). Following centrifugation for 15 minutes in a microcentrifuge at room temperature, the coloured supernatant was extracted repeatedly with an equal volume of water saturated isobutanol until the colour disappeared. The DNA was precipitated by the addition of 400µl of TE buffer, 50µl NaOAc pH 5.5 and

720µl of isopropanol, washed with 70% ethanol, dried and resuspended in 20µl of sterile distilled water. Typically 5µl (3µg), was used for sequencing (Saunders and Burke, 1990).

Alternatively, DNA isolated by the rapid method was digested with RNase A (80µg/ml for 2 minutes at 37°C), extracted with phenol and chloroform, ethanol precipitated, washed with 70% ethanol, dried and resuspended in 50µl of sterile distilled water. Typically 7µl (3µg) was used for sequencing.

Prior to sequencing, double stranded DNA was denatured according to the protocol supplied with the Sequenase Version 2 sequencing kit. Briefly, the DNA was denatured by the addition of 0.1 vol 2M NaOH, 2mM EDTA pH8.0 and incubated at 37°C for 30 minutes. Following this, the DNA was ethanol precipitated, washed with 70% ethanol, dried and resuspended in an appropriate volume.

Preparation of single stranded DNA template

Single stranded recombinant M13 DNA was prepared according to a standard protocol (Sambrook *et al.*, 1989).

An M13 recombinant plaque was picked with a sterile pasteur pipette into 2 ml L broth which had been inoculated 1:100 from an overnight culture of *E.coli* strain XL1-Blue, and grown for 6 hours, at 37°C, at 225rpm, on a shaking platform. The cells were pelleted in a microcentrifuge for 4 minutes at room temperature. The cell pellet was kept for the isolation of double stranded DNA. The supernatant, containing the phage, was decanted into 1/8 vol of a solution containing 20% (w/v) PEG 6000 and 2.5M NaCl. The mixture stood for 30 minutes at room temperature, then the phage were pelleted by centrifugation in a microcentrifuge for 5 minutes. The supernatant was discarded and any remaining traces of the supernatant removed by wiping the inside of the tube with tissue. The phage pellet was resuspended in 100µl TE buffer and extracted with phenol and chloroform. The DNA was ethanol precipitated, washed with 70% ethanol, dried and

resuspended in 30µl TE buffer. A 1µl aliquot was electrophoresed through a 1% (w/v) agarose gel in TAE buffer to determine the yield of DNA. Typically 5µl (1µg) was used for sequencing.

Double stranded recombinant M13 DNA was prepared from the cell pellet using the rapid boil method of Holmes and Quigley (section 2.5) and resuspended in 25µl. An appropriate digest was done to confirm the presence of an insert and the size of the insert was determined by agarose gel electrophoresis against the BRL 1kb size markers.

Nucleotide Sequencing Reaction

Sequencing was carried out using the dideoxy chain termination method of Sanger modified for T7 polymerase (Sequenase, USB).

Single stranded recombinant M13 DNA was sequenced using 0.5pmol of the -40 universal primer provided in the kit.

Double stranded DNA cloned in the vector pGEM-2 was sequenced using 0.5pmol of commercially prepared T7 or SP6 primer (Promega Biotech) complementary to the T7 and SP6 promoter regions which lie on either side of the multiple cloning site of the vector pGEM-2.

In some cases 0.5pmol of synthetic oligonucleotides complementary to specific regions of the PSG11 gene, were used to prime sequencing reactions.

Denaturing Polyacrylamide Gel Electrophoresis

Sequencing reactions were electrophoresed through a 6% (w/v) polyacrylamide (38:2 acrylamide:*bis*-acrylamide) gel containing 8M urea, in sequencing buffer (135mM Tris-HCl, 45mM Boric acid, 2.5mM EDTA) for 6 hours at 65 Watts constant power. Typically 3µl of sequencing reaction was loaded initially (long run) and a second 3µl

volume loaded after 4 hours electrophoresis (short run). Following electrophoresis, gels were transferred directly on to Whatman 3MM chromatography paper, dried in a slab gel drier (BioRad model 583) for 40 minutes at 85°C and exposed to X-ray film (Fuji RX or Kodak XAR) for at least 16 hours at room temperature. Typically 350bp of sequence could be read accurately.

2.11 Southern Blotting

Restriction enzyme digests of DNA were electrophoresed through a 0.7% (w/v) or a 1% (w/v) agarose gel in TAE buffer, at 30V/cm, for 16 hours, to achieve good separation of fragments. The gel was stained with ethidium bromide, destained for 15 minutes in water, photographed alongside a ruler under short wave UV light and transferred to nitrocellulose or nylon membrane by vacuum blotting.

To assemble the vacuum blotting apparatus (Pharmacia), two pieces of Whatman 3MM chromatography paper that had been wet with 2x SSC, were placed on top of the porous bed. The nitrocellulose or nylon filter cut to 1-2mm larger than the gel, was laid over the wet paper, and a plastic mask with a window 1-2mm smaller than the gel, was then laid on top of the filter. The gel was placed over the window, the sealing frame was fixed in place, and a vacuum of 5mmHg applied.

The DNA in the gel was first depurinated for 8 minutes by overlaying the gel with 2% (v/v) HCl, denatured with 1M NaOH, 0.5M NaCl, for a further 8 minutes and neutralised with 1M Tris-HCl, 3M NaCl, pH5.5, for 8 minutes. Following neutralisation, the gel was overlaid with transfer buffer (20x SSC), for 30 minutes. After transfer, the solution was discarded, the wells were marked through the gel on to the filter, and the apparatus was dismantled. To immobilise the DNA, nitrocellulose filters were then baked at 80°C, 80kPa for 1-3 hours. Nylon filters were exposed to short wave UV on a transilluminator for 1 minute.

2.12 Probe Making and Hybridisation

Nick translation

Plasmid DNA was radiolabelled according to the protocol of Rigby *et al.* (1977). Typically, 200ng of DNA was labelled in a reaction containing 0.1mM each of dATP, dGTP and dTTP, 10U DNA Polymerase 1, 1pg DNAse 1 (diluted from a 1mg/ml stock in glycerol), 50μCi α[^{32}P] dCTP (3000Ci/mol, 10μCi/μl,), 50mM Tris-HCl pH 7.5, 5mM $MgCl_2$, 10mM β-ME, in a total volume of 25μl, for 15 minutes at 15°C.

Following labelling, the reaction was diluted to 200μl with TEN buffer (TE containing 0.1M NaCl). The extent of incorporation of the radiolabel was determined by thin layer chromatography on phosphoethyleneimine paper developed in 2N HCl and was typically 60% of the input isotope. Unincorporated nucleotides were removed from the probe by centrifugation through a Sephadex G50 minispin column equilibrated with TEN buffer, for 3 minutes at 1600g (3000rpm, MSE bench centrifuge). The specific activity of the probe was determined by counting 1μl of spun probe and was typically $1x10^8$ cpm/μg DNA.

5'-endlabelling of oligonucleotides

Typically 50pmol of oligonucleotide DNA was labelled at the 5' end using 10U T4 Kinase in kinase buffer (50mM Tris-HCl pH7.5, 10mM $MgCl_2$, 10mM DTT) and 30μCi γ[^{32}P]ATP (3000 Ci/mol, 10μCi/μl) at 37°C for 1 hour.

Pre-hybridisation

Filters were prehybridised for 1-3 hours at the calculated hybridisation temperature in 6x SSC, 10x Denhardt's solution, for a nick translated probe, or in 6x SSC, 0.5% (w/v) SDS, 0.05% (w/v) sodium pyrophosphate, 10x Denhardt's solution, for an oligonucleotide probe.

Probe denaturation

Before the probe was added to the filter, it was diluted with 0.5 vol each of herring testes DNA (4mg/ml, sonicated) and TNES buffer (10mM Tris-HCl pH8, 10mM NaCl, 2mM EDTA, 0.1% (w/v) SDS) and denatured in a boiling water bath for 5 minutes.

Hybridisation conditions

For nick-translated probes, the melting temperature for the probe DNA was calculated was calculated using the formula:

$$\mathbf{T_m = 81.5^{\circ}C - 16.6(\log_{10}[Na^+]) + 0.41(\%G+C) - 0.63(\%formamide) - 600/l}$$

(Bolton *et al.*, 1962)

where l = length of the probe in bp and (%G+C) is the relative amount of guanosine and cytosine in the probe DNA. For DNA from the human genome, (%G+C) is 40%.

Hybridisations were carried out at 15°C below T_m (usually 65°C or 68°C) in 1M NaCl, 50mM sodium phosphate pH6.5, 2mM EDTA, 0.5% (w/v) SDS and 10x Denhardts) in a rotary oven (Bachoefer) for 16 hours.

Filters were washed twice, for 15 minutes each, at the hybridisation temperature in 2x SSC, 0.1% (w/v) SDS then twice, for 15 minutes each, at the hybridisation temperature in 1x SSC, before being air dried, covered in plastic wrap and exposed to X-ray film (Fuji RX or Kodak XAR) for at least 16 hours.

For oligonucleotide probes labelled at the 5' end, the melting temperature for the oligonucleotide DNA was calculated using the equation:

$$T_m = 2(A+T) + 4(G+C)^{\circ}C$$ (Itakura *et al.*, 1984)

where (A+T) is the number of adenosine and thymidine bases and (G+C) the number of guanosine and cytosine bases in the probe DNA.

Hybridisations were carried out 15°C below T_m, in 6x SSC, 0.05% (w/v) sodium pyrophosphate, 20x Denhardt's solution, for 16 hours. The filters were washed three times, for 10 minutes each, in 6x SSC, 0.05% (w/v) sodium pyrophosphate, at room temperature, with a final wash of 2 minutes at the hybridisation temperature. Filters were air dried, covered in plastic wrap and exposed to X-ray film for at least 16 hours.

2.13 Library Screening

Preparation of plating cells

A culture of *E.coli* C600 cells was grown overnight in 10 ml L broth to stationary phase. The cells were then pelleted by centrifugation for 10 minutes at 2400g (3800rpm in a Heraeus Christ centrifuge) at 4°C, resuspended in 5ml ice-cold 10mM $MgSO_4$ and stood on ice for 30 minutes.

Determination of Library titre

In order to determine the titre of the library, serial dilutions of the phage lysate were assayed. Typically 200µl of plating cells were mixed with the lysate in glass kimax tubes, stood for 30 minutes at 37°C, then plated with 3.5ml top agar on 88mm L plates, or with 9ml top agar on 150mm L plates. Following an overnight incubation at 37°C, plaques were counted and library titre was determined. Ten large plates (150mm), each containing 10^5pfu/plate, were then prepared for lifts.

Plaque lifts.

Duplicate lifts on to nitrocellulose membrane filters were taken from each plate. For the first lift, the filter was left in contact with the plaques for 30 seconds. A second lift was taken after a duration of 60 seconds. Filters were then denatured in 0.5M NaOH, 1M NaCl, for 2.5 minutes, neutralised in 0.5M Tris-HCl pH7.5, 3M NaCl for 5 minutes, washed by dipping in 2x SSC and finally air dried.

Hybridisations were carried out as described in section 2.12.

2.14 Computer Programs

The programs WORDSEARCH and LINEUP from the GCG package (1991) were used for sequence comparison and alignment.

SplitsTrees (Huson and Wetzel, 1994) and PHYLIP 3.5 (Felsenstein, 1993) were used to construct evolutionary trees.

The program PREPARE (Penny *et al.*, 1993) was used to quantify the variability patterns in the PSG data.

CHAPTER 3. RESULTS

3.1 Analysis of the Lambda Clone λC3.

A genomic library (from Dr.B.Mansfield) was plated at high density (10^5 plaques per plate) and screened with a PSG11s cDNA probe, (phBB5) (Brophy *et al.*, 1992). One hundred plaques hybridised with this probe from a total of 1 x10^6 clones screened. These were picked and transferred with a toothpick into a grid pattern, on to plates that had been freshly overlaid with top agarose containing a culture of C600 cells.

These plates were subjected to a second round of screening using the PSG11s cDNA and an oligonucleotide specific for the C-terminal region of PSG11s (C-terminal oligonucleotide: GAGACTCTGTCAGGTCTCCATGGC). All clones that hybridised with the cDNA were picked into SM buffer and stored as mixed eluates. One clone hybridised with both probes. This was plaque purified and a single positively hybridising clone, designated λC3, was isolated.

The DNA was extracted from λC3 and cut with the restriction enzymes, B*am*HI and S*al*I, which released the insert from the lamdba vector and cut the clone into several large fragments. The fragments were size fractionated by electrophoresis through an 0.7% (w/v) agarose gel, transferred to nitrocellulose membrane and hybridised with the PSG11s C-terminal oligonucleotide probe. An 8kb B*am*HI-S*al*I hybridising fragment was identified, which was subsequently isolated from a 1% (w/v) Sea Plaque preparative gel and ligated into a B*am*HI and S*al*I cut vector pGEM2. The resulting construct, designated pBS8, was anticipated to contain the C-terminal exon Cs of PSG11.

The subclone pBS8 was mapped with the restriction enzymes P*vu*II, H*ind*III and P*st*I. The digests were blotted on to nitrocellulose and hybridised with the C-terminal oligonucleotide. The results of this analysis are presented in FIG.8.

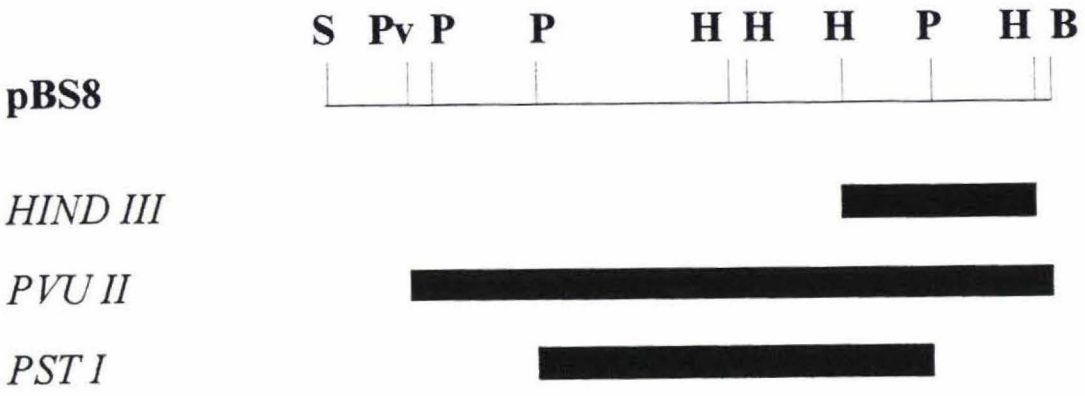


FIG. 8. Characterisation of the subclone pBS8, from λC3.
The subclone pBS8 was mapped using the enzymes H*ind*III (H), P*vu*II (Pv), P*st*I (P), B*am*HI (B), E*co*RI (E) and S*al*I (S). The relative positions of the enzyme sites are shown above. Single digests were blotted and probed with the C-terminal oligonucleotide. Hybridising fragments are shown as solid boxes beneath the map.

To determine the nature of the pBS8 sequence which was adjacent to the oligonucleotide binding site, the C-terminal oligonucleotide was used as a primer for double stranded sequencing. A short length of sequence (133 nucleotides) was determined initially. Comparison of this sequence to PSG11s cDNA sequence using the computer programme WORDSEARCH (Genetics Computer Group, 1991) failed to find any appreciable identity between the two sequences (results not shown). Also, the restriction map of λC3 did not appear to overlap with that of hC3.11 as was expected if λC3 was to contain the 3' region of the PSG11 gene. In the light of these results, further analysis of the lambda clone was not pursued.

3.2 Analysis Of The Cosmid Clones

A group from the Human Genome Center, Lawrence Livermore National Laboratory, California, are mapping the CEA/PSG region of chromosome 19 using Fluorescence *in situ* Hybridisation techniques (Brandriff *et al.*, 1992, Thompson *et al.*, 1992, Tynan *et al.*, 1992, Olsen *et al.*, 1994). Dr. Anne Olsen of this group, provided our group with five cosmids believed to span the PSG11 locus.

The cosmids came from a library made from a partial M*bo*I digest of chromosome 19 cloned into the B*am*HI site of the vector Lawrist 5. The inserts have an average size of 35kb (personal communication, Dr. A. Olsen). Lawrist 5 was designed for mapping purposes (De Jong *et al.*, 1989) and contains two *cos*-termini from lambda, which flank a 3kb fragment of DNA. During in vitro packaging, this fragment is lost (FIG.9, arrowed region). The vector also has two E*co*RI sites, between which are two S*fi*I sites, that in turn flank a single B*am*HI site (FIG.9). The S*fi*I recognition sequence is 5'...GGCCNNNN\NGGCC...3' and in this vector each site has a unique sequence. Therefore, oligonucleotides unique to each end of an S*fi*I cut recombinant clone can be used to map an insert essentially in the same way as λ *cos*-mapping (section 2.6).

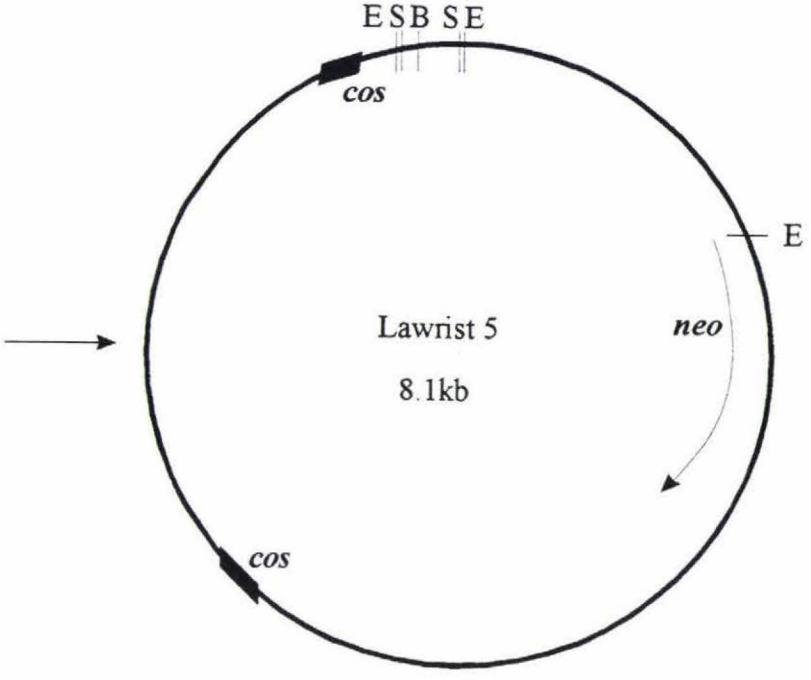


FIG. 9. Restriction map of the vector Lawrist 5.
The relative positions of restriction sites for the enzymes B*am*HI (B), E*co*RI (E), and S*fi*I (S), the ***cos*** sites and the neomycin resistance gene (***neo***) are shown. The smaller region between the two ***cos*** sites (arrowed) is eliminated during packaging of the cosmid construct.

The vector also contains a third E*co*RI site that yields vector fragments of 1.2 and 3.7kb upon digestion of recombinant clones. Used in conjunction with other enzyme digests, an E*co*RI digest can identify fragments containing vector arms.

For selection, Lawrist 5 carries the gene for neomycin resistance which is selected in bacterial hosts by the presence of 30ug/ml kanamycin. The *E.coli* host strains are DH1αmcr and ED8767.

DNA was isolated from each of the five clones provided (section 2.5). Initial restriction analysis with E*co*RI showed that three of the cosmids contained inserts of only 8-20kb. For the library construction to have been successful, the vector plus insert had to have been around 40kb otherwise the construct would not have been packaged. Therefore, it was concluded that these three clones had deleted parts of their inserts subsequent to packaging. The remaining clones (cosmids F11193 and F20478) contained inserts of around 35kb each and these were examined further.

3.2.1 Southern Analysis

Cosmids F11193 and F20478 and hC3.11were digested with the enzymes B*am*HI and E*co*RI. Digests using B*am*HI, gave three to five fragments ranging from 6kb to greater than 13kb in size (FIG 10, lanes 3, 5 and 7 from the left). Digests using E*co*RI, gave a much larger range of fragments (from less than 1kb to 10kb) for all cosmids (FIG 10, lanes 2, 4 and 6 from the left). The patterns of the E*co*RI digests showed that there was considerable overlap between the two cosmids, based on the numbers of same-sized bands they shared in common.

The PSG11s cDNA clone, phBB5, was used as a positive control for filter hybridisations. A double-digest with the enzymes S*ma*I and E*co*RI yielded a 2.9kb vector fragment, a 5' fragment of 0.9kb and a 3' fragment of 0.8kb. As only a small amount

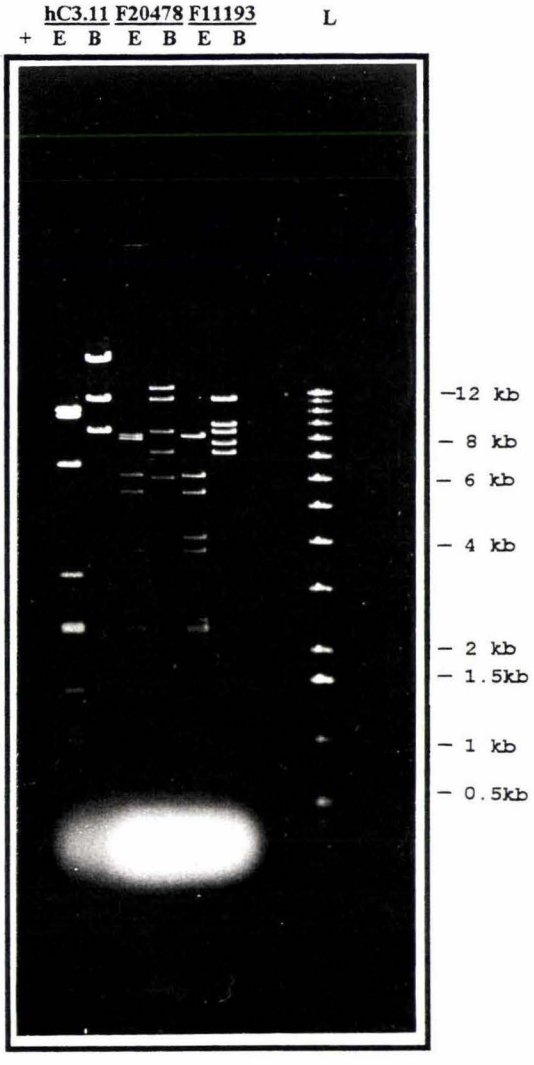


Fig. 10. Photograph of one of the replicate gels showing restriction digests of the cosmids hC3.11, F20478 and F11193.

Cosmids hC3.11, F20478 and F11193 were digested with the enzymes B*am*HI (B) and E*co*RI (E) and electrophoresed against a positive control (PSG11s cDNA, phBB5, cut with SmaI and EcoRI - +) and the BRL 1kb ladder molecular weight markers (L). Fragment sizes are given in kilobases (kb).

(3pg - 3ng of digested DNA) was loaded on to the gel, these fragments are not visible in FIG.10 (lane 1 from the left, marked +).

The digests discussed above were size fractionated through four replicate 0.7% (w/v) agarose gels at low voltage (30V/cm), for 16hrs (FIG.10) and the DNA was transferred on to a nylon membrane. Replicate blots were then hybridised with four different probes, to identify potential regions of the PSG11 gene (FIG. 11a, b, c and d). The C-terminal oligonucleotide was used as a probe for the Cs domain of PSG11 (FIG. 11a). The PSG11s cDNA, phBB5, was used to confirm the presence of the PSG11 gene in cosmids F11193 and F20478, by comparison with hC3.11 hybridisation patterns (FIG. 11b). An N domain oligonucleotide (GTACCAGATGTAGCCAGCAAG) was used to determine the presence of an N-terminal exon (FIG. 11c) and a 0.5kb BamHI-EcoRI fragment from an intron upstream of the PSG11 gene (Joe, 1994 and FIG. 6) was used to identify potential upstream regions (FIG. 11d).

The **C-terminal oligonucleotide probe** hybridised strongly with:

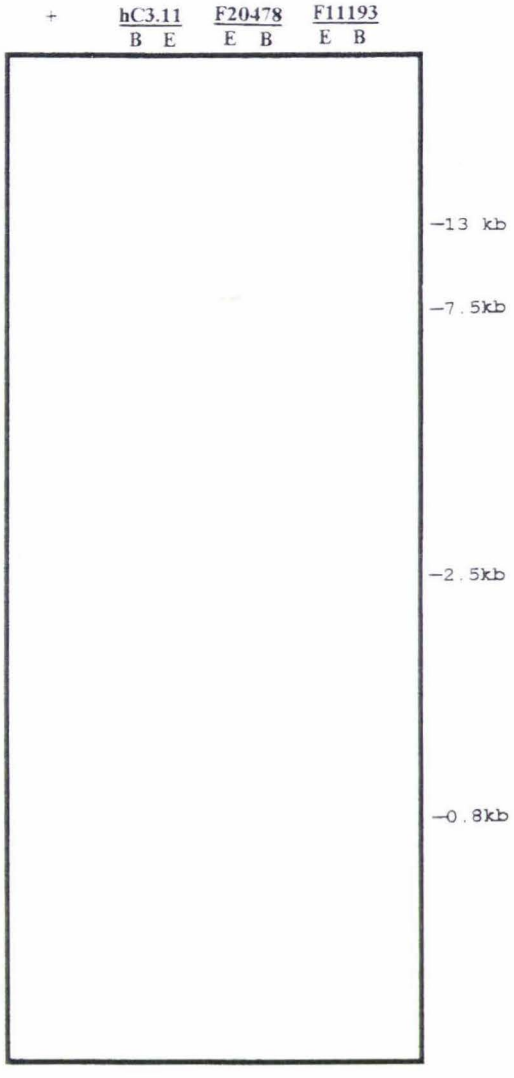
- the largest (>13kb) B*am*HI fragment and a 10.4kb E*co*RI fragment of cosmid hC3.11 (FIG11a, lanes 2 and 3 from the left)
- two B*am*HI fragments of 13kb and 7.5kb and two E*co*RI fragments of 8kb and 2.5kb of cosmid F20478 (FIG11a, lanes 4 and 5 from the left)
- a 7.5kb B*am*HI fragment and a 2.5kb E*co*RI fragment of cosmid F11193 (FIG11a, lanes 6 and 7 from the left)

Since the C-terminal oligonucleotide probe should recognise only the Cs domain of PSG11s, any cosmid containing this region would be expected to show a hybridisation pattern of one single strongly hybridising fragment in both the B*am*HI and E*co*RI digests. This pattern was found with cosmid F11193.

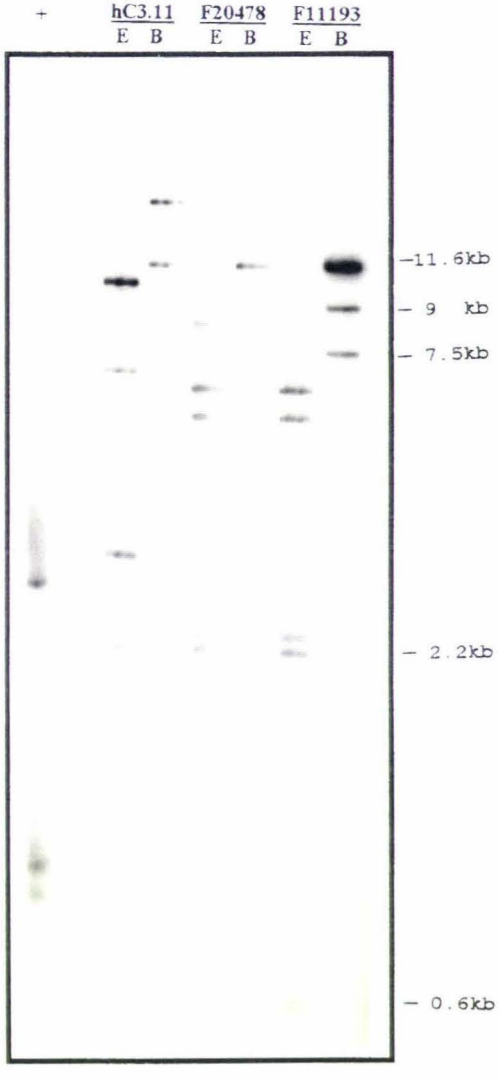
The presence of a fragment that hybridised to the C-terminal oligonucleotide probe in hC3.11, was unexpected. This cosmid had been extensively mapped and sequenced. The

FIG. 11. Autoradiographs of hybridisation results from blots of replicate gels using four different probes.

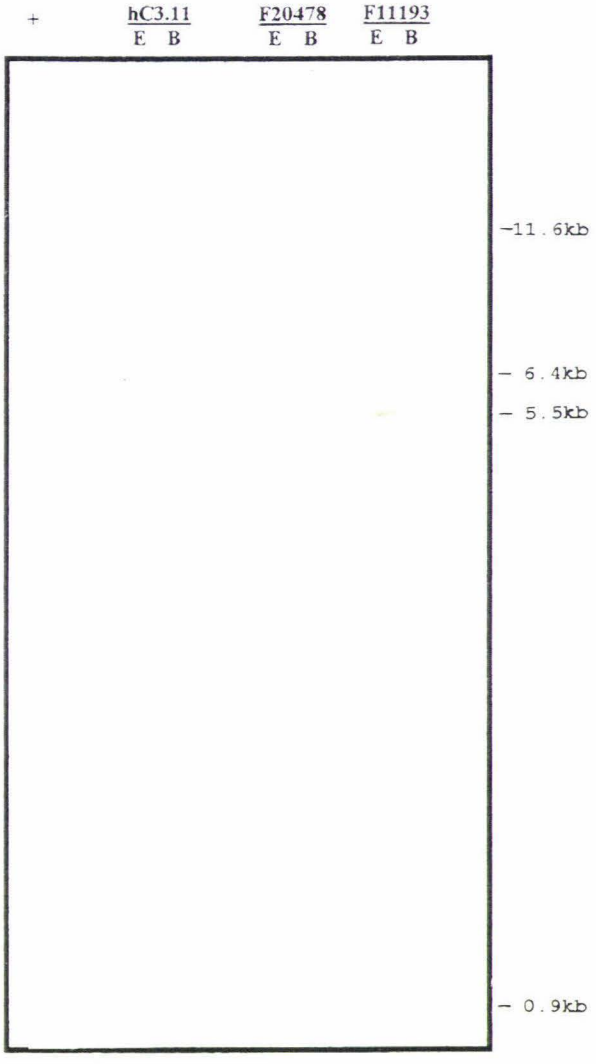
Replicates of the gel shown in FIG. 10 were blotted and probed with four different probes; the C-terminal oligonucleotide probe (a), the PSG11s cDNA probe (b), the N-terminal oligonucleotide probe (c) and the intron probe (d). Lanes are labelled as in FIG. 10. Fragment sizes are given in kilobases (kb).



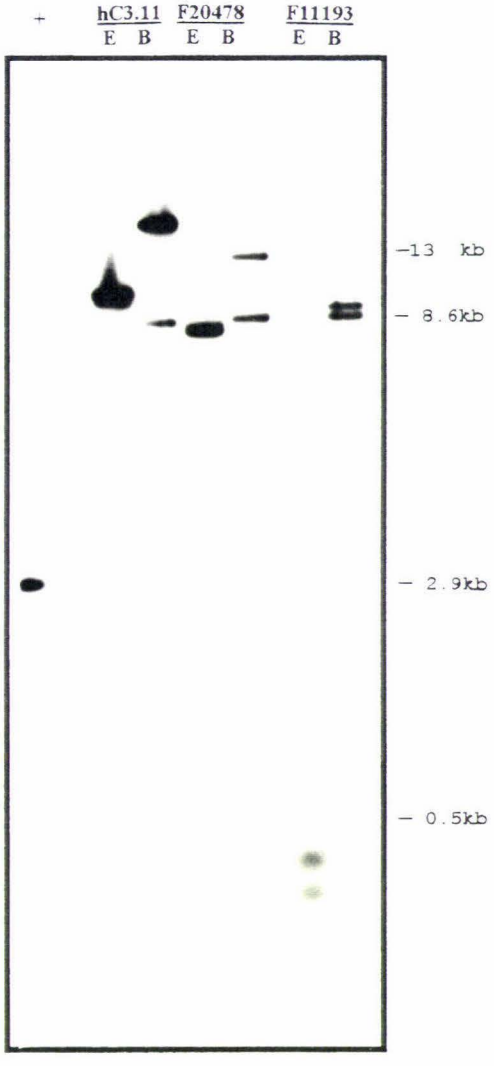
a. C-terminal oligonucleotide probe



b. PSG11s cDNA probe



c. N-terminal oligonucleotide probe



d. 5' intron probe

large hybridising B*am*HI and E*co*RI fragments were known to lie upstream of the PSG11 gene (the arrowed region in FIG.6). Two possibilities might explain why the C-terminal oligonucleotide probe was hybridising to that region. A second PSG11 locus, or a PSG with a similar C-terminal domain (such as PSG 6 or 12) might lie 5' to the PSG11 gene. Alternatively, the small oligonucleotide probe may have been hybridising to a region that has no similarity to the 3' end of PSG11, similar to that found in pBS8.

The C-terminal oligonucleotide probe hybridised with a 7.5kb B*am*HI and a 2.5kb E*co*RI fragment of cosmid F20478, the same pattern as found in cosmid F11193. However, the 13kb B*am*HI and the 8kb E*co*RI fragments hybridised equally well. Neither of the vector E*co*RI fragments of 1.8 and 3.7kb, hybridised with the oligonucleotide probe. This pattern could be consistent with hC3.11 and pBS8 patterns but needed further investigation.

The patterns of hybridisation for the cosmids with the **PSG11s cDNA probe**, phBB5, are as follows:

- the largest B*am*HI fragment (>13kb) and another of 11.6kb
  and the following E*co*RI fragments: 10.4kb, 6.4kb, 3.2kb, and 2.2kb of cosmid hC3.11 (FIG11b, 2 and 3 from the left)
- B*am*HI fragments of 13kb, 11.6kb, 7.5kb
  and the following E*co*RI fragments: 8kb, 6kb, 5.5kb, 2.5kb, and 2.2kb of cosmid F20478 (FIG11b, lanes 4 and 5 from the left)
- B*am*HI fragments of 11.6kb, 9kb, and 7.5kb
  and the following E*co*RI fragments: 8kb, 6kb, 5.5kb, 2.5kb, 2.2kb, 1kb and 0.6kb of cosmid F11193 (FIG11b, lanes 6 and 7 from the left)

All three cosmids had a strongly hybridising B*am*HI fragment of 11.6kb. In cosmid hC3.11 this fragment carries the N, A1, B1, A2, B2 domain exons (FIG.6). The N domain is present on a 6.4kb E*co*RI fragment in hC3.11 (FIG.6) that is not present in cosmids F11193 or F20478.

Cosmids F11193 and F20478 contain 7.5kb B*am*HI and 2.5kb E*co*RI fragments that are lacking in hC3.11. These fragments give strong positive signals with both the C-terminal oligonucleotide and PSG11s cDNA probes.

These preliminary results suggested there was extensive overlap between all three cosmids and that cosmids F11193 and F20478 were likely to contain at least the central domains and C terminus of the PSG11 gene.

The **N-terminal oligonucleotide probe** hybridises with:

- the 11.6kb B*am*HI fragment and the 6.4kb E*co*RI fragment of cosmid hC3.11 (FIG.11c, lanes 2 and 3 from the left)
- the 11.6kb B*am*HI fragments and the 5.5kb E*co*RI fragments of cosmids F20478 and F11193 (FIG.11c, lanes 4 - 7 from the left)

These results indicate that the N domain exon is likely to be present in the cosmids F11193 and F20478. This implies that these two cosmids are likely to contain the whole PSG11 locus. It would also appear that the fragment that contains the N domain in the cosmids F11193 and F20478 has a polymorphic restriction site.

The **0.5kb B*am*HI-E*co*RI intron probe** hybridised with:

- the largest (>13kb) and the 8.6kb B*am*HI fragments and the 10.4kb E*co*RI fragment of cosmid hC3.11 (FIG11d, lanes 2 and 3 from the left)
- the 13kb and the 8.6kb B*am*HI fragments and the two E*co*RI fragments of 7.8kb and 8kb of cosmid F20478 (FIG11d, lanes 4 and 5 from the left)
- the 9kb and the 8.6kb B*am*HI fragments and three E*co*RI fragments less than 0.5kb of cosmid F11193 (FIG11d, lanes 6 and 7 from the left).

These results, using the intron probe, were inconclusive at this stage but they were used later in conjunction with further restriction and mapping analysis (section 3.2.3).

3.2.2 Subcloning and Sequencing

On the basis of the initial hybridisation results, fragments were chosen from the cosmids for more extensive characterisation.

Four fragments were selected initially for subcloning, restriction mapping and sequencing since they hybridised with the C-terminal oligonucleotide and were therefore candidates for containing the C-domain/3'UTR of PSG11s. These were: the 2.5kb E*co*RI fragment from cosmid F11193, the 2.5kb and 8kb E*co*RI fragments from cosmid F20478 and the large fragment (>13kb) from hC3.11, which was cloned as a 9.5kb B*am*HI-E*co*RI fragment.

Characterisation of the 2.5kb EcoRI fragments from cosmids F11193 and F20478

The 2.5kb E*co*RI fragments from cosmids F11193 and F20478 were purified from sea plaque agarose following electrophoresis and ligated into E*co*RI site of the vector, pGEM-2, to create the subclones p11E2.5 and p20E2.5 (from cosmids F11193 and F20478 respectively). Short lengths of sequence (e.g. sequences ***no.4*** and ***no.1*** in FIG 12a) were obtained from each end of the insert using vector primer sites (SP6 and T7). Sequence was also obtained from these subclones using the C-terminal oligonucleotide as a primer (e.g. sequence ***no.42*** in FIG. 12a). The sequences were identical, suggesting that the two E*co*RI inserts were equivalent.

The subclone p11E2.5 was then mapped with other enzymes, to facilitate further cloning and sequencing. The enzymes P*vu*II, H*ind*II and P*st*I were found to cut the fragment internally (FIG. 12a). Digests using these enzymes were run on a gel and hybridised with the C-terminal oligonucleotide probe. Hybridisation identified a 0.9kb H*ind*III - E*co*RI fragment (FIG.12a) which was subcloned into the sequencing vectors and sequenced using standard single stranded methods. This sequence (sequence *no.1* and *no.11* in FIG 12a) overlapped with sequence from the SP6 priming site and also with sequence obtained using the C-terminal oligonucleotide as a primer (sequence *no.42*). Sequences

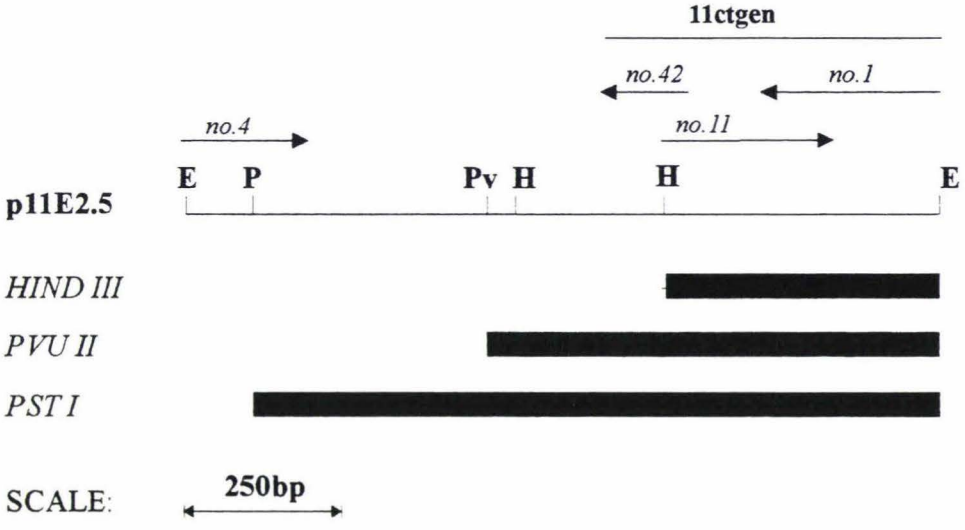


FIG.12a. Hybridisation, mapping and sequencing of p11E2.5.
A map of the insert of subclone p11E2.5 showing the relative positions of the H*ind*III (H), P*vu*II (Pv) and P*st*I (P) restriction sites. Single digests were blotted and probed with the C-terminal oligonucleotide probe. Hybridising fragments are shown as solid boxes beneath the map. Labelled arrows above the map indicate the position of preliminary sequence obtained from this subclone. Sequence *no.4* was determined using the vector primer site T7 and double stranded sequencing methods. Sequence *no.42* was obtained using the C-terminal oligonucleotide as an internal primer. The 3' H*ind*III-E*co*RI fragment was cloned into M13 sequencing vectors and sequenced using standard single stranded methods to give sequences *no.1* and *no.11*. The sequences *no. 42, no. 11* and *no. 1* were combined to give sequence **11ctgen** (FIG 12b).

FIG. 12b. The nucleotide sequence, 11ctgen, from p11E2.5 .
Restriction sites for H*ind*III and E*co*RI are labelled and shown in bold. The first 236 nucleotides have no homology to previously reported PSG sequence. At nucleoetides 237-250 there is a consensus splice acceptor sequence, shown in bold and underlined. Following this, is sequence that shares absolute homology to the PSG11s cDNA, highlighted by a shaded box. A short open reading frame that encodes the Cs domain is shown in single amino acid code beneath the nucleotide sequence. A polyadenylation sequence is present at nucleotides 589-594 (shown in bold). Homology to the PSG11s cDNA ceases at this point.

no.42, no.11 and *no.1* were combined to give the sequence **11ctgen** which is presented in FIG 12b.

The sequence **11ctgen** was analysed using the computer programme WORDSEARCH (Genetics Computer Group Inc.(1991) (FIG. 12b). The first 236 nucleotides show no homology to any reported PSG sequences. However, a splice consensus sequence (shown in bold and underlined in FIG 12b) is present at nucleotides 237-250 and from nucleotide 251, there is sequence that has absolute identity to the Cs domain (translated in single amino acid code in FIG 12b) and 3'UTR of the PSG11s cDNA. The polyadenylation consensus sequence at nucleotides 589-594 (shown in bold in FIG 12b) marks the end of the homology to the cDNA..

Next, the nature of the second strongly hybridising fragments of cosmid F20478 and hC3.11 was investigated.

Characterisation of the 8kb EcoRI fragment from cosmid F20478

The 8kb E*co*RI fragment from cosmid F20478 was cloned into pGEM-2 creating clone p20E8 and mapped with the enzymes H*ind*III, P*vu*II and P*st*I (FIG. 13a). Single digests of subclone DNA were used for Southern analysis with the C-terminal oligonucleotide probe and the hybridising fragments are shown as solid boxes beneath the maps in FIG. 13a. Initial sequence from p20E8 was obtained using the C-terminal oligonucleotide as a primer (sequence *no.40*, FIG13a).

The smallest fragment recognised by the C-terminal oligonucleotide probe was a 0.85kb H*ind*III-P*st*I fragment. This fragment was cloned into M13mp18 and M13mp19 and sequence obtained from both ends of the fragment (sequences *no.2* and *no.5* in FIG 13a). Sequence *no.2* from the HindIII site was found to overlap with sequence *no.40* from the C-terminal oligonucleotide priming site. The combined sequences (no.*2* and no.*40*) are shown in FIG 13b (sequence p20E8).

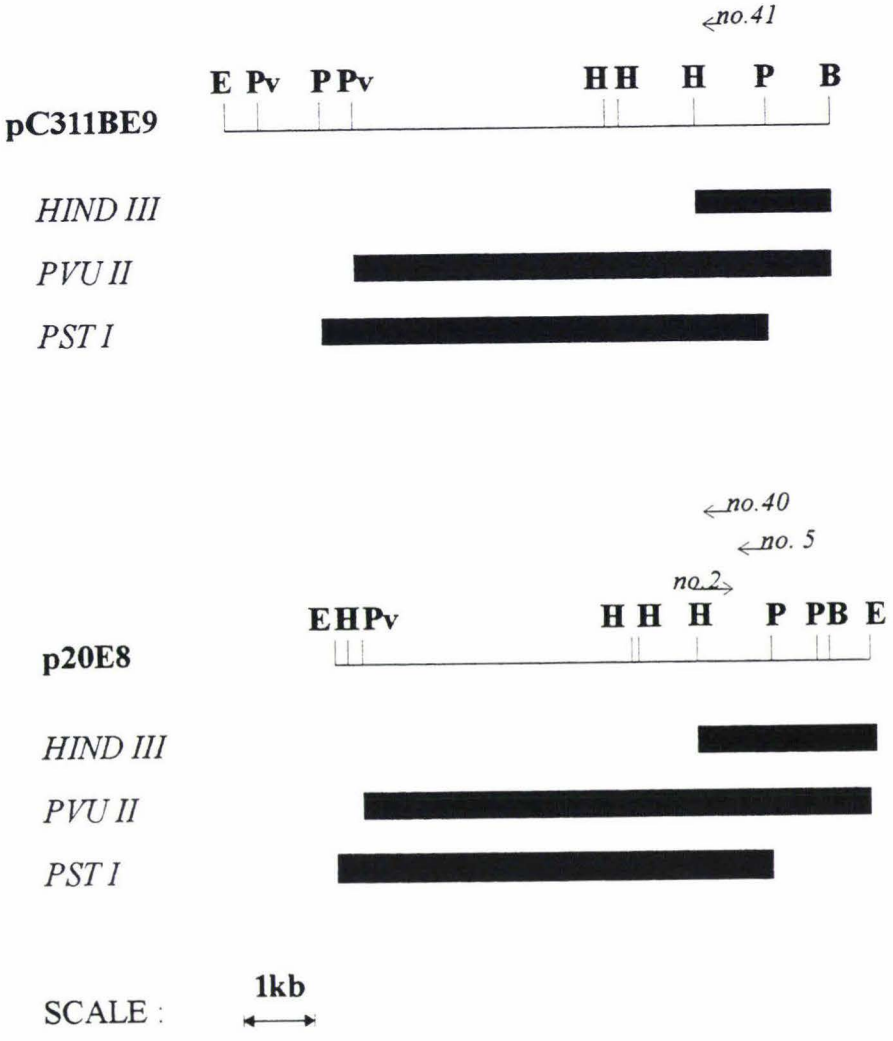


FIG. 13a. Further characterisation of p20E8 from cosmid F20478 and pC311BE9 from cosmid C3.11

Subclones pBS8 and pC311BE9 were mapped using the enzymes E*co*RI (E), P*vu*II (Pv), P*st*I (P), H*ind*III (H), and B*am*HI (B) and the relative positions of these enzyme sites are shown above. Single digests were blotted and probed with the C-terminal oligonucleotide. Hybridising fragments are shown as solid boxes beneath the maps. Labelled arrows above the maps indicate the positions of sequences determined from these clones. Sequences *no.40* and *no.41* were obtained using the C-terminal oligonucleotide as a primer in conjunction with double stranded sequencing methods. Sequences *no.2* and *no.5* were obtained by further subcloning of the 0.85kb HindIII-PstI fragment intoM13mp18 and 19 and the use of standard single stranded sequencing methods.

```
p20E8   AATGTGTTCT AGTAGGTATC AGTCTTCTAT CAATTTCATC CTGATCATCG

        51
p20E8   TCGCATCAGA TATTTTGAAG CCTTATGATG TGAGAAAGGC TGACGTCTAT

        101
p20E8   TTTCTGTGTC ATTAGAACTT TCTACCCTTT CATGGTTCGA TCTTTTTCTC

        151
        C-terminal oligonucleotide:GAGACT CTGTCAGGTC TCCATGGC
p20E8   AGTGTGTCAG TGTGGTAGGT ATGATAAGCT CTGTCAGGTC TCCATGGCAG

        201
p20E8   CTGAGTCTTG GTGGAGCCAT ATATGAGTAG GGCCGAGAAG TATAGGCTTA

        251
pC311BE9 .......... .........C CTAGGTAATT CCAGCACTCC CAGCTAATAA
  pBS8  .......... ......TTCC CTAGGTAATT CCAGCACTCC CAGCTAATAA
p20E8   TGTATGTTGA AAAAAATTCC CTAGGTAATT CCAGCACTCC CAGCTAATAA

        301
pC311BE9 TTGATTTATT AATAATCGAT TGACTTGATT GACAGAAGGG CCCAGATCAG
  pBS8  TTGATTTATT AATAATCGAT TGACTTGATT GACAGAAGGG CCCAGATCAG
p20E8   TTGATTTATT AATAATCGAT TGACTTGATT GACAGAAGGG CCCAGATCAG

        351
pC311BE9 TTGCATTCCA AATTGACA.. .......... .......... ..........
  pBS8  TTGCATTCCA AATTGACAAC CTACTTTGCA TAGAGTTCCT CACTTTCCT.
p20E8   TTGCATTCCA AATTGACAAC CTACTTTGCA TAGAGTTCCT CACTTTCCTG

        401
p20E8   AAGCTCCAGC AGGGATATGA AAGCAAGCCC AGTTCTATGA GCCTCTGTTT

        451
p20E8   CATTCCTGGG TGACTGGTGG GAGACTCCCC ATCAGTCTTG CCATGATCTC

        500
p20E8   TTGAACTGTA GCTAAATCTT CTTTCTCTTT CTCCTTCACA GGAATCAGAT

        550
p20E8   TTGATAGTGG TCCCAAAGCT T
```

FIG. 13b. Aligned nucleotide sequence from the subclones p20E8, pBS8 and pC311BE9.

Sequence from p20E8 was obtained by subcloning a H*ind*III-P*st*I fragment into M13mp18 and using standard single stranded protocols (see text). Sequence from pBS8 and pC311BE9 was obtained directly, using the C-terminal oligonucleotide as a primer. The priming site is underlined and the sequence for the C-terminal oligonucleotide is shown above. A H*ind*III site is shown in bold.

This sequence was analysed by computer using the programme WORDSEARCH (Genetics Computer Group, 1991). There was no appreciable similarity between this sequence and the PSG11s cDNA , except for 22 nucleotides that were identical to the C-terminal oligonucleotide (the underlined nucleotides 179-198, in FIG. 13b). This homology would be sufficient to allow the C-terminal oligonucleotide to bind this region of DNA, in the hybridisation and the sequencing experiments.

However, this region was found to have absolute identity to the sequence obtained from the lambda subclone pBS8 in section 3.1 (sequence pBS8 in FIG 13b).

Characterisation of the 9.5 BamHI-EcoRI fragment from cosmid hC3.11

The large 9.5kb B*am*HI-E*co*RI fragment from cosmid hC3.11 was subcloned into pGEM-2 to create clone pC311BE9. This was mapped and found to contain sites for P*vu*II, H*ind*III and P*st*I (FIG 13a). Single digests were used for Southern analysis, using C-terminal oligonucleotide as a probe and the hybridising fragments are shown in FIG. 13a, as solid boxes, beneath the map.

A short length of sequence (99bp) was determined from this subclone using the C-terminal oligonucleotide as a primer (sequence pC311BE9 in FIG 13b). This sequence was identical to that found in pBS8 and p20E8.

Other subclones

In order to investigate the cosmids F11193 and F20478 in more detail, it was necessary to create subclones with smaller inserts that could be characterised more easily. The enzyme B*am*HI cut each of the cosmids into five fragments ranging in size from 6.6kb to 13kb. Since these are manageable sizes for further restriction analysis, the B*am*HI fragments of both cosmids were cloned into pGEM-2 using a shotgun approach.

Table 3. A SUMMARY OF SUBCLONES FROM THE COSMIDS F11193, F20478 and hC3.11.

*Eco*RI fragments from cosmid F11193.

subclone	size	comments
p11E5.5	5.5kb	hybridises to the N-terminal oligonucleotide, sequence from end matches PSG11 N-terminal sequence (~700bp)
p11E4	4kb	contains a BamHI site,ends sequenced (~ 600bp) no known homology to available PSG sequence
p11E2.5	2.5kb	hybridises to the C-terminal oligonucleotide, mostly sequenced, contains Cs domain, sequence from ends matches p20E2.5 (see Fig. 12a and b)
p11E1	1kb	ends sequenced, matches 5' PSG11 sequence (~ 600bp)
p11E0.6	0.6kb	ends sequenced (~ 150bp) matches 5' PSG11 sequence.
p11E0.3	0.3kb	ends sequenced,(~ 300bp) no known homology to available PSG sequence.

*Eco*RI fragments from cosmid F20478.

subclone	size	comments
p20E8	8kb	hybridises to the C-terminal oligonucleotide, partially sequenced (see Fig.13b.)
p20E2.5	2.5kb	hybridises to the C-terminal oligonucleotide, sequence from ends matches p1120E2.5.

*Bam*HI fragments from cosmid F11193.

subclone	size	comments
p11B11.6	11.6kb	hybridises extensively to PSG11s, same restriction pattern as 11.6kb B*am*HI fragment from cosmid C3.11, assumed to contain central domains of PSG11.
p11B9	9kb	contains Lawrist vector fragments.
p11B8.6	8.6kb	contains 5.5kb E*co*RI fragment that hybridises to the N-terminal oligonucleotide, sequence from ends matches 8.6kb B*am*HI fragment from cosmid C3.11
p11B7.5	7.5kb	contains 2.5kbE*co*RI fragment and C-terminal domains of PSG11.

*Bam*HI fragments from cosmid F20478.

subclone	size	comments
p20B13	13kb	contains Lawrist vector fragments
p20B8.6	8.6kb	as for p11B8.6
p20B7.5	7.5kb	as for p11B7.5
p20B6	6kb	No E*co*RI sites, small amount of sequence (136bp)

From cosmid hC3.11

subclone	size	comments
pC311BE9	9.5kb	hybridises to the C-terminal oligonucleotide, some mapping done, small amount of sequence (see Figs13a and b.)

Four B*am*HI fragments from each clone were subcloned successfully by this method; they were the 9kb, the 8.6kb, the 11.6kb and the 7.5kb fragments from cosmid F11193 and the 13kb, the 8.6kb, the 7.5kb and the 5.5kb fragments from cosmid F20478. The 8.0kb fragment of cosmid F11193 and the 11.6kb fragment of cosmid F20478 were not successfully subcloned.

The ends of the 7.5kb B*am*HI fragments of cosmids F11193 and F20478 were sequenced by priming at the SP6 and T7 sites in the vector and found to be identical.

An EcoRI digest of the cosmid F11193 created many smaller fragments. Of these, the 6kb, 5.5kb, 1kb, 500bp, 300bp fragments were subcloned and used for restriction analysis, sequencing and as probes to confirm the mapping data.

All the subclones created from the cosmids F11193, F20478 and hC3.11 are summarised in Table 3.

3.2.3 Mapping

The cosmid hC3.11 had already been extensively mapped and sequenced by our group (section 1.6 and FIG.6).

The cosmid F11193 was mapped with B*am*HI using a modification of the S*fi*I-Linker mapping system (FIG.7). The results are summarised in Table 4 and the deduced map is given in FIG.14.

The sizes alone of the fragments on the mapping gel (Table 4) could not be determined accurately enough to identify with absolute confidence, the order of the five B*am*HI fragments (**11.6kb, 9.0kb, 8.6kb, 8.0kb and 7.5kb**). However, when used in conjunction with other results, the map shown in FIG. 14 could be established.

From the mapping data, it could be seen that a large fragment (16.8 on the left map, 12 on the right map, FIG 14) appeared to be adjacent to a 5' fragment (6.2 on the left map, 8 on the right map FIG 14) which abutted a fragment containing the vector. It was already known that the **8.6kb** fragment of F11193 hybridised with the N-domain oligonucleotide (section 3.2.1 and FIG. 11c). The hybridisation pattern of the **11.6kb** fragment with the PSG11s cDNA probe (section 3.2.1 and FIG. 11b) indicated it contained the central domains. It was concluded that the largest fragment on the maps (as above) corresponded to the **11.6kb** fragment and that the adjacent 5' fragment was actually the **8.6kb** fragment. This order was also consistent with the map for the cosmid hC3.11.

An E*co*RI digestion of the subcloned B*am*HI fragments (see section above) identified the **9.0kb** fragment of cosmid F11193 as the one that contained the vector arms. This fragment is cut twice by S*fi*I so it will not appear in the mapping gel data (Table 4 and FIG.14). It was likely then, that the order of the first three fragments was **9.0kb, 8.6kb and 11.6kb.**

Two other B*am*HI fragments remained to be ordered, namely the **7.5kb and an 8kb** fragment. The mapping gel did not distinguish between the two (8 and 8 on the left map, 10 and 10 on the right map, FIG 14), but did indicate that they both lay 3' with respect to the largest fragment. Initial sequence from one of the ends of the **7.5kb** fragment showed it contained part of the PSG11w domain sequence. The first 10 base pairs of the Cw domain are found at the 3' end of the **11.6kb** fragment and include the B*am*HI site (Chan *et al.*, 1991, McLenachan *et al.*, 1994). This suggested the **7.5kb** fragment lay immediately 3' of the **11.6kb** fragment, followed by the remaining **8kb** fragment.

The order, from the left S*fi*I site, of the B*am*HI fragments of cosmid F11193 is therefore: **9.0kb, 8.6kb, 11.6kb, 7.5kb, 8.0kb**.

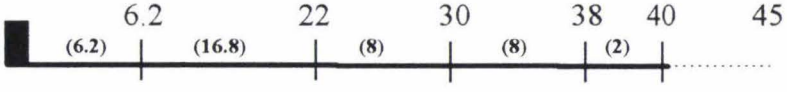
Using data from previous restriction digests, it was possible to construct a map for F20478 guided by the maps for cosmid F11193 and hC3.11. The 8.6kb, 11.6kb and

Table 4. MAPPING RESULTS FOR COSMID F11193.

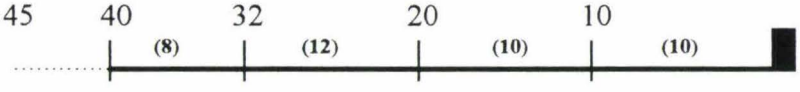
MARKERS		LEFT		RIGHT	
mobility (cm)	mwt (kb)	mobility (cm)	mwt (kb)	mobility (cm)	mwt (kb)
5.5	48.5	5.6	45	5.6	45
6.05	38.4	5.8	40	5.8	40
6.75	29.9	6.0	38	6.3	32
7.45	22.6	6.6	30	7.6	20
7.95	15.0	7.4	22	9.6	10
10.2	8.6	11.0	6.2		

Cosmid F11193 was mapped with BamHI using the Sfi-Linker mapping system. The results are presented as mobility of the mapped fragments in cm. Lanbda fragments of known molecular weight (markers) were used to calibrate the gel. The molecular weights of the cosmid fragments were determined from a graph (mobility vs mwt) drawn from the marker data. Left denotes cosmid fragments labelled with an oligonucleotide specific for the left *Sfi*I site of cosmid F11193; right denotes fragments labelled with a right *Sfi*I site -specific oligonucleotide.

Left Map.



Right Map.



Deduced map for cosmid F11193.

FIG. 14. Mapping results.
Overlapping maps constructed using data from Table 4. Mobility is marked in large numbers above the left and right maps. Deduced molecular weights of fragments from the mapping gels are indicated on the left and right maps in bracketed small bold numbers. Dotted lines indicate the position vector arms. Solid boxes indicate left and right hybridising oligonucleotides. A map, deduced in conjunction with other data (see text), is shown with the actual molecular weights of the fragments shown in bold.

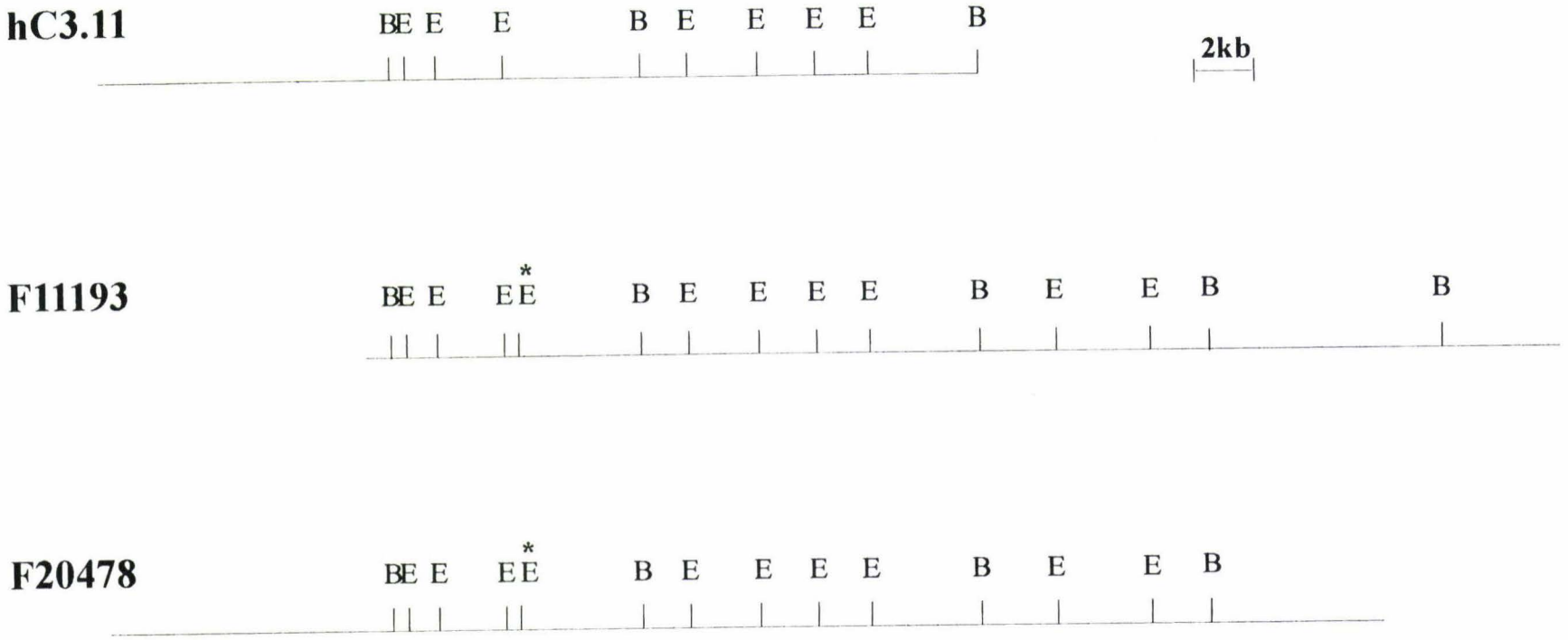


FIG. 15. Restriction maps of the cosmids F11193, F20478 and hC3.11.
Sites for the restriction enzymes B*am*HI (B) and E*co*RI (E) are labelled. A polymorphic E*co*RI site is marked with an asterix (*). Maps are aligned to indicate overlapping regions.

7.5kb B*am*HI fragments could be ordered with confidence, by comparison with the other maps. The 13kb fragment of cosmid F20478 was digested with E*co*RI and contained the vector fragments. The pattern of hybridisation of this fragment with both the C-terminal oligonucleotide probe and the 0.5kb B*am*HI-E*co*RI places it 5' with respect to the 8.6kb fragment. The position of the remaining 5.5kb fragment at the 3' end of the cosmid was confirmed also by hybridisation and mapping with E*co*RI.

Cosmid F20478 contains two E*co*RI fragments of 8kb and 7.8kb. The 8kb fragment hybridised with the C-terminal oligonucleotide probe and has a restriction map similar to that of the large 5' fragment of hC3.11 (FIG. 13a). None of the three central B*am*HI fragments contain a 7.8kb E*co*RI fragment. This fragment must therefore lie at the 3' end of the cosmid. The 5.5kb B*am*HI fragment does not contain any E*co*RI restriction sites and must lie within the 7.8kb E*co*RI fragment.

The order, from the left S*fi*I site, of the B*am*HI fragments of cosmid F20478 therefore is: **13kb, 8.6kb, 11.6kb, 7.5kb and 5.5kb**.

The restriction maps for the enzymes B*am*HI and E*co*RI, of the three cosmids F11193, F20478 and hC3.11, are shown in FIG.15.

3.2.4 Further sequencing

The hybridisation and the mapping results showed that the 2.5kb E*co*RI fragment containing the C-domain of PSG11s lay within the 7.5kb B*am*HI fragment. In order to characterise as much of the 3' end of the locus as possible, further sequences were obtained.

The HindIII and PstI digests of the 7.5kb BamHI fragment of cosmid F11193 were cloned into M13mp18 using a shotgun approach and sequenced. This method yielded a total of ten sequences (sequences ***2-19Hind*** and ***8-12Pst*** in FIG 16)

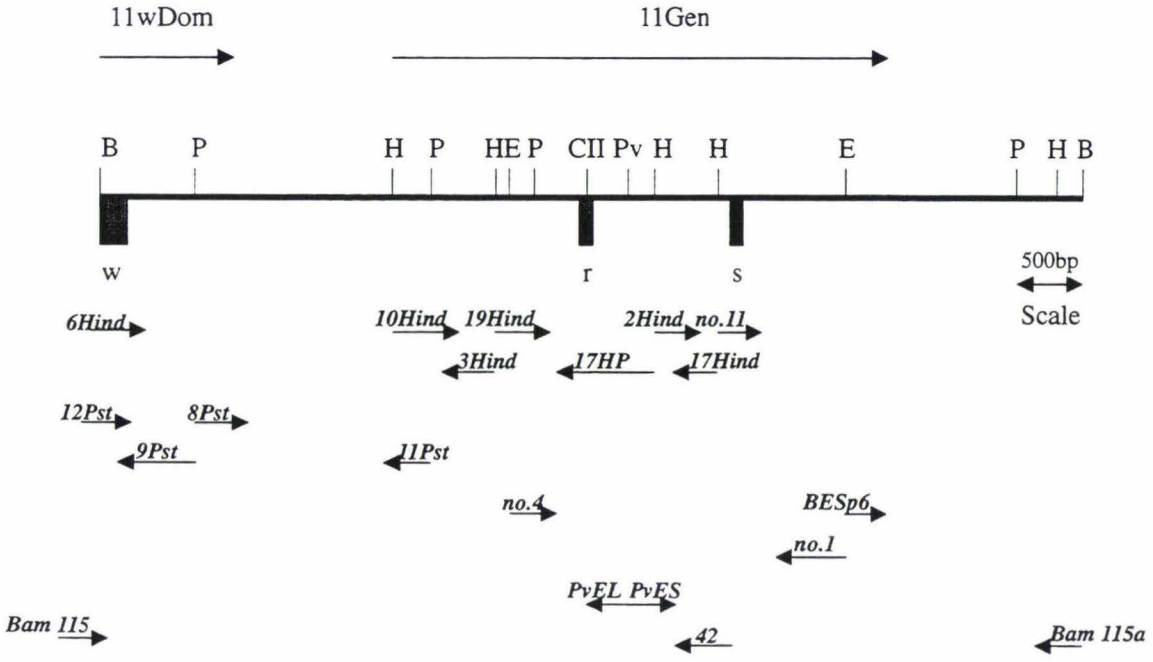


FIG. 16. Sequencing strategy for p11B7.5.

A map of the insert of p11B7.5 showing restriction sites for the enzymes B*am*HI (B), P*st*I (P), H*ind*III (H), E*co*RI (E), H*inc*II (CII), and P*vu*II (Pv). The positions of the predicted C-terminal domains Cw, Cr and Cs are indicated as shaded boxes beneath the map. Short lengths of sequence obtained from various subclones (see text) are labelled and marked by arrows beneath the map. Regions of continuous sequence that could be combined are labelled and marked by arrows above the map and are presented in FIG. 17a and 18.

The two P*vu*II-E*co*RI fragments were subcloned into M13mp18 and sequenced (sequences ***PvEL*** and ***PvES*** in FIG 16).

A 900bp H*ind*III-P*st*I fragment was gel purified, subcloned into M13mp18 and M13mp19 and sequenced (sequence ***17HP*** in FIG.16).

These sequences were combined with other sequences already determined (see FIG 16), to give two regions of continuous sequence. The 0.9kb from the B*am*HI site that lies adjacent to the 11.6kb B*am*HI fragment was named **11wDom,** while the 3.5 kb of sequence, about 0.9kb further downstream, was named **11Gen** (FIG 16).

Analysis of genomic PSG11 sequence 11wDom

The sequence 11wDom was analysed by computer using the programme WORDSEARCH (Genetics Computer Group Inc., 1991) and is presented in FIG.17a. The sequence has absolute identity to the Cw domain and 3'UTR of a previously reported clone, hPS91 (Chan *et al.*, 1991). An open reading frame, which actually begins 5' of the BamHI cloning site, is present and the translation of the predicted Cw domain is shown underneath the sequence (FIG 17a). A polyadenylation consensus sequence is present at nucleotides 967-972 (FIG 17a).

In addition, however, the sequence 11wDom also shares identity to reported carboxyl termini and genomic sequence of other PSG, namely PSG1,(Lei *et al.*, 1992) PSG4 (Thompson *et al.*, 1990), PSG5 (Khan and Hammarstrom, 1989), PSG8 (Oikawa *et al.*, 1988), and PSG12ψ (Lei *et al.*, 1993).

PSG 1,4, and 8 match PSG11 sequence for 148bp (nucleotides 7-153), as shown in FIG.17b. A Cd domain is encoded within this region and is alternatively spliced into transcripts of these subgroup 1 genes. In our sequence, the presence of a T at nucleotide 38 , disrupts the TAG termination codon present in other genes and allows readthrough to the next stop codon, some 205bp away. This creates the Cw domain which is

```
 1  GGATCCCAGCATCCTTGGCAGTAGGGTTTTTTGTGGAGTCTATCTGGCTTTCAGAGAAGA
  .... I  P  A  S  L  A  V  G  F  F  V  E  S  I  W  L  S  E  K  S

 61  GTCAGGAAAACATTTTTATTCCCAGCCTGTGTCCCATGGGCACAAGCAAATCCCAAATTC
       Q  E  N  I  F  I  P  S  L  C  P  M  G  T  S  K  S  Q  I  L

121  TCCTCCTGAACCCTCCCAATTTGTCTCTACAGACTCTCTTCTCCTTGTTTTTCTGTTTTC
       L  L  N  P  P  N  L  S  L  Q  T  L  F  S  L  F  F  C  F  L

181  TTATGGCTGACCTTGTGTCTGGCCTGAAAAAGGTAGGGAGGGGGCTTTATCAGCCCTGAG
       M  A  D  L  V  S  G  L  K  K  V  G  R  G  L  Y  Q  P  *

241  CCCTATGTGGTAGAAGAGGCTTCAGAGAGGGACAAGAAGGAGAGTCCAAGATCAAGTTGC

301  TTCTCGCTGTCACCGACACATCCCCTTCTGCCACGTCTTTTTTTTCTTGTACCTATTCCA

361  TGAGCTACAAGGGACAGCTGAGGCTTTGAAACAACGCTCACACTTTTCCCCCAGATGAGA

421  GGAGGAAGCCCCTTGGGTGAGGGAGGAGCAACGTCAGACTCTCGTTCCTCGTCTGCTCCG

481  GGCTCCTCTGGTGACTGCCCTGCCTCACTCAACCTGGGGTTGGGACTAGCGTGTGTGTAG

541  AAAGGCGCTGGTGGCCTGTCCTGAATTTTGGCAAATTCAGCTGCCAGTTAAGCCAAGCCT

601  CCCCCGGGCCAGGCTGCAGGGAAATAAGAAGAGAGGGAGCCTCAGGGCAGACTCCTGAGC

661  TGCGTCCTGGCTCTGAAGTCACCAGCTGTATGAGGCTTTGGGCACAGCACGTGGGACACA

721  GCACGGAGGACAGTGACTGATGCAGACTGGAGGAATAGGGAGATTCTCCGTTGGGGTTAT

781  TCATGGCAGGAAAGGGGCAGTGCAAAAAGTGTGTAATTATAGAGAGGGTAAGACTACCAG

841  ACACTTTATATATCTAATATAAGACTTACCATTAACTATTTCTATGTGTGCAATTTAGTG

901  TTGTGTAACCATCACACTATCCATTTCCAGAACTTTTTCCTCTTACCATATTAAACCTCT

961  GTACCCAATAAACAGTAACTCTCACTCCTTCTTCCCC  997
```

FIG. 17a. Nucleotide sequence 11wDom from p11B7.5.
The 5' B*am*HI site is shown in bold. The translation of the predicted Cw domain is shown underneath the sequence in single letter amino acid code. The polyadenylation sequence is shown in bold and underlined, at nucleotides 967 - 972.

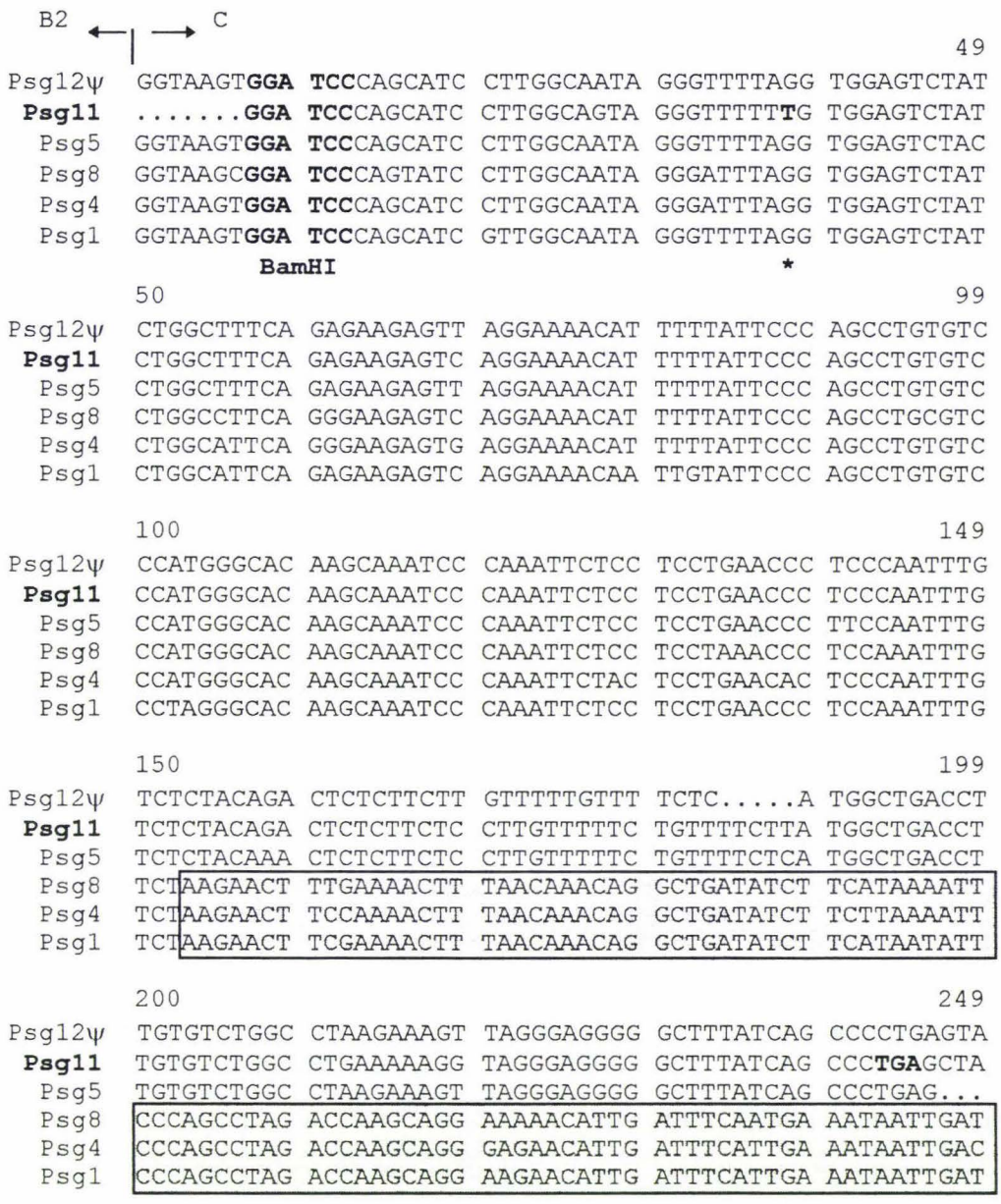


FIG. 17b. Aligned nucleotide sequence of six PSG, from the end of the B2 domains.

The sequences of six PSG are aligned to show similarity in the region adjacent to the end of the B2 domain. A B*am*HI site is marked in bold and labelled. The point mutation that creates the 81 amino acid Cw domain in PSG11 is shown in bold and marked (*). For PSG 1, 4 and 8, this also marks the stop codon for the Cd domain. Gaps (shown as dots) have been introduced to maximise alignment. The stop codon defining the end of the Cw domain of PSG11 is shown in bold (nucleotides 243-245). Dissimilarity between subgroup 1 sequences and those of subgroups 2 and 3 is indicated by a shaded box.

reportedly unique to PSG11. In all other PSG, the stop codon defines the end of the Cd domain. The identity between subgroup 1 sequences (PSG 1, 4, 8) and PSG11 sequence ceases abruptly at nucleotide 153 (FIG. 17b).

Identity between PSG5, a subgroup 2 gene and PSG11 continues throughout the available sequence for PSG5, which is only another 92bp. PSG5 appears to encode a potentially functional Cd domain but as yet, transcripts from subgroup 2 genes using Cd domains have not been detected (Chou and Plouzek, 1992).

PSG12ψ, a subgroup 3 gene and PSG11 share extensive similarity over all reported sequence. (about 3.5kb, Lei et al., 1993.)

Analysis of the 11gen sequence

The sequence was analysed by computer using the programme WORDSEARCH (Genetics Computer Group Inc., 1991) and is presented in FIG.18.

Two C-termini are encoded in this stretch of continuous sequence. The first shares identity with the C terminus of PSG6r and PSG12r (Zimmerman *et al.*, 1989 Lei *et al.*, 1993) and is the Cr domain and 3'UTR of PSG11. The other is the Cs domain and 3'UTR of PSG11. The relative positions of these domains can be seen in FIG. 16.

The Cr domain sequence lies about 3.5 kb downstream from the end of the B2 domain. An open reading frame of 22 amino acids is present and differs in just 1 nucleotide from PSG6r (nucleotide 1518, FIG. 18), resulting in an amino acid difference (S -PSG11 to F -PSG6). There is extensive similarity throughout the 3'UTR regions of PSG6r and our PSG11r, except over a 40bp region (nucleotides 1722-1762, FIG 18). A polyadenylation sequence ATTAAA is found at nucleotides 1815-1820 and is identical to the signal found in PSG6.

```
   1  GCCTTTACTTCTGATAATGTCTCGACATATAAAAGCTTGATTTTGTGAAGTCCGATTATC

  61  CATTTATTTGTTGCCTATGCTTTTGTTGTTACAACCAAGAAAATCATTGGGAAATCCAGT

 121  ATCATGAAGCTTTTCTTCTAAGAGTTGTGTAGTTTTTCTCTTACATTTAGATCTTTGATC

 181  TATTGTGGGTTAATTTTTGTACATGGTGTTAGGTAAAGGTTCCAATCTTCTTGCCCTTGG

 241  ATATCCAGCTTTCCCAATATCCTTTGGTGAGAACACTGTCCCTTCCCCATTGTAAATGCT

 301  GAAATCAGGAAGTGTGAGTCCTCCAGCTTCATTCTTCCTCTTCAAAGCTGTGTGTCTATT

 361  TAGAGTCATGAGATACAATATAAATTTTAGGACAGATTTTTCTTTTTCTGCAAAAATGCC

 421  ACTGAGAATCTGATAGGAATTGTATTGAATCTGCAGTCACTTTGGGTAGCACTGTTCTCC

 481  TAACAATATTGAGTCTTCCAATCTGAAAACAAATGTCTTTCAATTTATTGATGTCATCTT

 541  TCATTTCTTTCAGCAATATTTTGTAGATTTCAGGTATAATCATTTCACCTCTTTGATTAA

 601  ACTTATTCCTAAATATTTTATTCTTTTTGATGTTAATATAAATTGAAATTTTTTTCTTAA

 661  TTTCCCTTCAGATTGTTCATGGTTAGTGTATTGAAATACAACAGATGTTTGAATGTTGAT

 721  TTTGTAGTTGTCAACATTACTGAATTTATTTATTAATTCTAATTAGGTTGTCAATCTTTA

 781  GGATTTTCTACATATAAGTTCAAGTTATCTATAAACAGAAATAATTTTACTCCTTCCTTC

 841  CAATTTGAATGTCTTTTTTTCAAATTCTTGCCTAATTTTTCTGACTAGACCTTTCAATAC

 901  TACGTTGAATAAAAGTGTCAAAAGGCAGGCATCCTTGTCTTGTTACTGCTCATACAGGGA

 961  AAGCTTTTCAGTCTTTCTCCATTGAGTATGATGCTAGCATTGGGTTTTTCACATATTGCC

1021  TTTATGTTGAGGTGGTTTCCTTCCATTCTTAGGATGTTTTCATTATGAAAAAATATTGAA

1081  TTCATCAAATGCTTTTATTGATTCAATCTTATTACTGATTATAGTTATACTCATATTTTT

1141  GTGTGGTTCTAGGAGTTGGTCCATTTCATCTAGGTTATCCAATTTATTGGCATACAATTA

1201  TTTATAGTACTTTCATCATCATTATTTTATTAGAATTGGTAGTAATCGCTTCATTTTTCT

                |→
1261  TTCTTTCTTTCTTTTTTTTTTTTTTTTGGAAGAGAGAGAGACAGGCTCTCACTCTGTAGG

1321  CCAGCCCAGGATGGAATACAGTGGTGTGATTACGGCTCACTGCAGCCTTAACCTCCTGGG

1381  CTCAAGCAATTCTCCTTCTTCAGCCTCCCAAGATGCTAGGACTACAGGTGCATGTCACCA
1441  TGCCCAGCTAATTTTTTTTTATTTTTTTGTAGAGACAGCATCTCCCCAGGTTACCTATGC
PSG6                                ..............................
PSG11r                             E  T  A  S  P  Q  V  T  Y  A
PSG6r                              .  .  .  .  .  .  .  .  .  .

1501  TGGTCCAAACACCTGGTCTCAAGAAATCCTTCTGCTGTGACCTCCCAAAGTGCTAGGATT
PSG6  .................T.....................................
PSG11r G  P  N  T  W  S  Q  E  I  L  L  L  *
PSG6r  .  .  .  .  .  F  .  .  .  .  .  .  *

                        ←|
1561  AAAACATGACCCACCATGCTCAGAGTCCATTTTCATTTCTGATTTGAGTAATTTTAAACT
PSG6  ....................................A.................T-..

1621  TTTCTCTTTTTTTCTTAGTCAACTCAGTTAATGGTTGTCAATTTTGTTGATTT-ATTT-GAA
PSG6  .................C...TCT............................T....T...

1681  GAATCAACTTT-GGTT-CA-TAATTTCCTCTATTCTGTTTCCATCCATTTGAAGATCACTTTG
PSG6  ...........T....T..G.......................TCTCCATTTT..TTA.A.C
```


1741 GTTCATTATTCTCTATCTGTTCATCTCATTCACTGTGCTTGGGTTTAGTTTGTTCTTCTT
PSG6 CAC.C.A..A..TAT.A.T.C.--.....
1801 TCATATCCTGAAGTATTAAAGTAGGTTGTTGACCTGAGATCTTTCTTCTTTTTTAATGTA
PSG6 ..G.....
1861 AGAGTTTACAGTTATAAATTTCTTGCACAAAGACTCAACTTCTCTGAACCTCTGATTCCT
1921 TACCTGAAAATTGCAATGAGAGTACTTTCTTCTTACCATTGTTTTAAAGGTTTAATGCAG
1981 TCAATGAAACAATATGCCACACACAGCGGAACCAATGTCAGCTGCTATATTACTACCATC
2041 ATCATTAGCCTTGAGGTCAAATAGTCCTAGAATCAAATCTCAGATCCACCTGTCACTAGC
2101 CATATGACACCAGGAAAGTTTTTACACCATGCTAAGCTTTCTGTCTTTTCATCGGCAAAA
2161 TGGAAATAATGTCTATGTGACAGGGTTATTGTGTGGCTTAAATGAGATACAGGTAAAGTA
2221 TTGAGCACAGGGCCTGGCACATAGGAAGTGCACCTCAACAGTACCTACCTTTTTCCATAT
2281 ATATGGAAAAAGAGGTAACACATAAAACACTAGGACATGGTTACTGACTACTTGTGGGAG
2341 AGAAAAAAAAAAGCTAAGGGCAAAGAATCAACTGTGGTATGTTAGTTTTTACCAACTGAGA
2401 TGCATCCAAGATGGGATTAGACATACAAGGTAATTTATCAGGGAAGACACCTGTGAGGGA
2461 ATGTGGAGCAGGCATGAAGGTAGTATGGGAGAACCCACAGAACACTATGCAGAGCTGATT
2521 CCTGTGAAAAAGAAAGAGAAAGAAGTTTTAGGTACCAATGCAGTTCTAAGAGTATTTGCA
2581 AGGCTGATGGGGAATCCTCCAACCAGCCACCCATTAGAGTTAAATAGAGCCTCAGAGAAC
2641 TAGGCTTGCTTTCACACCCTTCCTGGGAGCCTGTAGGAAAGAAGCTTTCTGTGTAAAGGA
2701 GGTAGTGAATTTGAAATGCACTGACCTGGGCCTTCTGTCAATCAGGTCCCTGCCATGGAG
PSG11s G P C H G D
→ C REGION
2761 ACCTGACAGAGTCTCAGTCATGACTGCAACAACTGAGACACTGAGAA------AAAGAACAGGCTG
PSG11s L T E S Q S *
PSG8 /A...-...TT...AACTTT..CA.....
2821 ATACCTTCATGAAATTC-------AAGACAAAGAAGAAAAAAACTCAATGTTATTGGACTAA---ATAAT
PSG8 ...T.....A.....CCAGCCT....C....C..G.....CA.TG..T.C.A..A.A....TTG....
2881 CAAAAGGATAATGTTTTCATAATTTTTTATTGGAAAATGTGCTGATTCTTTGAATGTTTT
PSG8 A.TG.....T..G.-....C...T.....T.....A....G...
2941 ATTCTCCAGATTTATGAACTTTTTTTC--TTCAGCAATTGGTAAAGTAT---ACTTTTGTAAACA
PSG8 G..T..T.C...G.C.G.A.....C..TT..A.C.T..CT...GCT...AGC.G..CAA.....T
←
3001 AAA-------------------ATTGAAATATTTGCTTTTAGCTGTCTATCTGAATGCCCCAGAATTGTGAAACTATTC
PSG8 .T.CCGCAGTTTATTGAACTGTA.....A.....-...T..C....C.....G.C.....
3061 ATGAGTATTCATAGGTTTATGGTAATAAAGTTATTTGCACATGTTCCGTAAGAATCTGCT
PSG8 .....A...G...T.....C.CA.....A..A.A.C..C.....
3121 CTC-TTATAACA-GACACATTTGAAA-CATTGGTTATATTACCAAGGCTTTGACTGG-ATGTTT
PSG8 ...T..G.....G.....C...T.....G...-..
3181 ATTATTTTGAGAATATACATAGA--TAACCA-TAGGAA-TGCAGGCAAAGTCTGAAGTGGGCCT
PSG8 ..--A...A..G.....G.....ATG.....G..T...C.....CA....
3241 TGGTTTGGCTTCCTAGTCTCAAGAGGTTTTTGGAAGTTTCATCTGAGATTCTTATTAAAA
PSG8 .....T.....T...G.AA.....A....C.....C.TA.....
3301 ACTTCTAGCAAAGAGAAGTTTAAAA-AGAGCCTCTATGGTCCATTGCTACTCTTGCCGCAC
PSG8 ....AG..A.....A..T.....G.....AC.C.....T....

FIG 18. Nucleotide sequence 11Gen, from the subclone p11B7.5.

Splice acceptor sequences are shown in bold and underlined. Potential C-terminal domains are translated underneath the nucleotide sequence in single letter amino acid code, with stop codons marked (*). Polyadenylation sequences are marked in bold. Alignments to PSG6 and PSG8 are shown beneath the nucleotide and amino acid sequence for PSG11. Matching nucleotides are indicated with a dot (.). Spaces introduced to maximise alignment are indicated with a dash (-). The region of sequence that has homology to the Alu repeat sequence (nucleotides 1267-1586) is marked (|→.....←|). The region used for tree reconstruction is shaded and labelled (→ ***C REGION***).

The identity of PSG11 to PSG12ψ is similar to that of PSG6 over this region (Lei *et al.*, 1993, alignment not shown). Genomic sequence available for PSG12ψ ends at nucleotide 1836 in the PSG11 sequence (FIG.18).

Another feature of note is that the Cr domain falls within a 300 nucleotide sequence region that has 73% identity to the Alu family repeat sequence (Deninger *et al.*, 1981, see FIG 18, nucleotides 1267-1586).

The Cs domain lies 4.6kb from the end of the B2 domain.

At nucleotide 2796, just 12 bp past the end of the Cs domain, the sequence resumes identity with the subgroup 1 gene sequence. This is the point where identity was abruptly discontinued in 11wDom sequence (nucleotide 153, FIG. 17a). The similarity between PSG11 and the subgroup 1 genes now continues for over 1kb, to the end of the 11gen sequence. Within this sequence in subgroup 1 genes are the other alternative C domains, namely Cc, Ca and Cb.

The Cc domain present in subgroup 1 sequences is disrupted in the PSG11 sequence (nucleotides 2980-3004 in FIG. 18). In the PSG11 sequence, the splice acceptor site (AG) is mutated (AA). There are two deletions of 3 and 20bp and the sequence between these deletions does not match the subgroup 1 Cc domain sequence very well. Sequence following the 20bp deletion however, beginning at position 3004 in FIG. 18, resumes close homology to the subgroup 1 sequences. At nucleotides 3553 - 3570 there is a sequence that could encode a Ca domain. There is a good splice acceptor sequence and the predicted amino acid sequence (DWTLP) differs by one residue from that of PSG8 (DWTGP) and PSG1 (DWTVP). However, no PSG11a transcripts have been reported to date.

Downstream of the Ca domain in subgroup 1 genes lies an exon encoding the Cb domain (EAL in PSG8 and PSG1, ESL in PSG7). In the PSG11 sequence, a transversion disrupts the analogous splice acceptor site (AG →TG, nucleotide 3638, FIG. 18). A

potential splice site is present 3bp further down which would predict a two residue domain AI. To date, no such transcripts have been reported.

3.3 Evolutionary Analysis.

Optimal trees from the data.

In order to investigate historical relationships between the PSG genes, trees were constructed from the N, A1, B2 and C regions of seven PSG that shared all these regions in common (PSG1, 2, 3, 5, 6, 8, 11, Table 5). The sequence alignment used in this analysis for the C region, is given in FIG.19. Only blocks of sequence data bounded by constant columns were used for analysis. Sequence bounded by shaded boxes is excluded. For example, the Cc region, a potential coding region in subgroup 1 genes, is mostly excluded. This ensures only regions of unambiguous alignment were used.

Two methods of tree construction were used. PHYLIP 3.5 (Felsenstein, 1993) was used to construct the Neighbour-joining trees on Hamming distances (observed number of differences/total sequence length). Split decomposition on Hamming distances was carried out using SplitsTree V1.0 (Huson and Wetzel, 1994). The trees constructed for each domain are presented in FIG.20.

None of the trees are exactly the same and the C tree and the N tree are markedly different. Some of the groupings found in the C tree are seen in trees from the other domains. For example, PSG1 and PSG8 (subgroup 1 genes) group together in the C and the B2 trees. Similarly, PSG6 and PSG11 (subgroup 3 genes) come together in the C, the A1 and the B2 trees. The PSG2 and PSG5 (subgroup 2 genes) come together in the A1 and C trees.

The SplitsTree graphs in FIG. 20 show both the support in the data for each edge in the tree as well as the contradictions against those edges. In the absence of conflicting patterns in the data, the tree appears as a normal two dimensional binary plot. If the data

Table 5. GENBANK ACCESSION NUMBERS FOR PSG NUCLEOTIDE SEQUENCES USED FOR TREE BUILDING.

	Domain					
	N	A1	A1-B1 Intron	B1 ψ-Exon	B2	C
Gene	**Accession number**					
PSG1	M17908	M93701	M93702	M93703	M17908	M93705
PSG2	M20822	M20822	n.a.	n.a.	M20822	M20822
PSG3	M34420	M34420	n.a.	n.a.	M34420	M34420
PSG5	M32630	M32631	n.a.	M32632	M32634	M25384
PSG6	M31125	M31125	n.a.	n.a.	M31125	M31125
PSG8	M74106	M22311	M22311	M22311	M22311	M22312
PSG11	M69025	U04323	U04323	U04323	U04324	U04325

FIG. 19. Aligned nucleotide sequence from the C-terminal region of seven PSG genes.

Sequences obtained from the Genbank database (Table 5) and this study (FIG. 18) were aligned using the computer programme LINEUP (Genetics Computer Group, 1991). Spaces introduced to maximise alignment are indicated by dots (...) . Regions of sequence not used for tree building are indicated with shaded boxes. Examples of variable sites are highlighted with coloured boxes and labelled **P** for ***parsimony*** sites, **S** for ***singleton*** sites and **O** for ***other variable*** sites.

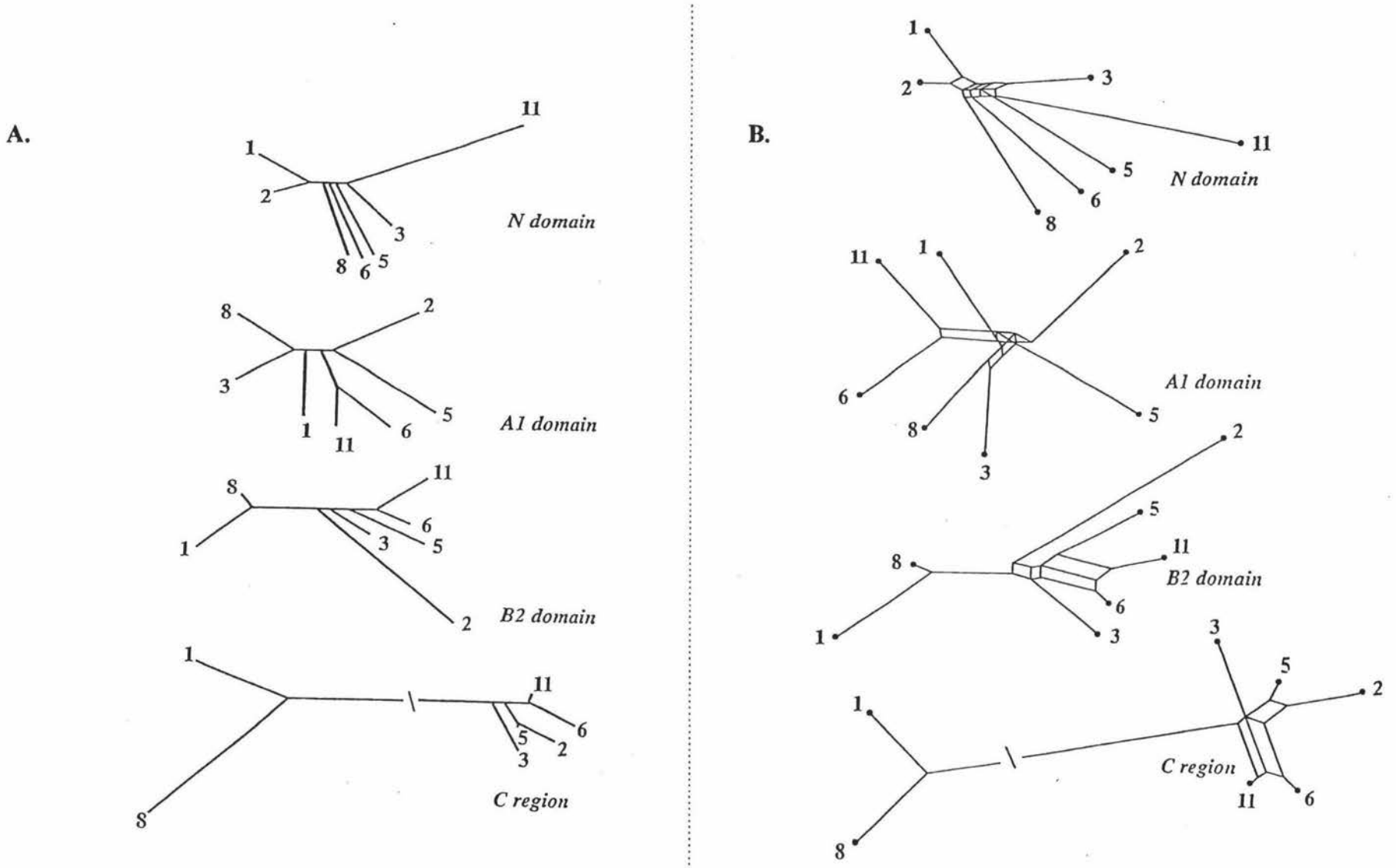


FIG. 20. Trees constructed from nucleotide sequence from the N, AI, B2 domains and the C-region of seven PSG using Neighbour-joining on observed distances (A) and split decomposition on observed distances (B).

contains many conflicting patterns, the edges of the tree become more box-like. The relative lengths of the edges of the box indicate the strength of the conflicting patterns. A square box then, for example, would indicate strongly contradictory patterns in the data.

The box-like starburst shape of the N domain SplitsTree graph is an indication that the data for the N domain contains many contradictory signals. These conflicts could not be resolved by transformation of the data using simple parameter correction formulae such as Jukes-Cantor (Jukes and Cantor, 1969), Kimura 2ST (Kimura, 1980), LogDet (Lockhart *et al.*, 1994), (results not shown).Conflicting patterns also occur with the data from the A1 domain. However, in the SplitsTree graphs for the B2 and C regions, the split separating the subgroup 1 genes and the subgroup 2 and 3 genes becomes progressively more resolved. A trend can be observed, in the support for this split as the gene is traversed from the N domain to the C region (FIG. 20, see also McLenachan *et al.*, 1995).

That the evolution of the N domain in particular cannot be described by a simple model is an interesting observation in itself. It suggests processes such as positive selection and/or gene conversion and/or unequal crossover between loci may have been involved in the evolution of the PSG genes. These possibilities were investigated by looking more carefully at the patterns of change in the different PSG domains.

Variability patterns.

The amount of variability in sequence data was investigated by characterising the differences between sequences using the computer programme PREPARE (Penny *et al.*, 1993). The patterns of difference for each aligned nucleotide position can be described in specific terms. *Parsimony sites* are columns where there are at least two occurrences of at least two different nucleotides (e.g. column **p** in FIG.19) *Two character state singleton sites* are patterns which show a single occurrence of a different nucleotide (e.g. column **s** in FIG.19). *Other variable sites* (e.g. column **o** in FIG.19) include all other different patterns of variability.

Table 6. SITES OF VARIABILITY IN DIFFERENT REGIONS OF SEVEN PSG GENES.

	Domain					
	N	A1	A1-B1[1] Intron	B1[1] ψ-exon	B2	C
Type of site	**Variability**					
Parsimony sites	13 (0.224)[2]	12 (0.245)	4 (0.105)	3 (0.300)	14 (0.304)	50 (0.714)
Singleton sites	41 (0.706)	36 (0.735)	33 (0.868)	10 (0.660)	31 (0.674)	12 (0.171)
Other variable sites	4 (0.069)	1 (0.200)	1 (0.026)	2 (0.130)	1 (0.022)	8 (0.114)
Total variable sites	58	49	38	15	46	70
Total sites	336	279	400	250	252	207
Frequency of variable sites	0.172	0.175	0.095	0.060	0.182	0.338

[1]Data available for PSG 1,5,8,11 only.

[2]Frequency of parsimony sites relative to the total number of variable sites.

Table 6 shows the pattern of variability for seven PSG genes in each of the N, A1, B2 domains and the C region and for four PSG genes in intron c (between the A1 and B1 domains) and in the BI pseudoexon. Parsimony sites are much more common in the C region than in the other domains. This is most striking in comparison with the N domain, where most of the variability occurs as singleton changes. The significance of these observed patterns in the data is discussed in Chapter 4.

CHAPTER 4. DISCUSSION

4.1 Alternative C-Terminal Domains for the PSG11 Gene

Evidence from hybridisation, mapping and sequencing studies presented here suggests that the cosmids hC3.11, F11193 and F20478 overlap extensively and span the PSG11 locus.

In addition, cosmids hC3.11 and F20478 and the lambda clone λC3 share a region of sequence in common that maps 5' to the PSG11 locus. The restriction maps of this region are somewhat different (FIGs 8 and 13a). This may be due to restriction site polymorphism in the DNA of different individuals, as cosmids hC3.11 and F20478 and λC3 are each derived from different sources. It may also mean that the lambda clone is not derived from the PSG11 gene locus, but comes from somewhere else within the gene family. This is not unlikely as PSG genes share high sequence similarity in both their introns and exons (Chou and Plouzek, 1992, Rudert *et al.*, 1989, Streydio *et al.*, 1990).

This is the first report of the complete gene sequence for a functional subgroup 3 gene. Analysis of sequence in the 3'region of the PSG11 locus showed three potential exons for alternative C domains.

The first exon encodes a complete PSG11w C terminus. Only PSG11 has been reported to contain a Cw domain which is quite unique in that it is 81aa long, unlike all other PSG C domains that vary between 3 and 22aa.

Moreover, two PSG11w cDNA, hPS91 (Chan *et al.*, 1991) and hPS2 (Zheng *et al.*, 1990) have been reported. The two cDNA are identical up to an E*co*RI site in hPS2, which lies 40 bp into the 3'UTR. The E*co*RI site is not present in hPS91 or in the genomic sequence (11wDom) and there is no homology between hPS2 and the othe two sequences (hPS91 and 11wDom) beyond the E*co*RI site. The sequence around the

*Eco*RI site does not resemble a splice consensus sequence which suggests clone hPS2 may be a cloning artifact.

A recent collaborative investigation has shown that PSG11w is expressed in the placenta, throughout pregnancy (Chen *et al.*, 1993). This study also showed that in contrast to other PSG, PSG11w is retained and degraded within the cell, in the endoplasmic reticulum (ER) and raised the possibility that PSG11w may be an integral membrane protein of the ER.

The second PSG11 C-terminal exon encodes a Cr domain with near identity to the Cr domain of PSG6. To date, PSG6r is the only reported transcript containing a Cr domain. There is an intact consensus splice acceptor site in the PSG11 sequence and the domain differs in just 1 nucleotide and 1 amino acid from PSG6. There is extensive similarity in the 3'UTR and the polyadenylation signal, ATTAAA, is apparently used in PSG6. On the basis of sequence data alone it would appear that PSG11 could produce an PSG11r transcript from this locus, although none have been reported to date. It would be interesting to see if such a transcript can be detected. That it has not been detected in placenta despite extensive characterisation, may mean that it is present in other tissues like, for example, PSG6r, which is found in hydatidiform mole tissue (Leslie *et al.*, 1990).

The PSG11 Cr exon lies within a 300bp region of sequence that is part of the Alu family of repeat sequences. Other genes in the CEA/PSG multi-gene family have Alu repeat sequences in their 3'UTR. The distribution and identification of these sequences across the CEA/PSG genes is currently being analysed (Olsen *et al.*, 1994).

The most common transcripts from subgroup 3 genes appear to be those that use a Cs C-terminal exon so it was surprising to find that this exon was the furthest from the end of the B2 domain, some 4.6kb away. The protein domain predicted by this sequence is 12 amino acids long and hydrophilic.

The 3'UTR of the Cs domain exon, contains over 1kb of continuous sequence that has strong identity with sequence from the 3' region of subgroup 1 genes, including sequence encoding the Cc, Ca and Cb domains. Consensus splice sites appear to be intact for the Ca and Cb domains. On the basis of sequence data alone, it may be possible for the PSG11 locus to produce PSG11a and PSG11b transcripts but as yet, none have been detected. These domains may not be used and the sequence may be a remnant of an evolutionary event. Alternatively, the transcripts may be extremely rare or may only be produced at a particular time during development and/or in a particular tissue.

The implication of this finding is that subgroup 1 and subgroup 3 genes are related by an insertion or deletion (indel) event as shown in FIG.21. An examination of the sequences bordering the site of the indel failed to show any small repeat sequences that might indicate that the event was mediated by a transposable element. Subgroup 2 (PSG 2,3,5) genes are similarly related to subgroup 3 genes by an independent insertion or deletion yet to be fully characterised.

4.2 Evolutionary Analysis

The genomic organisation of genes within particular subgroups appears consistent for all genes of that subgroup. This suggests that genes of given subgroup all shared a recent common ancestor. A previous attempt to reconstruct a phylogeny for the PSG gene family (Khan et *al.*, 1992) had suggested genes from the same subgroups were only distantly related. This inference followed from the observation that when parsimony and simple distance methods are used to reconstruct evolutionary trees with N domain sequences, the resulting trees place sequences from different subgroups as most closely related (e.g. PSG1 and 2, FIG.5). In the previous study, the appropriateness of the analysis method was not evaluated. I now present further results which suggest the inferences of Khan *et al.* (1992) were not justified.

I have compared reconstructed trees from different regions of the PSG loci and examined the substitution patterns of these regions.

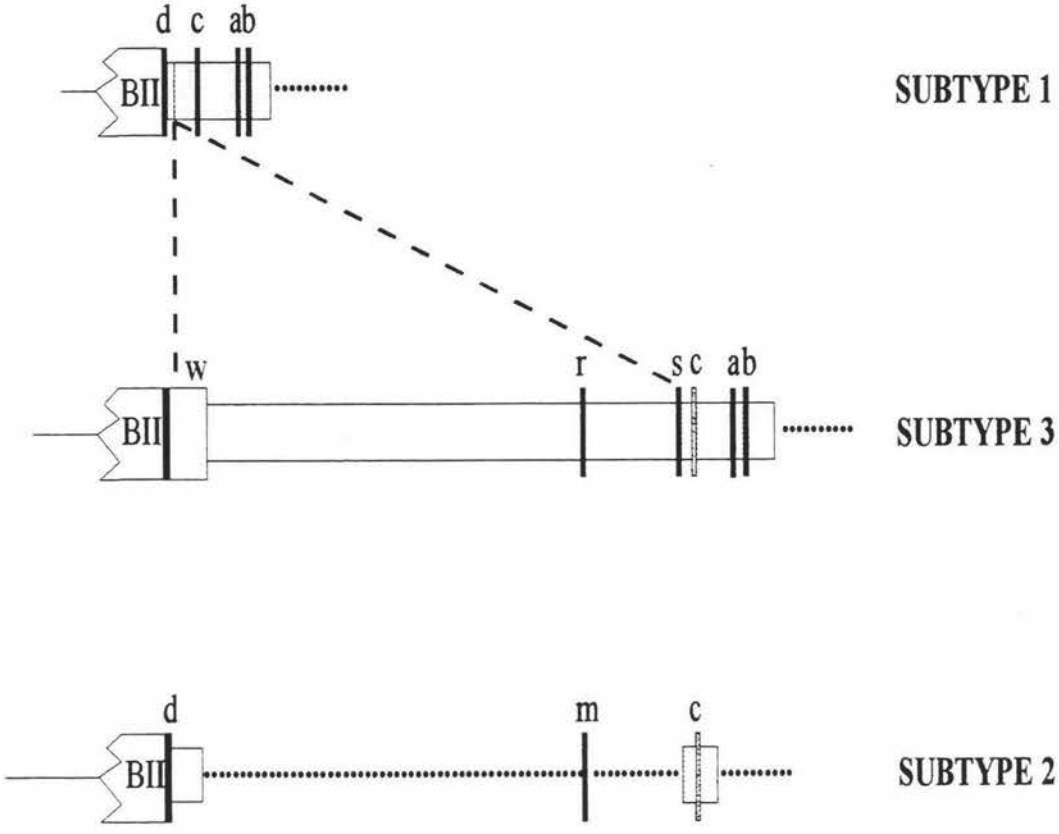


FIG. 21. A diagram showing the relationship between the C-terminal coding regions of subgroup 1, 2 and 3 PSG genes.
Insertion/deletion events relate subgroups 2 and 3 to subgroup1 genes. Alternate C-domains (Ca, b, c, d, m, r, s, w) are labelled and denoted by tall boxes. Related sequences are indicated by clear, solid and hatched boxes. Regions for which there is no sequence available are indicated by dotted lines.

Neighbour-joining trees

The Neighbour joining tree constructed for the N domain (FIG. 20) is very similar to that presented by Khan *et al.*, (1992) in FIG.5. An interesting feature of this tree is that the external edge for PSG11 is comparatively long. This means that PSG11 has undergone a lot of changes in the N domain. If a molecular clock is invoked as part of the mechanism of sequence change, then the presence of this long edge leading to PSG11 implies that PSG11 is the "oldest" locus. Further, if an outgroup is used to root this tree, it will tend to join on to this longest edge (Hendy and Penny, 1989).

In contrast, the longest edge in the C tree is the one that separates subgroup 1 sequences from those of subgroups 2 and 3. Also, the external edges leading to PSG1 and 8 are much longer than those leading to the subgroup 2 and 3 genes. If the evolution of these data is described by a molecular clock (an assumption not tested here) then the implication of these observations is that whilst subgroups 1, 2 and 3 have shared a common ancestral sequence, subgroup 1 genes diverged earlier and subgroups 2 and 3 have only relatively recently diverged from each other. This is consistent with the hypothesis that the RGD sequence present in the N domain is an original characteristic which has been lost in some genes of the subgroup 1 lineage.

SplitsTree graphs

Commonly used implementations of tree building methods are based on simple assumptions of sequence evolution, e.g. all changes are independent and identically distributed (Penny *et al.*, 1992, Lockhart *et al.*,1994). The reliability of both the method and the resulting tree is dependent on there being a good fit between such assumptions and the actual evolution of the biological sequences in question. It is important to be able to evaluate whether the assumptions of any particular tree building process are met by the data. Recently developed methods such as the Hadamard analysis (Hendy *et al.*, 1994) and split decomposition (Bandelt and Dress, 1992) make it possible to help evaluate the fit between a model (i.e.: a tree with edge lengths and the mechanism

describing sequence change) and the data. In the present study, split decomposition is used to evaluate inference from PSG sequences (FIG.20).

There are many contradictory signals in the data for the N domain and this can be seen clearly in the box-like, starburst nature of the SplitsTree graph. Because there is such a poor fit between the model and data, there can be no confidence that this tree is an accurate description of the evolutionary relationships among the PSG family.

The SplitsTree graph for the C-region however is much more two dimensional, particularly in the signal separating PSG1 and 8 from PSG2, 3, 5, 6 and 11. This signal is also consistent with genomic organisation data and this observation gives some confidence that the C-region sequences contain information about historical relationships in the PSG gene family.

The contradiction of this signal, most strongly in the N domain, suggests either that the domains have had different evolutionary histories or that historical signals have been obscured by other processes of evolution. The SplitsTree graphs for the N and A1 domains indicate that there are no clear relationships between any of the PSG and this observation supports the latter hypothesis.

Processes such as positive selection and/or gene conversion and/or unequal crossover between loci, have been suggested as having the effect of homogenising sequences and in some cases creating hypervariability in certain regions of other members of the immunoglobulin superfamily (Li and Graur, 1991, Ohta, 1991, Ohta, 1992, Hughes and Nei, 1988, Tanaka and Nei 1989). Such processes may have been operating also in the evolution of the CEA/PSG gene family. Further observations described below support this proposal.

Observed patterns in the data.

Three types of patterns can be defined in aligned sequences; they are *parsimony sites, two character state singleton sites* and *other variable sites*.

Parsimony sites provide direct support for possible internal edges of a tree. For example, the pattern shown in column **p,** FIG.19, simply interpreted, would provide support for an edge separating PSG1 and 8 from PSG2, 3, 5, 6, 11. The relative numbers of these types of patterns in the data are important for determining support for and against possible internal edges in the best tree.

Two character state singleton sites are patterns which show a change has occurred on an external edge of a tree. For example, in column **s** in FIG.19, a change has occurred on the edge leading to PSG1. The relative number of singleton changes in the data then, provides information about the lengths of the external edges.

Other variable sites (column **o**, FIG.19) can be used to aid in the interpretation of the parsimony patterns by providing information about the probability of multiple changes occurring at the parsimony sites.

The PSG gene family is thought to have arisen through the duplication of a single primordial gene (Streydio *et al.*,1990). At some point in this duplication, the three subgroups arose. By comparing the relative numbers of parsimony and two character state singleton patterns in the data it is possible to determine when changes have been accumulating in the PSG sequences. For example, many parsimony sites and few singletons indicates that most point changes have accumulated **before or as** the PSG sequences duplicated, whereas many singletons and few parsimony sites might indicate that the changes have accumulated **after** the sequences duplicated.

The patterns of variability across seven PSG genes is summarised in Table 6. Most strikingly, the ratio of the total number of variable sites to the total number of sites is

smaller for the N, A1 and B2 domains than it is for the C region. That is, the C region has accumulated more changes than the domains. Since the C region shows the best fit between model and data (fewer conflicting patterns), this suggests the N, A1, B2 domains have been constrained or homogenised in some way throughout their evolution. On this interpretation, there would be little expectation of finding historical signals in the N domains which could be used to reconstruct PSG phylogeny.

Further, the number of parsimony sites in each of the N, A1, B2 domains is much smaller than the number of singleton sites, whereas this trend is reversed in the C region. As is discussed above, this pattern most likely indicates that most of the observed changes in the N, A1 and B2 domains have occurred after the PSG genes duplicated. In contrast, changes in the C region have accumulated whilst the PSG genes were duplicating. The observed patterns of variability also suggest that after duplication of PSG loci, processes that were acting to constrain or homogenise the genes were removed / were relaxed or became inefficient.

Surprisingly, the pattern of variability for the intron and the B1 pseudoexon is the same as for the N, A1 and B2 domains. This observation is interesting as these regions should be free from selective constraint. This further suggests, that whole PSG gene units, have been subject to some kind of homogenisation during duplication.

Amino acid sequence variation

The PSG are members of the Immunoglobulin superfamily. It has been reported elsewhere that amino acid sequence variation was more marked in regions of the PSG thought to be analogous to the complementarity determining regions of the immunoglobulins (Khan *et al.*,1992).

When amino acid changes are plotted for the N domain, the changes cluster around the RGD tripeptide in the N-domain (Khan *et al.*, 1992, McLenachan *et al.*, 1995, in press). This region may have accumulated changes simply because it can, without disrupting the

protein structure of the molecule. However, the RGD tripeptide is involved in cell signalling (Ruoslahti and Piersbacher,1987) and it is interesting therefore to speculate that the variation in this region is in some way important with respect to interactions between the different PSG molecules and their cellular receptors.

Two different hypotheses have been proposed as mechanisms for generating hypervariability within the complementarity determining regions of immunoglobulins and in the antigen recognition sequences of the major histocompatibility complex. These are gene conversion and positive selection. Gene conversion occurs when small regions of sequence are "corrected" by recombinational events, from a template (Li and Graur, 1991). It is generally invoked as a mechanism to explain homogenisation such as in the small repeat sequences of the Alu family in the human genome or the existence of short sequences within a domain that appear to have a different history to the surrounding sequence. However, it is also thought to be able to generate hypervariability within members of the immunoglobulin family (Ohta, 1992, Wysocki and Gefter, 1989).

Positive selection is a descriptive term and is used when the number of non-synonymous amino acid changes in a coding region exceeds the number of synonymous amino acid changes. This is the case with the changes seen in the N-domain of the PSG and particularly around the RGD tripeptide region. Such a process is thought to be another means of generating hypervariability in members of the immunoglobulin superfamily (Tanaka and Nei, 1989, Hughes and Nei, 1988)

Specific statistical tests have been used to differentiate between gene conversion and positive selection (Hughes and Nei, 1988, Sawyer, 1989, Ohta, 1992), but they cannot be applied to the PSG data sets as the power of the tests is too low to obtain statistically meaningful results with these data. Hence, whilst processes such as gene conversion and positive selection are clearly implicated in the evolution of the PSG genes, it is not possible to evaluate their contribution to the hypervariability seen in the RGD region of the N domains.

CHAPTER 5. CONCLUSIONS.

Summary.

In summary, this thesis presents the isolation, mapping and sequence analysis of the 3' region of a functional PSG subgroup 3 gene, PSG11.

The exons for three alternative C-terminal domains have been determined. Two of these exons are known to be incorporated into cDNA transcripts (Cw and Cs). PSG11r has not been previously reported as a transcript. The presence of a good consensus splice acceptor sequence at the start of this putative domain and close similarity to the expressed PSG6r domain are indications that it could be expressed. PSG6r is thought to be a hydatidiform molar specific transcript (Leslie *et al.*, 1990). The expression of PSG11r may also be tissue specific. The sequencing of this region will enable the synthesis of specific oligonucleotide primers for the detection of PSG11r transcripts in a variety of tissues by reverse transcriptase PCR in further studies.

The genomic organisation of the 3' region of PSG11 has been completed with the Cs domain being found to lie 4.6kb from the end of the B2 domain. An interesting and exciting discovery was the determination of a region of sequence which is homologous to approximately 1kb of sequence within the 3'UTR of subgroup1 genes. Close examination of the 3'UTR shows juxtaposed regions of homology and non-homology and suggests the different subgroups are related by insertion/deletion events. Subgroup 2 and subgroup 3 genes are more similar to each other than to subgroup1 genes.

The discovery of sequences homologous to subgroup 1 C-terminal regions within the subgroup 2 and 3 genes gave a region of DNA from all three subgroups that was used to determine historical relationships between PSG members. Sequence data from the N, A1 and to a certain extent B2, domains was found to be unsuitable for use in reconstructing historical relationships. Such domains contain many contradictory patterns which would suggest contradictory histories (as can be seen in the split decomposition plots). Our

observations about the variability patterns in the data (consistent with those of others on patterns of change in domains of other members of the immunoglobulin gene family) suggest that the mechanism of sequence evolution for these regions is not neutral. Hence the patterns in these data are misleading for reconstructing evolutionary histories. In contrast, the data from sequence in common at the C region, resolves a tree that is consistent with the genomic organisation of the genes. Thus we suggest this region may contain the most useful information for reconstructing historical relationships for the PSG genes.

The mapping and sequencing data, together with the sequence analysis presented in this thesis, now provide a more comprehensive framework for further investigations into a potential biological role for the PSG in general and more specifically, for PSG11.

REFERENCES

Arakawa F, Kuroki M, Misumi Y, Matsuo Y, Matsuoka Y (1991). The nucleotide and deduced amino acid sequences of a cDNA encoding a new species of Pregnancy-Specific β-1 glycoprotein (PSβG). Biochimica et Biophysica Acta 1048:303-305.

Bandelt H-J and Dress AWM (1992). A canonical decomposition theory for metrics on a finite set. Advanced Mathematics 92:47-105

Barnett TR, Kretschmer A, Austen DA, Goebel SJ, Hart JT, Elting JJ, Kamark ME (1989). Carcinoembryonic Antigens: Alternative Splicing Accounts for the Multiple mRNAs that Code for Novel Members of the Carcinoembryonic Antigen Family. The Journal of Cell Biology 108: 267-276.

Barnett TR, Pickle II W, Elting JJ (1990). 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

Barnett TR and Zimmerman W (1990). Workshop Reports: Proposed Nomenclature for the Carcinoembryonic Antigen (CEA) Gene Family. Tumor Biology 11:59-63

Bates PA, Jingchu L and Stenberg MJE (1992). A predicted three dimensional structure for the carcinoembryonic antigen (CEA). FEBS Letters 301: 207-214

Beer AE and Billingham RE (1976). The Immunobiology of Mammalian Reproduction. Prentice-Hall.

Beggs KT (1990).Characterisation of the Genomic Sequence of a Member of the Pregnancy-Specific β1 glycoprotein Family. BSc(Hons) Thesis, Massey Universtiy.

Benchimol S, Fuks A, Jothy S, Beauchemin N, Shirota K, Stanners CP (1989). Carcinoembryonic Antigen, a Human Tumour Marker, Functions as an Intercellular Adhesion Molecule. Cell 57: 327-324.

Bergelson JM and Finberg RW (1993). Integrins as receptors for virus attatchment and cell entry. Trends in Microbiology 1 8:287-288

Birnboim HC and Doly J (1979). A Rapid Alkaline Extraction Procedure for Screening Recombinant Plasmid DNA. Nucleic Acid Research 7:1513 -1515

Bischof P (1984). Placental Proteins . Contributions to Gynaecology and Obstetrics 12:1-96

Bohn H (1971). Nachweis und Charakterisierungvon Schwangerschafts protein in der Menschlichen Plazenta, sowie ihre quantitative immunologische Bestimmung im serum schwangerer Frauen. Archives of Gynaecology 210:440-457

Bolton ET and McCarthy BJ (1962). A General Method for the Isolation of RNA Complementary to DNA. Proceedings of the National Academy of Science USA 48:1390.

Borjigin J, Tease LA, Barnes W and Chan W-Y (1990). Expression of the Pregnancy-Specific Beta 1-Glycoprotein Genes in Human Testis. Biochemical and Biophysical Research Communications 166 2:622-629.

Brandriff BF, Gordon LA, Tynan KT, Oslen AS, Mohrenweiser HW, Fertitta A, Carrano AV, Trask BJ (1992). Order and Genominc Distances among Members of the Carcinoembryonic Antigen (CEA) Gene Family Determined by Fluoresence *in Situ* Hybridization. Genomics 12: 773-779.

Brophy BK, MacDonald RE, McLenachan PA and Mansfield BC (1992). cDNA sequence of the Pregnancy-Specific β1-glycoprotein-11s (PSG11s). Biochimica et Biophysica Acta 1131:119-121

Buttice G, Kaytes P, D'Armiento J, Vogeli G, Kurkinen M (1990). Evolution of Collagen IV Genes from a 54-Base Pair Exon: A Role for Introns in Gene Evolution. Journal of Molecular Evolution 30:479-488

Chamberlin ME, Lei K-J, Chou JY (1994). Subtle Differences in Human Pregnancy-Specific Glycoprotein Gene Promoters Allow for Differential Expression. The Journal of Biological Chemistry 269, 25: 17152-17159.

Chan W-Y, Borjigin J, Shupert WL, Zheng Q-X (1988a). Characterization of cDNA Encoding Human Pregnancy-Specific β1-Glycoprotein from Placenta and Extraplacental Tissues and Their Comparison with Carcinoembryonic Antigen. DNA 7 8: 545-555.

Chan W-Y, Tease LA, , Borjigin J, Chan PK, Rennert OM, Srinivasan B, Shupert WL, Cook RG (1988b). Pregnancy-Specific β1 glycoprotein mRNA is present in placental as well as non-placental tissues. Human Reproduction 3 5: 677-685.

Chan W-Y, Tease LA, Bates Jr JM, Borjigin J, Shupert WL (1988c). Pregnancy-Specific β1 glycoprotein in rat: tissue distribution of the mRNA and identification of testicular cDNA clones. Human Reproduction 3 5: 687-892.

Chan W-Y, Zheng Q-X, McMahon J and Tease LA (1991). Characterisation of new members of the Pregnancy-Specific β1-glycoprotein family. Molecular and Cellular Biochemistry 106:161-170

Chen H, Chan W-Y, Chen C-L, Mansfield BC, Chou JY (1993). The Carboxyl-terminal Domain of the Human Pregnancy-Specific Glycoprotein Specifics Intracellular Retention and Stability. The Journal of Biological Chemistry 269 29: 22066-22075.

Chou JY and Plouzek CA (1992). Pregnancy-Specific β1-Glycoprotein. Seminars in Reproductive Endocrinology 10 2:116-126

Chu TM, Reynoso G, Hansen HJ (1972). Demonstration of Carcinoembryonic Antigen in Normal Human Plasma. Nature 238:152-153

Cohen SN, Chang ACY, Hsu L (1972). Non-chromosomal Antibioic Resistance in Bacteria: Genetic Transformation of E.*coli* by R-factor DNA. Proceedings of the National Academy of Science USA 69:2110

Deininger PL, Jolly DJ, Rubin CM, Friedmann T, Schmid CW (1981). Base Sequence Studies of 300 Nucleotide Renatured Repeated Human DNA Clones. The Journal of Molecular Biology 151: 17-33.

De Jong PJ, Yokobata K, Chen C, Lohman F, Pederson L, McNinch J, Vandilla M (1989). Human Chromosome-specific Partial Digest Libraries in K and Cosmid Vectors. Cytogenetics and Cellular Genetics 51:1-4

Fagnart OC, Cambiaso CL, Lejeune MD, Noel G, Maisin H, Masson PL (1985). Prognostic Value of Concentration of Pregnancy-Specific β-1 Glycoprotein (SP-1) in Serum of Patients with Breast Cancer. International Journal of Cancer 36:541-544.

Felsenstein J (1993). PHYLIP, Phylogeny Inference Package Version 3.5. The University of Washington, Seattle, W.A.

Genetics Computer Group Inc.(1991). Program manual for the GCG Package Version 7. University Research Park, Madison Wisconsin.

Gold P and Freedman S (1965). Specific Carcinoembryonic Antigens of the Human Digestive System. Journal of Experimental Medicine 122:467-481.

Guilbert L, Robertson SA, Wegmann TG (1993). The Trophoblast as an Integral Component of a Macrophage Cytokine Network. Immunology and Cell Biology 71 1:49-57

Hamilton WJ, Boyd JD, Mossman HW (1972). Human Embryology. Cambridge: W Heffer and Sons Ltd. 4th edition.

Heery DM, Gannon F, Powell R (1990). A Simple Method for Subcloning DNA Fragments from Gel Slices. Trends in Genetics 6 6: 173.

Heikinheimo M, Aula P, Rapola J, Wahlstrom T, Jalanko H, Sepala M (1982). Amniotic Fluid Pregnancy-Specific β-1 Glycoprotein (SP-1) in Meckels' Syndrome: a New Test for Prenatal Diagnosis? Prenatal Diagnosis 2:103-108.

Hendy MD and Penny D (1989). A framework for the quantative study of evolutionary trees. Systematic Zoology 38:297-309

Hendy MD, Steel MA and Penny D (1994) A discreet Fourier analysis for evolutionary trees. Proceedings of the National Academy of Science USA 91:3339-3343

Hertz JB and Schultz-Larsen P (1986). Placental Proteins in Threatened Abortion. In Hau J (ed) : Pregnancy Proteins in Animals. New York : Walter de Gruyter, pg 31-40.

Hinoda Y, Neumaier M, Hefta SA, Drzeniek Z, Wagener C, Shively L, Hefta LJF, Shively JE, Paxton RJ (1988). Molecular cloing of a cDNA coding biliary glycoprotein I: Primary Structure of a glycoprotein immunologically crossreative with carcinoembryonic antigen. Medical Sciences 85: 6959-6963.

Holmes DS and Quigley M (1981). A Rapid Boiling Method for the Preparation of Bacterial Plasmids. Analytical Biochemistry 114:193

Horne CHW, Reid IN, Milne GD (1976). Prognostic Significance of Inapproiate Production of Pregnancy Proteins by Breast Cancers. Lancet 2:279-282

Hughes AL and Nei M (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170

Huson D and Wetzel R (1994). SplitsTree Version 1.0 University of Bielefeld, Germany.

Huttenmoser JL, Weil-Franck C and Bischof P (1987). The disappearance rate of Schwangerschafts-protein 1 in normal and pathological pregnancies. British Journal of Obstetrics and Gynaecology 5:420-424

Itakura K, Rossi JJ, Wallace RB (1984). Synthesis and Use of Synthetic Oligonucleotides. Annual Review of Biochemistry 53:323

Joe TW (1994). Characterisation of the Genomic Region Upstream of PSG-11. Dissertation, MSc , Massey University.

Joe TW, McLenachan PA and Mansfield BC (1994). Sequence of a Novel Pregnancy-Specific β1-Glycoprotein C-Terminal Domain. Biochimica et Biophysica Acta 1219:195-197.

Jukes TH and Cantor CR (1969). Evolution of Protein Molecules p12-132 in Mammalian Protein Metabolism , HN Munro Ed. Academic Press, New York.

Kan M and Tatarinov Y (1990). Proceedings of the International Workshop on CEA, Montreal. pg11.

Khan WN and Hammarstrom S (1989). Carcinoembryonic Antigen Gene Family: Molecular Cloning of cDNA for a PSβG/FL-NCA Glycoprotein with a Novel Domain Arrangement. Biochemical and Biophysical Research Communications 161 2:525-535.

Khan WN, Teglund S, Bremer K, Hammerstrom S (1992). The Pregnancy Specific Glycoprotein Family of Immunoglobulin Superfamily: Identification of New Members and Estimation of Family Size. Genomics 12: 780-787.

Kimura M (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16:111-120

Kuroki M, Kuroki M, Ichiki S, Matsuoka Y (1984). Identification and partial characterisation of the unglycosolated peptide of carcinoembryonic antigen synthesised by human tumor cell lines in the prescence of tunicamycin. Molecular Immunology 21 8:743-746 .

Lei K-J, Sartwell AD, Pan C-J, Chou JY (1992). Cloning and Expression of Genes Encoding Human Pregnancy-Specific Glycoproteins. The Journal of Biological Chemistry, 267 23:16371-16378.

Lei K-J, Wang C, Chamberlin ME, Liu J-L, Pan C-J, Chou JY (1993). Characterization of Two Allelic Varients of a Human Pregnancy-Specific Glycoprotein Gene. The Journal of Biological Chemistry, 268:17528-17538

Leslie KK, Watanabe S, Lei K-J, Chou DY, Plouzek CA, Deng H-C, Torres J, Chou JY (1990). Linkage of two human pregnacy-specific β1-glycoprotein genes: One is associated with hydatidiform mole. Proceedings of the National Academy of Science USA 87: 5822-5826.

Li W-H and Graur D (1991). Fundamentals of Molecular Evolution. Sinauer Associates, Inc.

Lin TM, Halbert SP, Spellacy WN (1974). Measurement of Pregnancy-associated Plasma Proteins During Human Gestation. Journal of Clinical Investigations 54:576-582

Lockhart PJ, Steel MA, Hendy MD and Penny D (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. Molecular Biology and Evolution 11: 605-612

McLenachan PA, Lockhart PJ, Faber HR, Mansfield BC (1995). Evolutionary Analysis of the Multi-gene Pregnancy-specific β_1-glycoprotein Family: Separation of Historical and Non-historical Signals. (submitted: Journal of Molecular Evolution)

McLenachan PA, Rutherfurd KJ, Beggs KT, Sims SE and Mansfield BC (1994). Characterisation of the PSG11 Gene. Genomics 22:356-363.

McLenachan Tand Mansfield B (1989). Expression of CEA-Related Genes in the First Trimester Human Placenta. Biochemical and Biophysical Research Communications 162 3: 1486-1493.

Masson GM, Anthony F, Wilson MS (1983). Value of Schwangershaftsprotein (SP-1) and Pregnancy - associated Protein (PAPP-A) in the Clinical Management of Threatened Abortion. British Journal of Obstetrics and Gynaecology 90:146-149.

Miller JH (1972). Experiments in Molecular Genetics. Cold Spring Harbour Laboratory, Cold Spring Harbour N.Y.

Mosmann TR and Moore KW (1991). The Role of IL-10 in Crossregulation of T_H1 and T_H2 Responses. Immunology Today, March A48-A53.

Neumaier M, Zimmerman W, Shively L, Hinoda Y, Riggs AD, Shively JE (1988). Characterisation of a cDNA Clone for the Nonspecific Cross-reacting Antigen (NCA) and a Comparison of NCA and Carcinoembryonic Antigen. The Journal of Biological Chemistry 263 7:3202-3207.

Ohta T (1992). A Statistical Examination of Hypervariability in Complementarity-Determining Regions of Immunoglobulins. Molecular Phylogenetics and Evolution 1 4:305-311.

Ohta T (1991). Multigene Families and the Evolution of Complexity. Journal of Molecular Evolution 33:34-41

Olsen A, Georgescu A, Batzer M, Burgin M, Gordon L Lennon G and Carrano A (1994). Cosmid clone map of a 2.3MB region in 19q13.2 encompassing the CEA/PSG gene family. Proceedings from the Fifth International CEA/PSG Workshop, 18-20 July 1994, University of Freiberg, Germany.

Oikawa S, Nakazato H, Kosaki G (1987). Primary Structure of Human Carcinoembryonic Antigen (CEA) Deduced from cDNA Sequence. Biochemical and Biophysical Research Communications 142 2:511-518.

Oikawa S, Inuzuka C, Kosaki G, Nakazato H (1988). Exon-Intron Orginization of a Gene for Pregnancy-Specific β1- Glycoprotein, a Subfamily Member of a CEA Family: Implications for its Characteristic Repetitive Domains and C-Terminal Sequences. Biochemical and Biophysical Research Communications 156 1: 68-77.

Oikawa S, Inuzuka C, Kosaki G, Nakazato H, Kuroki M (1989a). Pregnancy-Specific β1-Glycoprotein, a CEA Gene Family Member, Expressed in a Human Promyelocytic Leukemia Cell Line, HL-60: Structures of Protein, mRna and Gene. Biochemical and Biophysical Research Communications 163 2:1021-1031.

Oikawa S, Inuzuka C, Kuroki M, Matsuoka Y, Kosaki G, Nakazato H (1989b). Cell Adhesion Activity of Non-specific Cross-reacting Antigen (NCA) and Carcinoembryonic Antigen (CEA) on CHO Cell Surface: Homophillic and Heterophillic Adhesion. Biochemical and Biophysical Research Communications 164 1:39-45

Paxton RJ, Mooser G, Pande H, Lee TD, Shively JE (1987). Sequence Analysis of Carcinoembryonic Antigen: Identification of Glycosolation Sites and Homology with the Immunoglobulin Superfamily. Proceedings of the National Academy of Science USA. 84:920-924.

Penny D, Hendy MD and Steel MA (1992). Progress with Evolutionary Trees. Trends in Ecology and Evolution 7: 73-79

Penny D, Watson EE, Hickson RE and Lockhart PJ (1993). Some recent progress with methods for evolutionary trees. New Zealand Journal of Botany 31:275-288

Petrocik E, Wassman ER, Lee JJ, Kelly JC (1990). Second Trimester Maternal Serum Pregnancy-Specific Beta-1-Glycoprotein (SP-1) levels in Normal and Down Syndrome Pregnancies. American Journal of Medical Genetics 37:114-118.

Promega Protocols and Applications Guide (1991). 2nd Edition. (David Titus ed.)

Powell SK, Cunningham BA, Edelman GM, Rodriguez-Boulan E (1991). Targeting of transmembrane and GPI-anchored forms of N-CAM to opposite domains of a polarized epithelial cell. Nature 353: 76-77.

Ramsey EM and Donner MW (1980). Placental Vasculature and Circulation. Georg Thieme Publishers, Stuttgart.

Rigby PWJ, Dieckmann M, Rhodes C, Berg P (1977). Labelling Deoxyribonucleic Acid to High Specific Activity by Nick Translation with DNA Polymerase 1. Journal of Molecular Biology 113:237-251.

Roitt IM (1980). Essential Immunology. Blackwell Scientific Publications. 4th edition.

Rojas M, Fuks A, Stanners CP (1990). Bilary Glycoprotein, a Member of the Immunoglobulin Supergene Family, Functions *in vitro* as a Ca^{2+}-Dependent Intercellular Adhesion Molecule. Cell Growth and Differentiation 1:527-533.

Rooney BC, Wilson Horne CH, Hardman N (1988). Molecular cloning of a cDNA for human pregnacy-specific β1 Glycoprotein: hormology with human carcinoembryonic antigen and related proteins. Gene 71: 439-449.

Rosen S (1986). New Placental Proteins: Chemistry, Physiology and Clinical Uses. Placenta 7:575-594.

Rudert F, Zimmerman W, and Thompson JA (1989). Intra- and Interspecies Analyses of the Carcinoembryonic Antigen (CEA) Gene Family Reveal Independent Evolution in Primates and Rodents. Journal of Molecular Evolution 29:126-134

Ruoslahti E and Pierschbacher MD (1987). New Perspectives in Cell Adhesion: RGD and Integrins. Science 238:491-497.

Sambrook J, Fritsch EF, Maniatas T (1989). Molecular Cloning, a Laboratory Manual. Cold Spring Harbour Press, 2nd Edition.

Saunders SE and Burke JF (1990). Rapid Isolation of Miniprep DNA for Double Strand Sequencing. Nucleic Acid Research 18 16:4948

Sawyer S (1989). Statistical Tests for Detecting Gene Conversion. Molecular Biology and Evolution 6 5:526-538

Schewe H, Thompson J, Bona M, Hefta LJF, Maruya A, Hausser M, Shively JE, von Kleist S and Zimmerman W (1990). Cloning of the complete gene for the carcinoembryonic antigen: Analysis of its promoter indicates a region conveying cell type-specific expression. Molecular and Cellular Biology 10: 2738-2748

Springer TA (1990). Adhesion Receptors of the Immune System. Nature 346: 425-434

Sterzik K, Rosenbusch B, Benz R (1989). Serum-specific Protein-2 and Beta-human Chorionic Gonadotrophin Concentration in Patients with Suspected Ectopic Pregnancies. International Journal of Gynaecology and Obstetrics 28:253-256

Streydio C, Lacka K, Swillens S, Vassart G (1988). The Human Pregnacy-Specific β1-Glycoprotein (PSβG) and the Carcinoembryonic Antigen (CEA)-Related Proteins are Members of the Same Multigene Family. Biochemical and Biophysical Research Communications 154 1:130-137.

Streydio C, Swillens S, Georges M, Szpirer C and Vassart G (1990). Structure, Evolution and Chromosomal Localization of the Human Pregnancy-Specific β1 Glycoprotein Gene Family. Genomics 6:579-592

Streydio C and Vassart G (1990). Expression of Human Pregnancy Specific β1 Glycoprotein (PSG) Genes During Placental Development. Biochemical and Biophysical Research Communications 166, 3: 1265-1273.

Tamsen L, Johansson SGO, Axelsson O (1983). Pregnancy-Specific β-1 Glycoprotein (SP-1) in Serum of Pregnant Women with Pregnancies Complicated by Intrauterine Growth Retardation. The Journal of Perinatal Medicine 11:19-25

Tanaka T and Nei M (1989). Positive Darwinian Selection Observed at the Variable- Region Genes of Immunoglobulins. Molecular Biology and Evolution 6 5:447-459.

Tatarinov YS (1978). Trophoblast-specific Beta-1 Glycoprotein as a Marker for Pregnancy and Malignancies. Gynaecologic and Obstetric Investigation 9:65-97

Tatarinov YS and Masyukevich VN (1970). Immunological Identification of a New Beta-1 Globulin in the Blood Serum of Pregnant Women. Byull Eksp Biol Med USSR 69:66-68.

Tawaragi Y, Oikawa S, Matsuoka Y, Kosaki G, Nakazato H (1988). Primary Structure of Nonspecific Crossreacting Antigen (NCA), a Member of Carcinoembryonic Antigen (CEA) Gene Family, Deduced from cDNA Seqeunce. Biochemical and Biophysical Research Communications 150 1:89-96.

Teglund S and Hammarstrom S (1994). The PSG gene region contains six genes (CGM13-CGM18) that form a new subgroup within the CEA family. Proceedings from the Fifth International CEA/PSG Workshop, 18-20 July 1994, University of Freiberg, Germany.

Thomas P and Toth CA (1990). Carcinoembryonic Antigen Binding to Kupffer Cells is via a Peptide Located at the Junction of the N-terminal and First Loop Domains. Biochemical and Biophysical Research Communications 170 1:391-396 .

Thompson J and Zimmerman W (1988). The Carcinoembryonic Antigen Gene Family: Structure, Expression and Evolution. Tumor Biology 9:63-83.

Thompson JA, Mauch E-M, Chen F-S, Hinoda Y, Schrewe H, Berling B, Barnert S, von Kleist S, Shively JE, Zimmerman W (1989). Analysis of the Size of the Carcinoembryonic Antigen (CEA) Gene Family: Isolation and Sequencing of N-terminal Domain Exons. Biochemical and Biophysical Research Communications 158 3:996-1004.

Thompson J, Koumari R, Wagner K, Barnet S, Schleussner C, Schrewe H, Zimmerman W, Muller G, Schempp W, Zaninetta D, Ammaturo D, Hardman N (1990). The Human Pregnancy-Specific Glycoprotein Genes are Tightly Linked on the Long Arm of Chromosome 19 and are Coordinately Expressed. Biochemical and Biophysical Research Communications 167 2: 848-859.

Thompson J, Grunert F, Zimmerman W (1991). The Carcinoembryonic Antigen Gene Family : Molecular Biology and Clinical Perspectives. Journal of Clinical Laboratory Analysis 5:344-366

Thompson J, Zimmerman W, Osthus-Bugat P, Schleussner C, Eades-Perner A-M, Barnert S, von Kleist S, Willcocks T, Craig I, Tynan K, Olsen A, Mohrenweiser H (1992). Long-Range Chromosomal Mapping of the Carcinoembryonic Antigen (CEA) Gene Family Cluster. Genomics 12 761-772.

Trask B, Fertitta A, Christensen M, Youngblom J, Bergmann A, Copeland A, de Jong P, Mohrenweiser H, Olsen A, Carrano A, Tynan K (1993). Flourescence *in Situ* Hybridisation Mapping of Human Chromosome 19: Cytogenetic Band Location of 540 Cosmids and 70 Genes or DNA Markers. Genomics 15:133-145.

Turbide C, Rojas M, Stanners CP, Beauchemin N (1991). A Mouse Carcinoembryonic Antigen Gene Family Member is a Calcium-dependent Cell Adhesion Molecule. The Journal of Biological Chemistry 266 1:309-315.

Tynan K, Olsen A, Trask B, de Jong P, Thompson J, Zimmerman W, Carrano A, Mohrenweiser H (1992). Assembly ans Analysis of Cosmid Contigs in the CEA-gene Family Region of Human Chromosome 19. Nucleic Acids Research 20 7:1629-1636.

Watanabe S and Chou JY (1988 a). Human Pregnancy-Specific β1-Glycoprotein: A New Member of the Carcinoembryonic Antigen Gene Family. Biochemical and Biophysical Research Communications 152 2:762-768

Watanabe S and Chou JY (1988 b). Isolation and Characterisation of Complementary DNAs Encoding Human Pregnancy-Specific β1-Glycoprotein. The Journal of Biological Chemistry 263 4:2049-2054

Williams AF (1987). A Year in the Life of the Immunoglobulin Superfamily. Immunology Today. 8 10:298-303.

Wright C, Angus B, Napier J, Wetherall M, Udagawa Y, Sainsbury JRC, Johnston S, Carpenter F, Horne CHW (1987). Prognostic Factors in Breast Cancer: Immunohistochemical Staining for (SP-1) and NCRC 11 Related to Survival, Tumor Epidermal Growth Factor and Oestrogen Receptor Status. Journal of Pathology 153:325-331.

Wu S-M, Bazar LS, Cohn ML, Cahill RA and Chan W-Y (1993). Expression of Pregnancy-Specific β-1 Glycoprotein Genes in Hematopoietic Cells. Molecular and Cellular Biochemistry 122:147-158

Wysocki LJ and Gefter ML (1989). Gene Conversion and the Generation of Antibody Diversity. Annual Reviews of Biochemistry 58:509-531

Zheng Q-X, Tease LA, Shupert WL, Chan W-Y (1990). Characterisation of cDNAs of the Human Pregnancy-Specific β1-Glycoprotein Family, a New Subfamily of the Immunoglobulin Gene Superfamily. Biochemistry 29:2845-2852.

Zimmerman W, Weiss M and Thompson JA (1989). cDNA Cloning Demonstrates The Expression Of Pregnancy-Specific Glycoprotein Genes, A Subgroup Of The Carcinoembryonic Antigen Gene Family, In Foetal Liver. Biochemical and Biophysical Research Communications 163 3:1197-1209.

Zoubir F, Khan WN and Hammarstrom S (1990). Carcinoembryonic Antigen Gene Family members in Submandibular Salivary Gland: Demonstration of Pregnancy-Specific Glycoproteins by cDNA Cloning. Biochemical and Biophysical Research Communications 169 1:203-216.

APPENDIX