CANDIDATE GENE ANALYSIS OF THE CYP11A1 GENOMIC REGION IN AND POLYCYSTIC OVARY SYNDROME

By

Michelle Lynn Gaasenbeek

A thesis submitted for the degree of

Doctor of Philosophy

in the

University of London

April 2005

Department of Paediatrics, Gynaecology and Obstetrics

Imperial College Faculty of Medicine

Hammersmith Hospital

London W12 0NN



To my husband, David

ABSTRACT

Polycystic ovary syndrome (PCOS) is one of the most common endocrinopathies affecting women of reproductive age. Previous studies have shown linkage at the *CYP11A1* locus and a significant association between a pentanucleotide repeat and testosterone levels in PCOS women. In this thesis, examination of the *CYP11A1* locus to study the relationship between *CYP11A1* and PCOS is described.

Examination of the *CYP11A1* locus, by computational and molecular analysis, has identified 2 novel exons (1a and 1b) located in intron 1 of *CYP11A1* and a tetranucleotide repeat 5' of exon 1a. Expression the new exons was demonstrated in placenta, ovary and adrenal. A novel gene has been identified in the genomic region upstream of *CYP11A1*.

Association studies examining *CYP11A1* and PCOS have been completed. The pentanucleotide and tetranucleotide repeats were genotyped in 230 PCOS cases, 331 population controls and 141 parent-offspring trios of UK origin, and 1589 women from a population-based birth cohort from Finland (NFBC1966). Though nominally-significant differences in allele and genotype frequencies at both loci were observed, these findings were not substantiated in the other analyses, and no discernable relationship was seen between variation at these loci and serum testosterone, or other intermediate traits.

The role of the repeat length of the pentanucleotide repeat in the regulation of *CYP11A1* transcription was investigated by transient transfection analysis. Three constructs, containing 4-, 6- or 9-repeats, were transfected into JEG-3 cells. Preliminary results of

this study have shown that transcription of *CYP11A1 in vitro*, is unaffected by repeat length.

In summary, significant evidence for the existence of alternatively spliced transcripts for *CYP11A1* and a novel gene have been shown, suggesting a more complex arrangement in this region than previously recognized. Furthermore, this study has demonstrated that the effect of the *CYP11A1* locus on PCOS susceptibility may be more modest than previously anticipated.

ACKNOWLEDGEMENTS

I would first like to thank my supervisors, Mark McCarthy, Steve Franks and Stephanie Halford for their guidance, assistance and knowledge. Thanks to Mark for giving me this opportunity to study in his laboratory. I would like to thank Stephanie for her incredible encouragement and friendship over the years; it means a great deal to me.

Many people made working at Hammersmith Hospital enjoyable. Special thanks to Neda Gharani Tasha Weerasekera, Poorvi Patel, Damian Smedley and Mandy Bennett for making me feel welcome, helping me in the lab and for their continued friendship and kindness. It was an extraordinary team to work with. Thanks also to Caroline Wallace, Helen Figueira, Lisa Lowry and Fiona Maciver for adopting me when the lab moved to Oxford!

I would also like to thank my parents, Peter and Irene, and brother Paul, for their continued support and love. Most of all, I would like to thank my husband, David, for his endless supply of patience, ideas, love, motivation and generosity.

DECLARATION

I hereby declare that all of the work presented in this thesis is my own except the following contributions: all DNA isolations for the association studies were performed elsewhere; all haplotype analysis in the UK datasets were performed by Chris Groves (University of Oxford); and all computer analysis for the NFBC1966 dataset was performed by Ulla Sovio (Imperial College, London).

CONTENTS

Title		1
Dedication		2
Abstract		3
Acknowledgments		5
Declaration		5
Table of Contents		6
List of Figures		13
List of Tables		15
Abbreviations		18
1.0	**INTRODUCTION**	19
1.1	**The Ovary**	20
	1.1.1 Overview of the Ovarian Cycle	21
	1.1.1.1 Follicular Phase	22
	1.1.1.2 Ovulatory Phase	23
	1.1.1.3 Luteal Phase	24
	1.1.2 The Steroidogenic Pathway in the Ovary	24
1.2	**Polycystic Ovary Syndrome**	26
	1.2.1 Overview	27
	1.2.2 History and Diagnosis	29
	1.2.3 Gonadotrophin Abnormalities	29
	1.2.4 Steroidogenesis	29
	1.2.5 Altered Folliculogenesis	31
	1.2.6 Insulin	32
	1.2.7 Treatment	34
	1.2.8 Genetics	35
1.3	**Analysis of Complex Traits**	36
	1.3.1 Overview	36
	1.3.2 Resources and Methodologies for the Analysis of Complex Traits	37

1.3.2.1 Human Genome Project 37
1.3.2.2 Linkage 38
1.3.2.3 Linkage Disequilibrium 38
1.3.2.4 Association Studies 39

1.4 Candidate Genes in PCOS 40
1.4.1 Overview 40
1.4.2 Insulin 41
1.4.3 Calpain-10 43
1.4.4 Follistatin 44
1.4.5 Resistin 46
1.4.6 *CYP17* 46
1.4.7 *CYP19* 47

1.5 *CYP11A1* 48
1.5.1 Function 49
1.5.2 Expression 49
1.5.3 *CYP11A1* and PCOS 51
1.5.4 Linkage and Association 52

1.6 AIMS 55

2.0 SUBJECTS, MATERIALS AND METHODS 56

2.1 Subjects 56
2.1.1 United Kingdom (UK) Cases 56
2.1.2 UK Population Controls 57
2.1.3 Northern Finland Birth Cohort of 1966 (NFBC1966) 58

2.2 Materials 60
2.2.1 Reagents 60
2.2.2. Enzymes 60
2.2.3 Radioisotopes 60
2.2.4 Electrophoresis and Blotting Materials 60

2.2.5 Photograph and Autoradiography 61
2.2.6 Oligonucleotides 61
2.2.7 PCR Reagents 61
2.2.8 DNA Sequencing Materials 61
2.2.9 Genotyping Materials 62
2.2.10 Statistical Programs 62
2.2.11 Bacterial Strains 63
2.2.12 BAC Clone 63
2.2.13 IMAGE Clones 63
2.2.14 Cell Culture Reagents 63

2.3 Buffers, Solutions, Media and Kits 64
2.3.1 General Stock Reagents 64
2.3.2 Preparation of Human Genomic DNA 64
2.3.3 Preparation of Plasmid DNA 65
2.3.4 Gel Electrophoresis 65
2.3.5 Northern Blot Screening 66
2.3.6 Kits Utilised 66

2.4 Methods 67
2.4.1 Human Genomic DNA Extraction from Whole Blood 67
2.4.2 Phenol/Chloroform DNA Extraction 67
2.4.3 Plasmid Miniprep 68
2.4.4 Plasmid Midiprep 69
2.4.5 Large Construct Maxiprep 69
2.4.6 Spectrophotometry 70
2.4.7 Isolation of Total RNA 71
2.4.7.1 Total RNA Isolation from Tissue 71
2.4.7.2 Total RNA Isolation from Cell Lines 71
2.4.7.3 Isolation of Total RNA using TRIZOL 71
2.4.8 Northern Analysis 72
2.4.8.1 Radioisotopic Labelling of DNA Probes 72
2.4.8.2 Probe Hybridisation and Washing of Filters 73
2.4.8.3 Stripping of Filters 73

2.4.8.4 Autoradiography 74
2.4.9 Size Fractionation Gel Electrophoresis 74
2.4.9.1 Agarose Gel Electrophoresis 74
2.4.9.2 RNA Denaturing Gel Electrophoresis 75
2.4.10 Restriction Endonuclease Digestion 75
2.4.11 Polymerase Chain Reaction (PCR) 76
2.4.11.1 Principles 76
2.4.11.2 Standard PCR Parameters 77
2.4.12 PCR Purification 77
2.4.13 Rapid Amplification of cDNA Ends (RACE) 78
2.4.14 Primer Design and Optimisation 79
2.4.14.1 Standard PCR 79
2.4.14.2 RACE PCR 80
2.4.15 Elution of Target PCR Fragments from Agarose Gels 80
2.4.16 Cloning 81
2.4.16.1 Cloning of PCR Fragments into Plasmid Vectors 81
2.4.16.2 Ligations 82
2.4.16.3 Transformations 82
2.4.17 Automated DNA Sequencing 83
2.4.18 Primary Granulosa Cell Culture 84
2.4.19 Cell Line Culture 85
2.4.20 Transient Transfections 86
2.4.21 Genotyping 88
2.4.22 *In Silico* Analysis 89

3.0 GENOMIC REGION of *CYP11A1* 90

3.1 Introduction 90

3.2 *CYP11A1* 90
3.2.1 Bioinformatic Analysis 91
3.2.2 Genomic Structure of *CYP11A1* 91
3.2.3 Identification of New *CYP11A1* Exons 94
3.2.4 Expression Analysis of *CYP11A1* Alternate Transcripts 95

3.2.5 Northern Analysis of *CYP11A1* Alternate Transcripts 103
3.2.6 5' RACE 105
3.3 *CYP11A1* Genomic Region 108
3.3.1 Isolation and Characterisation of *PS-1* 109
3.3.1.1 Expression Analysis of *PS-1* 109
3.3.1.2 Northern Analysis of *PS-1* 110
3.3.1.3 5' and 3' RACE 111
3.3.1.4 *In Silico* BLAST Searches 115

3.4 Discussion 121
CYP11A1 Alternate Transcripts 121
PS-1 124

4.0 ASSOCIATION STUDIES 127

4.1 Introduction 127

4.2 Pentanucleotide and Tetranucleotide Repeat Polymorphism 127
4.2.1 Pentanucleotide Repeat 127
4.2.2 Tetranucleotide Repeat 128

4.3 Linkage Disequilibrium 132

4.4 UK Sample Set Association Study 132
4.4.1 Case-Control Study 132
Pentanucleotide Repeat 133
Tetranucleotide Repeat 133
4.4.2 Family-Based Association Study 136
Pentanucleotide Repeat 137
Tetranucleotide Repeat 137

4.5 The Northern Finland Birth Cohort of 1966 (NFBC1966) 139
Pentanucleotide Repeat 139
Tetranucleotide Repeat 140

4.6 Intermediate Trait Analyses 142
4.6.1 Testosterone 143
4.6.1.1 UK Sample Set 143
Pentanucleotide Repeat 143
Tetranucleotide Repeat 144
4.6.1.2 NFBC1966 Sample Set 145
Pentanucleotide Repeat 145
Tetranucleotide Repeat 146
4.6.2 Other Intermediate Trait Variables 147
BMI 148
Weight 151
Fasting Insulin 151
LH 151

4.7 Discussion 157

5.0 PROMOTER FUNCTION 161

5.1 Introduction 161

5.2 Results 163
5.2.1 Transient Transfections 163
5.2.2 Nucleotide Repeat Constructs 164
5.2.3 Transient Transfection of Pentanucleotide Repeat Constructs 171
5.2.3.1 Human Embryonic Kidney 293T Cell Line 171
5.2.3.2 JEG-3 Cell Line 174

5.3 Discussion 178

6.0 FUTURE WORK 182

6.1 Association Studies 183

6.2 Functional Role of Pentanucleotide and Tetranucleotide Repeats 184

6.3 Novel *CYP11A1* Transcripts and *PS-1* 187

6.4 Conclusions 188

7.0 REFERENCES 189

8.0 PUBLICATIONS 208

LIST OF FIGURES

1.1	Steroidogenesis pathway in the human ovary.	26
1.2	Genomic organisation of the *CYP11A1* gene.	49
1.3	*CYP11A1* promoter region.	51
3.1	Genomic organisation of *CYP11A1* gene and EST clones	92
3.2	Sequence of *CYP11A1* exons 1a, 1b and 2 and primer sequences for RT-PCR RACE.	97
3.3	Photograph of PCR gel for identification of exons 1a and 1b.	98
3.4	Alignment of *CYP11A1* novel exons 1a and 1b with exon 2.	100
3.5	Photograph of PCR gel of transcript I, transcript II and transcript III in ovary and adrenal gland.	101
3.6	Panel of autoradiographs of Clontech Northern Blots.	104
3.7	Schematic diagram of possible ORFs for transcripts I, II and III.	107
3.8	Photograph of RT-PCR gel of *PS-1*, *CYP11A1* and *ACTB* in various steroidogenic tissues.	110
3.9	Autoradiograph of Northern Blot using *PS-1* specific probe.	111
3.10	Schematic diagram showing the location of *PS-1* and *CYP11A.1*	113
3.11	Sequence of *PS-1* 552bp 5' RACE product, the 1558bp 3' RACE product and C18438.	114
3.12	Schematic diagram of overlapping clones of *PS-1*.	119
4.1	Sequence of 5' promoter region of *CYP11A1* and position of PCR primers for pentanucleotide repeat genotyping.	129
4.2	Sequence of intron 1 region of *CYP11A1*, and position of PCR primers for tetranucleotide repeat genotyping.	131

5.1 Sequence of pentanucleotide repeat and zero repeat constructs. 166

5.2 Schematic diagram of pentanucleotide repeat constructs ligated into the promoterless expression vector, pGL3-Basic. 167

5.3 Sequence of tetranucleotide repeat constructs. 169

5.4 Schematic representation of the strategy for producing the zero repeat construct. 170

5.5 Transient transfection of the pentanucleotide repeat construct into the human embryonic 293T cells. 172

5.6 Transient transfection of the pentanucleotide repeat construct into 293T cells treated with cAMP or forskolin. 173

5.7 Transient transfection of the pentanucleotide repeat construct into JEG-3 cells, untreated and treated with cAMP or forskolin. 176

LIST OF TABLES

2.1	Patient data.	59
3.1	*CYP11A1* splice junctions including new exons 1a and 1b.	93
3.2	Primers for sequence of *CYP11A1* IMAGE clones.	95
3.3	PCR primers for analysis of alternate transcripts of *CYP11A1*.	96
3.4	EST clones containing *CYP11A1* exons 1a and 1b.	102
3.5	Primers used for *PS-1* analysis	109
3.6	Primers used for *PS-1* sequence analysis	112
3.7	EST clones found containing exon 1a and exon 1b of *CYP11A1* and *PS-1*.	120
4.1	PCR primer sequence and conditions for pentanucleotide and tetranucleotide repeat for UK and NFBC1966 sample sets.	128
4.2	Allele and genotype frequencies at the pentanucleotide repeat in the UK sample set.	134
4.3	Allele and genotype frequencies at the tetranucleotide repeat in the UK sample set.	135
4.4	Haplotype analysis at the pentanucleotide and tetranucleotide repeat in the UK cases and controls.	136
4.5	TDT analysis at the pentanucleotide repeat in the UK trios.	137
4.6	TDT analysis at the tetranucleotide repeat in the UK trios	138
4.7	Haplotype analysis at the pentanucleotide and tetranucleotide repeat in the UK trios.	138
4.8	Allele and genotype frequencies at the pentanucleotide repeat in the NFBC1966 cases and controls.	140

4.9 Allele and genotype frequencies at the tetranucleotide repeat in the NFBC1966 cases and controls. 141

4.10 Haplotype analysis at the pentanucleotide and tetranucleotide repeat in the NFBC1966 cases and controls. 142

4.11 Pentanucleotide genotype vs. testosterone in the UK cases. 144

4.12 Tetranucleotide genotype vs. testosterone in the UK cases. 145

4.13 Pentanucleotide genotype vs. testosterone in the NFBC1966 cases and controls. 146

4.14 Tetranucleotide genotype vs. testosterone in the NFBC1966 cases and controls. 147

4.15 Pentanucleotide repeat genotypes vs. BMI in the UK cases. 148

4.16 Tetranucleotide repeat genotypes vs. BMI in the UK cases 149

4.17 Pentanucleotide repeat genotypes vs. BMI in the NFBC1966 sample set. 150

4.18 Tetranucleotide repeat genotypes vs. BMI in the NFBC1966 sample set. 150

4.19 Pentanucleotide repeat genotypes vs. WT in the UK cases. 151

4.20 Tetranucleotide repeat genotypes vs. WT in the UK cases. 152

4.21 Pentanucleotide repeat genotypes vs. WT in the NFBC1966 sample set. 152

4.22 Tetranucleotide repeat genotypes vs. WT in the NFBC1966 sample set. 153

4.23 Pentanucleotide repeat genotypes vs. fasting insulin in the UK cases. 153

4.24 Tetranucleotide repeat genotypes vs. fasting insulin in the UK cases. 154

4.25 Pentanucleotide repeat genotypes vs. fasting insulin in the NFBC1966 sample set. 154

4.26 Tetranucleotide repeat genotypes vs. fasting insulin in the NFBC1966 sample set. 155

4.27 Pentanucleotide repeat genotypes vs. LH in the UK cases. 155

4.28 Tetranucleotide repeat genotypes vs. fasting insulin in the UK cases. 156

4.29 Allele frequencies at the pentanucleotide repeat in the UK, Afro-Caribbean and South Asian cases 160

5.1 PCR primers for the pentanucleotide and tetranucleotide repeat constructs. 165

5.2 Primers for sequencing of the pentanucleotide repeat constructs. 165

5.3 Transient transfection of the pCMV-*Renilla* into JEG-3 cells showing an increase in expression of luciferase light units. 177

ABBREVIATIONS

AP-2	Activating protein-2
Apo(a)	Apolipoprotein (a)
BAC	Bacterial artificial chromosome
DEPC	Diethylpyrocarbonate
bp	Base pairs
BMI	Body mass index (kg/m^2)
$\chi 2$	Chi square
cDNA	Complimentary DNA
cM	Centimorgan
cAMP	Cyclic adenosine monophosphate
CRE	cAMP-response element
CVD	Cardiovascular disease
dbEST	Database expressed sequence tag
DHEAS	Dehydroepiandrosterone-sulphate
DNA	Deoxyribonucleic acid
ds	Double stranded
E_2	Oestradiol
EST	Expressed sequence tag
FSH	Follicle stimulating hormone
GnRH	Gonadotropin-releasing hormone
g	Grams
HDL	High density lipoproteins
HGP	Human Genome Project
HGMP-RC	UK Human Genome Mapping Project Resource Centre
IBD	Identity by descent
IGF2	Insulin like growth factor II
IR	Insulin receptor
INSR	Insulin receptor gene
INS	Insulin gene
IPTG	Isopropyl β-D-thiogalactopyranosid
IRS	Insulin receptor substrate
kb	Kilobases

LH	Luteinizing hormone
LDL	Low density lipoproteins
LOD	Logarithmic odds
mm	Millimetre
MAPK	Mitogen-activated protein kinase
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
ncRNA	Non-coding RNA
NPL	Non parametric linkage score
nr	Non-redundant
P	Progesterone
PCOS	Polycystic Ovary Syndrome
PCO	Polycystic ovaries
PCR	Polymerase chain reaction
PI3K	Phosphatidylinositol 3-kinase
ORF	Open reading frame
RNA	Ribonucleic Acid
RT-PCR	Reverse transcription polymerase chain reaction
RF	Resting Follicles
SHBG	Sex hormone binding globulin
SNP	Single nucleotide polymorphism
StAR	Steroidogenic acute regulatory protein
SF-1	Steroidogenic factor -1
TH	Tyrosine hydroxylase
TDT	Transmission Disequilibrium Test
UTR	Untranslated region
VNTR	Variable number tandem repeats
WT	Weight
XGAL	5-bromo-4-chloro-3-indolyl-bD-galactosidase

1.0 INTRODUCTION

Polycystic ovary syndrome (PCOS) is a very common endocrine disorder that affects up to 10% of women of reproductive age. PCOS is characterized by hyperandrogenaemia and is responsible for over 70% of cases of anovulatory infertility (Franks, 1995). The main problems associated with PCOS involve anovulation, hyperandrogenism and a disturbance in insulin secretion and action. PCOS is a familial disorder with a complex genetic basis. Previous studies have shown both linkage and association at the *CYP11A1* locus, the location of the gene responsible for one of the rate-limiting steps in the synthesis of androgens (Gharani *et al.*, 1997). Mutation screening of the 5' promoter region identified a polymorphic pentanucleotide repeat and, in a case control study, found a significant association between this allelic variant and serum testosterone levels in PCOS women when compared with normal controls (Gharani *et al.*, 1997).

The work described in this thesis aims to identify and determine the significance of regulatory elements controlling *CYP11A1* expression in relation to the aetiology of PCOS and to determine the role (if any) of microsatellites in this locus in influencing this syndrome. This introduction will first outline how the normal ovary functions, discuss PCOS and the polycystic ovary, and finally include a detailed discussion of genes, more specifically *CYP11A1*, implicated in the aetiology of PCOS.

1.1 The Ovary

The human ovary contains from birth, in the form of oocytes, all the genetic material that is passed, from mothers, onto future generations. The primary function of the ovary is to generate a mature fertilisable ovum each menstrual cycle and this is facilitated by the contribution of various components of the hypothalamic-pituitary-ovarian axis. In

addition to the production of a mature oocyte at the time of ovulation, pituitary gonadotrophins act in concert to produce the sex steroids oestradiol (E_2) and progesterone (P). These hormones prepare the endometrium for implantation of a fertilised embryo.

1.1.1 Overview of the Ovarian Cycle

The hypothalamic hormone, gonadotrophin releasing hormone (GnRH) is essential for the control of cyclical ovarian function. In response to serotonin, corticotropin-releasing hormones and other neurotransmitters, the hypothalamic neurones secrete GnRH in a pulsatile manner. This in turn, stimulates the anterior pituitary to release luteinising hormone (LH) and follicle stimulating hormone (FSH) to stimulate the ovary to secrete oestradiol and progesterone (Conn and Crowley, Jr., 1991). A feedback loop ensures that after LH and FSH stimulate the ovary, both hormones loop back to modulate the relative amounts of LH and FSH at the pituitary and GnRH at the hypothalamus. Both LH and FSH function to stimulate follicular growth.

At birth, the human ovaries contain one to two million primordial follicles, each of which has the potential to mature, secrete oestrogens, and ovulate. A primordial follicle consists of an oocyte surrounded by a single layer of flattened granulosa cells and the basal lamina. Factors triggering initial recruitment and growth of follicles are not completely known but stem cell factor (Parrott and Skinner, 1999), basic fibroblastic growth factor (Nilsson *et. al*, 2001), growth differentiation factor –9 (Dong *et al.*, 1996) and anti-Mullerian hormone (Durlinger *et al.*, 1999) are thought to be involved. It is only when the follicle has been growing for several months, and has reached the antral stage, that it becomes dependent on gonadotrophins.

The relationship between sex steroids and folliculogenesis remains unclear; however, several studies have now highlighted effects of androgens on early follicular growth. Androgen treatment in intact monkeys have been shown to increase the number of preantral and small antral follicles up to 1 mm in diameter, and this effect is thought to be acting though the androgen receptor (Hillier *et al,* 1997; Weil *et al.,* 1998). An increase in both thecal cell and granulosa cell proliferation has been shown to be promoted by androgens and that this effect on folliculogenesis predominates in small follicles (Vendola *et al.,* 1998). Indeed, androgen receptor gene expression has been shown to be at the highest in preantral and antral follicles and this is positively correlated with granulosa cell proliferation and folllicular growth and negatively correlated with apoptosis and follicular aetresia (Weil *et al.,* 1998). In addition to the effect of androgens on thecal and granulosa cells, testosterone also acts to augment gonadotropin responsiveness by amplifying the actions of FSH in the ovary (Weil *et al.,* 1999). The ovarian cycle can be divided into three recognizable stages: the follicular, ovulatory and luteal phases.

1.1.1.1 Follicular Phase

The follicular phase begins from the first day of menses continuing until ovulation, with the release of a mature follicle. This stage is under the control of the hypothalamic-pituitary axis and lasts approximately 14 days. Androgens are produced by the theca cells of the antral follicles under the influence of LH and are converted to oestrogen by the action of FSH in granulosa cells (which cannot themselves synthesize androgens). Several genes are involved in the production of androgens, the precursors of oestrogen. In particular, enzymes encoded by the genes *CYP11A1* and *CYP17* (17 α- hydroxylase,

17-20 lyase) are expressed in greater amounts as the levels of LH rise and diffuse into the basal lamina of the granulosa cells.

At the beginning of the follicular phase, the ovary contains a pool of small antral follicles at a diameter of 0.2-0.4 mm, one of which typically increased in size in response to the increase in blood FSH levels to approximately 20–25 mm just before ovulation. Usually only one follicle is selected at this point for ovulation (Gougeon, 1996); (McNatty *et al.*, 1983).

Levels of FSH stimulate the granulosa cells of the antral follicle to activate the enzyme cytochrome P450 aromatase that, in turn, functions to convert testosterone from the theca cell to oestradiol in the granulosa cell. These levels of aromatase activity continue to increase such that by the late follicular phase, aromatase activity is hundreds of times higher in the preovulatory follicle than in any other follicle. Oestradiol levels continue to rise which sensitises the pituitary to GnRH and the release of LH (Filicori *et al.*, 2002).

1.1.1.2 Ovulatory Phase

The surge of LH from the pituitary marks the beginning of the ovulatory phase. The rise in LH results in the final growth and maturation of the follicle, ovulation and formation of the corpus luteum. Ten to twelve hours following the surge of LH, ovulation occurs with the rupture of the follicular wall, and the release of the oocyte (Baird *et al.*, 1975).

1.1.1.3 Luteal Phase

The LH surge stimulates luteinization of the granulosa cells and stimulates the synthesis of progesterone. The luteinised granulosa cells combine with the newly formed lutein-theca cells and surrounding stroma to form the corpus luteum. The corpus luteum is an endocrine gland and functions to secrete both oestrogen and progesterone (Findlay, 1986). Under the control of LH, levels of both hormones increase, and act by a negative feedback mechanism to suppress the expression of gonadotrophins, thereby preventing more follicles from developing until the next cycle (Benyo *et al.*, 1993). If pregnancy does not occur, the corpus luteum regresses, levels of progesterone and oestrogen fall leading to a rise in FSH and the start of the next ovarian cycle.

1.1.2 The Steroidogenic Pathway in the Ovary

Steroid hormones have a multitude of roles in the human body including the induction of male and female secondary sexual characteristics (androgens and oestrogens respectively) and endometrial receptivity (progestins). Central to this production of androgens, oestrogens and progestins is the utilisation of cholesterol, which is the starting point of the steroidogenic cascade. The early steps of steroid synthesis involve either the synthesis of cholesterol *de novo* from acetate or from either high-density lipoproteins (HDL) or low-density lipoproteins (LDL) (Miller, 1988). In humans, cholesterol enters the mitochondria of the thecal cell, the site of synthesis of many steroidogenic enzyme synthases, aided by the activity of the steroidogenic acute regulatory protein (StAR). In the thecal cell of the ovary, several enzymes function in a sequential manner in the biosynthesis of steroid hormones.

The *CYP11A1* gene encodes the P450 side chain cleavage enzyme (P450scc) and this enzyme is the first rate-limiting enzyme in the steroidogenesis pathway. P450scc functions to convert cholesterol to pregnenolone by three distinct chemical reactions; 22 hydroxylation, 20 hydroxylation and $C_{20,22}$ bond cleavage. It is this side chain cleavage of cholesterol that is the first rate-limiting step in steroidogenesis (Miller, 1988). Through a series of chemical reactions, pregnenolone is converted to testosterone, catalysed by the enzymes, P450C17, 3β hydroxysteroid dehydrogenase (3βHSD) and 17β hydroxysteroid dehydrogenase (17βHSD). P450scc is present in theca cells in small antral follicles, whereas levels in granulosa cells from these follicles are low but increase during preovulatory follicle development, induced by FSH. Levels of P450scc are at the highest after luteinization, when progesterone production is also high (Miller, 1988). Ovarian granulosa cells are where androstenedione and testosterone (after crossing the basal lamina from the theca cell into the granulosa cell) are converted to oestrogen/oestradiol respectively. This is through the action of *CYP19*, which encodes the enzyme aromatase. Figure 1.1 illustrates the androgen synthesis pathway that occurs in the ovary.

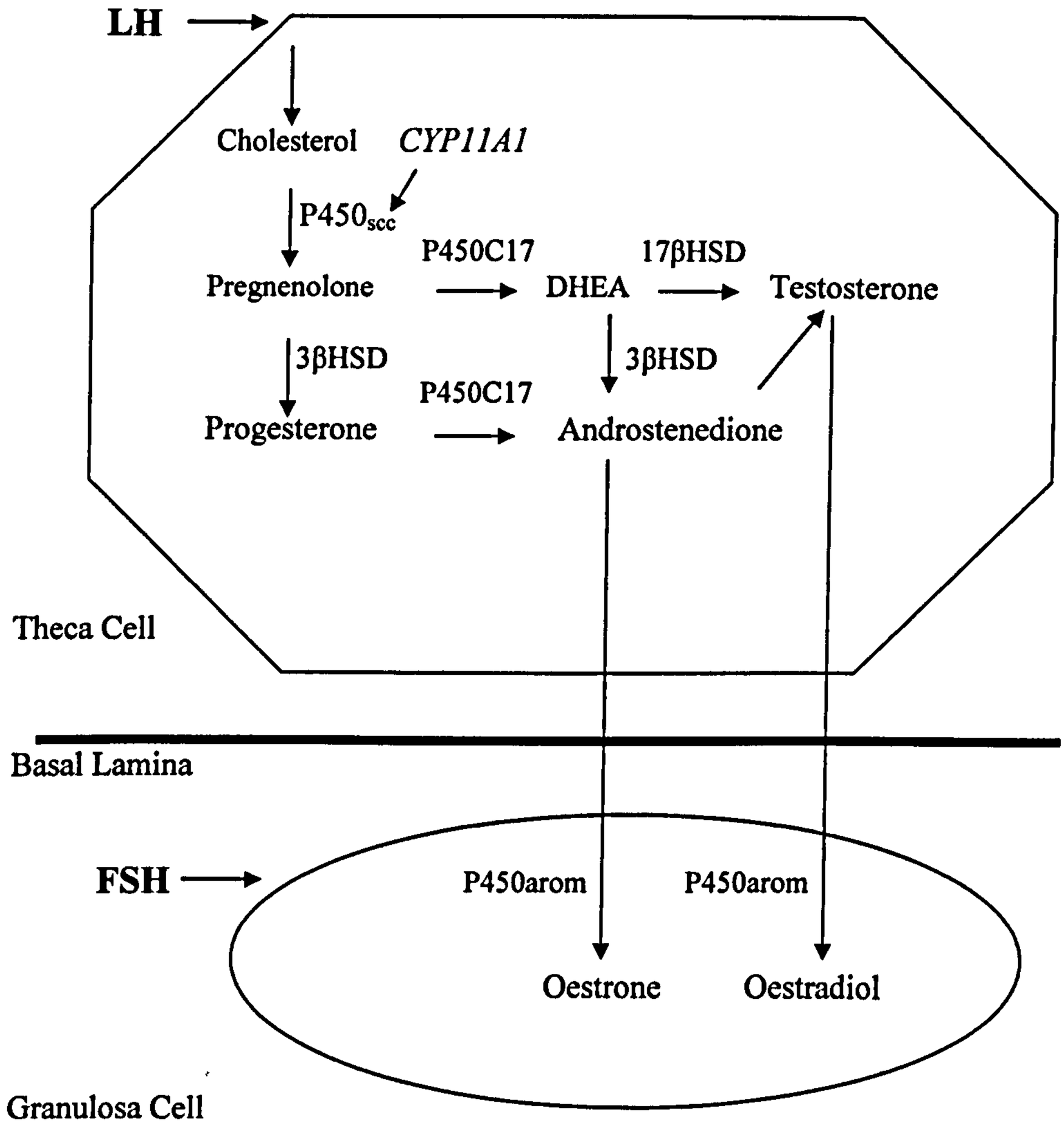


Figure 1.1: Steroidogenesis Pathway in the Human Ovary. DHEA: Dehydroepiandrosterone sulphate. See text for details.

1.2. Polycystic Ovary Syndrome

1.2.1. Overview

PCOS is characterized by hyperandrogenaemia, anovulation and polycystic ovaries and is a common cause of anovulatory infertility (Franks, 1995); (Rotterdam, 2004). It is estimated that the prevalence of PCOS in women of reproductive age is between 4%

and 12% (Knochenhauer *et al.*, 1998). The aetiology of PCOS is unknown; clinical and biochemical abnormalities may arise from dysregulation of a number of mechanisms including abnormalities of steroid hormone production, of gonadotrophin secretion, and insulin signalling or secretion. There is a strong familial aggregation of this disorder, with approximately 50% risk to siblings of affected patients, which supports the likelihood of a major genetic component underlying this disorder (Givens, 1988). Clinical features associated with PCOS include menstrual irregularities, hirsutism, obesity, hyperinsulinaemia, and peripheral insulin resistance. Studies have shown that there is an increased risk of type 2 diabetes (Franks, 1995) and cardiovascular events in women with PCOS (Christian *et al.*, 2003). In addition, there is evidence suggesting that the male phenotype manifests as male pattern baldness (Carey *et al.*, 1993), although this remains controversial (Legro, 2003). Elucidation of the cause(s) of PCOS has remained elusive due to the heterogeneity of this disorder.

1.2.2 History and Diagnosis

Early descriptions of the syndrome date back to 1721 and 1844 but the first systematically documented cases of PCOS came in 1935, with observations made by Stein and Leventhal and presented in their paper at the Central Association of Obstetricians and Gynaecologists (Stein and Leventhal, 1935). In this paper, they described 7 women with amenorrhoea, hirsutism, and enlarged ovaries. Diagnostic criteria at this time relied on detecting enlarged sclerocystic ovaries at surgery and by histological confirmation, in women who were either anovulatory or hirsute or both (Stein and Leventhal, 1935). Treatment usually meant wedge resection of the ovaries, which often resulted in the return of menstruation. It soon became apparent however, that certain biochemical abnormalities were often associated with polycystic ovaries. In

1957, abnormalities in LH secretion were detected in women with bilateral polycystic ovaries (McArthur *et al.*, 1958) and became the benchmark for diagnosis along with elevated levels of testosterone (Yen *et al.*, 1970); (Rebar *et al.*, 1976). These changes were often associated with normal to low levels of FSH and, together with abnormal LH secretion, became the typical endocrine profile associated with PCOS (Franks, 1995).

Because of the broad spectrum of clinical and biochemical presentation, difficulties have arisen in defining diagnostic criteria for PCOS. It has been reported that as many as 21-23% of young women have the characteristic polycystic ovarian morphology (on ultrasound) with only a proportion having clear endocrine abnormalities (Conway *et al.*, 1989). In addition, some women with PCOS have normal reproductive hormone levels and ovulatory menstrual cycles. However, 80-90% of women with oligomenorrhea have PCOS. It is clear that the more severe the biochemical disturbance, the more likely the woman is to be symptomatic (Prelevic, 1997). Until recently, the most widely accepted definition was that proposed by the 1990 National Institutes of Health Workshop on PCOS (which is the association of oligo-or anovulation with clinical and/or biochemical evidence of androgen excess) (Zawadski and Dunaif, 1992). However, at the recent ASRM/ESHRE meeting in Rotterdam (May 1-3, 2003), a revised definition for PCOS was decided. The revised diagnosis of PCOS requires the presence of two (or more) of three features: oligo or anovulation, signs of androgen excess and polycystic ovaries. In addition, other disorders such as congenital adrenal hyperplasia, Cushing's syndrome and androgen secreting tumours which may present with a similar clinical appearance should be excluded (Rotterdam, 2004); (Yen, 1980).

1.2.3 Gonadotrophin Abnormalities

One of the first biochemical defects noted in women with PCOS was disordered gonadotrophin secretion causing a rise in LH concentrations, with normal to low levels of FSH (Baird *et al.*, 1977). Gonadotrophin releasing hormone, synthesized by hypothalamic neurones, stimulates the production of both of these hormones through a frequent pulsatile release. LH acts on the thecal cell of the ovary to stimulate ovarian androgen production. It was originally thought that the abnormality of gonadotrophins in PCOS arose as a result of dysregulation of GnRH secretion leading to prolonged hypersecretion of LH, and increased steroidogenesis (Marshall and Eagleson, 1999). In addition, women with PCOS seemed to have a decreased sensitivity to feedback effects of gonadal steroids on GnRH release. However, these features are also found in other disorders of androgen secretion such as androgen-secreting ovarian neoplasms and nonclassical 21-hydroxylase deficiency, which suggests that instead of being the primary cause, they occur secondarily in response to the hormonal environment and/or ovarian dysfunction (Fauser *et al.*, 1992). Elevated serum LH levels can be found in both ovulatory and anovulatory women with polycystic ovaries. Again, this is not a consistent finding as some women with both clinical and biochemical features of PCOS have normal serum LH levels.

1.2.4 Steroidogenesis

Characteristically, as described above, women with PCOS have elevated LH levels as well as high serum testosterone. While levels of LH are highly variable, testosterone concentrations are more consistently raised (Gilling-Smith *et al.*, 1997). Serum concentrations of testosterone and androstenedione, elevated in women with PCOS, are

on average 50% higher than in control women, regardless of ovulatory status (Adams *et al.*, 1986); (Franks, 1995). There has been some debate as to whether the ovary is the principal source of excess androgens seen in PCOS. Studies have shown that in some women with PCOS, there is increased secretion of adrenal androgens and hyper-responsiveness to adrenocorticotrophic hormone (ACTH) (Rodin *et al.*, 1994). Nevertheless, the weight of evidence suggests that the ovarian contribution to hyperandrogenism is more important. Gilling-Smith and co-workers cultured theca cells from five normal ovaries and compared these with nine ovaries from both anovulatory and ovulatory women under both basal and LH-stimulated conditions (Gilling-Smith *et al.*, 1994). Interestingly, they were able to show a significant (20-fold) increase in both basal and LH stimulated production of androstenedione from the thecal cells of the polycystic ovaries over the normal thecal cells (Gilling-Smith *et al.*, 1994). Subsequently, the notion that there is an intrinsic alteration in steroidogenesis in theca cells from polycystic ovaries was further supported by observations that the mRNAs transcribed from the steroid genes *CYP11A1* and *CYP17*, along with associated enzyme activities, were higher in polycystic ovary theca cells compared to normal theca cells (Nelson *et al.*, 1999). It has been reported that over-expression of these genes is not due to an increase in theca cell population but due to the hyperstimulation of the theca cells in polycystic ovaries compared with normal (Jakimiuk *et al.*, 2001).

Microarray analysis comparing the gene expression profile of normal and PCOS theca cells revealed an increased abundance of mRNA of the transcription factor, *GATA-6* (Wood *et al.*, 2003). In addition, GATA-6 protein expression has also been shown to be elevated in the PCO theca cell. Previous work in the porcine ovary has shown that GATA-6 regulates *StAR* expression, which suggested a role in steroidogenesis (Gillio-Meina *et al.*, 2003). *CYP17* and *CYP11A1* transcription, but not *StAR* transcription,

were increased when the theca cells were transiently transfected with *GATA-6* (Wood *et al.*, 2003). The authors concluded that the increased expression of *GATA-6* in PCOS theca cells may be responsible for the increased steroidogenesis in the PCOS theca cell directly through the activation of *CYP11A1* and *CYP17* genes (Wood *et al.*, 2003).

Symptoms arising from high levels of these androgens include hirsutism and acne. There is great variation between women in severity of hirsutism and acne. For example, some non-hirsute anovulatory women have hyperandrogenaemia similar to that of hirsute women; whereas some hirsute women with polycystic ovaries can have normal serum androgen levels. While the former is probably due to variable sensitivity of hair follicles to androgens, the latter is probably due to an increased rates of androgen production and clearance by peripheral tissues (Barth, 1988).

Sex hormone binding globulin (SHBG) is a plasma glycoprotein, produced by the liver, which binds and regulates the biological availability of testosterone and oestradiol. Thus expression of androgen action can be affected by changes in circulating concentrations of SHBG. Diet is an important regulator of SHBG; obesity suppresses SHBG production by the liver and this effect is mediated by insulin (Plymate *et al.*, 1988). Conversely, caloric restriction increases SHBG concentrations and therefore lowers free, biologically available, testosterone levels (Franks *et al.*, 1991).

1.2.5 Altered Folliculogenesis

One of the main phenotypes in PCOS women is anovulation, which is a common cause of infertility. Early evidence demonstrated that the polycystic ovary contained twice the normal number of growing follicles relative to the normal ovary, suggestive of primary

ovarian dysfunction (Hughesdon, 1982). It is well established that growth of these antral follicles in polycystic ovaries is arrested at 5-8 mm in diameter; therefore, early on in follicular development. It has been hypothesised that the accumulation of this large number of antral follicles would lead to an increased production of androgens from the theca cells and therefore, hyperandrogenaemia (Hughesdon, 1982).

In a study of 32 women with polycystic ovaries (16 women with regular ovulatory cycles and 16 with oligomenorrhea) and 24 women with normal ovaries, the mean densities of these follicles were 6 fold greater in the anovulatory women with PCOS than in the normal ovaries ($p=0.009$). There was no difference between anovulatory and ovulatory polycystic ovaries ($p=0.38$) or between ovulatory polycystic and normal ovaries ($p=0.4$). Furthermore, a significant increase in early growing primordial follicles from polycystic ovaries compared with normal was detected (Webber *et al.*, 2003). The exact mechanism for this altered folliculogenesis is unknown. The authors hypothesized that these differences may lie in excessive early foetal follicular growth or by a decreased rate of atresia during late foetal life, childhood and or adulthood (Webber *et al.*, 2003).

1.2.6 Insulin

Insulin is a hormone that has both mitogenic properties (control of cell growth and differentiation) and metabolic effects (promoting synthesis and storage of carbohydrates, lipids and proteins). Insulin signals through the insulin receptor (IR), a member of a family of tyrosine kinase receptors. Binding of insulin to its receptor stimulates receptor autophosphorylation of specific tyrosine phosphates on one of the two β subunits (the receptor has 2 β and 2 α subunits that form a heterodimer) (Saltiel,

1996) and the signal is transmitted downstream via two pathways: the phosphatidylinositol 3-kinase (PI3K) and the Ras-mitogen-activated protein (MAP) kinase pathways (Saltiel, 1996); (White and Yenush, 1998).

It has been estimated that 50-60% of PCOS women are insulin resistant compared with just 10-25% in the normal population (Dunaif *et al.*, 1989); (Ferrannini *et al.*, 1997). Women with PCOS have an increased risk of developing type 2 diabetes mellitus (Legro *et al.*, 1999). However, in contrast to insulin resistance seen in type 2 diabetes, there is said to be normal hepatic insulin sensitivity in women with PCOS (Peiris *et al.*, 1989). Insulin resistance in peripheral tissues leads to compensatory hyperinsulinaemia, which is thought to contribute to hyperandrogenaemia. *In vitro* studies have documented that insulin may enhance thecal androgen synthesis (Barbieri *et al.*, 1986). Weight loss, which also leads to a reduction in insulin, results in a lowering of androgen levels (Nestler *et al.*, 1989). However, in both obese and non-obese women with PCOS, insulin secretion is inappropriately low for the degree of insulin resistance which also indicates a level of pancreatic β- cell dysfunction (Dunaif *et al.*, 1989).

In spite of the peripheral insulin resistance seen in PCOS, paradoxically the ovary seems to remain responsive to the actions of insulin. Moreover, excessive insulin action appears to contribute to ovarian hyperandrogenism and anovulation in PCOS (Dunaif, 1997). However, the mechanism for the preservation of insulin action of steroidogenesis in the ovary concurrent with insulin resistance in skeletal muscle remains unclear. Several studies of PCOS have shown that the insulin resistance is more likely to occur at the level of phosphatidylinositol 3-kinase (PI3K) signalling in muscle whereas the mitogen-activated protein (MAP) kinase of the ovary remains unaffected (Ek *et al.*, 1997) (Ek *et al.*, 2002); (Ciaraldi *et al.*, 1998). This finding has

been supported by a study in which it was shown that defects in insulin signalling in skeletal muscle taken from type 2 diabetes mellitus patients were predominately on the PI3K activity whereas the MAP kinase pathway was unaffected (Cusi *et al.*, 2000).

1.2.7 Treatment

Women diagnosed with PCOS are often seen in the clinic presenting with infertility and/or hirsutism. Due to the complexity of this syndrome, and the fact that the exact cause underlying this disorder is unknown, management usually focus on treating the symptoms directly, rather than a global treatment. There are several treatments for menstrual disturbances. Oral contraceptives remain a typical choice of treatment for irregular menstrual cycles and hirsutism (Morin-Papunen *et al.*, 2000). For women who are also obese (BMI of 26 or over), behavioural modification (in terms of weight loss through diet and exercise) is commonly recommended first. It has been well documented that a reduction in weight can improve hormonal disturbances resulting in ovulation in overweight women with PCOS (Kiddy *et al.*, 1989).

In women with PCOS who have anovulatory infertility, the anti-oestrogen clomiphene citrate is often used as a first choice of treatment for induction of ovulation. This oral oestrogen antagonist functions to raise circulating concentrations of FSH and to induce follicular growth in most women. Of subjects treated this way, 75-80% will ovulate and there is a high resultant pregnancy rate (Hull, 1992); (Franks *et al.*, 1985). There is however, a subset of patients who are resistant to clomiphene citrate, and are therefore prescribed exogenous gonadotrophin preparations. These have been used for more than 30 years with variable success. Using a conventional step-up dosage regimen results in an overall pregnancy rate of less than 30%, and is associated with significant adverse

effects such as a high incidence of multiple pregnancies (25-30%) and ovarian hyperstimulation (Lunenfeld and Insler, 1978); (Wang and Gemzell, 1980). This has led to the development of a chronic low-dose gonadotrophin regimen that allows the induction of a single follicle at ovulation and thus reduces the rate of multiple pregnancies (<7%) and hyperstimulation syndrome while achieving conception rates of up to 45% (Brown, 1978); (Polson *et al.*, 1987); (White *et al.*, 1996). Pulsatile administration of gonadotrophin-releasing hormone is another (less successful) method that may be used to induce unifollicular ovulation in women with PCOS (Jacobs, 1994).

Insulin sensitising agents such as metformin and thiazolidinediones have been used with varying success in treatment of hyperinsulinaemia, insulin resistance and hyperandrogenaemia. Several studies have looked at the effects of these treatments on women with hirsutism, hyperandrogenaemia, oligomenorrhea, polycystic ovaries and impaired glucose tolerance (Nestler *et al.*, 1989); (Dunaif *et al.*, 1996). Subjects were given an oral glucose tolerance test before and after treatment with either metformin or troglitazone. Most studies reported a drop in plasma glucose levels and improved insulin sensitivity in women taking the insulin-sensitising agents (Diamanti-Kandarakis *et al.*, 1998); (Ehrmann *et al.*, 1997b) with one study reporting no effect (Ehrmann *et al.*, 1997a). Some women resumed normal menstruation and pregnancy after administration of either metformin or troglitazone (Diamanti-Kandarakis *et al.*, 1998).

1.2.8 Genetics

There is plentiful evidence for an underlying genetic basis of PCOS. It has been shown in a study of first-degree relatives that up to 46% of ascertainable sisters of women with PCOS have hyperandrogenaemia (Legro *et al.*, 1998). This is consistent with a major

genetic component contributing to its aetiology. It has been shown that insulin resistance also clusters in families (Legro *et al.*, 2002). In this study, 307 sisters of PCOS probands (the sisters all with PCOS and hyperandrogenaemia) were shown to have high levels of insulin and decreased levels of glucose/insulin ratios suggestive of insulin resistance. Unaffected sisters of the PCOS probands were recorded to have normal insulin action (Legro *et al.*, 2002).

1.3 Analysis of Complex Traits

1.3.1 Overview

The hunt for genes responsible for simple Mendelian traits has been hugely successful due to the advancement and refinement of genetic tools such as sequencing and high-throughput genotyping. Using such tools, the identification of DNA variants such as microsatelites, variable number of tandem repeats (VNTRs), insertion/deletion polymorphisms and single nucleotide polymorphisms (SNP) have been shown to directly cause a number of diseases such as Huntington's disease and cystic fibrosis. However, unlike monogenic traits in which the presence or absence of a disease causing allele usually guarantees the disease will be present or absent, in complex traits, the presence of a susceptibility allele may only contribute mildly to the susceptibility of disease. The difficulty arises from the fact that each factor (allele, environment) contributes, at most, a modest amount to the total variation in the trait observed in the entire population. In addition, multiple genetic and environmental factors may interact with each other in unpredictable and non-linear ways; there is rarely a 1:1 relationship between genotype and phenotype in complex traits.

The ascertainment of susceptibility alleles that predispose individuals to disease has been difficult. There has been an overwhelming inconsistency in the literature with respect to replication of previously identified associations. It is well accepted that there are several reasons for this including poor study design, incorrect assumptions about the underlying genetic structure and over-interpretation of the data (Ioannidis *et al.*, 2001); (Lohmueller *et al.*, 2003); (Colhoun *et al.*, 2003).

1.3.2 Resources and Methodologies for the Analysis of Complex Traits

1.3.2.1 Human Genome Project

The Human Genome Project (HGP) has provided scientists with the basic knowledge about our genome, along with that of many other nonhuman genomes. Started in 1990, goals of the project were to sequence all the 3 billion bases that make up the genome, to identify all of the genes in the human genome, now estimated to be approximately 25 000, and to devise genetic tools with which to analyse this information with hopes of contributing to diagnosis, treatment and prevention of the diseases that affect us.

Furthermore, the HGP has since identified approximately 10 million SNPs, amongst which are many good strong candidates for disease susceptibility for the more complex disorders. Genetic studies can be divided into whole genome approaches or candidate gene approaches. Both methods then use linkage and association studies as tools to aid in the discovery of causative genes.

1.3.2.2 Linkage

The genetic distance that separates two loci on a chromosome is measured in terms of the probability a resultant gamete contains a combination of paternal and maternal gamete haplotypes, i.e. is a recombinant. This is called the recombination fraction (θ). Two loci will obey Mendel's law of independent assortment if they are located on different chromosomes or far apart on the same chromosome. This means 50% of the offspring will be recombinant (contain a mixture of both paternal and maternal haplotypes) and 50% will be non-recombinant for the parental gamete haplotypes and the recombination fraction will be 0.5. However, if the two alleles are sufficiently close together on the chromosome (and are therefore inherited together), the recombination fraction will be <0.5. This provides the basis of linkage analysis. The unit of genetic distance is the Morgan (or the centimorgan, cM) where 1cM corresponds to 1 recombination per 100 meiosis or 1% recombination fraction.

Linkage analysis can therefore be used to establish the co-inheritance of chromosomal regions with a specific disease or trait of interest. There are two approaches to testing for linkage, parametric and nonparametric analysis. Parametric analysis relies on fitting the inheritance pattern on a specific model for the disease-causing trait whereas nonparametric analysis compares the inheritance pattern with what would be expected under independent assortment.

1.3.2.3 Linkage Disequilibrium

Linkage disequilibrium is the non-random association between a particular allele, such as a disease allele, with a specific allele at a nearby locus more often than expected by chance. LD between two alleles can be generated in finite populations by founder

effects, random drift, mutation, selection and non-random mating. The measurement of LD most commonly used is the correlation coefficient Δ (Hill and Weir, 1994). The correlation coefficient has a maximum value of +1 and a minimum value of –1, where a positive Δ (coupling) indicates association between common alleles of the two loci and a negative value (repulsion) indicates association between the common allele at one locus and the rare allele at the other. Absence of LD is indicated by Δ=0.

1.3.2.4 Association Studies

An association study aims to determine if a genetic variant is correlated with a disease phenotype. There are two broad strategies for association studies- a case-control study design and a family based study design. The case control study design has been widely implemented in the analysis of complex traits. Here, the frequencies of alleles/genotypes are compared between two groups of individuals: (a) a defined set of cases diagnosed with the disease; and (b) controls, a set of individuals not known to be affected, or randomly selected from the population. Any increase in either the allele or genotype frequency of the particular polymorphism studied in cases versus controls implies that presence of this polymorphism may increase the risk of disease.

One of the impediments with a case-control association study is the selection of an appropriate control population. One of the dangers of inappropriate selection is population stratification in which allelic differences between cases and controls are due to differences in genetic ancestry and not due to a true association with disease. If one of these groups has a higher disease prevalence than the others, stratification occurs, because that group will be over-represented in the cases and under-represented in the

controls. Any polymorphism associated with ethnic origin may then appear to be associated with disease, giving a false positive result.

The second methodology for association studies is the family-based approach, which involves using individuals from family members of affected probands as controls. This methodology was developed to circumvent the issue of population stratification. The TDT (transmission disequilibrium test) is an example of one such test of the association between sequence variation and disease (Clayton, 1999). The aim of the TDT is to detect instances of excess transmission of particular alleles from heterozygous parents to affected offspring. The alleles not transmitted are effectively being used as controls for the study.

One intrinsic feature of TDT analysis is that due to the usage of only heterozygous parents, some genotype information is lost. It follows that in order to obtain sufficient power to detect an association, a greater number of patients and families are needed. Therefore, this type of study may be difficult to implement due to the requirement for parental genotypes, which may not be available in late onset disorders.

1.4 Candidate Genes in PCOS

1.4.1 Overview

The difficulties in ascertaining the underlying genetic cause of PCOS are numerous. Due to the fact that PCOS strikes women of reproductive age, it is difficult to obtain family information spanning more than one generation for segregation or linkage studies. In addition, there is no accepted male phenotype. Further, the lack of

universally accepted diagnostic criteria for the syndrome makes comparisons between different groups difficult. However, by using both linkage and association data to identify susceptibility polymorphisms in genes for complex disorders, a growing list of candidate genes have now been identified in PCOS. Some of these studies are briefly examined below to give a general overview of genes studied to date and to highlight some of the problems connected with studies such as these.

1.4.2 Insulin

Insulin is a pancreatic hormone, secreted by the β-cells, which when bound to its receptor, results in a series of molecular events that promote glucose metabolism (see section 1.2.6) (Le Roith and Zick, 2001); (Burks and White, 2001). The development of all forms of diabetes are due in part to the lack of insulin production or secretion.

Examination of the insulin gene in relation to PCOS has focussed on the insulin VNTR, a polymorphic region consisting of a 14-15 bp unit consensus sequence with anywhere from 30 to several hundred repeats and located 500 bps upstream of the insulin gene (Bell *et al.*, 1982).

Depending on the number of repeats, the *INS*-VNTR has been clustered into class I repeats (average 40 repeats) and class III repeats (average 157 repeats) and class II repeats (average 100 repeats) (Bell *et al.*, 1982). Both the insulin gene and the VNTR are situated between the genes for tyrosine hydroxylase (*TH*) and the insulin like growth factor II (*IGF2*). Several studies have implicated the VNTR in susceptibility to type 2 diabetes (Huxtable *et al.*, 2000), type 1 diabetes (Bennett and Todd, 1996), and obesity (Le Stunff C *et al.*, 2000); (Le Stunff C *et al.*, 2001). It has been hypothesized that the VNTR might confer this susceptibility through changes in the regulation of transcription

of the insulin gene (Bennett and Todd, 1996); (Vafiadis *et al.*, 1997); (Huxtable *et al.*, 2000). It is therefore a strong biological candidate potentially contributing to the aetiology of PCOS due to the presence of abnormal insulin secretion, action or production in these women, similar to that seen in type 2 diabetes patients. Waterworth and colleagues found evidence for linkage to the insulin VNTR in 17 families (each of which had several affected members) through non-parametric and parametric analysis (Waterworth *et al.*, 1997). The highest linkage score was observed at marker D11S922, which is located 1.5 centimorgans (cM) distal to the insulin gene. Furthermore, it was reported that approximately 60% of the families were linked to this locus (maximum LOD score= 2.27, α= 0.63). Association was observed between the class III homozygote genotype and anovulatory PCO using a case control study involving 25 anovulatory PCOS cases, 34 ovulatory PCOS cases and 54 normal controls. This association was substantiated in a third dataset comprising of 52 PCOS cases for which the parents were available (Waterworth *et al.*, 1997). These results showed a significant transmission from heterozygous fathers to daughters with PCOS (χ^2= 7.54, *P*= 0.006). Linkage disequilibrium mapping showed that the region of association corresponded to class III VNTR homozygotes and an *Hph1* site, located at-23bp with respect to transcription initiation site. This study, providing linkage to the region and verification of the association results in several populations provided strong support for the insulin VNTR as a susceptibility locus for PCOS (Waterworth *et al.*, 1997). The authors hypothesized that the class III alleles functioned to alter insulin secretion causing insulin resistance and hyperinsulinaemia. One subsequent study performed by Urbanek and colleagues failed to find any significant IBD (identity by descent) in this region (Urbanek *et al.*, 1999). Furthermore, no association was found between class III alleles and PCOS or hyperandrogenaemia, possibly due to low sample size numbers (Urbanek *et al.*, 1999). This was again substantiated in a more recent

study of 255 parent-offspring trios, 185 separate cases and 1062 controls of British/Irish origin and 1599 cases and controls from the NFBC1966 (see section 2.1.3). In this study, no allelic association was detected between the previously reported class III alleles and PCOS (UK population, P=0.4; NFBC1966 P=0.3) (Powell *et al.*, 2005). Moreover, analysis of the parent-offspring trios revealed no excess of transmission of either allele (P=0.6) regardless of parent of origin (maternal, P=0.7; paternal, P=0.7) (Powell *et al.*, 2005).

1.4.3 Calpain-10

In a genome wide scan conducted in Mexican-Americans who have a high incidence of type 2 diabetes, the region of chromosome 2q37 was found to be significant for linkage to type 2 diabetes (Hanis *et al.*, 1996). This chromosomal region contained the gene coding for a cysteine protease, Calpain-10 (*CAPN10*). Several variants within this gene were subsequently identified, with one variant within this gene being strongly associated with type 2 diabetes (Horikawa *et al.*, 2000).

Due to the fact that both type 2 diabetes and PCOS share common metabolic abnormalities, genes found to be associated with one might contribute to the aetiology of the other. Therefore, *CAPN10* has been studied in both a case control and family based association study (Haddad *et al.*, 2002). Four variants analysed in the previous diabetes studies were examined (SNP-43, SNP-44, SNP-19 and SNP-63). Analysis of the 146 parent-offspring trios revealed evidence of over-transmission of the common allele, SNP-63; however, this was not substantiated in the case control study involving 185 PCOS cases and 525 population controls at both the single locus and haplotype

level. Therefore, the authors concluded that variants in *CAPN10*, while showing association in type-2 diabetes, were not associated with PCOS (Haddad *et al.*, 2002). In another study which looked at three SNPs in *CAPN10* (SNP-43, -19, -63) alone or in haplotype combination in a set of 124 nondiabetic white probands of European ancestry, no association was found (Ehrmann *et al.*, 2002). In addition, no association for each individual SNP was found in a set of 57 nondiabetic African Americans. However, one haplotype combination, 112/121 in the African American group, was shown to be associated with higher insulin levels in response to an oral glucose challenge. Furthermore, this haplotype combination was shown to be associated with a 2-fold risk increase of PCOS in both African Americans and whites (Ehrmann *et al.*, 2002).

A third study of the association of *CAPN10* and PCOS in a Spanish population of 148 women with PCOS showed an association between the SNP-44 allele and polycystic ovaries (Gonzalez *et al.*, 2003). Associations between PCOS phenotypes such as hyperandrogenaemia and hypercholesterolaemia and SNP-44 were obtained in this study (P=0.05 and P=0.005 respectively with haplotype 1121) (Gonzalez *et al.*, 2003). Such a variation in results between different studies highlights many of the issues surrounding analysis of complex traits such as PCOS. Furthermore, two of the three studies here consisted of small sample sizes and this factor has been described as being a predictor of unreliable association studies (Ioannidis *et al.*, 2001).

1.4.4 Follistatin

Follistatin is a logical biological candidate gene as its role is to bind to activin, thereby inhibiting its function. Activin is a protein that promotes ovarian follicular development, inhibits androgen production by the theca cell, increases follicle-

stimulating hormone secretion by the pituitary, and increases insulin secretion by the pancreatic β-cells (Guo Q. *et al.*, 1998); (Mather *et al.*, 1997). Therefore, the inhibition of activin by follistatin might be expected to lead to a reduction in FSH levels and arrested follicular development. This inhibition of activin has been shown in a study of transgenic mice overexpressing follistatin (Guo Q. *et al.*, 1998). In principle, this inhibition would leave androgen synthesis unchecked, driven by LH. Follistatin is expressed in many tissues including the ovary, pituitary, adrenal cortex, and pancreas. A study published by Urbanek and colleagues showed linkage to the follistatin gene locus, in a small set of affected sisters (39 sibpairs studied; 72%, χ^2=12.97; nominal $P= 3.2 \times 10^{-4}$) (Urbanek *et al.*, 1999). Linkage was performed using the affected sib pair (ASP) method, which tests for increased marker similarity in affected siblings. After correction for multiple testing, the results were still significant at P=0.01 (Urbanek *et al.*, 1999).

However, in a follow-up study Urbanek and colleagues sequenced the follistatin gene in 85 members of 19 families to identify variants contributing to the linkage described above (Urbanek *et al.*, 2000). Out of the 17 variants identified from sequencing, only 1 variant (a single base pair change in the last exon that is not translated) was common enough to possibly make a major contribution to susceptibility. When this variant was tested for linkage and association with PCOS in a sample set of 249 families, the findings were not significant. In addition, mRNA levels of follistatin between cultured fibroblasts from women with PCOS and cultured fibroblasts from normal control women were not significantly different. The authors therefore concluded that any contribution made by follistatin in the aetiology of PCOS must be, at best, modest (Urbanek *et al.*, 2000).

1.4.5 Resistin

Previous studies have identified a region on chromosome 19p13.3 that showed strong evidence for linkage and association to PCOS (Urbanek *et al.*, 2003). In this study, a 14-Mb region flanking the insulin receptor gene (*INSR*) was shown to have an excess of IBD in affected sisters (63% IBD, χ^2= 8.79, *P*=0.003 at marker D19S884). Further examination using TDT found an association between allele 8 of this marker and PCOS (Urbanek *et al.*, 2003). The authors hypothesized that since the marker was ~900 kb centromeric to the INSR gene, the receptor itself could not possibly contain the variant responsible for the linkage and association. Further analysis of the region identified the gene for the hormone, resistin, located 400 kb downstream of *INSR*. The resistin protein is thought to regulate insulin action and glucose tolerance. Studies conducted in mice have shown that resistin impairs glucose tolerance and inhibition of resistin action improves insulin sensitivity (Steppan *et al.*, 2001a). Indeed, resistin has been found to be downregulated in mice treated with thiazolidinediones, an insulin sensitising agent (Steppan *et al.*, 2001b). However, when Urbanek and colleagues tested for an association between a marker in the promoter region of the resistin gene and three phenotypes (PCOS, obesity and insulin resistance) using 258 families with 323 affected offspring, no association was found (Urbanek *et al.*, 2003).

1.4.6 *CYP17*

Due to the deregulation of androgen production seen in women with PCOS, the steroid synthesis pathway has been extensively studied. Several key enzymes that have been analysed such as CYP17 (cytochrome P450 17 α-hydroxylase/17,20-desmolase; which converts progesterone to androstenedione) and CYP19 (cytochrome P450 aromatase;

which converts androgens to oestrogens), as discussed in section 1.1.2. Gilling-Smith and colleagues showed that androgen production is a stable phenotype in primary theca cells from PCO women and is increased when compared with normal theca cells *in vitro* (Gilling-Smith *et al.*, 1997). *CYP17* initially seemed a strong candidate when a single nucleotide polymorphism (T to C) found in the promoter region was shown in a small study to contribute to an increased susceptibility to PCOS (Carey *et al.*, 1994). *CYP17* expression (as well as *CYP11A1* expression) was seen to be increased in forskolin-stimulated PCOS theca cells and this increase in expression has been shown to be a stable property of these cells in culture (Nelson *et al.*, 1999). Forskolin is a drug that raises cyclic andenosine monophosphate (cAMP) levels, a known modulator of *CYP17* (and *CYP11A1*) expression. However, when the original study by Carey and colleagues was expanded to 96 cases and controls (from the original 68) and the T to C polymorphism was examined for association with PCOS, no association was found (Gharani *et al.*, 1996).

1.4.7 *CYP19*

Interest in *CYP19* arose from studies that demonstrated that granulosa cells taken from anovulatory polycystic ovaries produced greater amounts of oestrogen than those from normal ovaries *in vitro*. Additionally, these cells seemed to be hypersensitive to the actions of FSH (Mason *et al.*, 1994). Subsequent studies revealed an association with aromatase deficiency and hyperandrogenaemia (Harada *et al.*, 1992); (Ito *et al.*, 1993). Taken together, it was hypothesized that mutations in the *CYP19* gene could be contributing to the aetiology of hyperandrogenaemia and or PCOS. A linkage study involving 20 pedigrees (145 family members) was conducted and the results analysed by both parametric and non-parametric means. Analysis using the GENEHUNTER

program resulted in clear exclusion of the *CYP19* locus as there was no evidence for excess allele sharing across the entire region studied (Gharani *et al.*, 1997).

1.5 *CYP11A1*

One of the most promising candidate genes for PCOS codes for the enzyme that catalyses one of the initial rate limiting steps in the steroid hormone biosynthesis pathway. This step, the conversion of cholesterol to pregnenolone, is catalysed by the enzyme cholesterol side chain cleavage (P450 scc) and is encoded by the gene *CYP11A1*. This gene was the focus of the present PhD thesis and is discussed in detail below.

CYP11A1 has been reported in the literature to consist of 9 exons separated by 8 introns and to span 29.9 kb (Morohashi *et al.*, 1987). The genomic structure of *CYP11A1* was first described in 1986 and subsequently mapped more precisely to chromosome 15q23-q24 (Sparkes *et al.*, 1991); (Mount, 1982). A 'CAT' box and 'TATA' box are present 129 and 91 bp upstream from the initiation codon. All exons follow the GT/AG rule for intron/exon boundaries except for exon 6, which begins with a GC at the 5' end. The polyadenylation signal is located approximately 193 bp downstream from the termination codon (Morohashi *et al.*, 1987). Figure 1.2 shows a schematic diagram of the reported genomic structure of *CYP11A1*.

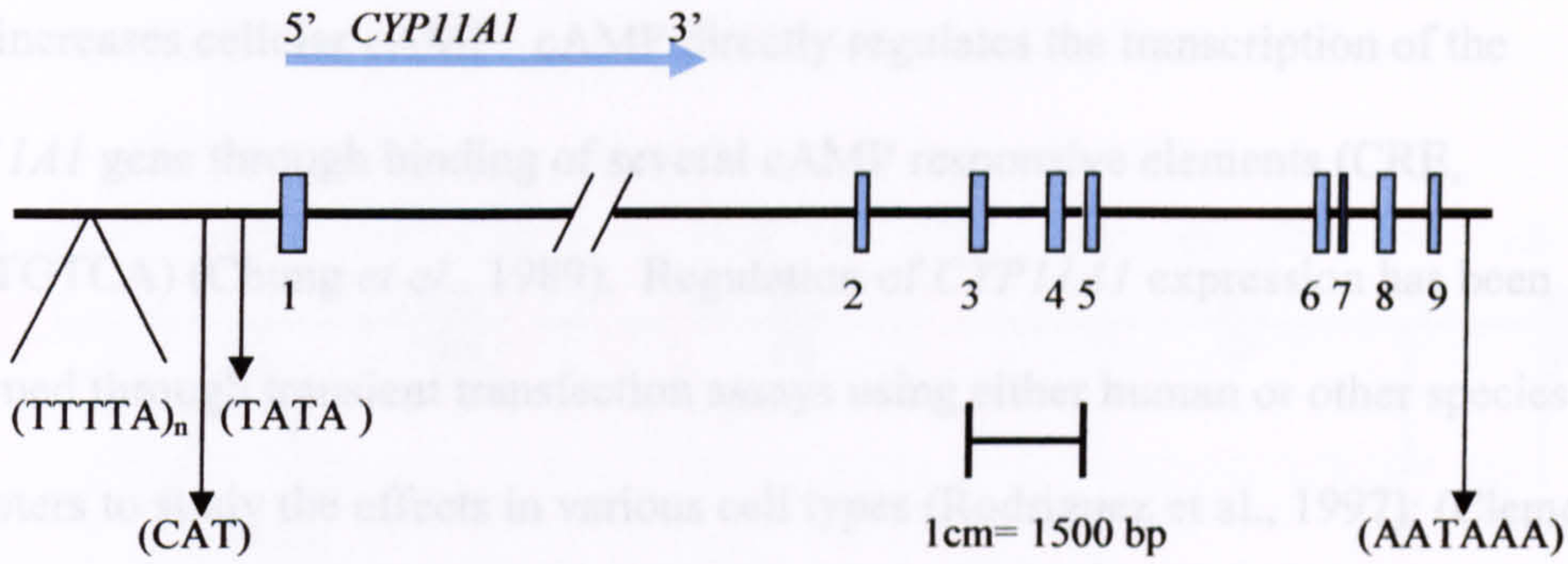


Figure 1.2: Genomic organisation of *CYP11A1* gene: Schematic representation of the *CYP11A1* genomic structure depicting 9 exons separated by 8 introns. Blue arrow indicates the direction of *CYP11A1* transcription. Exons are represented by blue boxes and are to scale relative to the introns. Numbers beneath the boxes represent the number of each exon. Pentanucleotide repeat $(TTTTA)_n$, CAT box, TATA box and polyadenylation signal (AATAAA) as shown.

1.5.1 Function

As discussed in see section 1.5.1, CYP11a catalyses the first rate limiting step in the synthesis of androgens in the mitochondria. The steroidogenic acute regulatory protein, StAR, facilitates the movement of cholesterol into the mitochondria. CYP11a mediates the 20- and 22-hydroxylations and the cleavage of the C20-C22 bond of cholesterol on the inner mitochondrial membrane, to produce pregnenolone (see figure 1.1 and section 1.1.2). The transcriptional regulation has been described and many tissue and species-specific differences identified (Moore *et al.,* 1992); (Rodriguez *et al*., 1997); (Ben Zimra *et al*., 2002); (Lui and Simpson, 1997).

1.5.2 Expression

CYP11A1 expression is regulated by gonadotrophins via cAMP in the ovary and by ACTH (adrenocorticotrophic hormone) via cAMP in adrenal cortical cells. In general, the stimulatory hormone binds the cell surface receptor, activates a G-protein, which in

LONDON. LIBRARY

turn, increases cellular cAMP. cAMP directly regulates the transcription of the *CYP11A1* gene through binding of several cAMP responsive elements (CRE, TGATGTCA) (Chung *et al.*, 1989). Regulation of *CYP11A1* expression has been discerned through transient transfection assays using either human or other species' promoters to study the effects in various cell types (Rodriguez et al., 1997); (Clemens *et al.*, 1994). However, this has lead to discrepancies between specific element usages in the different cell types. However, a general pattern of the major elements has emerged and is depicted in figure 1.3.

In transfection experiments using the transformed human placental cytotrophoblast cell line, JEG-3, sequences from –79 to +49 relative to the transcriptional start site conferred basal expression (Moore *et al.*, 1990). In addition, four possible basal elements lie between sequences –79 to –152, with a powerful element lying between –152 and –142 and a repressor lying between –177 and –152 (Moore *et al.*, 1990). cAMP responsiveness was mediated by sequences between –108 and –89 (Moore *et al.*, 1990); (Moore *et al.*, 1992). A highly conserved *cis*-acting sequence in the rat, mouse, porcine, ovine, bovine and human P450scc gene located between –60 and –32 binds the transcription factor activating protein-2 (AP-2) and is required for promoter activity in the placenta (Ben Zimra *et al.*, 2002).

However, P450scc is differently regulated in adrenal cells as compared with placental cells. As mentioned above, ACTH secreted by the pituitary regulates the expression of this gene, again through the increase of intracellular cAMP. Several regions also confer cAMP adrenal specific expression located at –1633 and –1626 (Chung *et al.*, 1989), {Moore, 1990 2 /id}. Analysis of *CYP11A1* expression in the human ovary has not been studied.

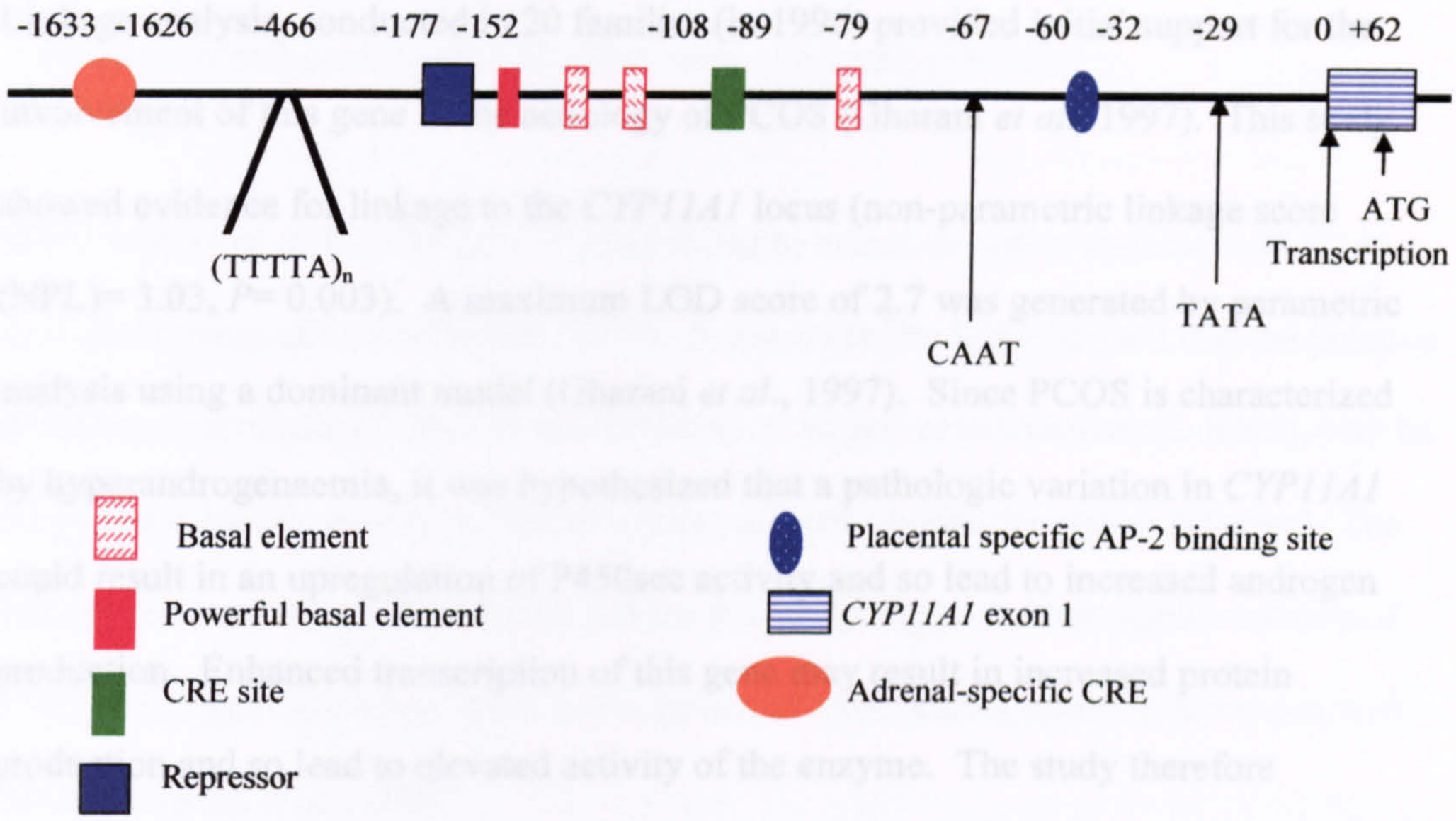


Figure 1.3: Summary of *CYP11A1* promoter region (not to scale): Numbering is relative to start of transcription. Pentanucleotide repeat $(TTTTA)_n$ is as shown.

1.5.3 *CYP11A1* and PCOS

CYP11A1 remains a strong biological candidate gene based on the function of this enzyme in the cell. As mentioned in section 1.2.6, CYP11a converts cholesterol to pregnenolone in the thecal cell of the ovary, as well as in the adrenal cortex and the placenta. Furthermore, studies have revealed an increase in androgen production from thecal cells of PCOS patients in addition to increased *CYP11A1* mRNA expression when compared with normal ovarian cells (Gilling-Smith *et al.*, 1997); (Jakimiuk *et al.*, 2001); (Daneshmand *et al.*, 2002). This overexpression has been shown to be a stable property of PCOS theca cells in long term culture and this increase has been hypothesized to contribute to the excess androgen production (Wood *et al.*, 2003).

1.5.4 Linkage and Association

Linkage analysis, conducted in 20 families (in 1996) provided initial support for the involvement of this gene in the aetiology of PCOS (Gharani *et al.*, 1997). This study showed evidence for linkage to the *CYP11A1* locus (non-parametric linkage score (NPL)= 3.03, *P*= 0.003). A maximum LOD score of 2.7 was generated by parametric analysis using a dominant model (Gharani *et al.*, 1997). Since PCOS is characterized by hyperandrogenaemia, it was hypothesized that a pathologic variation in *CYP11A1* could result in an upregulation of P450scc activity and so lead to increased androgen production. Enhanced transcription of this gene may result in increased protein production and so lead to elevated activity of the enzyme. The study therefore concentrated on screening the promoter region to identify any such elements that would enhance transcription. A polymorphic pentanucleotide $(TTTTA)_n$ microsatellite marker present at position –466 from the start of transcription was identified. It was demonstrated that variation at this site was associated with hirsutism and total serum testosterone levels in a study comparing subjects with PCOS and matched normal controls (Gharani *et al.*, 1997). There were four alleles observed in the data set, alleles 216bp (4 repeat), 226bp (6 repeat), 236bp (8 repeat) and 241bp (9 repeat). To minimize multiple allele comparisons for infrequent alleles, two groups were formed containing the most common allele (the four repeat unit allele designated 216+) and those lacking this common allele (designated 216-). An association was found between the 216-polymorphism and testosterone levels (*P*=0.002 in the complete dataset) (Gharani *et al.*, 1997). Furthermore, the association between this genotype and testosterone was only found in women with PCOS (*P*=0.009) and not in normal individuals (*P*=0.68) or in the asymptomatic PCO group (aPCOS, asymptomatic women were classified as non-hirsute women with regular menstrual cycles; *P*=0.95). However numbers in these latter

groups were small, therefore raising the possibility of a type 2 error. The results were also significant when *CYP11A1* genotype distributions between the PCOS group and a combined normal control/aPCOS group were compared (P=0.03).

Initial screening of 1.85 kb of *CYP11A1* failed to detect any other mutations likely to have pathogenic effects (Gharani, 2000). It was therefore hypothesised that the putative mutation, contributing to PCOS susceptibility or variation in testosterone levels, may be the pentanucleotide repeat or another variant located outside the region screened. The linkage and association data obtained from this study suggested that allelic variants of *CYP11A1* might contribute to the development of hyperandrogenaemia in women with PCOS (Gharani *et al.*, 1997).

Subsequent studies have further substantiated these results, although numbers in these data sets were also small (Diamanti-Kandarakis *et al.*, 2000); (Urbanek *et al.*, 1999); (Daneshmand *et al.*, 2002). In a Greek study, 80 PCOS patients, so defined by the 1990 National Institutes of Health-National Institute of Child Health and Human Development conference on PCOS, and 90 healthy controls with normal menses, were compared for association between *CYP11A1* genotype and testosterone levels (Diamanti-Kandarakis *et al.*, 2000). Subjects were again assigned into either the 216+ and 216- genotype groups and results showed that patients in the 216- group had higher testosterone levels than those in the 216+ group (100.0 $\pm$ 23.3 ng/dL vs. 78.0$\pm$ 19.8 ng/dL, P=0.004) (Diamanti-Kandarakis *et al.*, 2000).

Urbanek and colleagues used 39 sibpairs (from 28 PCOS pedigrees) to carry out non-parametric linkage analysis on 37 candidate genes. Nominal linkage to *CYP11A1* (67% identity by descent (IBD, λ^2= 5.34, P= 0.02) was obtained, but once corrected for

multiple testing, this was no longer significant (Urbanek *et al.*, 1999). Daneshmand and colleagues also reported a reduction in the frequency of the 216+ allele in PCOS (14% in 51 PCOS women *vs*.44% in 280 controls) associated with a striking compensatory difference in the frequency of the 9-repeat allele (56% vs 7%) (Daneshmand *et al.*, 2002). One Spanish study failed to find any association with the pentanucleotide repeat and serum testosterone levels in their population. In this study, 92 women presenting with hirsutism, some of whom also had corresponding menstrual disturbances (n=34) were compared for association between testosterone levels and *CYP11A1* genotype. When grouping as in the earlier studies, these authors failed to find any associations between the 216- allele group and higher serum testosterone levels in women with hirsutism, or with women with both hirsutism and menstrual disturbances compared with 30 healthy control women (χ^2= 0.166, *P*=0.983) (San Millan *et al.*, 2001).

While three out of four of these studies show statistically significant replication of the original association of *CYP11A1* and PCOS, all studies have been small and therefore run the risk of providing false negative results. This is a frequent problem with association studies and is often the reason for the lack of replication seen in the literature (Hirschhorn *et al.*, 2002). To effectively reduce the false negative, or false positive results, it is prudent to have large enough sample sizes, in the magnitude of hundreds or thousands of samples.

1.6 Aims

Many questions remain as to the significance of the pentanucleotide repeat and its association with several phenotypes seen in PCOS. The significant association between the *CYP11A1* $(TTTTA)_n$ polymorphism and serum testosterone levels may indicate either that this polymorphism itself is functional and has a direct role in the regulation of expression of the *CYP11A1* gene: or that this association is due to the presence of linkage disequilibrium (LD) between this marker and a putative aetiological mutation located elsewhere, either within the *CYP11A1* gene or possibly even within another gene in the region. Conversely, the significant associations may be spurious due to type 1 error or excessive data mining. *CYP11A1* remains a strong biological candidate gene with a reasonable amount of prior evidence for association between *CYP11A1* variation and PCOS but this has not been conclusive. The aim of this thesis was to investigate the role of *CYP11A1* in the aetiology of polycystic ovary syndrome. In particular, goals of this thesis were to ascertain if aberrant function/expression of *CYP11A1* in the human ovary contributes to the progression of PCOS. This was to be accomplished by:

1) Identifying other regulatory polymorphisms in the 5' flanking region and the surrounding genomic region of *CYP11A1* that might be contribute to the association results presented in the literature.
2) Examining the significance of the association of the pentanucleotide repeat polymorphism and other variants identified in (1) and PCOS by screening larger and more well characterised population data sets.
3) Determining the functional significance of the pentanucleotide repeat and other polymorphisms found with respect to the transcriptional regulation of *CYP11A1*.

2.0 SUBJECTS, MATERIALS AND METHODS

2.1 Subjects

2.1.1 United Kingdom (UK) Cases

Three approaches to association testing were taken in the UK samples: (i) case- control analysis: this used 230 PCOS cases of British/Irish origin and these were compared with 331 controls (see section 2.1.2); (ii) family-based association methods: this used 141 parent-offspring trios (ascertained through an offspring with PCOS). Since family based methods are not vulnerable to the effects of population stratification, families of diverse ethnic origin were included here. Thus, included in the 141 British/Irish families, there were 4 Afro-Caribbean, 1 Middle East, 2 other white, 5 South Asian and 12 Other European). Family relationships in the trios were confirmed by genotyping with a panel of highly polymorphic microsatellite markers (combined heterozygosity >99%). (iii) quantitative trait analyses (see section 4.6) were performed in all PCOS cases but excluded any woman taking oral contraceptive treatment, pregnant at the time of sampling, or with previous oophorectomy. A small number of women (4%, 14 of 371) who had received metformin were included in these analyses. Cases were defined as an affected and were ascertained independently of the relatives that were diagnosed at the infertility and endocrine clinic at St. Mary's and the Middlesex Hospitals London. All subjects gave fully informed consent. The essential criterion for the diagnosis of PCOS in this study was the presence of bilateral polycystic ovaries on ultrasound in a patient presenting with anovulation (oligo- or amenorrhoea) and/or hyperandrogenism. A polycystic ovary was defined as one which contains 10 or more cysts 2-8mm in diameter arranged peripherally in the ovary with an increase in the ovarian stroma (Adams *et al.*, 1985); (Adams *et al.*, 1986). Ovarian volume was also calculated from

the measurements obtained from the ultrasound scan. A diagnosis of hyperandrogenism required clinical (presence of hirsutism or acne) and/or biochemical (serum total testosterone >2.7nmol/l) features. Other potential endocrine and neoplastic causes of hyperandrogenaemia were excluded (London, UK) as previously described (Haddad *et al.*, 2002). These include Cushing's syndrome (excess glucocorticoid secretion of adrenal or pituitary origin), congenital adrenal hyperplasia (impaired production of cortisol from the adrenal glands), hyperprolactinaemia (a pituitary disorder) and androgen-secreting tumours (of the ovary or adrenal gland).

A full screening procedure for the family members was completed. Weight, height and waist/hip ratios were taken. Blood samples were obtained from all family members and probands for DNA analysis as well as for measurement of various serum indices. Subjects had a blood sample taken after a 12 hour fast for an extensive range of hormone and lipid assays including luteinizing hormone (LH), total testosterone, and total insulin.

2.1.2 UK Population Controls

All controls (n= 331) used for the case control analysis were taken from a random UK population of blood donors and DNA was isolated by CAMR, the Centre for Applied Microbiology and Research, part of the Health Protection Agency (Salisbury, UK) for case control studies such as this. All subjects were of British/Irish origin. Use of population controls (rather than subjects known not to have PCOS) is generally associated with only a modest reduction in power which can readily be compensated by increasing the number of control subjects typed (as here) (Colhoun *et al.*, 2003)

2.1.3 Northern Finland Birth Cohort of 1966 (NFBC1966)

In addition to the UK sample set, samples were available from the Northern Finland Birth Cohort of 1966. This cohort is based on 12058 live-born babies born in the northernmost two provinces of Finland arising from pregnancies with expected dates of delivery during the calendar year 1966. Of 5926 cohort members (3088 women) who had undergone clinical review and DNA sampling at age 31, 1589 patients were genotyped for this study (Taponen *et al.*, 2003). These included 527 women who had reported symptoms of hirsutism and/or oligo/amenorrhoea on the 31-year review questionnaire (symptomatic 'cases') and 1062women (cohort controls) from the same cohort without such symptoms (generating an approximate 1:2 ratio of cases to controls, all matched for age given the birth cohort design). The cohort cases have been shown as a group to have biochemical features consistent with PCOS (e.g. hirsutism and or oligomenorrhea consistent with hyperandrogenaemia) (Taponen *et al.*, 2003) but only a proportion have proven PCOS. Ovarian ultrasound (USS) examination of these "cases" is ongoing, with 67 (out of 178) so far examined having PCO status confirmed on USS. These 67 individuals are referred to as "USS-confirmed cohort cases". For the quantitative trait analyses, which were the main emphasis of analysis in the Finnish subjects, the group size was reduced to 1435 (475 cases; 960 controls) by exclusion of 138 women who were pregnant at the time of the blood sampling and 16 in whom no testosterone measures were available. The single subject taking metformin at the time of recruitment is included. Clinical features of all sample sets are provided in table 2.1. All clinical investigations were conducted in accordance with the guidelines in The Declaration of Helsinki, and the study was approved by the relevant ethics committees in UK and Finland. All subjects provided fully informed consent.

	UK subjects			Finnish subjects	
	PCOS cases	PCOS probands	Population controls	Symptomatic cases	Cohort controls
N	230	141	331	527	1062
% female	100	100	55.5	100	100
Age (years)	33.8 (6.9)[b]	31.4 (6.1)[b]	38.6 (8.1)[a]	31	31
BMI (kgm^{-2})	27.4 (7.8)	27.0 (7.4)	-	25.2 (5.4)	24.2 (4.6)
Waist:hip ratio	0.80 (0.07)	0.80 (0.09)	-	0.82 (0.08)	0.81 (0.08)
Testosterone (ng/ml)[c]	0.64 (0.26)	0.68 (0.34)	-	0.67 (0.29)	0.59 (0.25)
Glucose (mg/dl)[d]	86 (10)	85 (18)	-	90 (18)	88 (12)
Symptoms hirsutes oligomenorrhoea	 105 (46%) 145 (63%)	 87 (62%) 115 (82%)	-	 295 (56%) 312 (60%)	 0 0
Ovarian status	USS confirmed PCOS	USS confirmed PCOS	-	USS underway, 67/178 confirmed	Not known but asymptomatic

Table 2.1: Patient data. Quantitative data are presented as mean (SD); [a]Age information is available from only 38% of these control subjects. [b]Ranges for subject ages were as follows: UK PCOS cases 19.3-57.6 years, UK PCOS probands 16.5-47.7 years. All Finnish women were aged 31

2.2 Materials

2.2.1 Reagents

The majority of reagents used were supplied by Sigma (UK), BDH (Poole, Dorset), Invitrogen (UK) and Pharmacia (UK). Reagents for bacteriological media were from Oxoid.

2.2.2 Enzymes

All restriction endonucleases were supplied along with their appropriate buffer by Invitrogen (UK).

2.2.3 Radioisotopes

[α-^{32}P] dCTP (10mCi/mmol, 370MBq/ml) radioisotope was supplied by Amersham Pharmacia Biotech.

2.2.4 Electrophoresis and Blotting Materials

Agarose and low melting temperature (LMT) agarose were from Invitrogen. Hybond-N nylon membranes were supplied by Amersham and 3MM filter paper was from Whatman Ltd. The 1kb ladder used as a size marker was from Invitrogen and ethidium bromide was supplied by Fluka BioChemika (UK). RNA markers were supplied by Promega (UK).

2.2.5 Photography and Autoradiography

Photographic images were recorded using two methods. Firstly, photographs were taken using a UVP transilluminator and a Polaroid MP +4 camera with a Wratten gelatin filter. PLUS-X pan Polaroid film and X-ray film were supplied by Kodak. Both types of film were developed using a Fuji automated X-ray film processor (FujiRGII) with appropriate reagents from Fuji Corporation Ltd. In addition, digital photographs were taken using a Syngene, Gene Genius Bio Imaging System with a UVP transilluminator (Synoptics Ltd.).

2.2.6 Oligonucleotides

Commercial oligonucleotides for standard PCR were supplied by MWG, UK. Fluorescent-labelled oligonucleotides for automated genotyping were also from MWG, UK.

2.2.7 PCR Reagents

10X NH_4 reaction buffer, 10mM dNTPs and 50mM $MgCl_2$ were supplied by Promega. BioTaq DNA polymerase, PCR buffer and magnesium chloride were supplied by Bioline (UK).

2.2.8 DNA Sequencing Materials

The ABI sequencing gel for sequence analysis using an ABI 373a automated DNA sequencer was made up of 40ml SequaGel^{TM} ready-to-use 6% sequencing gel solution (National Diagnostics), 60ml SequaGel^{TM} complete buffer reagent (National Diagnostics) and 0.04g ammonium persulphate (Sigma). Sequence analysis using an

ABI 3100 Genetic Analyser was carried out using components supplied by Applied Biosystems: 50cm capillary array, 10X buffer, 3100 Performance Optimized Polymer-6™ (POP-6™) sieving medium and Hi-Di™ Formamide.

2.2.9 Genotyping Materials

All reagents used for genotyping were purchased from Applied Biosystems (UK).

2.2.10 Statistical Programs

Linkage disequilibrium relationships were determined in UK and Finnish control subjects using the LDMAX option within the GOLD (Graphical Overview of Linkage Disequilibrium) package, which implements the EM (Expectation maximization) algorithm (Abecasis and Cookson, 2000). This program allows for straightforward pair-wise measurements of LD amongst loci or populations with a moderate number of markers.

Between-group comparisons of genotype and allele frequencies used standard contingency table methods implemented in STATXACT (Cytel, Cambridge, MA), which allows exact probabilities (or Monte-Carlo estimates thereof) to be generated. Haplotype frequencies were derived by standard maximum likelihood methods implemented in TRANSMIT, with group comparisons assessed by likelihood ratio testing and permutation methods (at least 1000 permutations) (Clayton, 1999).

In the UK trios, family-based association tests were undertaken using TRANSMIT: alleles and haplotypes with frequency <2% were pooled for their respective analyses.

Quantitative trait analyses of serum testosterone levels were performed by analysis of variance on log-transformed values, using SPSS (ver 10) and SAS (ver 8.2). For analyses of testosterone levels, all women taking oral contraceptive medication at the time of sampling were excluded. Power calculations were perfomed using the program QUANTO (version 1.0) (Gauderman, 2002)

2.2.11 Bacterial Strains

The JM109 and DH5α bacterial strain competent cells for transformation reactions were supplied by Promega and Invitrogen respectively.

2.2.12 BAC Clone

The BAC clone, CTD-2311M21 was purchased from Research Genetics, a division of Invitrogen (UK).

2.2.13 IMAGE Clones

All IMAGE clones were purchased from the MRC UK HGMP Resource Centre in Hixton (UK).

2.2.14 Cell Culture Reagents

Cell culture reagents for the primary cultures were purchased from Invitrogen (UK). The ovarian cell line, HGL5, was a generous gift from Dr. Bill Rainey (Texas, USA), JEG-3 from Dr Graham Williams (Imperial College, UK), BeWo from Dr Dave Carling (Imperial College, UK) and 293T cells were from Prof. Malcolm Parker (Imperial

College, UK). For all the stable cell lines (HGL5, JEG-3, and 293T), reagents were supplied by Invitrogen (DMEM F-12, fetal bovine serum, glutamine and penicillin-streptomycin) whereas the gentamicin was purchased through Sigma-Aldrich, ultroser were supplied by Biosepra (France) and the ITS+ supplement was purchased though Collaborative Biomedical Products (MA, USA).

2.3 Buffers, Solutions, Media and Kits

2.3.1 General Stock Reagents

0.5M EDTA (ethylenediaminetetraacetic acid, pH 8.0): adjust to pH 8.0 by addition of 5M NaOH.

5M NaOH (sodium hydroxide): prepared with dH_2O.

10% SDS (w/v) (sodium dodecyl sulphate): prepared with dH_2O.

20X SSC (standard saline citrate): 3M NaCl, 0.3M sodium citrate

2.3.2 Preparation of Human Genomic DNA

Sucrose Lysis Buffer: 0.32 M sucrose, 10mM Tris pH 7.5, 5mM $MgCl_2$, 1% triton-X100.

Nuclei Lysis Buffer: 10mM Tris-HCl, 400mM NaCl, 2mM Na_2EDTA

10mg/ml proteinase K

6M NaCl

2.3.3 Preparation of Plasmid DNA

LB Broth: 1% (*w/v*) Bacto-tryptone, 0.5% (*w/v*) yeast extract, 1% (*w/v*) NaCl, autoclave before use.

LB Agar: LB broth containing 1.5% (*w/v*) bacteriological agar, autoclave before use.

Ampicillin (50mg/ml stock): Ampicillin (Sigma) dissolved in dH_2O, filter sterilized and used at a final concentration of 50µg/ml (ie. 1µl per 1ml media).

Chloramphenicol: Chloramphenicol (Sigma) was dissolved in distilled water, filter sterilized and used at a final concentration of 12.5 µg/ml

X-GAL: 5-bromo-4chloro-3-indolyl-bD-galatosidase (XGAL) (Sigma) dissolved in dimethyl formamide and used at a final concentration of 20µg/ml.

IPTG: isopropyl β-D-thiogalactopyranosid (IPTG) (Sigma) dissolved in dH_2O, filter sterilized and used at a final concentration of 25µg/ml.

SOC Media : 2% (*w/v*) Bacto-tryptone, 0.5% (*w/v*) yeast extract, 0.2% (*w/v*) NaCl, 2.5mM KCl, 20mM glucose and, after autoclaving, 1mM sterile $MgCl_2$.

2.3.4 Gel Electrophoresis

50X TAE Electrophoresis buffer (per litre): 24.2% *(w/v)* Tris base, 57.1ml glacial acetic acid, 0.05M EDTA (pH8.0).

10X TBE Electrophoresis buffer: 1M Tris base, 0.83M Boric acid, 10mM EDTA (supplied by National Diagnostics).

ABI 373a sample loading dye: 25mM EDTA (pH8.0), deionised formamide and dextran blue (50mg/ml) in a ratio of 5:1 formamide to EDTA.

Ficoll loading dyes (10X): 25% (*w/v*) Ficoll, 0.25M EDTA, 0.25% (*w/v*) Orange G or Bromophenol blue in 1XTAE.

Formamide loading buffer (10X): 95ml formamide, 50mg Bromophenol blue, 50mg xylene cyanol and 4ml 0.5M EDTA.

MOPS buffer: 3-n-morpholino propanesulphanic acid

2.3.5 Northern Blot Screening

Pre/Hybridisation Buffer: ExpressHyb (Clontech)

Wash solution 1: 2XSSC, 0.05%SDS

Wash solution 2: 0.1XSSC, 0.1%SDS

2.3.6 Kits Utilised

- Purgene DNA Extraction Kit – Gentra Systems, Minneapolis, Mn. USA
- QIAprep Miniprep Spin Kit – Qiagen, UK
- Wizard™ Minicolumn Kit – Promega
- Oligolabelling Kit – Amersham Pharmacia Biotech
- ProbeQuant™ G-50 Microcolumn Kit – Amersham Pharmacia Biotech
- ABI PRISM™ Big Dye™ Terminator Cycle Sequencing Kit with AmpliTaq® DNA Polymerase, FS - v.3.0 and v.4.0 (Applied Biosystems)
- Large construct kit- Qiagen
- Midi kit- Qiagen
- PCR purification- Qiagen
- Gel purification kit- Qiagen
- RACE kit- Clontech
- PCR advantage polymerase kit- Clontech
- pGEM T-Easy Kit- Promega
- Luclite transfection kit- Perkin Elmer

2.4 Methods

2.4.1 Human Genomic DNA Extraction From Whole Blood

DNA was extracted by colleagues in the laboratory from the UK and NFBC1966 population datasets from whole blood using the PURGENE kit (Gentra Systems). Between 5 to 10 mls of blood (fresh, or thawed from frozen) was decanted into 50 ml Nunc tubes and ice cold lysis buffer was added to the 30 ml mark. The tube was inverted to mix and spun at 2000 *g* for 10 minutes. The supernatant was poured off and the pellet resuspended in 10 mls lysis buffer. After a 5 minute incubation on ice, the tubes were spun at 2000*g* for 5 minutes. The supernatant was poured off to leave a whitish pellet. The following solutions were added to the pellet and mixed by inversion after adding each solution: 3 mls T20 solution, 200 µl 10% SDS and 500 µl Proteinase K solution (10 mg/ml stock). The samples were incubated overnight at 4°C. One ml of saturated NaCl solution was added to each sample and mixed vigorously for 15 seconds. They were then spun at 2500*g* for 30 minutes at 4°C. The supernatant was carefully removed to a fresh Nunc tube and the DNA precipitated using 2 volumes of ethanol at room temperature. After mixing, the DNA was visible and was gently spooled out using an inoculation loop. After a 70% ethanol wash, the pellet was air dried for a few minutes and dissolved in 1ml dH_2O. The DNA was stored at 4°C.

2.4.2 Phenol/Chloroform DNA Extraction

The addition of phenol to an aqueous solution of nucleic acids helps to remove any contaminating proteins. This purification process is improved by the addition of chloroform/isoamyl alcohol ($CHCl_3$/IAA, 24:1), which denatures proteins and enhances the separation of the organic and inorganic phases. Extractions were performed by the

addition of an equal volume of phenol or chloroform/IAA and vigorous mixing followed by centrifugation to separate the phases. Phenol/chloroform extractions were performed in the same manner but with the addition of 0.5 volumes of phenol followed by 0.5 volumes of chloroform/IAA.

2.4.3 Plasmid Miniprep

For the isolation of DNA from an overnight bacteriological cell culture, a QIAprep® Miniprep Spin Kit (QIAGEN®) was used as per the manufacturer's instructions. Ten mls of selective LB broth, which was inoculated with a single bacterial colony was incubated with shaking at 37°C overnight. Three mls of the culture was spun in an eppendorf tube in a microcentrifuge at 13,000rpm for 2 minutes. The supernatant was removed to leave the cell pellet, which was resuspended in 250 µl of P1 solution and inverted several times to mix. Two hundred and fifty µl of solution P2 was added and inverted to mix. Solution N3, 350 µl, was then added immediately and the tube inverted to mix. The mixture was spun at 13,000rpm for 10 minutes, resulting in the separation of cellular debris and the DNA-containing aqueous solution. The aqueous DNA solution was added to a pre-equilibrated Qiagen miniprep spin column and spun at 13,000rpm for 1 minute. The eluted solution was discarded. The column resin was washed with 750 µl of ethanol wash buffer and spun twice at 13,000rpm for 1 minute each time, with the eluted solution being discarded after each spin. The column was then transferred to a clean eppendorf tube and the DNA was eluted with 50 µl of dH_2O applied to the centre of the resin. This was incubated at room temperature for 1 minute and centrifuged for 1 minute at 13,000rpm to collect the plasmid DNA. Approximately 6 µg of plasmid DNA was routinely recovered.

2.4.4 Plasmid Midiprep

For the isolation of larger amounts of DNA from overnight bacteriological cell cultures, a QIAprep® Miniprep Spin Kit (QIAGEN®) was used as per the manufacturer's instructions. A bacterial overnight culture was set-up using 10mls of selective LB broth, inoculated with a single bacterial colony and incubated with shaking at 37°C overnight. The next day, 250 µl of this overnight culture was used to inoculate 25 mls of selective LB media, which was incubated with shaking at 37°C overnight. The bacterial cells were harvested by centrifugation at 6000*g* for 15 minutes at 4°C. The supernatant was removed to leave the cell pellet, which was resuspended in 4 mls of P1 solution and vortexed to mix. Four mls of solution P2 was added and inverted to mix. Solution P3, 4 mls, was then added immediately and the tube inverted to mix. The QIAGEN-tip was equilibrated with 4 mls of buffer QBT and the column allowed to empty by gravity. The cell lysate was filtered though the QIAfilter and loaded onto the previously equilibrated Qiagen-tip. This was then washed with 2x 10 mls of buffer QC and the DNA was eluted from the column with 5 mls of buffer QF. Three and a half mls of room temperature isopropanol (0.6 volumes) was used to precipitate the DNA and was mixed and the solution centrifuged immediately at 15000*g* for 30 minutes at 4°C. Seventy percent ethanol was used to wash the DNA pellet and centrifuged again at 15000*g* for 10 minutes. The pellet was air dried and redissolved in an appropriate amount of water.

2.4.5 Large Construct Maxiprep

For isolation of DNA from BAC clones, the QIAGEN® Large Construct Kit was used as per manufacturer's instructions. A bacterial overnight culture was started using

10mls of selective LB broth, inoculated with a single bacterial colony and incubated with shaking at 37°C overnight. The next day, 1ml of the overnight culture was used to inoculate 500 mls of selective LB broth, which was incubated with shaking at 37°C overnight. The bacterial cells were harvested by centrifugation at 6000*g* for 15 minutes at 4°C. The supernatant was removed to leave the cell pellet, which was resuspended in 20 mls of P1 solution and vortexed to mix. Twenty mls of solution P2 was added and inverted to mix. Chilled solution P3 was then added immediately at a volume of 20 mls, the tube inverted to mix and incubated for 10 minutes on ice. The supernatant was removed following centrifugation at ≥ 20 000*g* for 30 minutes at 4°C. The resulting lysate was filtered through a folded filter prewetted with dH_2O. The DNA was precipitated by adding 0.6 volumes (approximately 36 mls) of room temperature isopropanol, mixed and recovered by centrifugation at ≥ 15 000*g* for 30 minutes at 4°C. Five mls of 70% ethanol was used to wash the DNA pellet followed by centrifugation at 15000*g* for 15 minutes and the pellet was air dried and redissolved in water.

2.4.6 Spectrophotometry

To determine the integrity and yield of DNA (or RNA) samples were diluted with dH_20 and the absorption measured at 260nm and 280nm using a UV500 visible spectrophotometer (Unicam). Purity of DNA was determined by the ratio of absorbance at 260nm to that at 280nm. High quality, pure DNA has a 260nm: 280nm ratio of ≥1.8, RNA a ratio of 2.

The following equation was applied to determine yield:

$$\text{DNA concentration } \mu g/ml = Abs_{260nm} \times \text{dilution factor} \times y$$

Where y= 50 for DNA, 40 for RNA.

2.4.7 Isolation of Total RNA

2.4.7.1 Total RNA Isolation from Tissue

Placental tissue was removed from the –80°C and placed in a pre-chilled mortar containing liquid nitrogen. Using the pestle, the tissue was crushed until it became a fine powder (time depends on tissue type/size). A pre-chilled spatula was used to scrape the powder into a 50 ml Falcon tube containing an appropriate amount of TRIZOL (Invitrogen); the amount of TRIZOL used depended on the size of the placental slice. In general, 5 mls of TRIZOL was used per 1g of tissue. The tissue was homogenized for 35-45 seconds at a polytron setting of 24-36, and incubated at 15°C-30°C for 5 minutes to completely dissociate the nucleoprotein complexes. The protocol as listed in section 2.4.7.3 was then followed.

2.4.7.2 Total RNA Isolation from Cell Lines

Cells from culture were trypsinised and pelleted as described in section 2.4.19. Approximately 1 ml of TRIZOL was used per 5-10 x 10^6 cells.

2.4.7.3 Isolation of Total RNA using TRIZOL

Total RNA was isolated using TRIZOL reagent according to the manufacturers instructions. Following either homogenisation of tissue or resuspension of cells, 0.2 ml of chloroform was added per 1 ml of TRIZOL and the tubes were shaken vigorously by hand for 15 seconds. Following an incubation period of 3 minutes at room temperature, the samples were centrifuged at 12000*g* for 15 minutes at 4°C. The aqueous phase containing the total RNA was then transferred to a fresh tube and the RNA was

precipitated using 0.5 mls of isopropanol per 1 ml of TRIZOL used, incubated at room temperature for 10 minutes and centrifuged for 10 minutes at 12000*g* at 4°C. The supernatant was removed to leave the white RNA pellet, which was then washed with 75% ethanol and left to dry on the benchtop for 5 minutes. Finally, the RNA was dissolved in diethylpyrocarbonate (DEPC)-treated water and stored at -80°C.

2.4.8 Northern Analysis

2.4.8.1 Radioisotopic Labelling of DNA Probes

An Oligolabelling Kit (Amersham Pharmacia Biotech) was used to label DNA probes for Northern blot analysis. The method was based on the "random priming" procedure developed by Feinberg and Vogelstein in which random hexadeoxyribonucleotides anneal specifically to denatured DNA template strands and serve as primers for DNA synthesis by the action of the Klenow Fragment of *E.coli* DNA PolymeraseI (Feinberg and Vogelstein, 1983). By incorporation of [α-^{32}P] dCTP in the synthesis reaction a high specific activity DNA probe is generated. Briefly, 25-50ng of DNA template was denatured by heating to 95°C for 5 minutes. To this template, 10µl reagent mix, 1µl Klenow (5U/µl), 40-50µCi [α-^{32}P] dCTP (Amersham Pharmacia Biotech) and dH_20 to a final volume of 50µl was added. The labelling mix was incubated at 37°C for 10 minutes. After incubation the labelling mix was applied to a ProbeQuant™ G-50 Micro Column (Amersham Pharmacia Biotech) to remove unincorporated labelled nucleotides as per manufacturers' instructions. Briefly, the column was originally vortexed to resuspend the resin and centrifuged for 1 minute at 735*g*. The radioactively labelled probe was added directly to the centre of the resin and centrifuged for 2 minutes at 735g. The purified sample was collected at the bottom of the tube.

2.4.8.2 Probe Hybridisation and Washing of Filters

The hybridisation buffer (ExpressHyb Solution) used was purchased from Clontech (UK). The solution was heated to 68°C before use. The Northern blots were pre-hybridised in 5 mls of ExpressHyb solution for 30 minutes with constant shaking at 68°C. Probes used were dentured at 95°C-100°C for 5 minutes and were quickly chilled on ice. The prehybridisation solution was discarded and replaced with 5 mls of fresh hybridisation solution containing 1-2 x 10^6 cpm/ml of probe and incubated at 68°C with continuous shaking for 1 hour. The blots were next rinsed several times with Wash Solution I (2x SSC, 0.05% SDS) at room temperature, followed by a further 40 minute wash with continuous shaking, changing the wash solution several times. The blots were next washed with Solution 2 (0.1x SSC, 0.1% SDS) at 50°C for 40 minutes, with one change of solution. Blots were removed from the wash solution and immediately covered with clingfilm, to prevent drying out of the blot, mounted on Whatman paper and further wrapped in clingfilm.

2.4.8.3 Stripping

Northern blot filters were stripped of hybridised probe by incubating for 10min in 0.5% SDS heated to 90-100^{0}C. The solution containing the blot removed from heat and left to cool for a further 10 minutes and then air dried. Blots were stored at room temperature.

2.4.8.4 Autoradiography

Hybridised membranes sealed in clear plastic film were placed in an autoradiography cassette with two signal intensifying screens and were subject to autoradiography using material described in section 2.2.5. Depending on the result upon development of the film, the membranes were re-exposed for longer or shorter periods of time.

2.4.9 Size Fractionation Gel Electrophoresis

2.4.9.1 Agarose Gel Electrophoresis

DNA preparations, DNA fragments resulting from restriction enzyme digestion, cloned DNA and PCR products were size-separated by agarose gel electrophoresis. In order to optimise the resolution, the percentage (*w/v*) agarose used was varied according to the size of the DNA fragments to be separated. In general, the larger the expected fragment size of DNA the lower the gel percentage of agarose used. Agarose (ordinary agarose or low melting temperature (LMT) agarose; BioRad) was dissolved in 1 X TBE buffer by heating in a microwave oven. Once the agarose solution had cooled to below 60°C, ethidium bromide (Fluka BioChemika) was added to a final concentration of 5µg/ml and the solution was poured into a sealed gel casting tray. Before loading into the agarose gel, 1/10 volume of 6x blue loading dye was added to each sample. A commercial 1kb DNA ladder marker (Invitrogen, UK) was loaded alongside samples as a molecular size standard. Unless otherwise stated, agarose gels were run in 1X TBE electrophoresis buffer (section 2.3.5). During electrophoresis, the ethidium bromide in the gel becomes intercalated with the DNA thus allowing the DNA fragments to be visualised on a UV transilluminator (302nm) and photographed (section 2.2.5).

2.4.9.2 RNA Denaturing Gel Electrophoresis

RNA preparations, dissolved in DEPC-treated water were size-separated by agarose gel electrophoresis. Agarose was dissolved in distilled water by heating in a microwave oven. Once the agarose solution had cooled to below 60°C, 10 mls of 10x MOPS was added (to final concentration of 1x), 3 mls of formaldehyde (final concentration 3%) and 3 µl of 10µg/µl ethidium bromide was added and the solution was poured into a sealed gel casting tray. Before loading, 10 µl of formamide (final concentration 50%), 3.75 µl formaldehyde (final concentration 0.015M), and 1x MOPS to a total volume of 20 µl was added to each RNA sample including the RNA ladder (Promega, UK) and heated for 10 minutes at 65°C to denature the samples. Following denaturation, RNA loading dye (1 mM EDTA, 0.25% bromophenol blue, 50% glycerol and 1 mg/ml ethidium bromide) was added to each sample and loaded onto the gel.

2.4.10 Restriction Endonuclease Digestion

Restriction endonucleases are bacterial enzymes, which cleave double stranded DNA into discrete pieces, resolvable by gel electrophoresis. They cleave at specific recognition sequences within the DNA, the length of which (usually 4-8bp) determines the frequency of cleavage and therefore the size of resulting DNA fragments.

In principle, 1 unit of restriction enzyme is required to digest 1µg of DNA in 1 hour. Crude DNA preparations often require more enzyme and/or more time for complete digestion. Genomic DNA and cloned DNA were digested with a number of restriction endonucleases (Invitrogen). Briefly, 10X concentrated buffer (as supplied) was used with 1-2 units of endonuclease to digest 1µg of DNA and made up to volume with

dH_2O. The volume of enzyme used did not exceeded 1/10 of the total volume to avoid inhibition of digestion by the enzyme storage buffer, glycerol. Reactions were incubated at the optimum temperature recommended by the manufacturer (usually 37°C) for 1-3 hours and the products resolved by gel electrophoresis as described in section 2.4.9.1.

2.4.11 Polymerase Chain Reaction

2.4.11.1 Principles

The polymerase chain reaction (PCR) allows the amplification of specific DNA sequences using two oligonucleotide primers, which are designed to hybridise to opposite strands of the DNA and flank the target DNA region. The target DNA is denatured in the presence of a large excess of the two primers and then returned to a temperature, which will allow the primers to anneal to the DNA. A heat-stable DNA polymerase (that isolated from the thermophilic bacterium *Thermus aquaticus* and therefore called *Taq* polymerase), which synthesises DNA in the 5' to 3' direction, and all four nucleoside triphosphates are included in the reaction and the sample is incubated at a temperature optimal for elongation by the enzyme. In the first cycle two copies of the target sequence with indeterminate 3' termini are produced. The three steps of denaturing, annealing and extension are repeated, producing another four copies of the target sequence. Two of the new copies will have indeterminate 3' termini but two will now have termini dictated by the 5' terminus of the other primer. In subsequent cycles of reactions there will be further exponential amplification of DNA flanked by each primer (*n* cycles of PCR amplify the target 2^n-fold).

2.4.11.2 Standard PCR Parameters

Standard PCR reactions were performed in a final volume of 25µl in a 0.5ml eppendorf tube as follows:

1 x NH_4 buffer
1mM $MgCl_2$
200µM each dNTP
10 pmols each of forward and reverse primers
200ng-1µg DNA
1U Bio*Taq*
µl dH_20 to 25 µl total volume

The DNA underwent an initial denaturation stage at 95°C for 3 minutes. Standard thermal cycling conditions were 35 cycles of:

Denaturation: 95°C for 30 seconds
Annealing: X °C for 30 seconds
Extension: 72°C for 30 seconds.

For annealing, X is determined by sequence of the primers, see section 2.4.14.1. The cycling was followed by a final extension at 72°C for 5 minutes and cooling to 10°C. Any changes to this standard protocol are described in the text.

2.4.12 PCR Purification

For isolation of fragments generated by PCR, the QIAquick® PCR purification kit from Qiagen was used as per manufacturers' instructions. QIAquick® utilises a silico-gel membrane with a maximum binding capacity of 10µg of DNA. This kit can be used for the purification of double- or single-stranded PCR products ranging from 100bp to 10 kb in size. Briefly, 5 volumes of buffer PB was added to 1 volume of the PCR reaction sample and mixed. The resultant mixture was placed into the QIAquick spin column with a 2ml collection tube and centrifuged for 60 seconds at 13000 rpm. The flow-through was then discarded. To wash, 750µl of buffer PE was added to the column and

centrifuged for a further 60 seconds at 13000 rpm, the flow-through discarded and centrifuged for a further 60 seconds to get rid of the residual PE buffer. The QIAquick spin column containing the DNA bound to the silica-gel was then placed into a clean eppendorf tube and 50µl of dH_2O was added, the tube was incubated at room temperature for 1 minute and the DNA eluted by centrifugation for 1 minute at 13000 rpm.

2.4.13 Rapid Amplification of cDNA Ends (RACE)

Identification of both 5' and 3' ends of genes was performed using premade libraries of adapter-ligated double stranded (ds) cDNA from Clontech (UK) as templates for amplification according to the manufacturer's instructions. Preparation of these Marathon-Ready ™ cDNAs begins with first strand cDNA synthesis from poly A^+ RNA. First strand synthesis utilizes a modified lock-docking oligo (dT) primer with two degenerate nucleotide positions at the 3' end. Second strand synthesis is performed as per standard protocol using *E.coli* DNA polymerase 1, RNase H and *E. coli* DNA ligase. Blunt ends are created by T4 DNA polymerase and the resulting ds cDNA is ligated to the Marathon cDNA Adapter. The 5' end of the adapter is phosphorylated to facilitate blunt end ligation to both ends of the double stranded cDNA by T4 DNA ligase.

A PCR master mix containing the Clontech Advantage 2 polymerase kit containing 10x reaction buffer, dNTP (10mM) and TaqStart Antibody was used as per manufacturer's instructions. To this mix, 5 µl of Marathon-ready cDNA (0.5ng) was added with 10 µM of adapter primer (AP1) and 10µM of a gene-specific primer (GSP) to a total volume of 50 µl.

The thermal cycling parameters were as follows:

94°C for 30 seconds

5 cycles of:
94°C for 5 seconds
72°C for 3 minutes

5 cycles of:
94°C for 5 seconds
70°C for 3 minutes

20-25 cycles of:
94°C for 5 seconds
68°C for 3 minutes

and cooling to 10°C.

Five µl of the reaction was then analysed on a 1.2% agarose gel with the 1kb size ladder from Invitrogen. Both extension times and numbers of cycles were changed to optimise the reactions of each gene specific primer.

2.4.14 Primer Design and Optimisation

2.4.14.1 Standard PCR

When designing primer pairs for PCR amplification of known sequences, several rules were followed as far as possible to ensure optimal results: random base distribution and similar GC content for both primers (preferably below 50% of the total); an anchoring single or double C and/or G at either end (but not complementary); minimal secondary structure (i.e. no self-complementarity); low complementarity to each other (to reduce the incidence of 'primer dimer' formation); no greater than 4°C difference between the T_m values of both primers as far as possible given the above constraints; and a primer length of at least 20 nucleotides to increase the sequence specificity.

The optimal annealing temperature for specific primer pairs was approximated by first calculating the melting temperature (T_m) of each primer which is dependent on the nucleotide sequence and was derived using the following formula:

$$4(G+C) + 2(A+T) = T_m$$

and then by assigning an annealing temperature 4-6°C lower than the value obtained. Primer pairs were then tested by PCR of several samples of human genomic DNA, including a 'no DNA' control, and electrophoresed on agarose gels (section 2.4.9.1) to assess the adequacy of the PCR conditions for subsequent experiments. Additional bands to the authentic PCR product suggested cross-hybridisation of the primers to sequences within the genomic DNA that bear some degree of homology to the intended target sequence. These extra products could usually be eradicated by increasing the annealing temperature by 1-2°C or by altering the $MgCl_2$ concentration in the reaction.

2.4.14.2 RACE PCR

At least one gene-specific primer is needed for both 5' RACE and for 3' RACE. All primers for RACE were between 23-28 nucleotides in length with a G-C content of approximately 50-70% and a T_m of $\geq$65°C. Other standard primer design principles apply here as described in section 2.4.14.1.

2.4.15 Elution of Target PCR Fragments from Gels

Following visualisation by electrophoresis (see section 2.4.9.1), PCR products of interest were excised from the gel under long-wavelength ultraviolet light, with the aid of scalpel blades and DNA extracted from the gel using the Qiagen Qiaquick gel

extraction kit (Qiagen, UK). For purification of DNA, the agarose slice was placed in a pre-weighed eppendorf tube and the tube containing the gel slice was re-weighed. Three volumes of buffer QG were then added to 1 volume of gel (eg. 100 mg = 100 µl). The gel slice was next incubated at 50°C for 10 minutes until it had completely dissolved (the gel slice was vortexed every 2-3 minutes to help dissolve the gel). For DNA fragments ≥ 4kb in size, 1 gel volume of isopropanol was added to the mixture (eg 100mg= 100 µl of isopropanol). The mixture was next placed into a QIAquick spin column and centrifuged for 1 minute at 13000 rpm. The flow-through was next discarded, placed back into the collection tube and another 500µl of buffer QG was added to the column, which was centrifuged for 1 minute at 13000 rpm. To wash the DNA bound to the column, 750 µl of buffer PE was added to the column and centrifuged for 1 minute at 13000 rpm. With the flow-through discarded, another centrifugation step of 1 minute at 13 000rpm was done to remove any residual buffer PE. The QIAquick spin column was then placed into a clean eppendorf tube and 50 µl of dH_2O was added to the column and incubated at room temperature for 1 minute. The DNA was eluted by centrifugation for 1 minute at 13000 rpm.

2.4.16 Cloning

2.4.16.1 Cloning of PCR Fragments into Plasmid Vector

The plasmid vector used for cloning of PCR products was pGEM®-T Easy (Promega) which is a pUC-derived vector and contains dT overhangs at the cloning site in order to generate a sticky end. The sticky end is compatible with the terminal dA overhangs produced on any PCR product due to the template-independent action of certain *Taq* polymerases. This vector contains T7 and SP6 RNA polymerase promoters flanking a multiple cloning site (MCS) with the α-peptide coding region of the enzyme β-

galactosidase. Insertional inactivation of the α-peptide allows recombinant clones to be directly identified by colour screening on indicator plates (LB agar plus X-gal, IPTG and ampicillin). pGEM®-T Easy also contains a gene for ampicillin resistance for vector-positive clone selection and the *lacZ* promoter.

2.4.16.2 Ligations

Fifty ng of vector was ligated to an appropriate amount of purified PCR product in the pGEM®-T Easy buffer provided (Promega, UK) such that the ratio of molar ends of vector to PCR product was 1:3. One unit of T4 DNA ligase was used in a total reaction volume of 10 µl. Ligations were incubated at room temperature for 1 hour. Reactions were stored at -20°C following withdrawal of an aliquot for transformation of competent *E.coli* cells, JM109 (Promega, UK). Appropriate controls were also set up as per the manufacturers' instructions.

Ligation into pGL3 Basic vectors were performed using T4 DNA ligase at 1 unit in 20 µl total volume and incubated at room temperature for 5 minutes. Competent *E. coli* cells were used in the transformations. Reactions were again stored at –20°C.

2.4.16.3 Transformations

JM109 competent cells (Promega, UK) were used and transformed according to the manufacturers' recommendations. Briefly, 2µl of the ligation reaction (as in section 2.4.16.2) was mixed with 50µl of competent cells and incubated on ice for 20 minutes. The reactions were heat shocked at 42°C for 50 seconds and then replaced on ice for 2 minutes. Nine hundred and fifty µl of SOC medium was added to the reaction, which

was then incubated, with agitation at 37°C for 90 minutes to allow expression of the plasmid. Aliquots of 150 µl of the transformation reactions were plated onto LB agar plates containing ampicillin and X-gal and IPTG for colour selection.

For transformations using DH5α cells, 2 µl of ligation reaction was routinely used and was mixed with 100 µl of DH5α cells. The reaction mixture was incubated on ice for 30 minutes, heat shocked for 45 seconds at 37°C and returned to ice for a further 2 minutes. Pre-warmed SOC media was added to the cells at a volume of 950 µl and left to shake for 1 hour at 37°C at 225 rpm. Two hundred µl was then plated onto Ampicillin/X-Gal agar plates and placed at 37°C overnight.

2.4.17 Automated DNA Sequencing

The ABI 373a automated DNA sequencer and ABI 3100 Genetic Analyser were used routinely in this research. Sequencing reactions were carried out using ABI PRISM™ Big Dye™ Terminator Cycle Sequencing Kit with AmpliTaq® DNA Polymerase, FS versions 3.0 and 4.0 (Applied Biosystems). All sequencing was completed by the MRC Sequencing facility (Hammersmith Hospital, Imperial College, UK).

Essentially, the protocol provided by the manufacturer was followed. Three to ten ng of purified template DNA was cycle sequenced using the Big Dye™ AmpliTaq® dideoxy kit and an appropriate sequencing primer. The Big Dye™ reaction mix contains all of the required reagents: dye terminators; deoxynucleoside triphosphates; AmpliTaq® DNA Polymerase, FS; *rTth* pyrophosphatase; magnesium chloride and buffer. 4µl of Big Dye™ version 3.0 or 2µl of Big Dye™ version 4.0 reaction mix was used in a total reaction volume of 10µl, the remainder of which is made up with 3-10ng template

DNA, 1.6μM sequencing primer made up to volume with dH_20. The combined reaction volume of 10μl was then cycled on the Perkin Elmer 9600 PCR machine with the following cycling profile for 25 cycles:

Max. Ramp
96^0C 10 seconds

Max. Ramp
50^0C 5 seconds

Max. Ramp
60^0C 4 minutes

The cycle sequenced samples then underwent ethanol precipitation involving the addition of 60% ethanol, pelleting of the precipitate by centrifugation, washing of the pellet in 70% ethanol and finally air drying to remove any remaining ethanol. For analysis on an ABI 373a automated sequencer, the resultant pellet was resuspended in 3μl ABI loading dye and heated to 95°C for 3 minutes to separate the strands of DNA prior to loading 1.5μl of this product into the ABI gel. Samples for analysis on an ABI 3100 automated sequencer were resuspended in 10μl formamide (Applied Biosystems), 1-5μl of which was added to a microtitre plate for loading in a total volume of 10μl (made up to volume with additional formamide).

2.4.18 Primary Granulosa Cell Culture

Primary granulosa cells were isolated from follicular fluid, obtained from fully consented women undergoing *in vitro* fertilization at Queen Charlotte's Hospital, London, UK. Follicular fluid obtained from the second flush after the egg is expelled is put into universals (20 mls), leaving any large clumps behind and centrifuged for 5 minutes at 200*g* (1000rpm). After the clear supernatant was discarded, the pellets containing the granulosa cells were pooled together and the volume increased to 8 mls

with warmed PBS (at 37°C). An autoclaved glass pipette was filled with the cells and PBS mixture and overlaid onto a gradient of 45% percoll (2 percoll gradients are made up with 1% penicillin/streptomycin, and L-glutamine, 10 mls M199 with 1% FBS, 100 µl $NaHCO_3$ and 8.9 mls of percoll). The resulting mixture was centrifuged at 400*g* for 30 minutes. The pure granulosa cells were then aspirated from the percoll/PBS interface and placed into a tube containing more M199 media (pellets were pooled at this point) and the volume increased to 10 mls of M199 media. The cells were centrifuged for 5 minutes at 200*g* and the supernatant removed. The cells were gently resuspended in 1 ml of M199 with clumps of cells being broken up by putting them through a 1 ml syringe. The cells were again resuspended in 10 mls of M199 media and plated onto a large petri dish and placed into a 37°C incubator for no more than 30 minutes. The supernatant was then discarded and cells were made up to 1 ml with M199 media and spun in a microcentrifuge placed in the cold room for 5 minutes at 200*g*. Supernatant was removed and the cells were snap frozen in an ethanol/dry ice bath. Total RNA was isolated following the protocol described in section 2.4.7.2.

2.4.19 Cell Culture

HGL5 is a transformed human luteinized granulosa cell (GC) line immortalized from primary granulosa cells and was a kind gift from Dr. William Rainey at the University of Texas, USA (Rainey *et al.*, 1994). This cell line retained the ability to increase progesterone production and convert androstenedione to oestradiol in response to treatment with forskolin and dibutryl cyclic AMP and therefore represents an excellent model system with which to study *CYP11A1* expression (Rainey *et al.*, 1994).

The HGL5 cell line was maintained in complete media containing DMEM/F12, 1% penicillin/streptomycin, 0.01% gentamicin, 2% ultroser G and 1% ITS + supplement. Normal passaging of cells requires 1.4% trypsin per T75 flask of confluent cells, followed by 2 mls of Nu serum which functions to inactivate the trypsin. The cells and media were placed in a Falcon tube and spun at 200*g* for 5 minutes, the supernatant was discarded and the cells were resuspended in 10 mls of complete media. Alternatively, if cells were needed for total RNA, following the resuspension in 10 mls of complete media, the cells were centrifuged for 5 minutes at 200*g* at 4°C. The supernatant was removed and the tube containing the cells was placed in a bath of dry ice and ethanol to flash freeze the cells, followed by storage at -80°C.

2.4.20 Transient Transfections

Transient transfections were performed using both the pentanucleotide and the tetranucleotide repeats located in the promoter region and intron 1 respectively of the *CYP11A1* gene, to ascertain their function in cells. JEG-3 or 293 cells were plated in 96-well plates one day before the transfection experiment such that growth of 50-80% confluency was achieved. To determine transfection efficiency, pCMV-*Renilla* was co-transfected into the cells. In general, 20ng of pGL3 Basic vector, 5 ng of pCMV-*Renilla* vector and 25 ng of pCI vector was added to 0.15 µl of LIPOFECTAMINE ®reagent (Invitrogen, UK) was used in all transfections and complexes were formed following 15 minutes incubation at room temperature in a serum-free media (Optimem, Sigma, UK). All DNA was purified using the Qiagen kit as outlined in section 2.4.4. Five µl of the resulting complexes were added to each of the wells containing cells and returned to the incubator. Cells were incubated overnight at 37°C, in 5% CO_2. The next day, the media was removed and re-placed with fresh, charcoal stripped serum

containing media and either forskolin, 8-Br-cAMP or left untreated and returned to the incubator for an additional 24 hours.

Charcoal was used to strip the foetal bovine serum of endogenous steroids as so not to interfere with the exogenous steroids added. To do this, 100g of neutralized charcoal was added to double distilled water and allowed to mix completely for 10 minutes. The charcoal mixture was then placed at 4°C for one hour to settle the grains. Excess water was then poured off and 10g of T-80 dextran was added and stirred. Subsequently, 2 ml of 1M Tris (unbuffered) was added to achieve a pH of ~7.7. The mixture was then brought to a volume of 1 litre with double distilled water and incubated, with shaking in an environmental shaker, at 55°C for one hour. The solution was then poured into 50 ml centrifuge tubes (Invitrogen, UK) containing 25 mls of foetal bovine serum and centrifuged at 1650*g* for 60 minutes. The supernatant was removed and filtered twice under sterile conditions through a 0.45-micron syringe or bottle top filter (Invitrogen, UK). This was followed by filtration through a 0.22-micron filter using a peristaltic pump or another bottle top filter to ensure removal of any residual charcoal. Stripped serum was stored in 50 ml falcon tubes at -20°C.

Measurement of light production by the luciferase-DNA transfected cells lines was done using luclite luminescence reporter gene assay system from PerkinElmer the following day. 100 µl of substrate was added to each well containing 100 µl of cells/phenol red-free culture media and the plate was agitated gently under subdued light conditions. Plates were then sealed with the self-adhesive TopSeal-A™ and dark-adapted for 10 minutes before counting in a luminometer.

2.4.21 Genotyping

Microsatellites found in *CYP11A1* genomic region (pentanucleotide located at 466 bp from the start of transcription and a tetranucleotide located 841 bp into intron 1) were genotyped in all subjects (as described in section 2.1). The pentanucleotide was genotyped by PCR amplification in a 15µl volume containing 30ng genomic DNA; 10pmol of dNTPs (dTTP, dATP, dCTP and dGTP); 1.0mM $MgCl_2$; 1U *Taq* polymerase (Bioline, UK); and 8µM of each of the forward (FAM-labelled 5'-CCTGGGCAACAAGAATGAAA-3' for the UK subjects and FAM-labelled 5'-GGTGAAACTGTGCCATTGC-3' for the NFBC1966) and reverse (5'-GCCCAAGAGAAGGTCAGAGC-3' for the UK subjects and 5'-GTTTGGGGGAAATGAGGGGC-3' for the NFBC1966) primers. Sequence analysis of the *CYP11A1* gene revealed a second (tetranucleotide) microsatellite lying in intron 1 ~18.4kb upstream of exon 2 (see section 3.2.3). This microsatellite was amplified in a separate reaction (primers: HEX-labelled 5'-AGTAGCTGGGATTATAGACCC-3'(forward); 5'-GGAGACTGGTGAGGCTAAGT-3' (reverse); 1.5 mM $MgCl_2$ and the rest of the reagents as listed above for the pentanucleotide PCR). Resulting PCR products were diluted 1:4 in water, and 0.5 µl of each diluted PCR reaction mixed with deionised formamide and 450bp ROX size standards (Applied Biosystems, Warrington, UK), followed by separation and sizing on an ABI PRISM 3700 DNA sequencer (Applied Biosystems, Warrington, UK). Genotypes were derived using Genotyper software (Applied Biosystems, Warrington, UK). Where possible, Mendel-checking methods were used to confirm genotyping accuracy. In addition, approximately 10% of samples were re-genotyped. From these data, we estimate that the overall genotyping error is <1%. For convenience, alleles are designated by prefix "P" or "T" reflecting

pentanucleotide or tetranucleotide locus respectively, and a number reflecting the number of repeat units (e.g. P4 includes four repeats at the pentanucleotide).

2.4.22 *In silico* Analysis

Characterisations of differentially expressed sequences were carried out by BLAST searches http://www.ncbi.nlm.nih.gov/BLAST, by reference to the UniGene database http://www.ncbi.nlm.nih.gov/UniGene and the Human Genome Mapping Project http://www.esembl.org/. The genome browsers used were http://genome.ucsc.edu/cgi-bin/hgGateway and http://www.esembl.org/. The NCBI SNP database can be found on http://www.ncbi.nlm.nih.gov/SNP/.

3.0 GENOMIC STRUCTURE of *CYP11A1*

3.1 Introduction

CYP11A1 is a well-supported biological candidate gene for involvement in the aetiology of PCOS due to the linkage to the *CYP11A1* locus and association between a pentanucleotide repeat in the promoter region of this gene and serum testosterone levels in women with PCOS (Gharani *et al.*, 1997). The defects in steroidogenesis seen in cultured theca cells from PCOS women further supports a defect in the function of enzymes such as P450scc in the over production of androgens (see section 1.2.4) (Gilling-Smith *et al.*, 1997). However, it is possible that the pentanucleotide repeat identified by Gharani *et al* (Gharani *et al.*, 1997) is not the pathogenic mutation; the association may be due to the presence of linkage disequilibrium (see section 1.3.2.3) between the marker and a putative aetiological mutation located elsewhere, either within the *CYP11A1* gene, or in another gene in the region.

3.2 *CYP11A1*

Chung and colleagues first described the cloning of *CYP11A1* and expression in placenta in 1986 (Chung *et al.*, 1986). They mapped the gene to chromosome 15q23-q24 using a panel of 15 mouse-human somatic cell hybrids and demonstrated that it was expressed in placenta at mid-gestation (Chung *et al.*, 1986). The cDNA sequence described by Chung *et al.* (NM_000781) was 1821bp and encoded a protein of 521 amino acids (Chung *et al.*, 1986). A further study of *CYP11A1* by Morohashi *et al.*, in 1987 reported that the gene spanned approximately 29 kb and consisted of 9 exons separated by 8 introns (Morohashi *et al.*, 1987). As the complete sequence of the region

of chromosome 15q containing *CYP11A1* had become available (the finished sequence was available in late 2003), a bioinformatic analysis was initiated.

3.2.1 Bioinformatic Analysis

The nucleotide sequence of *CYP11A1* (NM_000781) was used in a BLAST search of both the non-redundant (nr) and human expressed sequence tag (dbEST) databases of the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/blast/). This identified a bacterial artificial chromosome (BAC), CTD-2311M21 (GenBank accession number AC090826), which contained the entire *CYP11A1* genomic sequence and a cDNA, AK056794, which was comprised of exons 2-9 of *CYP11A1* and 493bp of novel sequence at the 5' end. Nineteen clones representing the published *CYP11A1* sequence were also identified. The 493bp of novel sequence from the 5' end of AK056794 was then used in a BLAST search of dbEST and two uterine specific IMAGE clones, CM180-I22 (IMAGE number 3504861, GenBank accession number BE278699) and CM1250-I20 (IMAGE number 4553251, GenBank accession number BG337715) were identified.

3.2.2 Genomic Structure of *CYP11A1*

Comparison of the sequence of AC090826 with NM_000781 confirmed that the *CYP11A1* gene consists of 9 exons separated by 8 introns. Figure 3.1 (a) depicts a schematic diagram of the reported structure. In addition, both ends of all the introns conform to canonical GT/AG rule of intron exon boundaries, apart from intron 6 which begins with a GC at the 5' end, (Mount, 1982) as shown in black in table 3.1. However the splice donor site sequence AAG-gcaagc compares well to the consensus sequence (C/A)AG-gt(a/g)agt.

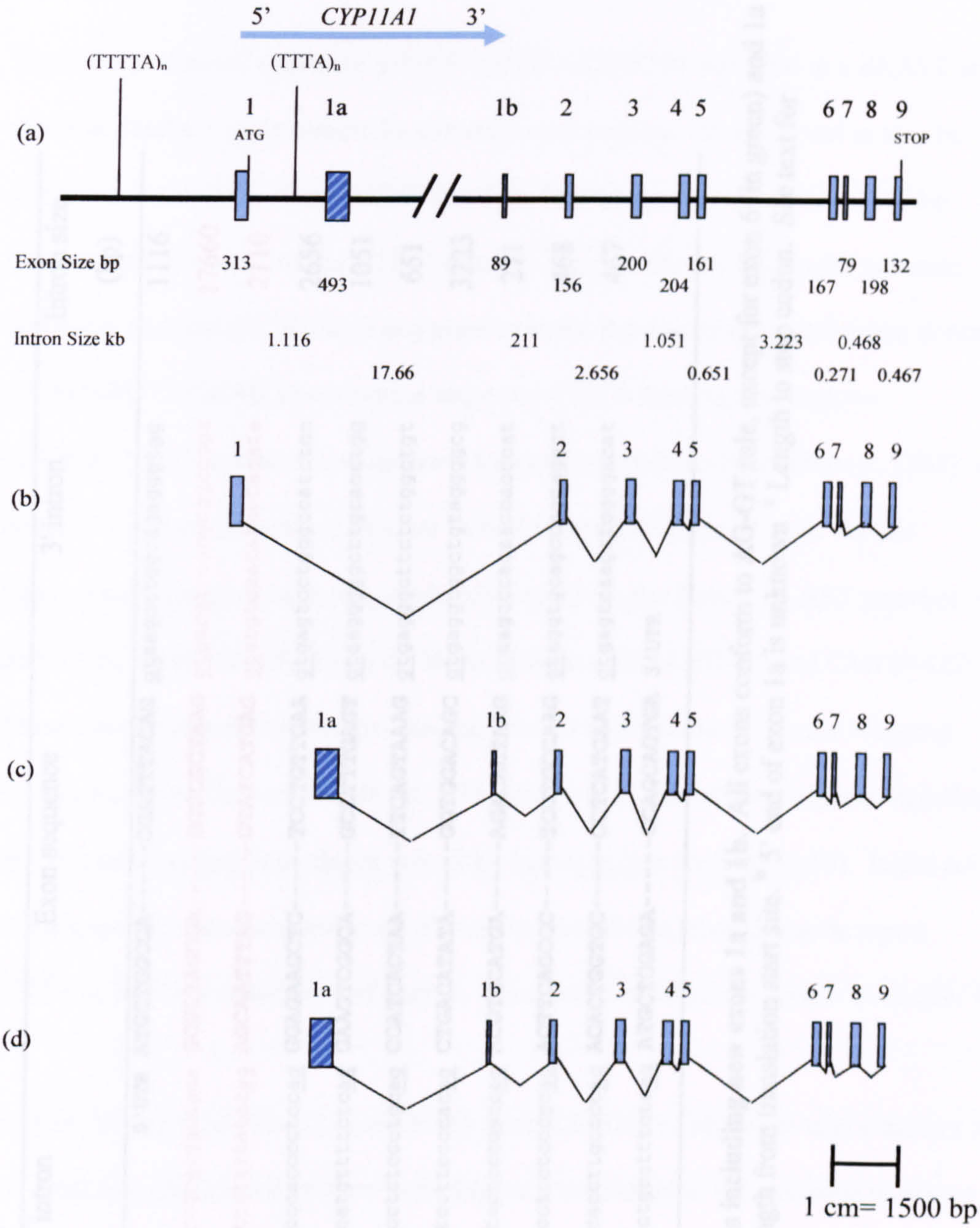


Figure 3.1: Genomic organisation of *CYP11A1* gene and EST clones. (a) Schematic representation of the *CYP11A1* genomic structure depicting 11 exons separated by 10 introns. Blue arrow indicates the direction of *CYP11A1* transcription. Exons are represented by blue boxes and are to scale relative to the introns; sizes of exons and introns listed below the exon boxes. Numbers above the boxes represent the number of each exon. Position of pentanucleotide repeat $(TTTTA)_n$ and tetranucleotide repeat $(TTTA)_n$ are shown. (b) Schematic diagram of *CYP11A1* clone NM_000781 (transcript I). (c) Schematic diagram of EST clone AK056794 (transcript II) and 2 identical IMAGE clones (CM180-I22 and CM1250-I20) identified though BLAST searches. (d) Schematic diagram of transcript III..

Exon	Exon size (bp)	5' intron	Exon sequence	3' intron	Intron size (bp)
1	313[a]		5'UTR ATGCTGGCCA-----CGATTTACAG	gtaagcctggcagagggtgg	1116
1a[b]	493	gtgacgtccgcgcggccgga	GCGCTAGCGA-----GCGCGCTAAG	gtgacgtccgcgcggccgga	17660
1b	89	gctgttcttgttttatacag	AGCAATTTAG-----GTATCATGAG	gtatgtgaatatactacata	2110
2	156	cctctcccccacccctccag	GGAGAAGCTC-----TCCTGTTGAA	gtgagtcctgggccatctcc	2656
3	200	gtgggcccatgttttctcag	GAAGTCGGCA-----GCCTTTGAGT	gtaaggggcttgcacctgg	1051
4	204	gcctcttcctctccctgcag	CCATCACTAA-----TTCAGTAAAG	gtgagggcttctctggctgt	651
5	161	tctttcctccttccccacag	CTGACATATA-----GGTGGACAGC	gtgaggtggctgtaggggcg	3223
6	167	tataccctactccccaccag	ACGTCCATGA-----AGACACTAAG	gcaagcccacaaccacccat	271
7	79	tgagtgccctcctcccgcag	ACTTCACCCC-----TCCTGCCAAG	gtaggtgcagccagcaggct	468
8	198	ctctgatcaccttgccccag	ACACTGGTGC-----CCTCATCAAT	gtgagtcaagctggggacat	467
9	132[c]	ctaccatctgcctttctcag	ATGCTGGAGA-----CCAGCAGTGA	3'UTR	

Table 3.1: *CYP11A1* Splice Junctions including new exons 1a and 1b. All exons conform to AG-GT rule, except for exon 6 (in green) and 1a where the sequence is unknown. .[a] Length from translation start site. [b] 5' end of exon 1a is unknown [c] Length to stop codon. See text for description of exons 1a and 1.

3.2.3 Identification of New *CYP11A1* Exons

The 493 bp of novel sequence at the 5' end of AK056794 was used in a BLAST search of the nr database to determine its chromosomal location. It was found to also be contained in the BAC clone AC090826 described above and was located 840 bp downstream of exon 1 of the *CYP11A1* gene. Due to this position in the genomic sequence, and that the genomic sequence contains the universally established donor (C(A)AG/GTA(G)A**GT**) consensus sequences (but is missing the acceptor (T(C)NC(T)**AG**) consensus sequences) for intron/exon boundaries (Mount, 1982), the novel 493 bp was designated as exon 1a in this thesis (Figure 3.1 (c)). As this represented a putative exon not previously described, the further BLAST searches were performed which identified two the IMAGE clones (CM1250-I20 and CM189-I22). These clones were subsequently obtained from the UK Human Genome Mapping Project Resource Centre (HGMP-RC) (Figure 3.1 (c)). Restriction enzyme digestion analysis showed that both clones were 2 kb in length (see section 2.4.10). Initial review of the genomic sequence upstream of exon 1a identified a tetranucleotide repeat $(TTTA)_n$, 824 bp downstream of exon 1. This repeat is discussed further in chapter 4.

As only 541 bp (CM180-I22) and 689 bp (CM1250-I20) of sequence was available on the database, sequencing primers were designed to determine the complete sequence of both clones and to ascertain any further sequence not found in AK056794. The complete list of these primers is presented in table 3.2. The resulting sequence was then compiled using the sequence analysis program LASERGENE (DNA STAR, Wisconsin. See section 2.1.11). Upon sequence analysis, no additional sequence to the 493 bp identified from clone AK056794 was discovered and both clones were found to be identical.

Primer Name	Sequence 5'→3'
IMAGE 1F	ggagtcggtttatgtcat
IMAGE 1R	ggagtggcccagctgagct
IMAGE 2F	ggaccagctgtagcatcgtt
IMAGE 2R	ccatctaccagatgttcca
IMAGE 3F	agacccagcattcctcc
IMAGE 4F	cgacccggtatgcgttt
IMAGE 5F	cccagaacgattcctcat
IMAGE 6F	cctcaaagccagcatcaa
IMAGE 7F	cctcattctgatgcctga
IMAGE 3R	tcttgtctttgctcagcc
IMAGE 4R	ccagcatctctgtgacgt
IMAGE 5R	ttcttggtggcctctgg

Table 3.2: Primers for sequencing of *CYP11A1* IMAGE clones. Sequencing conditions are as described in section 2.4.17. See text for details.

3.2.4 Expression Analysis of *CYP11A1* Alternate Transcripts

Information on the NCBI dbEST reveals that the cDNA AK056794 containing exon 1a was isolated from a placental library. To determine the authenticity of this expressed sequence in placenta, PCR was performed using primers designed in exon1a and exon 2 (called exon 1aF and exon 2R respectively). PCR was carried out using placental cDNA purchased from Invitrogen (UK). These primers were expected to produce a fragment of 150 bp (sequences shown in table 3.3 and their positions shown in figure 3.2). Primers were also designed in exon 1 of *CYP11A1* (exon 1F and exon 1R) to use as a positive control for PCR and these were expected to produce a fragment of 307 bp. PCR was performed as described in section 2.4.11 with specific cycling conditions outlined in table 3.3. The resultant products were run out on an agarose gel (figure 3.3).

PrimerName	Sequence 5'→3'	$MgCl_2$ mM	Annealing° C
Exon1 F	gctgaagtggagcaggtaca	1.5	52
Exon1 R	gcatttagcccgatatgaac	1.5	52
Exon1a F	gacagtggcgcgctaag	1.5	52
Exon1a R	gaatcgcgcggtgacag	1.5	52
Exon1a 2F	tattccgaaattgggtccgc	1.0	59
Exon1a 2R	tctttgcgtatggcccagcg	1.0	59
Exon2 R	cattatgaccactatccgc	1.0	52
Exon1bF	gcaatttagaaaatacaga	2.0	50
Exon1bR	gagtactatgaatttacaat	2.0	50
Exon1RRACE	cgtgagtctttccagaccgtcggaa	1.0	68
Exon1a RACE	acctggagcgggactccatttgcgg	1.0	68
Exon1b RACE	ccttaactaacatagactgaagtta	1.0	68
Exon1bR NRACE	acaaagagacataaaagatttaacg	1.0	68
Exon2R RACE	cctgaggatacccagagaccattat	1.0	68

Table 3.3: PCR primers for analysis of alternate transcripts of *CYP11A1*. Primer sequences, with the associated concentration of $MgCl_2$ and annealing temperatures, are listed. PCR cycling conditions are described in section 2.4.11.

All the expected bands described above were obtained (figure 3.3, lanes 1 and 3); however, an extra band of 240 bp was obtained with exon 1aF and exon 2R primers (figure 3.3, lane 3). Therefore, the 150 bp and 240 bp PCR fragments (obtained using the exon1aF and exon2R primers) were cloned into pGEM®-T easy vector and fully sequenced. As expected from the sequence, the 150 bp fragment was found to consist of exon 1a and exon 2 spliced together. However, the 240 bp fragment was found to contain the expected sequence of exon 1a and exon 2 but also 89 bp of novel sequence.

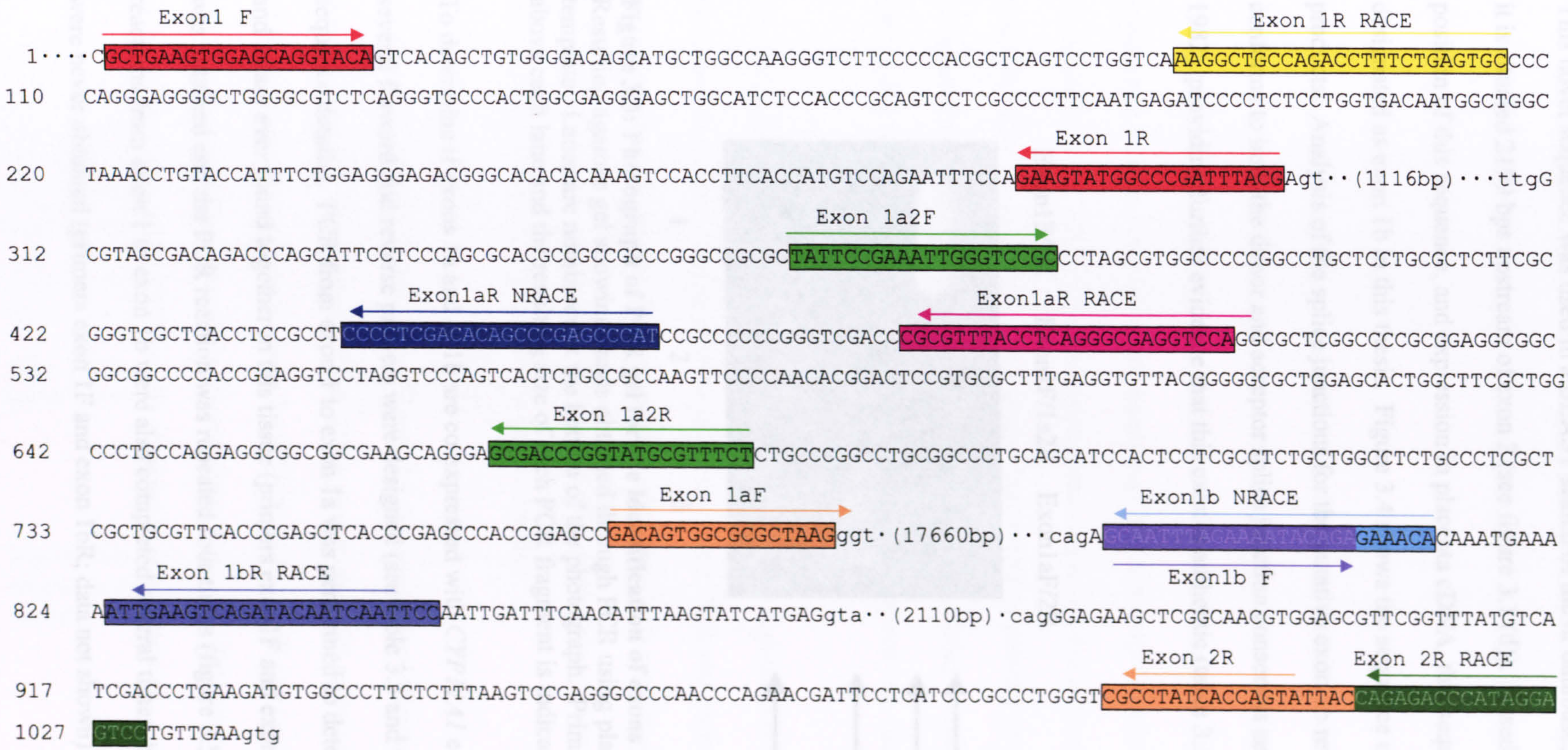


Figure 3.2: Sequence of *CYP11A1* Exons 1,1a, 1b and 2 for RT-PCR and RACE strategy. All PCR primers are as listed. Arrows are in the direction of PCR and sequences are shaded with names of primers directly above boxes. Sequence of exons (in order 1, 1a, 1b and 2) are in uppercase, partial intron sequence in lower case and the intron sizes are represented numerically in brackets.

This novel sequence was used in a BLAST search of the nr database, which showed that it is situated 2199 bps upstream of exon 2 (see figure 3.1 (d)). Based on the genomic position of this sequence, and expression in placenta cDNA, this sequence was designated as exon 1b in this thesis. Figure 3.4 shows the sequence of the two PCR products. Analysis of the splice junctions for the putative exon 1b reveals this exon conforms to both the donor and acceptor splice junction consensus sequences (Mount, 1982), providing further evidence that this exon is authentic (table 3.1).

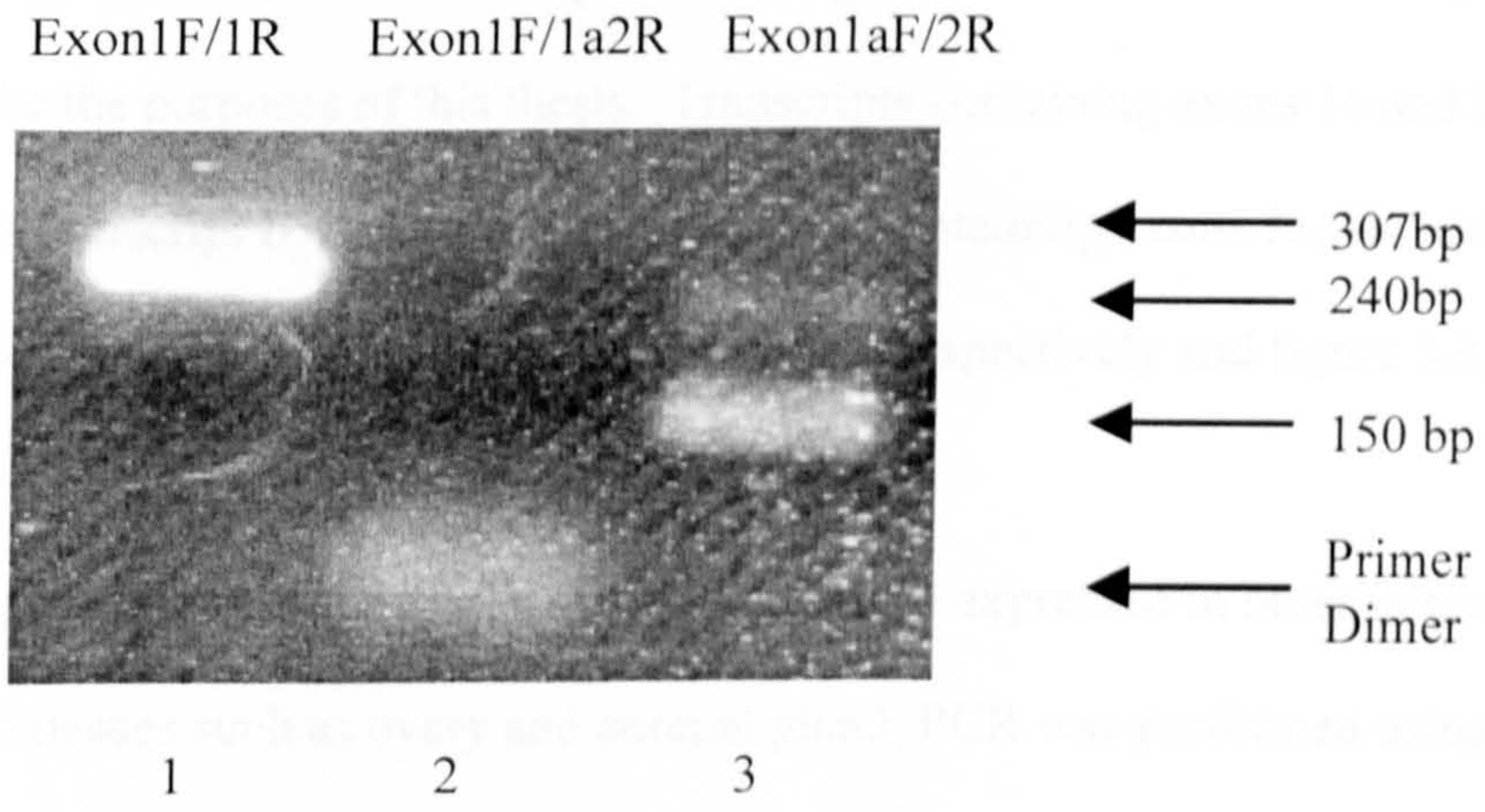


Figure 3.3: Photograph of PCR gel for the identification of exons 1a and 1b: Resulting agarose gel showing bands obtained through PCR using placental cDNA as a template. Lanes are numbered at the bottom of the photograph. Primer pairs used listed above each lane and the resulting size of each PCR fragment is indicated by the arrows.

To determine if exons 1a and or 1b are co-expressed with *CYP11A1* exons in placenta, several forward and reverse primers were designed (see table 3.3 and figure 3.2 for sequence details). PCR from exon 1 to exon 1a was performed to determine if exon 1 and 1a are ever found together in this tissue (primers exon1F and exon1a 2R). No band was obtained and the PCR reaction was repeated four times (figure 3.3, lane 2). PCR reactions from exon 1 to exon 1b were also completed several times, however, bands were never obtained (primers exon 1F and exon 1bR; data not shown). Therefore, the

PCR evidence indicates that two distinct transcripts exist containing exon 1a with or without exon 1b. Taken together, these data imply that *CYP11A1* encodes several transcripts.

Therefore, based on the work presented thus far, initial BLAST searches of the nr database and subsequent PCR analysis revealed the existence of two putative novel exons (1a and 1b) for *CYP11A1*. These data indicate the presence of two alternate transcripts, over and above the previously reported transcript containing exons 1-9 (Morohashi *et al.*, 1987). The transcript containing exons 1-9 is therefore designated as transcript I for the purposes of this thesis. Transcripts containing exons 1a and 2-9 are designated as transcript II whereas those transcripts containing exons 1a, 1b, and 2-9 are designated as transcript III (see figure 3.1 (c) and (d) respectively and figure 3.4).

To determine if exon 1a and exon 1b of *CYP11A1* were expressed in other important steroidogenic tissues such as ovary and adrenal gland, PCR was performed using cDNA purchased from Invitrogen (UK) with the resultant products separated out on an agarose gel (see figure 3.5). Again, as with placental PCR fragments, expected sizes of products were 150 bp (transcript II) and 240 bp (transcript III).

As shown in figure 3.5, both ovary (figure 3.5 (a)) and adrenal (figure 3.5 (b)) expressed exons 1a and 1b. The integrity of the cDNA used in the PCR reaction was verified by PCR using specific primers for exon 1 of *CYP11A1* (exon 1F and exon 1R) as this transcript is normally expressed in these tissues. Figure 3.5 (a) and (b), lane 1, show the expected PCR fragment of 307 bp was obtained in these reactions.

```
Exon 1a-2 (150bp)       CGCTGGAGCACTGGCTTCGCTGGCCCTGCCAGGAGGCGGCGGCGAAGCAGGGAGCGACCCGGTATGCGTTTCTCTGCCCGGCCTGCGG
Exon 1a-1b-2 (240bp)    CGCTGGAGCACTGGCTTCGCTGGCCCTGCCAGGAGGCGGCGGCGAAGCAGGGAGCGACCCGGTATGCGTTTCTCTGCCCGGCCTGCGG

Exon 1a-2               CCCTGCAGCATCCACTCCTCGCCTCTGCTGGCCTCTGCCCTCGCTCGCTGCGTTCACCCGAGCTCACCCGAGCCCACCGGAGCCGACA
Exon 1a-1b-2            CCCTGCAGCATCCACTCCTCGCCTCTGCTGGCCTCTGCCCTCGCTCGCTGCGTTCACCCGAGCTCACCCGAGCCCACCGGAGCCGACA

Exon 1a-2               GTGGCGCGCTAAG----------------------------------------------------------------------
Exon 1a-1b-2            GTGGCGCGCTAAGagcaatttagaaaatacagagaaacacaaatgaaaaattgaagtcagatacaatcaattccaattgatttaaca

Exon 1a-2               ----------------GGAGAAGCTCGGCAACGTGGAGTCGGTTTATGTCATCGACCCTGAAGATGTGGCCCTTCTCTTTAAGTCCGAGG
Exon 1a-1b-2             tttaagtatcatgagGGAGAAGCTCGGCAACGTGGAGTCGGTTTATGTCATCGACCCTGAAGATGTGGCCCTTCTCTTTAAGTCCGAGG

Exon 1a-2               GCCCCAACCCAGAA
Exon 1a-1b-2            GCCCCAACCCAGAA
```

Figure 3.4: Alignment of *CYP11A1* novel exons 1a and 1b with exon 2: Exons 1a and 1b were identified through PCR using placental cDNA and were cloned and completely sequenced. Exons 1a shown in uppercase; exon 1b shown in lowercase and shaded yellow, exon 2 is boxed. Sizes for each PCR product are shown in parenthesis.

BLAST searches of human dbEST revealed the existence of 10 ESTs that contain exon 1a and one clone that contains exon 1a and exon 1b and no clone that contains only exon 1b (table 3.4).

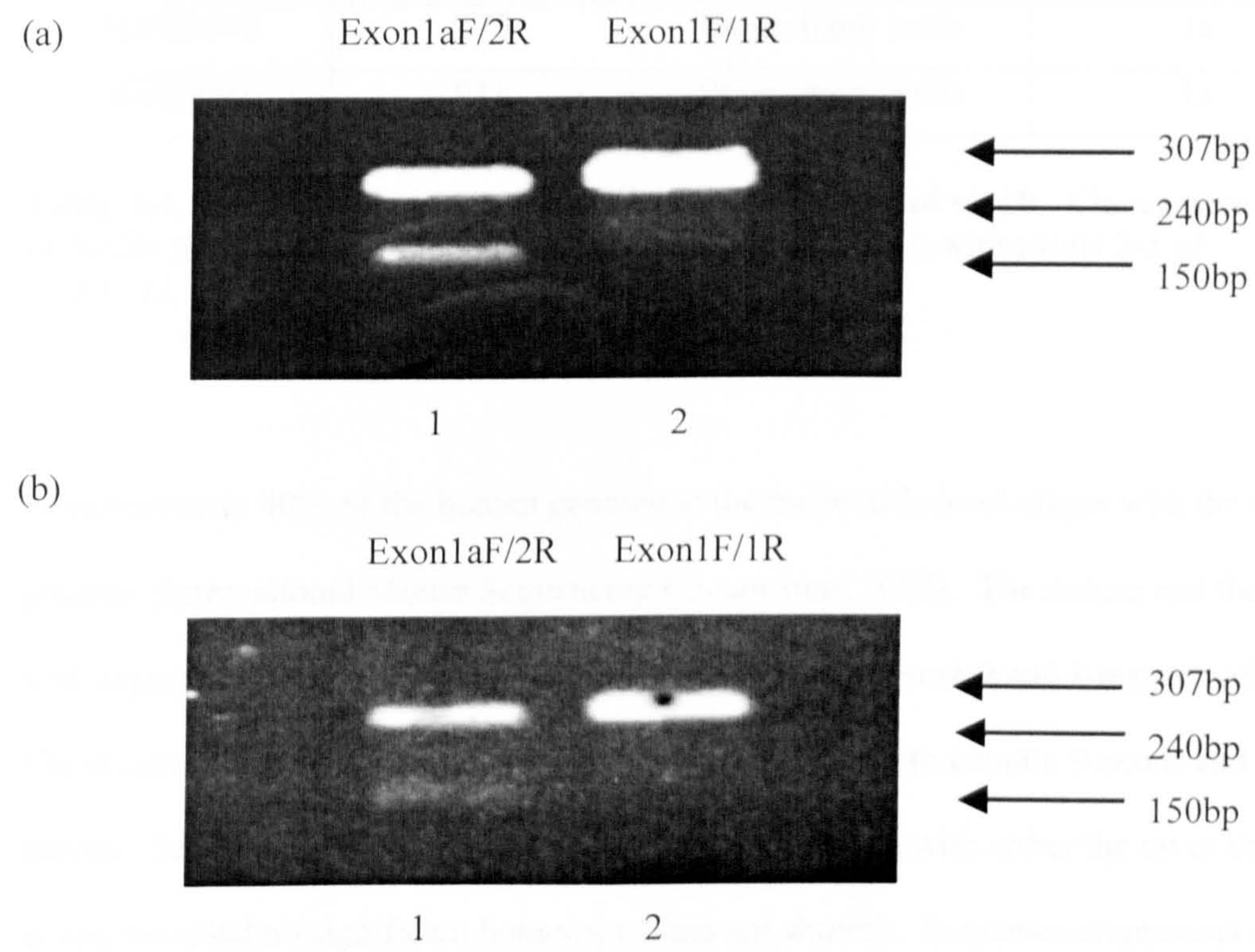


Figure 3.5: Photograph of PCR gel of transcripts I, II and III in ovary and adrenal cDNA. (a) Ovary cDNA; Lane numbers listed below the photograph; primer pairs used listed above each lane. (b) Adrenal cDNA; details as in (a). Arrows indicate size of transcripts.

EST Name	Size (bp)	Tissue	Exon(s)
CB118908	540	Liver	1a
BG337715	850	Uterus (tumour)	1a, 2, 3
BX379767	1042	Placenta	1a, 1b, 2, 3,4
AL555611	857	Hela	1a, 2, 3
AL540502	899	Pre-eclamptic placenta	1a, 2, 3, 4
BM834462	392	Stomach	1a
BI829814	209	Brain	1a
BI756987	732	Brain	1a, 2, 3, 4, 5
BI489642	775	Brain, lung, testis	1a
BI820991	911	Brain, lung, testis	1a

Table 3.4. EST clones containing *CYP11A1* exons 1a and or 1b. Clones were found on NCBI EST database to contain either exons 1a and or 1b with exons 2-5 of *CYP11A1*.

Approximately 40% of the human genome at the nucleotide level aligns with the mouse genome (International Mouse Sequencing Consortium, 2002). The mouse and the rat both express *CYP11A1* with the gene mapped to chromosomes 9 and 8 respectively. The mouse and the rat genomic sequence have been found to contain 9 exons and 8 introns. Sequence comparisons of human exons 1a and 1b with either the rat or the mouse revealed no significant homology (data not shown). Sequence alignments of exon 1 of human and mouse reveal an 82% homology whereas exon 2 shows 77% homology. Furthermore, the structure of intron 1 in the mouse is just over 1 kb whereas human intron 1 extends over 17660 bp. This suggests that transcript II and transcript III may not exist in mouse or in rat, indicating a more complex organisation in humans.

3.2.5 Northern Analysis of *CYP11A1* Alternate Transcripts

To determine the size of the novel *CYP11A1* transcripts, probes were made for Northern analysis using sequence within exon 1a. To do this, PCR was performed and the 328 bp fragment obtained from the reaction using primers exon 1a 2F and exon1a 2R (see table 3.3 for primer sequences and section 2.4.11 for methodology) was excised from an agarose gel and purified (section 2.4.16). This fragment was then cloned into pGEM®-T easy vector and fully sequenced to ensure correct sequence of the clone, as detailed in section 2.4.16 and 2.4.17 respectively. The fragment was then cut out of the pGEM®-T easy vector using *Eco*R1 restriction enzymes and purified. Probes were prepared using the oligolabelling kit from Amersham Pharmecia Biotech (UK), based on the random priming technique as described in detail in section 2.4.8. The PCR exon 1a specific fragment was then used as a probe on a human multiple tissue Northern Blot (Clontech, UK). This Northern Blot contains 2μg of polyA+ RNA from 8 human tissues transferred onto a positively charged nylon membrane. This type of blot can normally detect transcript sizes ranging from 0.5 kb to 10 kb. Blots were hybridised in Clontech ExpressHyb for 1 hour at 68°C, followed by a series of washes and exposed to a film for 24 hours (as outlined in detail in section 2.4.8). The resultant autoradiograph shows that the exon1a probe detects a transcript of 2.5 kb which is expressed in placenta but not in heart, brain, lung, liver, skeletal muscle, kidney or pancreas (figure 3.6 (a)). The blot was then stripped according to the manufacturers' instructions (see section 2.4.8.3) and exposed to film to ensure the previous probe had been completely removed. The blot was then re-probed with a probe from exon 1 of *CYP11A1* (transcript I). The exon 1 probe (307 bp in size, see figure 3.6 (b)) was made in the same way as described above using specific primers listed in table 3.3 (Exon1F and Exon1R). This exon 1 probe detected a transcript of 2.2 kb, which confirms the size reported by Chung *et al.*, (Chung

et al., 1986) (figure 3.6 (b)). Therefore, there is a difference of approximately 300 bases between the transcripts identified in these two blots (2.5 kb and 2.2 kb). As exon 1 and exon 1a are 313 bp and 493 bp in length respectively, this also corresponds to a difference of 180bp. Therefore, the difference between the transcript sizes detected on the Northern Blot (~300bp) and the exon sizes reported here (180bp), indicates there is a difference of approximately 120 bp of exon sequence. This sequence may represent an unidentified exon or further exon 1a sequence. Figure 3.6 (c) shows the same blot hybridised with a probe specific for the housekeeping gene, *ACTB*, to show loading of polyA RNA on the blot.

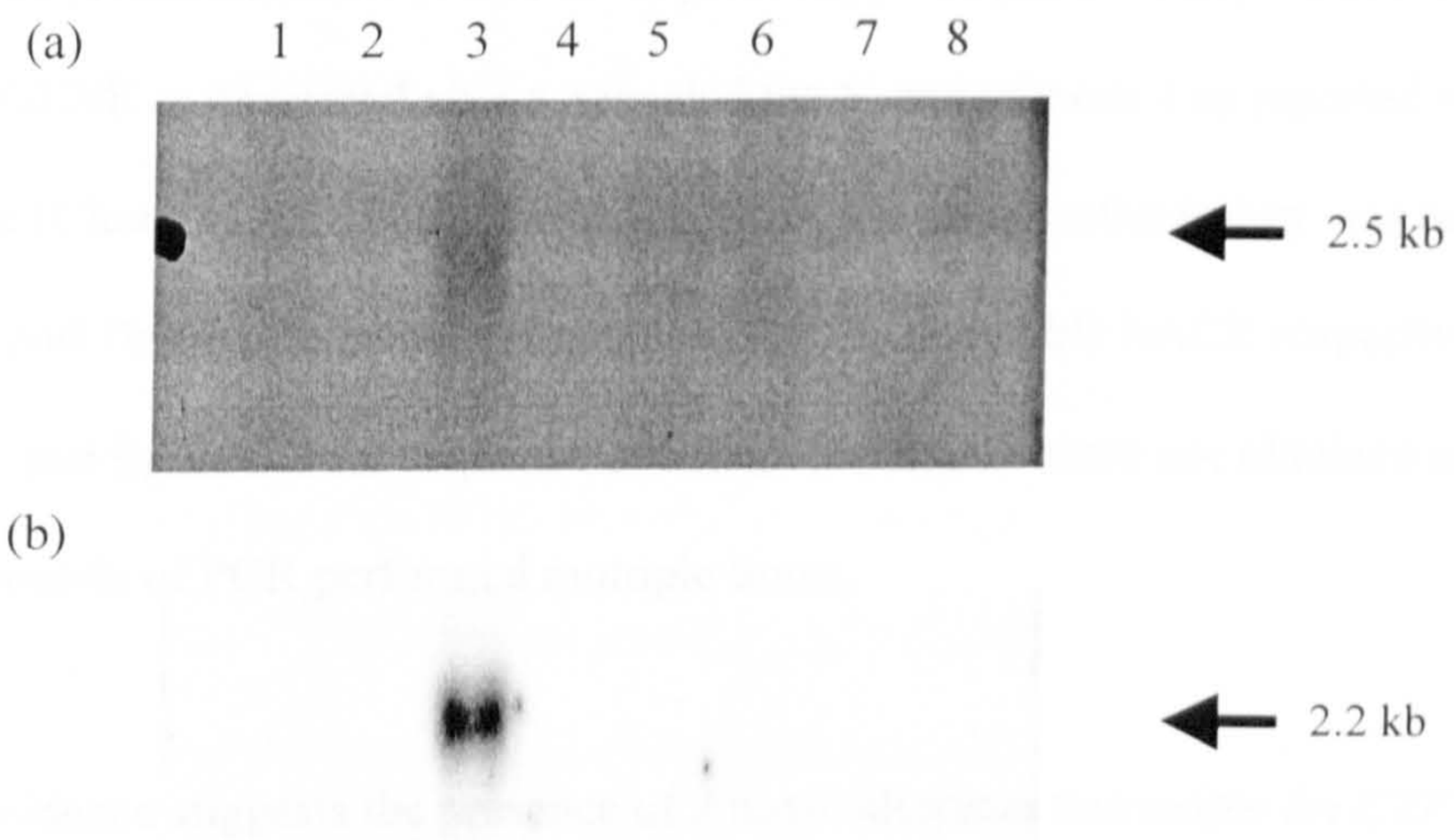


Figure 3.6: Panel of Autoradradiographs of Clontech Northern Blots. Blots contain 2μg of polyA RNA in each lane. a) *CYP11A1* exon 1a specific probe; b) *CYP11A1* exon 1 specific probe c) *ACTB* specific probe. Arrows indicate size of transcripts. Lane 1: heart; lane 2:brain (whole); lane 3: placenta; lane 4: lung; lane 5: liver; lane 6: skeletal muscle; lane 7: kidney; lane 8: pancreas.

3.2.6 5' RACE

Based on the Northern results and the respective sizes of exons 1 and 1a, approximately 120bp of sequence is unknown. Therefore, primers were designed in exons 1a and 1b for 5' rapid amplification of cDNA ends (RACE) as described in section 2.4.13. RACE is a method to identify the 5' and 3' ends of putative genes using gene specific primers and an adapter primer for PCR. Initially, to ensure that the methodology was working, 5' RACE using primer exon1R RACE (sequence in table 3.3, figure 3.2) was performed using ovary as template (Clontech Marathon Ready Ovarian RACE kit, Clontech, UK). The resulting PCR products were cloned into pGEM®-T easy plasmid and fully sequenced (section 2.4.16 and 2.4.17 respectively). Sequence analysis, utilising LASERGENE as described above, revealed the 5' end of exon 1 as reported in the literature (Chung *et al.*, 1986). However, when the same methodology was used with exon 1a and 1b specific primers (exon 1aR RACE, exon 1bR RACE respectively; see table 3.3 and figure 3.2 for sequence details), clean bands were not obtained after several rounds of PCR performed multiple times.

As the evidence suggests the presence of 2 novel alternate transcripts for *CYP11A1*, examination of the sequence obtained to date was performed to identify putative open reading frames (ORFs). For CYP11a (transcript I), the reported translational start site begins at +62 of exon 1 and the protein is 521 amino acids in length (Morohashi *et al.*, 1987) (figure 3.7 (a)). For transcript II, the longest ORF starts in exon 3 to the end of *CYP11A1* (see figure 3.7 (b)). No ORF was found that runs from the 5' end of exon 1a to exon 9 as all frames contain stop codons. The ORF beginning in exon 3 (figure 3.7 (b)) is preceded by an in-frame stop codon located in exon 1a. This means that exon 1a may be a non-coding exon and that the protein encoded by transcript II has a translation

start site somewhere else in the gene i.e. at the methionine found in exon 3. This would encode a protein of 363 amino acids. BLAST searches of the protein nr database of NCBI did not reveal significant homology of this hypothetical protein to any on the database other than that of CYP11a. As the 5' end of transcript II is unknown, a functional methionine may be upstream of the present sequence and as a result, code for a protein to this stop codon. Analysis of 1110 bp of putative promoter region located upstream of exon1a (using the program GENOMATRIX) did not reveal any evidence for transcription factor binding sites or other elements necessary for transcription (data not shown).

Sequence analysis of transcript III revealed that the longest ORF begins in exon 1b (figure 3.7(c)) as an in-frame methionine was found 5 bases from the end of this exon. BLAST searches of this translated sequence of 434 amino acids to the NCBI nr protein database did not reveal significant homology of this hypothetical protein to any on the database other than to CYP11a. An in-frame methionine was found in exon 1a; however, this was followed by an in-frame stop codon in exon 1b (figure 3.7 (c)). This could imply that a protein is being coded for beginning in unknown 5' exon 1a sequence to this stop codon.

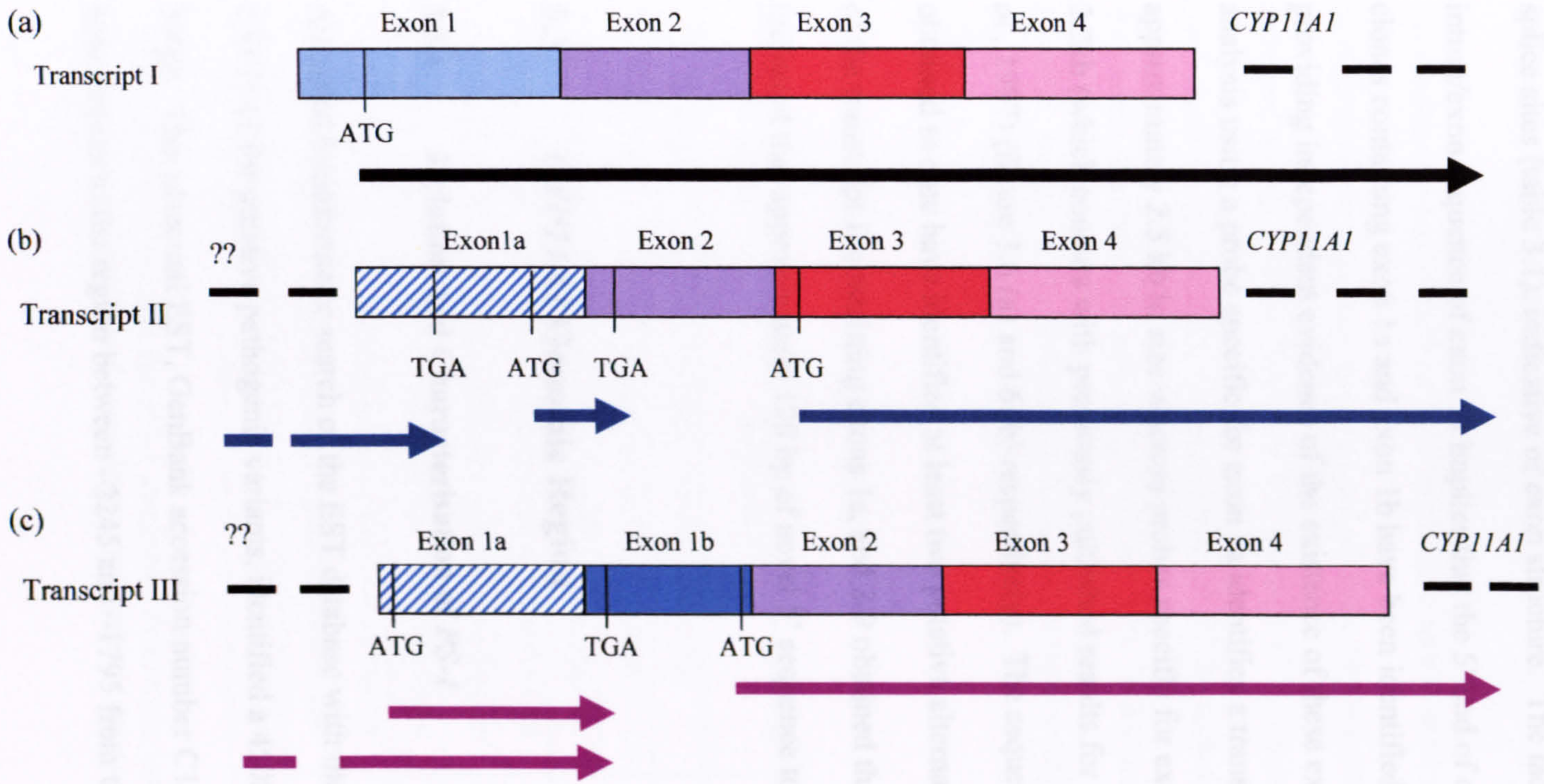


Figure 3.7: Schematic diagram of possible ORFs for transcript I, transcript II and transcript III. Possible ORFs are shown for each alternate transcript of *CYP11A1*. (a) Transcript I. Exons are represented by boxes with names listed above each box. Arrows are in the direction of translation. ATG is the translation start codon; TGA is the translation stop codon as written. (b) Transcript II. Details as in (a) (c) Transcript III. Details as in (a).

Therefore in summary, two novel *CYP11A1* exons have been identified which indicates the genomic structure consists of 11 exons, separated by 10 introns. Expression has been demonstrated in cDNA and in several steroidogenic tissues such as placenta, adrenal gland and the ovary (figure 3.3 and figure 3.5). Furthermore, exon 1b has both acceptor and donor intron/exon splice sites, while exon 1a has only donor intron/exon splice sites (table 3.1), indicative of exon structure. The lack of the acceptor intron/exon sequence of exon 1a implies that the 5' end of exon 1a is missing. Several clones containing exon 1a and exon 1b have been identified on public databases, providing independent evidence of the existence of these exons (table 3.4). Northern analysis using a probe specific for exon 1a identifies a transcript in placenta of approximately 2.5 kb in size whereas probes specific for exon 1 produce a band of 2.2kb (which concurs with previously published results for *CYP11A1*) (Morohashi *et al.*, 1987) (figure 3.6 (a) and 6 (b) respectively). The sequence and expression data obtained to date have identified at least two putative alternate transcripts. The total size of the transcript II containing exons 1a, and 2-9 obtained through Northern analysis indicated that approximately 120 bp of novel 5' sequence to exon 1a is missing.

3.3 *CYP11A1* Genomic Region

3.3.1 Isolation and Characterisation of *PS-1*

An initial bioinformatic search of the EST database with the 5' flanking region of *CYP11A1* for putative pathogenic variants, identified a 450bp placental EST (Gharani, 2000). This placental EST, GenBank accession number C18438, was shown to be 99% homologous to the region between –2245 and –1795 from the reported transcription site

of *CYP11A1* (Morohashi *et al.*, 1987). The clone was named placental-specific-1 (*PS-1*) for this thesis.

3.3.1.1 Expression Analysis of *PS-1*

To confirm the expression of *PS-1* in placenta and to ascertain expression in other steroidogenic tissues such as ovary and adrenal, internal primers PS-1F, PS-1R were designed within C18438 (see table 3.5, figure 3.11 for sequence details). PCR was performed using placental, ovary and adrenal cDNA as template and the resultant fragments run out on an agarose gel (figure 3.8). The expected band size of 256bp based on sequence was obtained in both placenta and ovary, but not adrenal (see figure 3.8).

Primer Name	Sequence 5'→3'	$MgCl_2$ mM	Annealing °C
PS-1F	cctcatcctctccaatcgc	1.5	52
PS-1R	tgggagccttcgacacac	1.5	52
PS-1 F RACE	attatctcacagcttccgagggtgc	1.0	68
PS-1 R RACE	agcataggccagcgattcagtttgg	1.0	68
PS-1 F NRACE	acaaattatcccaaaatttagtggc	1.0	68
PS-2F RACE	ccagtagtcactccgttttgtccga	1.0	68
PS-2 F NRACE	ccaagtactgactactccatcacca	1.0	68
PS-3 F RACE	ggctccactctggacacactctaa	1.0	68
PS-3 F NRACE	caaatctgggtcacctggggagaga	1.0	68

Table 3.5: Primers used for *PS-1* analysis. Primer sequences, with the associated concentration of $MgCl_2$ and annealing temperatures, are as listed. See text for details.

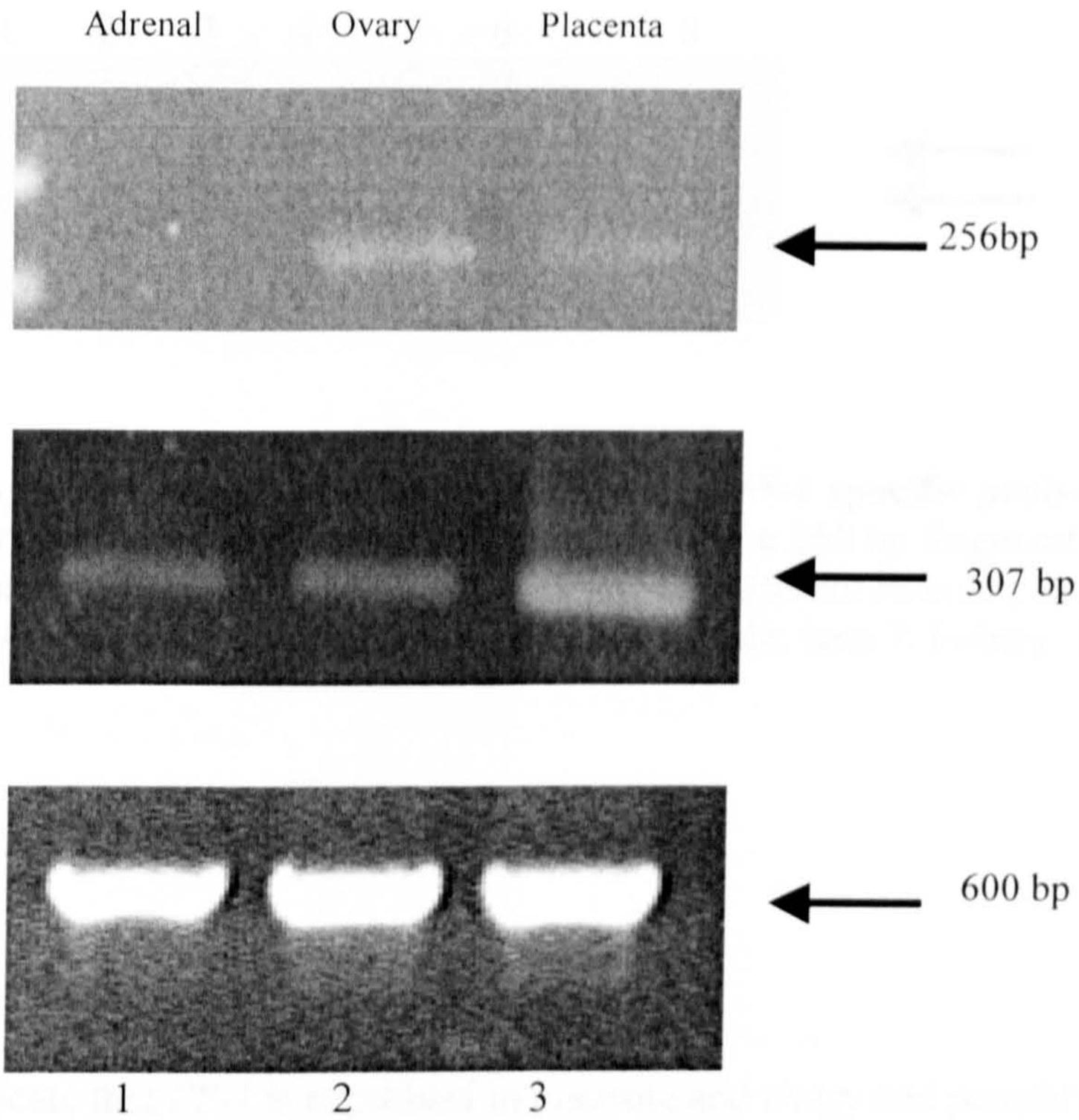


Figure 3.8: Photograph of RT-PCR gel of *PS-1*, *CYP11A1* and *ACTB* in various steroidogenic tissues. Primers and conditions listed in table 3.5. Lanes listed at bottom of the photograph. Tissues used listed above each lane. Arrow indicates size of band obtained.

3.3.1.2 Northern Analysis of *PS-1*

To determine the approximate size of the *PS-1* transcript, a probe for Northern analysis was generated. The original 256 bp PCR product obtained from placenta (see figure 3.8) was isolated and subcloned into pGEM®-T easy plasmid to use as a probe, as outlined in section 2.4.16. A Clontech Northern blot was hybridised with the probe specific for *PS-1*, incubated for 1 hour, as described in sections 3.2.3 and 2.4.8.2, and exposed to film for 7 days. The probe detects two transcripts of 7.5kb and 6.5kb in placenta but not in heart, brain, lung, liver, skeletal muscle, kidney or pancreas (figure 3.9).

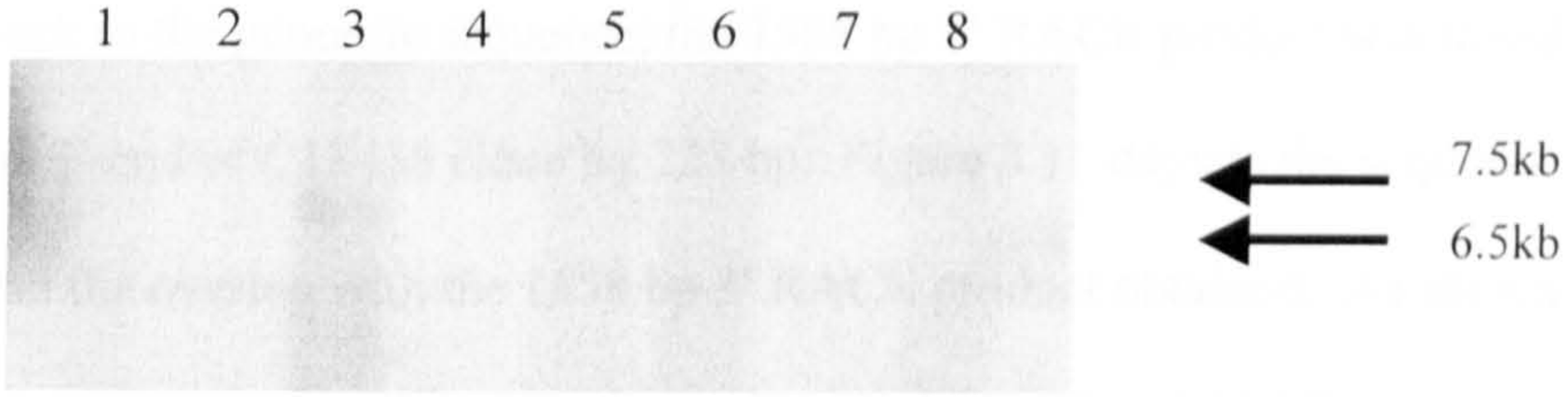


Figure 3.9: Autoradiograph of Northern Blot using *PS-1* specific probe: Blot contains 2 µg polyA RNA in each lane. Hybridised with a 256 bp fragment specific for *PS-1* and exposed to film for 7 days. Lane1: heart; lane 2:brain (whole); lane 3: placenta; lane 4: lung; lane 5: liver; lane 6: skeletal muscle; lane 7: kidney; lane 8: pancreas.

3.3.1.3 5' and 3' RACE

These data indicate that *PS-1* is expressed in placenta and ovary and possibly represents a novel gene, not previously known. The only gene annotated on the genomic databases in the region is *SEMA7A*, located 42 kb further 5' than *CYP11A1*. The product of this gene is thought to play a role in the nervous system and in modulating immune function and is not expressed in placenta. This implies that *PS-1* is unlikely to be part of *SEMA7A*. As only 450 bp of sequence was known of *PS-1*, and Northern analysis indicates the existence of two transcripts of 7.5 and 6.5 kb, more sequence needed to be obtained. Therefore, both 5' and 3' RACE were performed using the Clontech Marathon Ready Ovarian RACE kit (Clontech, UK) as before (see section 2.4.13 and section 3.2.6 respectively). As described earlier, RACE is a methodology that can be used to identify 5' and 3' ends of genes. For 3' RACE, a primer was designed in the 3' end of C18438 called PS-1R RACE (see table 3.5 and figure 3.11 for sequence details). A 1558 bp fragment was obtained from the 3' RACE reaction. This was subsequently cloned into pGEM®-T easy vector and completely sequenced using vector primers, M13F and M13R and other internal primers listed in table 3.6. When the sequence was

mapped back to the genomic sequence, the 1558 bp 3' RACE product was found to overlap the 3' end of C18438 clone by 223 bp. Figure 3.11 depicts the sequence of C18438 and the overlap with the 1558 bp 3' RACE product obtained. As shown in the diagram, the sequence contains a perfect polyadenylation signal (AATAAA) located 18 bp upstream of the polyadenylation tail. A schematic diagram showing the genomic location of the RACE products is shown in figure 3.12.

Primer Name	Sequence 5'→3'
PS-2F	ggctgtttcaggactattga
PS-3F	gatggtccaagctgctg
M13F	gtaaaacgacggccag
M13R	caggaaacagctatgac
PS-4F	ccaagtgggctgtaattt
PS-5F	gcagcacagatcaagtgtgt

Table 3.6: Primers used for *PS-1* sequence analysis. Sequencing conditions are as described in section 2.4.17. See text for details.

In addition to using RACE to identify the 3' end of this putative gene, the technique was also used to amplify the 5' end. A RACE primer was designed at the 5' end of C18438 clone (PS-1F RACE, table 3.5, figure 3.11 for sequence). In addition, a nested RACE primer was also designed, located 18 bp from the end of PS-1 F RACE primer (PS-1F NRACE, figure 3.11 for sequence). Nested RACE primers are used to PCR the pool of resultant fragments obtained in the first round of RACE PCR to increase the fidelity and specificity of the PCR reaction. Following successive rounds of RACE PCR, a 552 bp product was obtained. This product was cloned into pGEM®-T easy vector and sequenced using primers listed in table 3.6. When the sequence of this clone was mapped to the genomic sequence, the 552 bp fragment overlapped with C18438 by 333 bp, which gives an extra 219 bp of novel 5' sequence (figure 3.11).

Therefore, to summarise the findings described above, initial examination of the genomic region surrounding *CYP11A1* identified an expressed 450 bp placental transcript (C18348). This transcript was shown to be expressed in placenta and ovary cDNA by RT-PCR, but not in adrenal (figure 3.8). In addition, two transcripts of 7.5kb and 6.5kb were identified through Northern analysis in placenta (figure 3.9). To date, the most 5' end of *PS-1* obtained is located 900 bp from the 5' end of exon 1 of *CYP11A1*. Figure 3.11 depicts a schematic diagram of the genomic location of *PS-1* based on sequences detailed thus far in this thesis.

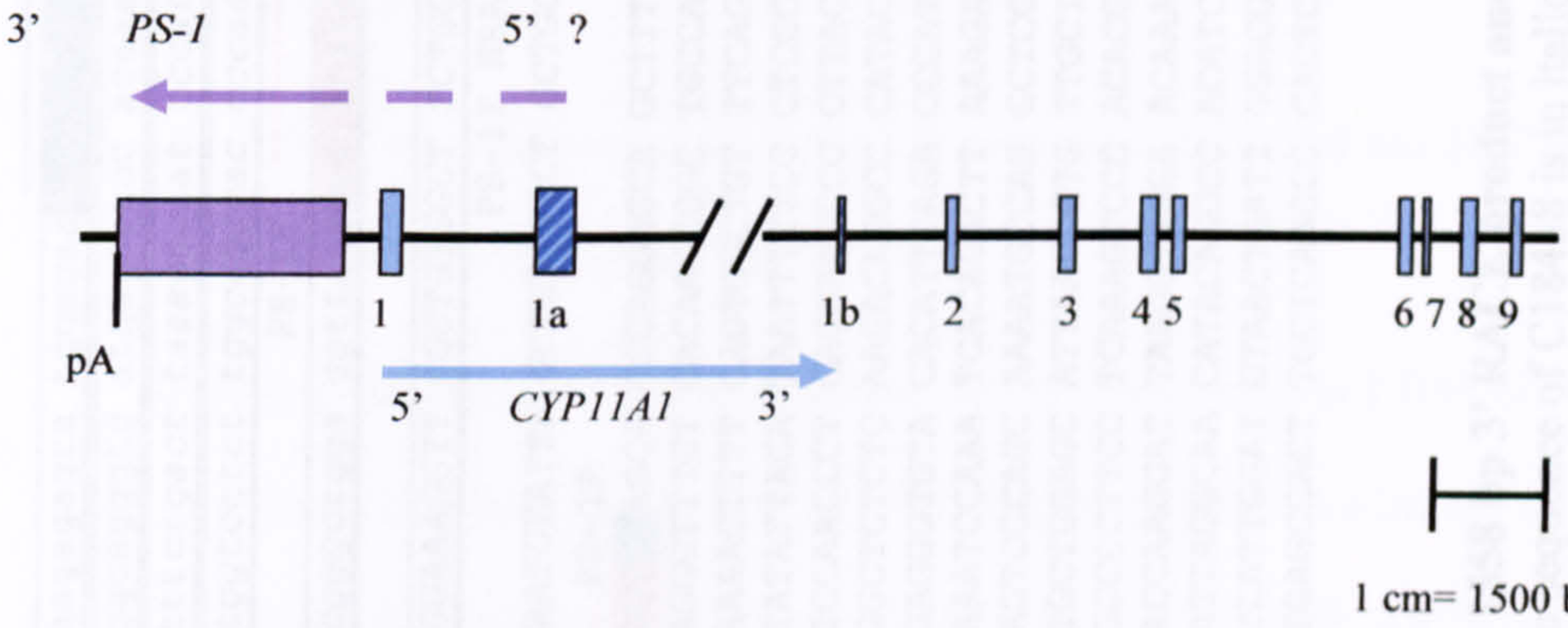


Figure 3.10 : Schematic diagram showing the location of *PS-1* and *CYP11A1*. Directions of transcription of *PS-1* is represented by the purple arrow; the direction for *CYP11A1* transcription is indicated by the blue arrow. *CYP11A1* exons are represented by the blue boxes. Known sequence for *PS-1* is represented by the purple box. The 5' end of *PS-1* is unknown.

To continue the search for the 5' end of *PS-1*, another 5' RACE primer (PS-2F RACE, table 3.5, figure 3.11 for sequence details) was designed within the 552 bp 5' RACE product. No clean product was obtained so a nested primer, PS-2F NRACE, was designed 24 bases from the 3' end of PS-2F RACE primer. However, after successive rounds of PCR, no clean PCR product was obtained. Several further attempts also failed to amplify a product.

PS-2F NRACE

PS-2F RACE

106436 gtgac cactacctca tcagtcatga accttctgat aaagagagta tggagaaagc ctgttttgcc tcactgatga ccttgagcct

106521 ggaatgacat cagccaagga aggcagtgtc tcacgagcag cggcagggtg atgagcttcc acaaagaaaa gcagtcttcc aaggactagt

106611 ttttcctcca agggtctcag cactgaggtg acatttataa ctttctgact taaagtcaat cccattgttc tacgcagccc cacattcctc

106701 cctcagaagg agccctaatc gatgtggcat ggaagaactt gtggtcctct tggcagccac cccaacttcc gttcaggcct ccgcacctcc

PS-1F

106791 cccagtgttg tggggcactg accgcctccc cagaagtggc ccgggctgga ggttattgcc tcatcctctc caatggccat ggtcacatca

PS-1R Race

106881 cAGCATAGGC CAGCGATTCA GTTTGGGCCA ATGAAACACA GGGAAAGGTT TGCTGTGGCT TCTGGGAGAA TAGCTCCCTC ACTTCTCTGA

PS-1F NRACE

106971 AAGAGCTTCC AGAAACACAT CACTCACTCC AAACAAGGCT GAATGTATTA GCTGTTTACT GCTGCTTAAC AAATTATCCC AAAATTTAGT

PS-1F RACE

PS-1R

107061 GGCTTAAAAC AATAAATGTT TATTATCTCA CAGCTTCCGA GGGTGCAGCA CCCAGGAGCA GCTTTTCTGG GTGACTGGGT GGCTGGGTGG

107151 CTGGGTGACT GGGTGGCTGG GTGGCTGAGT GGCTCATGCT CAGGGTTTCT CACAAGCGGC TGCCATCCCT GAAGCTTGAC CAGGGTGTAG

107241 GATCAGCTTC CAAGCTCAGT GGATTAAACT GGCCCTGCTT CAAAACTTTT CAGTGGCTGT TTCAGGACTA TTGAGAAGAT TAATGAATGA

107331 TACATACATT TCAAAATATA TATATTACAC ATATACATGT GTATATTACA TAATTTTTCC CTCCCTCTCT CTGTTCTTGC CAGTGCCCAA

107421 ACTGAAATTT TCACCCTCTC AAAAGTGATG GGAAGTTAGG GTCCAACCCT CACCAAGCCC CTTACGATGG GCAAGTGAAT CCTCTGCTAG

107511 GTGAGGTTAG CCTAGTGCTC GGCTATGCCA AAGTCCTTCT TGGCTCTCTC AAGACAGGCC CATACTTATA GTGCTTTCTC CTCCCACTCC

107601 TGCCTCCCAC ATTCTGTAAC TATCCTATCA TTCTCAGAGT AGAGGGTGTA CACATTTAGA CCCATCTAAA CGTTGCCTAG GACGGGGAAC

107691 ATGATTTTAT TCATTGCAGG CAGCTTGGAT GGATAAACAG AAAATCCAAA TCACAGACTT AAAGAAGATA GAAGTGTATT CTCTCTCCCA

107781 TAAAGGATGT CCAAGGCTGG GCAGTCCAGG GTGAGTGTGT CAGTCCCAGC AAAATGCCAG GCTCCATCTT CCTGCTTCAC CACGAAGACA

107871 TTTGGTTGGC ATTTCTAAGG CCACCCGATG GTCCAAGCTG CTGCTGGAGC ATTAACCGTG TTGCTGGCCA GAAGGCGGAA GAACTCAAGA

107961 AGACCAAAAG GGGCCCTCCC TGAAGAGCCA GCTTCATTTA AGCCCCTTCC TCAAAGTCCC ACACGAGTCC AACCATATCT GATTGGTGGG

108051 AACTTAACCA CATTGAGGCA GGAGAATAGG GAATTAGGGC AACCAAGGAT TAAGGTAGGA ACAAAAGAAG AGCAGGTGCA GCCAGTTCAC

108141 ATAAGCAAAA GGCACAGCAG GTACAGCCAA TTCTAGGCAG GATTAGGCAA CATACAGGCC ACATCTTCAC TTCTGTGATA ACCACAGGCC

108231 AAGTCGCCAC TTAAGCCTCT AATTGGTCAC GGATCAATCC TTCATTGGAT GTAACTGATT GGAGGCCTCT GGAGGGCACC GGGGGTGTTC

108321 ATGTTCTTTA GCTTAATAAA ATCTCTAATT GAAGGGGCTC CTGAGCCACT TGCTCAAGCC CACTCTTGCT CTGTGAATTG TACTTGCACA

108411 TCTTCAATAA ATCTGTGCCT TCATTACTCC $(A)_n$

Figure 3.11: 3' Sequence of *PS-1* 552 bp 5' RACE product, the 1558 bp 3' RACE product and C18438: The 552 bp 5' RACE product is boxed. The 1558 bp 3' RACE product is in uppercase. The known sequence of C18438 is in italics. Primers are as listed and in colour. The polyadenylation signal (AATAAA) is highlighted in yellow. Nucleotides are numbered from 5' end of the 5' RACE product.

3.3.1.4 *In Silico* BLAST Searches

To ascertain any additional 5' sequence through the identification of overlapping clones (see section 2.4.22), the sequence of C18438 and both the 5' and 3' RACE products were used in BLAST searches of the human EST and the nr genomic databases. Various clones were identified:

- Another placental EST, GenBank accession number BF996753 was identified. This clone, when mapped to the genomic sequence, overlapped at the 5' end of C18438 by 85 bases.
- The 219 bp of novel sequence 5' RACE sequence identified one EST clone, GenBank accession number CB989394.
- Searching the NCBI UniGene database with the sequence of CB989394 identified a UniGene cluster (Hs.371799) that contained, in addition to CB989394, four other EST clones (AW197444, AW594588, AV753134 and BG015590).

The 85 bp of extra sequence generated from clone BF996753 was then used in further BLAST searches of human dbEST but no further clones were identified. Two of the clones identified in the UniGene cluster, AW197444 and AW594588 were documented as 5' reads and so were ordered from HGMP-RC in anticipation of obtaining more 5' sequence. Sequencing of these clones (using primers listed in table 3.6, PS-4F and PS-5F), revealed that they were both 750 bases in length and completely within C989394. Therefore, no additional 5' sequence was generated. Analysis of the third clone in this cluster, AV753134, identified a further 282 bases of 5' sequence. In addition, clone BG015590, was found to overlap completely with clone AV753134. This therefore did

not provide any more 5' sequence. A schematic diagram of all clones and RACE products identified for *PS-1* is presented in figure 3.12.

In the continued search for the 5' end of *PS-1*, a set of 5' RACE primers were designed using the sequence obtained from the EST AV753134 (called PS-3F RACE and PS-3F NRACE, table 3.5). Several rounds of RACE PCR were completed; however no specific PCR products were obtained. Therefore to date, based on the experiments conducted using the methodology of both 3' and 5' RACE, and building on the initial sequence of C18438, a total of 2997 bp of novel *PS-1* sequence has been identified. In addition, the presence of a polyadenylation signal and poly adenosine tail, both features of mammalian genes, support the classification of *PS-1* as an authentic gene. In the examination of the translated sequence obtained thus far, no single ORF has been identified using the program ExPASy(UK, data not shown).

Based on the Northern results presented in figure 3.10, which identified two transcripts of 6.5 and 7.5 kb in size, another approximately 3500 and 4500 bp of *PS-1* sequence is unknown. In addition, the 5' end of transcript II and transcript III of *CYP11A1* is also unknown. Evidence to date from the 5' and 3' RACE experiments, suggests that the transcription of *PS-1* proceeds in the opposite direction to that of *CYP11A1*. It is therefore possible that the remaining *PS-1* sequence is either interspersed with *CYP11A1* or is further downstream of this gene. Therefore, in continuation of the *in silico* analysis of this region, 5000 bases from the 3' end of exon 1 of *CYP11A1* (and inclusive of *PS-1*) was used to search dbEST. The results are presented in table 3.7. As shown in this table, there are 19 EST clones on the NCBI database, half of which have an alignment in the transcriptional direction of *CYP11A1*, and half in the putative

transcriptional direction of *PS-1*. Furthermore, several exon 1a containing clones are shown to contain sequence in both directions, indicating that exon 1a of *CYP11A1* and *PS-1* may share this sequence (table 3.7, highlighted in yellow). BLAST searches of the dbEST using *CYP11A1* exon 1 sequence and *PS-1* sequence obtained to date failed to identify any sharing of exon 1

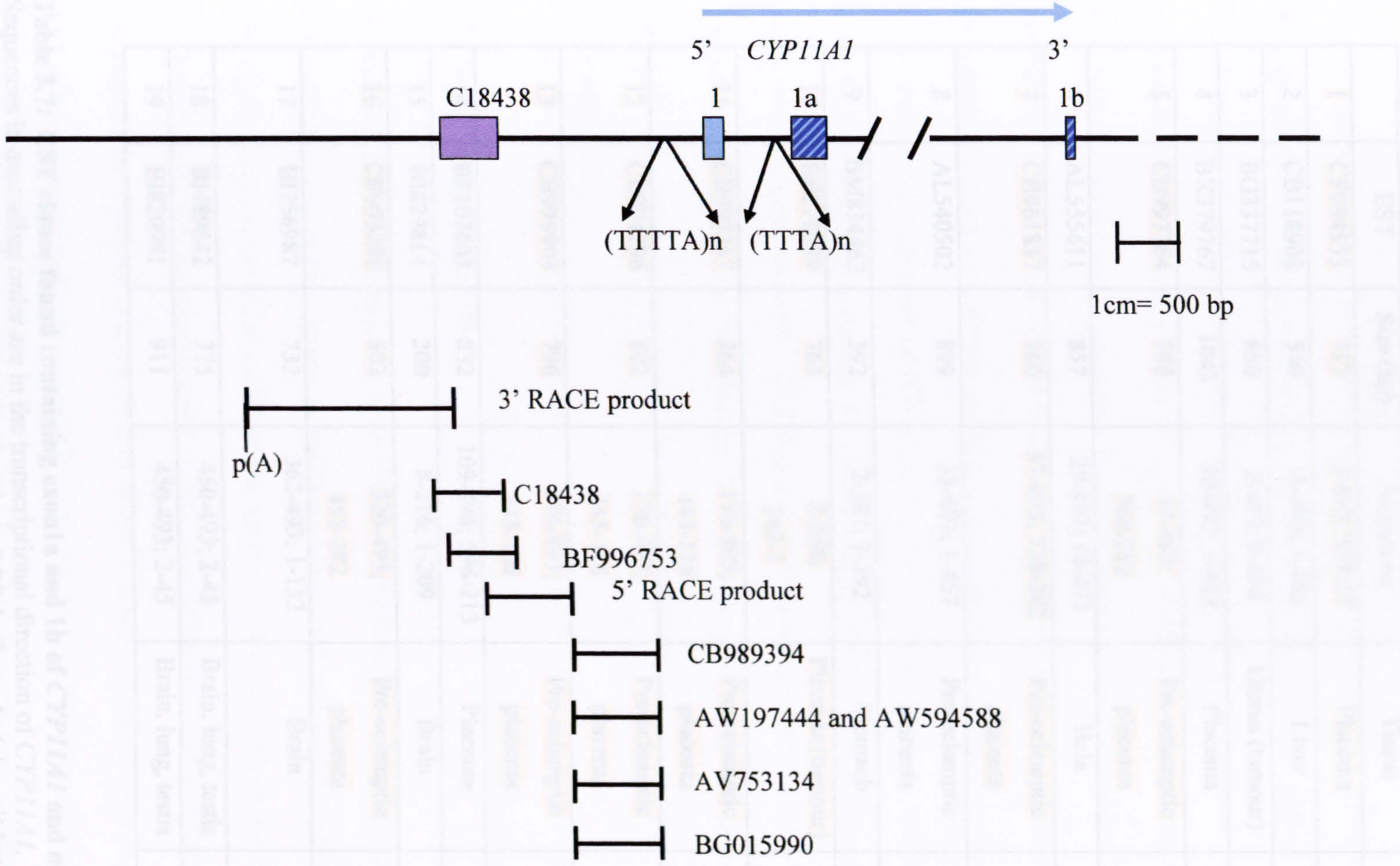


Figure 3.12: Schematic diagram of overlapping clones of *PS-1*: Clones were identified by BLAST searches of genomic nr and EST databases. 5'and 3' RACE were done using ovary as a template, sequences of primers as listed in table 3.5. p(A)= polyandenosine tail

	EST	Size (bp)	Alignment	Tissue	Exons
1	CF994553	717	2-493; 558-67	Placenta	1a
2	CB118908	540	11-493; 1-483	Liver	Just 1a
3	BG337715	850	2-493; 9-498	Uterus (tumour)	1a, 2,3
4	BX379767	1043	30-493; 1-463	Placenta	1a, 1b,2,3,4
5	CB997764	768	11-493; 768-282	Pre-eclamptic placenta	1a
6	AL555611	857	29-493; 10-471	Hela	1a, 2, 3
7	CB961837	760	37-493; 738-282	Pre-eclamptic placenta	1a, 2, 3, 4
8	AL540502	899	33-493; 1-457	Pre-eclamptic placenta	1a, 2, 3, 4
9	BM834462	392	2-387; 7-392	Stomach	1a
10	BE278699	763	2-366; 365-1	Placenta (tumour)	1a
11	CB997357	864	176-493; 443-126	Pre-eclamptic placenta	1a
12	CB994296	882	176-493; 443-126	Pre-eclamptic placenta	1a
13	CB995993	796	176-493; 443-126	Pre-eclamptic placenta	1a
14	BF107603	832	109-494; 596-213	Placenta	1a
15	BI829814	200	8-216; 1-209	Brain	1a
16	CB995289	502	320-493; 455-282	Pre-eclamptic placenta	1a
17	BI756987	732	362-493; 1-132	Brain	1a, 2, 3, 4, 5
18	BI489642	775	450-493; 2-45	Brain, lung, testis	1a
19	BI820991	911	450-493; 2-45	Brain, lung, testis	1a

Table 3.7: EST clones found containing exon1a and 1b of *CYP11A1* and of *PS-1*. Sequences in ascending order are in the transcriptional direction of *CYP11A1*, descending order in the transcriptional direction of *PS-1*. Exon 1a 'shared' between *CYP11A1* and *PS-1* highlighted in yellow.

3.4 Discussion

There is a growing body of evidence in support of the involvement of *CYP11A1* in the aetiology of PCOS (Gharani *et al.*, 1997); (Diamanti-Kandarakis *et al.*, 2000); (Daneshmand *et al.*, 2002). However, the pathogenic mutation has not been conclusively determined (see sections 3.1 and 1.5.4). Studies presented in this thesis of the *CYP11A1* genomic region have revealed the existence of novel alternative transcripts for *CYP11A1*, and a gene of unknown function (*PS-1*). It is therefore believed that there are several extra levels of complexity in the genomic region of 15q23-24 than originally reported (Chung *et al.*, 1986). Results from the annotation of the whole genome sequence through the Human Genome Project have estimated the genome comprises approximately 25 000 genes, much lower than previously anticipated. The genome has been found to be highly complex, with more alternate splicing than genomes of other species (International Human Genome Sequencing Consortium, 2001).

CYP11A1 Alternate Transcripts

The location of 1a and 1b sequences within intron 1 of *CYP11A1* and expression in placenta cDNA leads to the hypothesis that these were novel, previously unidentified, exons of this gene. As with the published transcript of *CYP11A1*, both exons are expressed in ovary and the adrenal gland (figure 3.5). There is evidence on the NCBI databases of other ESTs containing exon 1a, exon 1b (transcript III) or exon 1a (transcript II), together with exons 2-9 (table 3.4). Taken together, these data suggest that there are at least three possible transcripts for *CYP11A1*, containing putative novel first exons. Other genes from the cytochrome P450 gene family such as *CYP19* have also been discovered to have transcripts containing alternate first exons. The

organisation of *CYP19* is highly complex, with 10 distinct tissue specific promoters, each with distinct regulatory sequences and transcription factors regulating expression (Bulun *et al.*, 2004). It is therefore possible that transcript II and transcript III function to regulate expression, possibly in a tissue or developmental manner.

Attempts at defining the 5' end of the novel *CYP11A1* transcripts using 5' RACE were unsuccessful. RACE involved using a premade library (a human ovarian library was used here) of adapter-ligated double stranded (ds) cDNA from Clontech (UK) as template for amplification to identify the 5' end of genes. It is possible that the cRNA pool supplied by Clontech did not contain a high amount of these transcripts, relative to all other transcripts. The cRNA pool used here is not from a normalised pool of transcripts; therefore lowly expressed sequences may be harder to identify using this method. Inferences that both exons 1a and 1b identified here are lowly expressed are based on the faint bands obtained through Northern analysis and the scarcity of clones on the public databases. The same cDNA was used concurrently to successfully amplify exon 1 of *CYP11A1* (figure 3.5 (a) and (b)) implying that the starting material was both intact and of sufficient amount. Experiments such as quantitative PCR need to be completed to ascertain the relative expression levels of both novel transcripts in these tissues. Experiments such as primer extension and 5' RACE with increased cycles or new primer design will need to be completed in order to identify the 5' end of transcripts II and III. Conversely, screening of a placental or ovarian cDNA library by plaque hybridisation could be carried out to isolate clones containing the full sequence of both novel transcripts.

The completed sequence of the chimpanzee has recently been published. BLAST searches using the sequence of clone AK056794 (containing *CYP11A1* exons 1a and 2-

9) identified one chimpanzee clone, Genbank accession number XM_510666.1. This clone, when compared to AK056794, was found to contain 467 bps of exon 1a sequence. This is the first evidence of another mammal containing *CYP11A1* exon 1a other than human. Clones containing exon 1a or exon 1b in mouse, rat or bovine were not identified in these searches.

The translational start site in exon 1 of *CYP11A1* (transcript I) has been reported to start at +62 bases to produce a protein of 521 amino acids in length (Morohashi *et al.*, 1987). Analysis of putative ORFs in transcript II did not reveal a methionine in exon1a that remains in-frame to exon 9 (figure 3.7 (b)). Furthermore, the presence of a stop codon in exon 1a in this frame leads to several possibilities. Firstly, this may indicate that exon 1a is non-coding, with translation beginning at the methionine located in exon 3. Many genes are being discovered that make transcripts containing alternate non-coding exons. For example, the gene encoding a cytoskeletal liner protein, plectin, has been shown to have 14 alternate spliced exons with 11 directly splicing into exon 2. Furthermore, the alternatively spliced exons were found to be expressed in different tissues (Rezniczek *et al.*, 2003). This tissue specific distribution of plectin isoforms alludes to putative differences in function in each tissue. It has been shown that the different 5' UTRs and alternate non-coding exons confer a role in regulating protein expression and function (Rezniczek *et al.*, 2003). Furthermore, it is well known that a mechanism of differential regulation of genes comes from the utilisation of alternate promoters, which are often found upstream of untranslated or alternately spliced exons. Several untranslated exons of the cytochrome P450 aromatase gene in humans have been shown to modulate the different promoter use in placenta compared to that utilized in adipose tissue (Mahendroo *et al.*, 1999). Analysis of the sequence flanking the methionine codon in exon 3 does not appear to contain the Kozak consensus sequence

(ACCATGG), which functions to facilitate the binding of mRNA to the small ribosomal subunit for initiation of translation (Kozak, 1986). This implies that the codon may not function for translation initiation.

Secondly, as the 5' end of transcript II is unknown, it is possible that a functional methionine may lie within this unknown sequence and code for a protein to the stop codon in exon 1a. Based on the Northern results for transcript II and transcript I, (figure 3.6 (a) and (b) respectively) it is hypothesized that approximately 120 bp of exon 1a sequence in transcript II is missing. This unknown sequence may contain this methionine. Analysis of transcript III for putative ORFs produced similar hypotheses as discussed for transcript II. Therefore, the identification of the 5' end of both transcript II and transcript III needs to be identified for these hypotheses to be tested.

Primer extension and S1 nuclear protection assays could be performed to determine promoter utilization and function once the full sequence is ascertained. It is possible that these alternate transcripts may be present in tissues other than those studied here. A quantitative comparison of possible isoform expression on transcript levels in different tissues needs to be performed to determine the extent of expression, the results of which may allude to possible function.

PS-1

Initially, *in silico* BLAST searches of the 5' promoter region of *CYP11A1*, identified an expressed sequence (clone C18438). This clone was interesting due to the close proximity of this sequence to *CYP11A1*. Validation of the authenticity of this transcript has come from both PCR and Northern Blot analysis, and from data collated from the

public databases. Moreover, 3' and 5' RACE experiments resulted in the sequence of these products mapping to the same genomic region as C18438. The 3' RACE product was also found to contain a perfect polyadenylation signal, 18 bases from the end of the clone, and a string of poly adenosines, which is characteristic of human gene structures (Proudfoot and O'Sullivan, 2002). Taken together, these data imply that this transcript may represent a putative gene.

After successive rounds of 5' and 3' RACE PCR with both primary and nested primers, the PCR generated a total product of 2110 bp. However, using 5' RACE primers designed within the new 5'RACE product sequence was not successful. There are a number of reasons why the 5' RACE experiment failed. As with the novel exons of *CYP11A1* (section 3.2.3), if *PS-1* is lowly expressed, it may be difficult to isolate this gene using RACE PCR as the pool of cRNA pool used here in the Clontech Marathon Ready RACE kit is not normalised. There is no indication of this gene on any of the genomic databases searched (University of California, Santa Cruz (USCS) Genome Browser, Ensembl and NCBI); only EST clones. Therefore, redesigning the RACE primers for more subsequent nested reactions needs to be completed and/or the number of cycles needs to be increased for each reaction to obtain the 5' end. In addition, primer extension experiments could be performed. As with exons 1a and 1b, probes could also be made within *PS-1* to screen either a placental or an ovarian cDNA library by plaque hybridisation in order to isolate more *PS-1* sequence. As shown in table 3.7, clones have been identified containing C18438 sequence and *CYP11A1* exon 1a sequence, but in opposite transcriptional directions. This is due to the way the libraries containing the clones are constructed. Primers are designed to amplify transcripts that have a polyA tail. It is therefore possible that both *CYP11A1* and *PS-1* may share sequences, which may be difficult to decipher. There are several genes that are known

to share exon sequences in opposite transcriptional directions. For example, two genes located in chromosome 1q43, *OPN3* (panopsin) and *KMO* (*panopsin* kyurenine monooxygenase), have been shown to overlap at their 3' ends but are transcribed in opposite directions (Halford *et al.*, 2001). Therefore, it is possible that *CYP11A1* and *PS-1* are two genes that are transcribed in opposite directions on chromosome 15q24.

Based on the Northern results, transcripts representing *PS-1* are approximately 7.5 kb and 6.5 kb in size. The contiguous sequence obtained thus far (approximately 3kb) indicates that the sequence may still represent the 3' untranslated region (UTR). This is substantiated by the fact that no ORF has been identified. As mentioned previously, the remaining sequence of *PS-1* may be further upstream or interspersed with *CYP11A1*. In contrast, the contiguous sequence identified thus far may indicate that *PS-1* may represent an intronless gene. There are many examples of intronless genes, including a proapoptotic gene, BCC-2 (Broustas *et al.*, 2004), and many other mitochondrial genes (Onishi *et al.*, 2004).

Any possible function for *PS-1* remains highly speculative. Elucidation of other tissues in which this gene is expressed may allude to possible function. It is plausible that the pentanucleotide in the promoter region of *CYP11A1*, or indeed the tetranucleotide, may be influencing the expression of *PS-1* rather than *CYP11A1*. Elucidation of the entire genomic structure and regulation of the expression of *PS-1* needs to be completed to answer these questions fully.

4.0 ASSOCIATION STUDIES

4.1 Introduction

A pentanucleotide repeat in the 5' promoter region of *CYP11A1* has been found in several studies to be associated with serum testosterone levels in women with PCOS (Gharani *et al.*, 1997); (Daneshmand *et al.*, 2002); (Diamanti-Kandarakis *et al.*, 2000) (see section 1.5.4 for detailed discussion). However, the sample sizes in these studies have been modest. The following section describes investigation of this pentanucleotide repeat and a neighbouring tetranucleotide repeat for association to PCOS and polycystic ovary related symptoms. The numbers of subjects tested here exceeded the sum of all previous studies to date.

4.2 Pentanucleotide and Tetranucleotide Repeat Polymorphism

4.2.1 Pentanucleotide Repeat

The pentanucleotide repeat, as it has been shown in the above mentioned studies (section 4.1 and section 1.5.4), was also found to be polymorphic in this study. PCR primers used for genotyping are listed in table 4.1, with the sequence presented in figure 4.1. Variation at the pentanucleotide repeat in the present study revealed seven alleles ranging from 169 bp (pentanucleotide (P) 4 to 199bp (P10), with the P4 allele commonest in the UK sample set (frequency ~58% in controls) (Gharani *et al.*, 1997). The 216 bp allele described in the previous studies is equivalent to the 169 (P4) allele in the current study, the difference in size representing alternate PCR primers used. As described in section 2.2.10, Mendel-checking methods were used to confirm genotyping

accuracy. In addition, approximately 10% of the samples were re-genotyped and from these data, the overall genotyping error was <1%.

Primer Name	Sequence 5'→3'	$MgCl_2$ mM	Annealing °C
Penta UK F	*cctgggcaacaagaatgaaa	1.0	62
Penta UK R	gcccaagagaaggtcagagc	1.0	62
Penta Fin F	*ggtgaaactgtgccattgc	1.0	62
Penta Fin R	gtttgggggaaatgaggggc	1.0	62
Tetra All F	^agtagctgggattatagaccc	1.5	62
Tetra All R	ggagactggtgaggctaagt	1.5	62

Table 4.1: PCR primer sequences and conditions for pentanucleotide and tetranucleotide repeat for both UK and NFBC1966 samples. Primer sequences, with the associated concentration of $MgCl_2$ and annealing temperatures, are as listed. See text for details. *= FAM-labelled; ^= HEX labelled. UK- primers used in the UK dataset; Fin-primers used in the NFBC1966 dataset, ALL- used in both the UK and NFBC1966 sample sets.

4.2.2 Tetranucleotide Repeat

As described in section 3.2, examination of the genomic region of *CYP11A1* identified a previously undetected microsatellite, a tetranucleotide repeat $(TTTA)_n$, located 256 bp upstream of exon 1a (see section 3.2.3 and figure 3.1). Due to the location of this repeat and the linkage to the region as described by former colleagues (Gharani *et al.*, 1997), this repeat was examined in the studies described below.

63541 tgttttaaat gccttatttg ccagaaaata agattttggg gaaaaaaagc aagaagttgt accttccaat gataagtata aggtatagtg
63631 tagattgctg tcatgaaatg ctctacatgg cacatgtatg ttttctgacc tacacatttt taccaagtgg gctgtaattt tgttgaaatg
63721 acttgaaaac tctacaggtg actgaccctt gttgtttggg aaagtggtag gtgcagggtg atgggggtg ggtgtggggg ctgacctggg
63811 ctggaaggat gggtctgggg atatgatatg atgggagggg gtatggaagc agctctgagg ggaaagcagc acagatcaag tgtgtggaca
63901 gggaagagct gacatcctga gttccatggc aaccagattt gccaaggtct tagagtgtgt ccagagtgga gcctgaccac agacctcagc
63991 tcaagggacc cagagcccct ctgagtcagc tgtactgaat tacagcccca aatctgggtc aactggggag agacgacgag gattagggtt

Penta Finn F → UK Penta F →

64081 ccaaggtgaa actgtgccat tgcgctccag CCTGGGCAAC AAGAATGAAA ctctcttAAA ATAAAATAAA ATAAAATAAA ATAAAATAAA

64171 ATAAAATAAA Tagcctaagg atgcatttct cagaacttat ccctgttgtt caatgatgtg tgtctataca gtggggccat aactaagacg

← UK Penta R ← Penta Finn R

64261 tatgttgccc aagctggcaa gataGCTCTG ACCTTCTCTT GGGCccctca tttcccccaa acacaggttg tctgcagtct tgaccaatgg

64351 ctgccagggc atggactccg ctgcaggggc cagtgggagg ccccagctca ggcaaaagca caggcagata tttcaggagt ctgctagggc
64441 tggcactgag ggcagagaca gaggggtctc cctgtccttt ggagaacctc acgctgcaga aattccagac tgaaccttca taccgagtag
64531 gggaggagct gtctgcgggt ttgagcctgc agcaggagga aggacgtgaa cattttatca gcttctggta tggccttgag ctggtagtta
64621 taatcttggc cctggtggcc cagggctaca gtcatcctag CAGTCCCCGC TGAAGTGGAG CAGGTACAGT CACAGCTGTG GGGACAGCAT
64711 GCTGGCCAAG GGTCTTCCCC CACGCTCAGT CCTG

Figure 4.1: Sequence of 5' promoter region of *CYP11A1* showing position of PCR primers for pentanucleotide repeat genotyping: UK specific primers boxed and coloured in pink; NFBC1966 specific primers boxed and coloured in blue. Arrows for the primers are in direction of 5'→3'. Pentanucleotide repeat in bold uppercase and coloured in purple. *CYP11A1* exon 1 sequence underlined in black and uppercase. Nucleotides are numbered from 5' end of the genomic DNA clone

To determine whether the tetranucleotide repeat was polymorphic, a random set of 10 control DNAs were selected from the UK sample set (see section 2.1) for genotyping. Primers for PCR (see section 2.4.11 for detailed methodology) were designed flanking the tetranucleotide repeat with the sequences presented in table 4.1 and figure 4.2. The resultant PCR fragments were genotyped on an ABI PRISM 3700 sequencer using a ROX size standard and were analysed using the program Genotyper (see section 2.4.21). Seven alleles were observed at the tetranucleotide ranging from 276bp (tetranucleotide (T) 6) to 304 bp (T13).

Both the tetranucleotide and the pentanucleotide repeat were genotyped in the UK sample set (230 cases; and 331 controls; and 141 parent-offspring trios, section 2.1.1 and 2.1.2) and in the Northern Finnish Birth Cohort 1966 sample set (NFBC1966; 527 women who reported symptoms of hirsutism and/or oligo/ amenorrhoea designated as symptomatic 'cases' and 1062 women as cohort controls) (see section 2.1.3).

Tetra All F

65551 tcaaacaaaa caaaacaacc tcctcggtct gaggaggga gactggtgag gctaagttt gaattcctgg ccctgtagct gtcagcccag

65641 ccagcgagtt ttaaaaattg ggagaaggcc gggtgcggtg gctcacgctt gtaatcccag cactttggga ggctgaggca ggcggatcac

65731 ttgaggacag gagttcgaga tcaacccggc caacatggtg aaaccctgtc tctactacaa **AAATAAATAA ATAAATAAAT AAATAAATAA**

Tetra All R

65821 **At**aaaataaa aaaaattagc cgggcatggt ggcgggggtc tataatccca gctactcggg aggctgaggc aggacactcg cccgaaccca

65911 ggaggtggag gttgcagtga gccgagatca tgtcactgca ctcccgccct gggcaacaaa agcgaaactc tgtctcaaaa aaaaaaaaaa

66001 aaaaaaaatc gggggaggga gatttgaagg tcagagcaca gagtgaaaag gtcgcagcgt ggtcaagttt ggcgctagAG ACCCAGCATT

66091 CCTCCCAGCG CACGCCGCCG CCCGGGCCGC GCTATTCCGA AATTGGGTCC GCCCTAGCGT GGCCCCCGGC CTGCTCCTGC GCTCTTCGCG

66181 GGTCGCTCAC CTCCGCCTCC CCTCGACACA GCCCGAGCCC ATCCGCCCCC CGGGTCGACC CGCGTTTACC TCAGGGCGAG GTCCAGGCGC

66271 TCGGCCCCGC GGAGGCGGCG GCGGCCCCAC CGCAGGTCCT AGGTCCCAGT CACTCTGCCG CCAAGTTCCC CAACACGGAC TCCGTGCGCT

66361 TTGAGGTGTT ACGGGGGCGC TGGAGCACTG GCTTCGCTGG CCCTGCCAGG AGGCGGCGGC GAAGCAGGGA GCGACCCGGT ATGCGTTTCT

66451 CTGCCCGGCC TGCGGCCCTG CAGCATCCAC TCCTCGCCTC TGCTGGCCTC TGCCCTCGCT CGCTGCGTTC ACCCGAGCTC ACCCGAGCCC

66541 ACCGGAGCCG ACAGTGGCGC GCTAAGgtga cgtccgcgcg gccggacctc tcctgtccaa

Figure 4.2: Sequence of intron 1 of *CYP11A1* showing tetranucleotide repeat genotyping PCR primers: Primers boxed and coloured in blue. Arrows for the primers are in the direction of 5'→3'. Tetranucleotide repeat in bold uppercase and in pink. *CYP11A1* exon 1a is in uppercase. Nucleotides are numbered from 5' end of the genomic DNA clone.

4.3 Linkage Disequilibrium

Linkage disequilibrium is the non-random association between a particular allele, such as a disease allele, with a specific allele at a nearby locus more often than expected by chance. Linkage disequilibrium relationships were determined in UK and Finnish control subjects using the LDMAX option within the GOLD (Graphical Overview of Linkage Disequilibrium) package, which implements the EM (Expectation maximization) algorithm. It was found that the pentanucleotide and the tetranucleotide repeats were in strong linkage disequilibrium in the UK controls (D',~0.86, P<0.001). The most marked departures from equilibrium were positive associations between the P4 and T8 alleles (56.5% observed vs. 47. 6% expected) and between P6 and T12 alleles (11.3% vs. 3.1%); and the complementary negative associations between P6 and T8 (10.8% vs. 21.7%) and P4 and T12 (0.4% vs. 6.8%). In the NFBC1966 controls, the extent and direction of linkage disequilibrium between alleles at the two loci were similar (D',~0.87, P<0.0001).

4.4 UK Sample Set Association Studies

4.4.1 Case Control Study

Case control studies are designed to compare the allelic and genotypic frequencies of a particular variant between a group of individuals affected with the disease under study, and a separate set of individuals who are known to be either unaffected or alternatively, randomly selected from the population. An increased frequency of a particular variant in the affected population compared with controls indicates an increased risk for the disease. In the first study, both allelic and genotypic frequencies were examined in the UK cases and controls and results of which presented below. All PCOS subjects were

defined as described in section 2.1. Initially, all subjects obtained through the infertility and endocrine clinic at St. Mary's and the Middlesex Hospitals London, were typed which included both UK and non-UK ethnicities. However, the latter were not incorporated in the following case-control study.

Allelic and genotypic frequencies were compared in the UK sample set by standard contingency methods, which allows for exact probabilities, implemented using the statistical program STATXACT (Cytel, Cambridge, MA, section 2.1.11). Haplotypes were derived by standard maximum likelihood methods implemented in the program TRANSMIT (section 2.2.10), with group comparisons assessed by likelihood ratio-testing and permutation methods (at least 1000 permutations).

Pentanucleotide

At the pentanucleotide, there was a difference in allelic frequency comparisons between UK PCOS cases and population controls (representing an excess of P4 alleles, exact global P=0.005, table 4.2). Genotype frequency comparisons revealed significant differences in genotype frequencies (Monte-Carlo estimate, P= 0.029, 99% confidence interval 0.027-0.030), largely attributable to an excess of P4 homozygotes in the cases (47.2% vs. 34.8%, table 4.2).

Tetranucleotide

At the tetranucleotide repeat, allele frequency comparisons indicated an excess of T8 alleles in the UK cases (284bp; global P=0.0097; 88.4% vs 82.4% in controls,

table 4.3). However, analysis of the genotype frequencies showed no significant differences between cases and the population controls (Monte-Carlo estimate, global *P*=0.139; 99% confidence interval 0.133-0.150; table 4.3).

Locus	Pentanucleotide	UK Cases	UK Controls
		N (%)	N (%)
Allele counts	P4 (169bp)[a]	261 (67.6)	391 (58.2)
	P6 (179bp)	68 (17.6)	176 (26.2)
	P7 (184bp)	0	2 (0.3)
	P8 (189bp)	16 (4.2)	43 (6.4)
	P9 (194bp)	40 (10.4)	59 (8.8)
	P10 (199bp)	1 (0.3)	1 (0.2)
Genotype counts	P4/P4 (169/169)	91 (47.2)	117 (34.8)
	P4/P6 (169/179)	47 (24.4)	97 (28.9)
	P4/P8 (169/189)	9 (4.7)	23 (6.8)
	P4/P9 (169/194)	23 (11.9)	34 (10.1)
	P6/P6 (179/179)	4 (2.1)	22 (6.5)
	P6/P8 (179/189)	4 (2.1)	16 (4.8)
	P6/P9 (179/194)	8 (4.1)	19 (5.7)
	the rest[b]	7 (3.6)	8 (2.4)

Table 4.2: Allele and genotype frequencies at the pentanucleotide repeat in the UK sample set. There is an excess of P4 homozygotes (*P*=0.029). [a]The P5 allele was not seen in the UK cases and controls; [b]all genotypes <1% frequency are pooled for the purposes of this table: contingency table analyses were performed on unpooled counts.

Locus	Tetranucleotide	UK cases	UK Controls
		N (%)	N (%)
Allele Counts	T6 (276bp)	1 (0.3)	0
	T8 (284bp)	352 (88.4)	567 (82.4)
	T9 (288bp)	0	9 (1.3)
	T10 (292bp)	3 (0.8)	4 (0.6)
	T11 (296bp)	3 (0.8)	12 (17.4)
	T12 (300bp)	35 (8.8)	82 (11.9)
	T13 (304bp)	4 (1.0)	14 (2.0)
Genotype Counts	T8/T8 (284/284)	155 (77.9)	233 (67.7)
	T8/T9 (284/288)	0	6 (1.7)
	T8/T11 (284/296)	3 (1.5)	9 (2.6)
	T8/T12 (284/300)	33 (16.6)	72 (20.9)
	T8/T13 (284/304)	4 (2.0)	12 (3.5)
	The rest[a]	4 (2.0)	12 (3.5)

Table 4.3: Allele and genotype frequencies at the tetranucleotide repeat in the UK sample set. There is an excess of T8 alleles (global P=0.0097) but no significant difference with T8 homozygotes (global P=0.139); [a]all genotypes <1% frequency are pooled for the purposes of this table: contingency table analyses were performed on unpooled counts.

Global comparison of haplotype frequencies (by maximum likelihood testing, 1000 permutations) revealed no significant differences between the UK cases and controls (global exact P= 0.15), though, as expected given the allele frequencies stated above, the P4-T8 haplotype was more frequent in cases (65.5% vs 56.5% in controls, P=0.005 uncorrected, table 4.4).

Locus	Pentanucleotide	Tetranucleotide	UK Cases	UK Controls
			Frequency	Frequency
Haplotype	P4 (169bp)	T8 (284bp)	0.655	0.565
	P6 (179bp)	T8 (284bp)	0.085	0.108
	P9 (194bp)	T8 (284bp)	0.103	0.086
	P8 (189bp)	T8 (284bp)	0.041	0.064
	P4 (169bp)	T12 (300bp)	0.006	0.002
	P6 (179bp)	T12 (300bp)	0.078	0.004
	P9 (194bp)	T12 (300bp)	0.000	0.113
	P8 (189bp)	T12 (300bp)	0.000	0.000
	P4 (169bp)	T11 (296bp)	0.000	0.002
	P6 (179bp)	T11 (296bp)	0.001	0.000
	P9 (194bp)	T11 (296bp)	0.000	0.002
	P8 (189bp)	T11 (296bp)	0.000	0.020

Table 4.4: Haplotype analysis at the pentanucleotide and tetranucleotide repeat in the UK cases and controls. There is no significant difference in haplotype frequencies (global P=0.15); however, there is an excess of P4-T8 haplotypes (P=0.005).

4.4.2 Family-Based Association Study

For the parent offspring trios available (n=141, see section 2.1.1), a family-based association method was utilized (the transmission disequilibrium test, TDT, see section 2.2.10). Family-based association studies such as these examine the proportions of each allele transmitted from heterozygous parents to their offspring. This method essentially uses the parental untransmitted alleles as controls for the study. The program TRANSMIT was used to implement this analysis (section 2.2.10). In this dataset, as the use of the TDT protects against the effects of population stratification, all ethnicities in our sample set were included in the following analyses.

Pentanucleotide

In the analysis of the pentanucleotide repeat in the parent-offspring trios, TDT revealed no evidence of over-transmission from heterozygous parents to the affected offspring (global analysis, $\chi 2$= 4.15, df= 4, *P*= 0.39, P4 transmissions $\chi 2$=0.25, df=1, *P*=0.62. table 4.5).

Locus	Pentanucleotide	Observed	Expected	$\chi 2$
Allele	P4 (169bp)	183	186.2	0.25
	P6 (179bp)	90	86.7	0.33
	P7 (184bp)	0	0	0
	P8 (189bp)	23	18.8	1.95
	P9 (194bp)	28	31.9	1.04
	P10 (199bp)	0	0	0

Table 4.5: TDT analysis at the pentanucleotide in UK trios: There is no evidence of excess transmission at the pentanucleotide repeat. Global $\chi 2$ pentanucleotide repeat= 4.15, df= 4, *P*=0.39, P4 transmissions $\chi 2$=0.25, df=1, *P*=0.62.

Tetranucleotide

When the tetranucleotide repeat was analysed, it was found that there was no suggestion of excess transmission (global analysis $\chi 2$= 5.26 df=5, *P*=0.39, T8 transmissions $\chi 2$= 0.007, df=1, *P*=0.93, table 4.6).

Locus	Tetranucleotide	Observed	Expected	$\chi 2$
Allele	T6 (276bp)	0	0	0
	T8 (284bp)	269	269.4	0.01
	T9 (288bp)	1	25	1.88
	T10 (292bp)	0	0.5	1.03
	T11 (296bp)	10	8.1	1.04
	T12 (300bp)	44	41.9	0.22
	T13 (304bp)	2	3.5	1.36

Table 4.6: TDT analysis at the tetranucleotide in UK trios. There is no evidence of excess transmission at the tetranucleotide repeat. Global $\chi 2$ tetranucleotide repeat= 5.26 df=5, *P*=0.39 T8 transmissions $\chi 2$= 0.007, df=1, *P*=0.93.

Global haplotype frequencies using the program TRANSMIT were also examined for departures from the expected for the transmission of haplotypes in the trios. As shown in table 4.7, there was no evidence of excess haplotype transmissions (global haplotype analysis $\chi 2$ = 4.41 df= 5, *P*=0.49, P4/T8 haplotype χ^2= 0.017, df=1, *P*=0.90).

Locus	Pentanucleotide/ Tetranucleotide	Observed	Expected	$\chi 2$
Haplotype	P4/T8 (169/284)	180.9	180.1	0.02
	P6/T8 (179/284)	37.6	40.4	0.48
	P6/T12 (179/300)	36.0	34.6	0.14
	P9/T8 (194/284)	26.5	27.3	0.06
	P8/T8 (189/284)	23.0	18.3	2.51

Table 4.7: Haplotype analysis at the pentanucleotide and the tetranucleotide in UK trios. There is no evidence of excess transmission of haplotypes in the UK trios. Global $\chi 2$ haplotype analysis = 4.41, df= 5, *P*=0.49, P4/T8 haplotype $\chi 2$ = 0.017, df=1, *P*=0.90.

4.5 The Northern Finnish Birth Cohort of 1966 (NFBC1966)

Pentanucleotide Repeat

Allelic and genotypic frequencies of both microsatellites were examined in the NFBC 1966. As outlined in section 2.1.3, all subjects were part of an ongoing study established in 1966 examining foetal mortality of babies with expected dates of delivery in this year. At age 31 years, women who remained in the northern areas of Finland and Helsinki, were asked to fill in a questionnaire to determine the extent of PCOS-related symptoms (hirsutism and irregular menses) in this cohort. DNA was collected from these women. Those who reported PCOS related symptoms were then classified as PCO-symptomatic women and are utilized in this study as such (symptomatic 'cases'). Those women who did not report PCOS-related symptoms were classified as cohort controls for this study. For each case, two controls were obtained from the same birth cohort.

The program STATXACT was used to perform comparisons of the allelic and genotypic frequencies of both the pentanucleotide and tetranucleotide repeat (see section 2.2.10). Analyses of the allelic frequencies at the pentanucleotide repeat between the symptomatic 'cases' and cohort controls did not reveal any significant differences (Monte-Carlo estimate, global P= 0.269, 99% confidence interval 0.265-0.273, table 4.8). There were also no significant differences between genotypic frequencies at the pentanucleotide repeat (Monte-Carlo estimate, global P= 0.075, 99% confidence interval 0.073-0.077, table 4.8).

Locus	Pentanucleotide	Symptomatic 'Cases'	Cohort Controls
		N (%)	N (%)
Allele counts	P4 (212bp)	661 (64.4)	1384 (66.3)
	P6 (222bp)	262 (25.4)	504 (24.1)
	P7 (232bp)	80 (7.8)	130 (6.2)
	P8 (237bp)	29 (2.8)	70 (3.4)
Genotype counts	P4/P4 (212/212)	215 (41.7)	470 (45.0)
	P4/P6 (212/222)	162 (31.4)	332 (31.8)
	P4/P7 (212/232)	56 (10.9)	68 (6.5)
	P4/P9 (212/237)	13 (2.5)	44 (4.2)
	P6/P6 (222/222)	34 (6.6)	56 (5.4)
	P6/P7 (222/232)	19 (3.7)	43 (4.1)
	P6/P9 (222/237)	13 (2.5)	17 (1.6)
	The rest[a]	4 (0.8)	14 (1.4)

Table 4.8: Allele and genotype frequencies at the pentanucleotide repeat in the NFBC1966 cases and controls. There are no differences in allelic or genotypic frequencies at the pentanucleotide repeat (global P=0.269 and P= 0.075 respectively). [a]all genotypes <1% frequency are pooled for the purposes of this table: contingency table analyses were performed on unpooled counts.

Tetranucleotide Repeat

Examination of the tetranucleotide repeat in the NFBC1966 also did not reveal any differences in either allelic counts (global exact P=0.289, 99% confidence interval 0.285-0.293) or genotypes as shown in table 4.9 (Monte-Carlo estimate, global P= 0.186, 99% confidence interval 0.183-0.189).

Locus	Tetranucleotide	Symptomatic 'Cases'	Cohort Controls
		N (%)	N (%)
Allele Counts	T6 (276bp)	0	5 (0.3)
	T8 (284bp)	807 (81.2)	1586 (79.7)
	T9 (288bp)	27 (2.7)	49 (2.5)
	T10 (292bp)	0	2 (0.1)
	T11 (296bp)	42 (4.2)	77 (3.9)
	T12 (300bp)	115 (11.5)	268 (13.5)
	T13 (304bp)	3 (0.3)	5 (0.3)
Genotype Counts	T8/T8 (284/284)	327 (65.8)	624 (62.7)
	T8/T9 (284/288)	20 (4.0)	41 (4.1)
	T8/T11 (284/296)	39 (7.8)	66 (6.6)
	T8/T12 (284/300)	91 (18.3)	223 (22.4)
	T12/T12 (300/300)	7 (1.4)	15 (1.5)
	The rest[a]	13 (1.2)	27 (2.3)

Table 4.9: Allele and genotype frequencies at the tetranucleotide repeat in the NFBC1966 cases and controls. There was no significant difference of allelic or genotypic frequencies at the tetranucleotide repeat (global P=0.289 and global P=0.186 respectively). [a]all genotypes <1% frequency are pooled for the purposes of this table: contingency table analyses were performed on unpooled counts.

Global haplotype frequencies were calculated using the program TRANSMIT for the pentanucleotide and the tetranucleotide in the symptomatic 'cases' and cohort controls of the NFBC1966 sample set and were not found to be significantly different (P=0.27, table 4.10).

Locus	Pentanucleotide	Tetranucleotide	Symptomatic 'Cases'	Cohort Controls
			Frequency	Frequency
Haplotype	P4 (169bp)	T8 (284bp)	0.628	0.635
	P6 (179bp)	T8 (284bp)	0.079	0.065
	P9 (194bp)	T8 (284bp)	0.104	0.097
	P4 (169bp)	T12 (300bp)	0.002	0.004
	P6 (179bp)	T12 (300bp)	0.114	0.130
	P9 (194bp)	T12 (300bp)	0.000	0.000
	P4 (169bp)	T11 (296bp)	0.009	0.015
	P6 (179bp)	T11 (296bp)	0.058	0.053
	P9 (194bp)	T11 (296bp)	0.005	0.002

Table 4.10: Haplotype analysis at the pentanucleotide and tetranucleotide repeat in the NFBC1966 symptomatic cases and controls. There is no significant difference in haplotype frequencies (P=0.27).

4.6 Intermediate Trait Analyses

A cardinal feature of PCOS is the disturbance of ovarian steroidogenesis, with excessive androgen production responsible for many of the principal symptomatic manifestations (Franks, 1995). As the initial, smaller studies examining the pentanucleotide repeat and serum testosterone levels in PCOS women identified a positive association, (Gharani *et al.*, 1997); (Diamanti-Kandarakis *et al.*, 2000); (*Daneshmand et al.*, 2002), both the pentanucleotide repeat and the neighbouring tetranucleotide repeat were examined in the UK and NFBC1966 sample sets. As individuals with PCOS display metabolic abnormalities such as hyperinsulinaemia and abnormal LH secretion, abnormalities that are often exacerbated by obesity, these traits were also examined for association at both microsatellites in the UK and NFBC1966 sample sets. In the UK sample set, intermediate trait information was unavailable for the randomly selected population controls, so these analyses were performed in all PCOS cases (i.e. a combination of the

"cases" and the parent-offspring "probands") not receiving oral contraception or who were not pregnant at the time of sampling in the UK sample set. Furthermore, the all PCOS cases used in the UK sample set were restricted to those of UK origin only.

4.6.1 Testosterone

The relationship between both microsatellite genotypes and testosterone levels were analysed in the UK and the NFBC1966 sample sets. Quantitative trait analysis was performed by ANOVA on log-transformed values using SPSS (version 10) and SAS (version 8.2) (see section 2.1.10).

4.6.1.1 UK Sample Set

Pentanucleotide Repeat

There was no association between genotype at the pentanucleotide repeat and testosterone levels (pentanucleotide, global P=0.98, table 4.11). In particular, given previous findings with the P4 allele (Gharani *et al.*, 1997), testosterone levels were compared in subjects with one or more P4 allele (P4/-vs. the rest) (Gharani *et al.*, 1997), (Diamanti-Kandarakis *et al.*, 2000) with all other subjects and no appreciable difference was found (P4/-vs. the rest, geometric mean (SD range) 2.10 (1.3-3.3) nmol/l; non-P4, 1.9 (1.3-2.9) nmol/l, P=0.29).

Genotype	(N)[a]	Testosterone (nmol/l)
P4/P4	112	2.1 (1.3-3.2)
P4/P6	60	2.2 (1.4-3.5)
P4/P8	13	2.1 (1.5-2.9)
P4/P9	30	2.1 (1.2-3.6)
P6/P6	8	2.0 (1.5-2.8)
P6/P8	6	2.1 (1.4-3.0)
The rest	16	1.8 (1.2-2.8)
Significance tests		
Global	245	P=0.98
P4/-vs rest	245	P=0.29

Table 4.11:Pentanucleotide repeat genotypes vs testosterone in the UK subjects. Results show no appreciable difference between genotype and testosterone (P=0.98) or between P4/-vs. the rest and testosterone (P=0.29). Testosterone values are expressed as geometric mean (SD range) nmol/L. [a] PCOS subjects includes both cases and trios probands. These analyses only include women not receiving oral contraceptive therapy at the time of sampling.

Tetranucleotide Repeat

Comparison of the tetranucleotide genotypes with testosterone levels revealed no association, as shown in table 4.12 (global P=0.87). Groupings similar to the P4 allele described above (P4/-vs. the rest) was not completed for the T8 allele in the UK sample set as the groups not containing the T8 allele were too few to be obtain an accurate measurement.

Genotype	(N)[a]	Testosterone (nmol/l)
T8/T8	192	2.1 (1.4-3.4)
T8/T11	6	1.7 (0.9-3.4)
T8/T12	42	2.1 (1.3-3.4)
The rest	9	2.5 (2-3.1)
Significance tests		
Global	249	*P*=0.87

Table 4.12:Tetranucleotide repeat genotypes vs testosterone in the UK subjects. Results show no appreciable difference between genotype and testosterone (*P*=0.87). Testosterone values are expressed as geometric mean (SD range) nmol/L. [a] PCOS subjects include both cases and trios probands; These analyses only include women not receiving oral contraceptive therapy at the time of sampling.

4.6.1.2 NFBC1966

Pentanucleotide Repeat

As shown in table 4.13, there was no significant relationship seen between pentanucleotide genotypes and testosterone levels (symptomatic 'cases' P=0.89; cohort controls P=0.86). Furthermore, comparison of testosterone levels in P4/- individuals and those individuals with other genotypes revealed no significant difference (symptomatic' cases': P4/-, geometric mean (SD range) 2.0 (1.4-3.1) nmol/l vs non-P4, 2.1 (1.4-3.3) nmol/l, *P*=0.49; cohort controls: P4/-, 1.8 (1.2-2.8) nmol/l vs. non-P4, 1.9 (1.3-2.8) nmol/l, *P*=0.39, table 4.13).

	Symptomatic 'Cases'		Cohort controls	
Genotype	(N)[a]	Testosterone (nmol/l)	(N)[a]	Testosterone (nmol/l)
P4/P4	195	2.1 (1.3-3.2)	426	1.8 (1.2-2.8)
P4/P6	150	2.0 (1.4-3.5)	299	1.9 (1.2-2.9)
P4/P8	50	2.0 (1.4-3.0)	64	1.8 (1.3-2.6)
P4/P9	13	2.1 (1.6-2.7)	38	1.8 (1.2-2.8)
P6/P6	29	2.1 (1.4-3.2)	48	1.8 (1.2-2.8)
P6/P8	16	2.3 (1.5-3.3)	39	1.9 (1.3-2.6)
The rest	14	2.0 (1.2-3.4)	30	2.0 (1.4-3.0)
Significance tests				
Global	467	*P*=0.89	944	*P*=0.86
P4/- vs. rest	467	*P*=0.49	944	*P*=0.39

Table 4.13: Pentanucleotide repeat genotypes vs testosterone in the NFBC1966 symptomatic cases and controls. Results show no appreciable difference between genotype and testosterone for symptomatic 'cases' (*P*=0.89) or cohort controls (*P*=0.86). No difference was also seen between P4/-vs. the rest and testosterone in symptomatic 'cases' (*P*=0.49) or cohort controls (*P*=0.39). Testosterone values are expressed as geometric mean (SD range). [a] These analyses only include women not receiving oral contraceptive therapy at the time of sampling;

Tetranucleotide Repeat

When the tetranucleotide repeat was examined in the NFBC1966, there was also no association between this repeat and testosterone in the symptomatic 'cases' or in the cohort controls (global symptomatic 'cases' *P*=0.27 and cohort controls *P*=0.56 respectively, table 4.14). When grouped for T8 alleles/-vs. the rest, no significant difference was seen in either the symptomatic 'cases' (*P*=0.67) or cohort controls (*P*=0.18) as shown in table 4.14.

	Symptomatic 'Cases'		Cohort controls	
Genotype	(N)[a]	Testosterone (nmol/l)	(N)[a]	Testosterone (nmol/l)
T8/T8	298	2.1 (1.4-3.1)	565	1.8 (1.2-2.7)
T8/T12	79	2.1 (1.4-3.1)	201	1.9 (1.2-3.0)
T8/T11	36	2.0 (1.3-3.1)	61	1.8 (1.2-2.7)
T8/T19	20	2.1 (1.5-2.9)	37	1.8 (1.2-2.7)
T12/T12	6	1.7 (1.2-2.4)	13	1.8 (1.2-2.6)
The rest	10	2.8 (2.2-3.6)	23	1.9 (1.2-2.8)
Significance tests				
Global	449	*P*=0.27	900	*P*=0.56
T8/- vs rest	449	*P*=0.67	900	*P*=0.18

Table 4.14:Tetranucleotide repeat genotypes vs testosterone in the NFBC 1966 symptomatic 'cases' and cohort controls. Results show no appreciable difference between genotype and testosterone for symptomatic cases (*P*=0.27) or cohort controls (*P*=0.56) or between T8/-vs. the rest and testosterone in symptomatic 'cases' (*P*=0.67) or cohort controls (*P*=0.18). Testosterone values are expressed as geometric mean (SD range). [a] These analyses only include women not receiving oral contraceptive therapy at the time of sampling;

4.6.2 Other Quantitative Trait Variables

As mentioned above, other clinical features of PCOS include hyperinsulinaemia, gonadotropin abnormalities and obesity. Therefore, the relationships between microsatellites and BMI, weight, insulin and LH were analysed by ANOVA using SPSS (version 10) and SAS (version 8.2) (see section 2.2.10). Analyses were performed using both the UK and NFBC1966 sample sets where the corresponding clinical information was available.

BMI

When subjects were examined for an association between genotype at either locus and BMI in the UK sample set, no association was found (pentanucleotide, P=0.20, table 4.15; tetranucleotide P=0.48, table 4.16). Again as above, when BMI in subjects were compared with one or more P4 allele, no appreciable difference was found (P4/-vs rest, P=0.89, table 4.15).

Pentanucleotide	(N)[a]	BMI (kgm^{-2})
P4/P4	126	26.3 (20.4-33.9)
P4/P6	86	26.3 (20.4-33.9)
P4/P8	18	24.0 (20.8-27.7)
P4/P9	30	28.2 (21.9-36.3)
P6/P6	11	24.0 (20.7-27.8)
P6/P8	10	26.3 (19.5-35.5)
The rest	21	28.9 (21.7-38.5)
Significance tests		
Global	302	P=0.20
P4/-vs the rest	302	P=0.89

Table 4.15: Pentanucleotide repeat genotypes vs. BMI in the UK subjects. There was no difference between pentanucleotide repeat genotypes and BMI in the UK subjects (P=0.20; grouped P4/-vs. the rest, P=0.89). BMI is represented as the geometric mean (SD range) kgm^{-2}. [a] PCOS subjects include both the UK cases and the trios probands.

Tetranucleotide	(N)[a]	BMI (kgm^{-2})
T8/T8	229	26.3 (20.4-33.9)
T8/T11	8	26.9 (20.9-34.7)
T8/T12	59	26.3 (20.4-33.9)
The rest	12	26.2 (19.0-36.2)
Significance tests		
Global	308	P=0.48

Table 4.16: Tetranucleotide repeat genotypes vs BMI in the UK subjects. There was no difference between tetranucleotide repeat genotypes and BMI in the UK subjects (P=0.48). BMI is represented as the geometric mean (SD range) kgm^{-2}. [a] PCOS subjects include both the UK cases and the trios probands.

In neither the group of symptomatic 'cases' nor the cohort controls was any significant relationship seen between pentanucleotide (symptomatic 'cases' P=0.99; cohort controls P=0.78, table 4.17) or tetranucleotide genotype (symptomatic 'cases' P=0.86, cohort controls P=0.66, table 4.18) and BMI. Again as above, BMI in subjects were compared with one or more P4 allele vs. the rest and no appreciable difference was found (P4/-vs the rest symptomatic 'cases' P=0.50, cohort controls P=0.91, table 4.17). In addition, comparison of subjects with one or more T8 allele with all other subjects revealed no significant difference in BMI (symptomatic 'cases' P=0.56, cohort controls P=0.29, table 4.18).

Genotype	Symptomatic 'Cases' (N)	BMI (kgm^{-2})	Cohort Controls (N)	BMI (kgm^{-2})
P4/P4	194	24.5 (20.7-28.9)	421	23.9 (20.0-28.5)
P4/P6	152	24.6 (20.1-30.2)	296	23.8 (20.0-28.3)
P4/P8	50	24.6 (19.9-30.3)	64	23.6 (19.9-27.9)
P4/P9	13	25.3 (20.8-30.6)	37	23.8 (20.1-28.2)
P6/P6	29	25.0 (19.7-31.6)	46	23.0 (19.8-26.7)
P6/P8	16	25.1 (20.7-30.5)	39	24.1 (20.7-28.1)
The rest	13	25.0 (18.8-33.2)	29	24.5 (19.6-30.7)
Significance tests				
Global	467	P=0.99	932	P=0.78
P4/- vs rest	467	P=0.50	932	P=0.91

Table 4.17: Pentanucleotide repeat genotypes and BMI in the NFBC1966 sample set. Results show no appreciable difference between genotype and BMI for symptomatic 'cases' (P=0.99) or cohort controls (P=0.78) or between P4/-vs the rest and BMI in symptomatic 'cases' (P=0.50) or cohort controls (P=0.91). BMI values are expressed as geometric mean (SD range).

Genotype	Symptomatic 'Cases' (N)	BMI (kgm^{-2})	Cohort Controls (N)	BMI (kgm^{-2})
T8/T8	300	23.9 (20.4-29.8)	560	23.9 (20.0-28.6)
T8/T12	81	25.0 (20.0-31.1)	198	24.0 (20.1-28.5)
T8/T11	36	24.0 (19.8-29.1)	60	23.3 (20.2-26.9)
T8/T9	20	23.8 (20.4-27.8)	38	23.8 (19.5-29.0)
T12/T12	5	24.4 (20.8-28.7)	12	22.7 (19.2-26.7)
The rest	10	23.9 (19.3-29.5)	22	23.1 (19.8-26.9)
Significance tests				
Global	452	P=0.86	892	P=0.66
T8/- vs rest	452	P=0.56	892	P=0.29

Table 4.18: Tetranucleotide repeat genotypes and BMI in the NFBC1966 sample set. Results show no appreciable difference between genotype and BMI for symptomatic cases (P=0.86) or cohort controls (P=0.66) or between T8/-vs the rest and BMI in symptomatic 'cases' (P=0.56) or cohort controls (P=0.29). BMI values are expressed as geometric mean (SD range).

The equivalent analyses for the traits listed below also failed to show any evidence for association with microsatellite genotypes:

- *Weight (WT):* For data in the UK sample set, please see table 4.19 for the pentanucleotide, and table 4.20 for the tetranucleotide. Furthermore, for similar comparisons in the NFBC1966 sample set, see table 4.21 for the pentanucleotide and table 4.22 for the tetranucleotide.
- *Fasting Insulin*: For data in the UK sample set, please see table 4.23 for the pentanucleotide, and table 4.24 for the tetranucleotide. Furthermore, for similar comparisons in the NFBC1966 sample set, see table 4.25 for the pentanucleotide and table 4.26 for the tetranucleotide.
- *Luteinising Hormone (LH):* For data in the UK sample set, please see table 4.27 for the pentanucleotide, and table 4.28 for the tetranucleotide. LH measurements for the NFBC1966 sample set were not available.

Pentanucleotide	(N)[a]	Weight (kg)
P4/P4	126	72.0 (55.3-93.6)
P4/P6	86	70.8 (55.0-91.2)
P4/P8	18	66.1 (55.3-78.9)
P4/P9	30	75.9 (57.5-100.0)
P6/P6	11	67.6 (57.2-80.0)
P6/P8	10	74.1 (55.0-100.0)
The rest	21	80.2 (58.7-109.4)
Significance tests		
Global	302	P=0.31
P4/-vs the rest	302	P=0.87

Table 4.19:Pentanucleotide repeat genotypes vs WT in the UK subjects. There was no difference between pentanucleotide repeat genotypes and WT in the UK subjects (P=0.31; grouped P4/-vs the rest, P=0.87). WT is represented as geometric mean (SD range). [a] PCOS subjects include both the UK cases and the trios probands.

Tetranucleotide	(N)[a]	Weight (kg)
T8/T8	230	72.4 (55.0-95.5)
T8/T11	8	70.7 (56.9-88.1)
T8/T12	59	70.8 (53.7-93.3)
The rest	12	71.6 (52.9-96.9)
Significance tests		
Global	309	P=0.61

Table 4.20: Tetranucleotide repeat genotypes vs WT in the UK subjects. There was no difference between tetranucleotide repeat genotypes and WT in the UK subjects (P=0.61). WT represented as geometric mean (SD range) kg. [a] PCOS subjects include both the UK cases and the trios probands.

Genotype	Symptomatic 'Cases' (N)	Weight (kg)	Cohort Controls (N)	Weight (kg)
P4/P4	194	65.8 (55.3-78.3)	423	64.7 (53.5-78.2)
P4/P6	152	66.7 (54.2-82.1)	296	64.1 (53.3-76.9)
P4/P8	50	65.0 (52.0-81.4)	64	64.6 (53.6-77.8)
P4/P9	13	71.2 (58.8-86.1)	37	66.2 (54.4-80.5)
P6/P6	29	66.9 (52.5-85.4)	46	63.6 (54.1-74.7)
P6/P8	16	66.1 (54.3-80.4)	39	65.1 (55.7-76.1)
The rest	13	67.5 (48.7-93.5)	29	66.7 (52.6-84.6)
Significance tests				
Global	467	P=0.85	935	P=0.81
P4/- vs rest	467	P=0.37	935	P=0.89

Table 4.21: Pentanucleotide repeat genotypes and WT in the NFBC1966 sample set. Results show no appreciable difference between genotype and weight for symptomatic cases (P=0.85) or cohort controls (P=0.81) or between P4/-vs. the rest and WT in symptomatic cases (P=0.37) or cohort controls (P=0.89). WT values are expressed as geometric mean (SD range).

Genotype	Symptomatic 'Cases' (N)	Weight (kg)	Cohort Controls (N)	Weight (kg)
T8/T8	300	66.5 (54.5-81.1)	562	64.8 (53.5-78.5)
T8/T12	81	66.1 (52.9-82.6)	198	64.8 (54.0-77.8)
T8/T11	36	64.6 (52.9-78.8)	60	62.5 (53.7-72.7)
T8/T9	20	64.9 (55.2-76.2)	38	65.9 (53.2-81.5)
T12/T12	5	62.9 (52.7-75.2)	12	63.3 (53.1-75.3)
The rest	10	65.1 (52.0-81.5)	22	64.1 (56.0-73.4)
Significance tests				
Global	452	P=0.92	892	P=0.77
T8/- vs rest	452	P=0.48	892	P=0.67

Table 4.22: Tetranucleotide repeat genotypes and WT in the NFBC1966 sample set. Results show no appreciable difference between genotype and weight for symptomatic cases (P=0.92) or cohort controls (P=0.77) or between T8/-vs. the rest and WT in symptomatic cases (P=0.48) or cohort controls (P=0.67). WT values are expressed as geometric mean (SD range).

Pentanucleotide	(N)	Insulin (pmol/l)
P4/P4	97	33.1 (9.1-120.2)
P4/P6	55	25.7 (10.0-66.1)
P4/P8	16	16.6 (4.1-67.6)
P4/P9	29	28.2 (9.8-81.3)
P6/P6	6	20.0 (9.6-41.7)
P6/P8	7	43.7 (15.5-123.0)
The rest	18	11.0 (5.2-23.3)
Significance tests		
Global	228	P=0.22
P4/-vs the rest	228	P=0.15

Table 4.23: Pentanucleotide repeat genotypes vs. fasting insulin in UK subjects. There was no difference between pentanucleotide repeat genotypes and insulin in the UK subjects (P=0.22; grouped P4/-vs. the rest, P=0.15). Insulin is represented as geometric mean (SD range). [a] PCOS subjects include both the UK cases and the trios probands.

Tetranucleotide	(N)[a]	Insulin (pmol/l)
T8/T8	177	31.6(8.9-112.2)
T8/T11	4	12.6 (4.2-38.0)
T8/T12	42	22.9 (8.9-58.9)
The rest	8	11.0 (54.8-24.2)
Significance tests		
Global	231	*P*=0.33

Table 4.24: Tetranucleotide repeat genotypes vs. fasting insulin in UK subjects. There was no difference between tetranucleotide genotypes and insulin in the UK subjects (*P*=0.33). Insulin is represented as geometric mean (SD range). [a] PCOS subjects include both the UK cases and the trios probands.

Genotype	Symptomatic 'Cases' (N)	Insulin (pmol/l)	Cohort controls (N)	Insulin (pmol/l)
P4/P4	189	52.9 (36.9-76.0)	402	51.1 (36.5-71.5)
P4/P6	145	53.9 (34.5-84.1)	290	51.5 (36.6-72.5)
P4/P8	49	55.9 (38.4-81.3)	62	50.3 (35.4-71.5)
P4/P9	13	52.6 (35.7-77.4)	38	49.7 936.9-67.0)
P6/P6	27	59.2 (37.5-93.5)	46	50.8 (33.1-78.0)
P6/P8	14	52.7 (35.5-78.1)	38	53.3 (39.1-72.6)
The rest	14	66.0 (37.3-116.7)	29	56.9 (38.7-83.6)
Significance tests				
Global	451	*P*=0.47	905	*P*=0.72
P4/- vs rest	451	*P*=0.10	905	*P*=0.26

Table 4.25: Pentanucleotide repeat genotypes and fasting insulin in the NFBC1966 sample set. Results show no appreciable difference between genotype and insulin for symptomatic cases (*P*=0.47) or cohort controls (*P*=0.72) or between P4/-vs. the rest and insulin in symptomatic cases (*P*=0.10) or cohort controls (*P*=0.26). Insulin values are expressed as geometric mean (SD range).

Genotype	Symptomatic 'Cases' (N)	Insulin (pmol/l)	Cohort controls (N)	Insulin (pmol/l)
T8/T8	292	53.2 (36.3-77.9)	541	51.1 (36.1-72.3)
T8/T12	80	53.7 (35.0-82.4)	193	54.2 (39.3-74.8)
T8/T11	29	56.5 (34.0-93.9)	61	48.6 (34.4-68.7)
T8/T9	19	55.4 (34.4-89.4)	35	52.8 (37.8-73.9)
T12/T12	6	49.7 (35.2-70.1)	13	53.6 (28.8-99.7)
The rest	10	70.4 (41.1-120.7)	21	42.6 (33.5-54.3)
Significance tests				
Global	441	P=0.41	870	P=0.02
T8/- vs rest	441	P=0.13	870	P=0.22

Table 4.26: Tetranucleotide repeat genotypes and insulin in the NFBC1966 sample set. Results show no appreciable difference between genotype and insulin for symptomatic cases (P=0.41) but significant in the controls (P=0.02). The interpretation of this may be adjusted in light of the number of tests performed. There is no difference between T8/-vs. the rest and insulin in symptomatic cases (P=0.13) or controls (P=0.22). Insulin values are expressed as geometric mean (SD range).

Pentanucleotide	(N)[a]	LH (IU/l)
P4/P4	111	5.3 (2.1-12.9)
P4/P6	59	6.3 (2.9-13.8)
P4/P8	13	6.0 (2.3-15.9)
P4/P9	30	6.8 (3.0-15.1)
P6/P6	7	6.0 (4.1-8.9)
P6/P8	6	6.3 (2.9-13.8)
The rest	16	5.2 (4.3-6.3)
Significance tests		
Global	242	P=0.93
P4/-vs the rest	242	P=0.54

Table 4.27: Pentanucleotide repeat genotypes vs LH in UK subjects. There was no difference between the pentanucleotide repeat genotypes and LH in the UK subjects (P=0.93; grouped P4/-vs. the rest, P=0.54). [a] PCOS subjects include both the UK cases and the trios probands. LH represented as geometric mean (SD range).

Tetranucleotide	(N)[a]	LH (IU/l)
T8/T8	191	5.8 (2.4-13.8)
T8/T11	6	5.0 (2.4-10.5)
T8/T12	40	6.3 (3.0-13.2)
The rest	9	7.5 (4.1-13.6)
Significance tests		
Global	246	*P*=0.86

Table 4.28: Tetranucleotide repeat genotypes vs. LH in UK subjects: There was no difference between the tetranucleotide repeat genotypes and LH in the UK subjects (*P*=0.86). [a] PCOS subjects include both the UK cases and the trios probands. LH represented as geometric mean (SD range).

4.7 Discussion

This study attempts to address the role of *CYP11A1* in PCOS by using the considerable resources available with respect to DNA samples and clinical information. Both microsatellites have been typed in 2 different populations, representing the largest association study in PCOS of this locus to date. The first population group of samples represent 331 controls, taken from a random UK population and 371 PCOS cases have been collected section 2.1). Of these 371 cases, 141 were from a set of parent/offspring trios. The second group represents DNA and other clinical information collected from 527 PCOS cases and 1062 population controls from the Northern Finnish Birth Cohort 1966 (see section 2.1.3) (Taponen *et al.*, 2003).

The major finding of this study was that there was no association detected between the pentanucleotide located in the 5' promoter region of *CYP11A1* and serum testosterone levels in PCOS women, which is contrary to what has been previously reported (Gharani *et al.*, 1997); (Diamanti-Kandarakis *et al.*, 2000); (Daneshmand *et al.*, 2002) (table 4.12). In addition, no association between the novel tetranucleotide located 840 bp into intron 1 of *CYP11A1* and serum testosterone levels were identified in this study (table 4.13). Due to the large sample size, the lack of reproducibility of the association results cannot be attributed to lack of power (the probability of detecting a true association).

Furthermore, it is unlikely that the differences between these results and those in the literature are due to sample set differences in linkage disequilibrium relationships, for example, between the pentanucleotide and some, as yet unidentified aetiological variant. The close similarities in haplotype structure observed between the Finnish and UK

sample sets suggest that there are no profound differences in linkage disequilibrium structure, at least amongst populations of European origin (in whom the majority of the previous studies were conducted). LD analysis of both variants in the UK control group and the NFBC 1996 control group showed that they are in strong linkage disequilibrium (D'=0.86 and D'= 0.87 respectively). In addition, the failure to detect association was not due to genotyping error as approximately 10% of the samples were re-genotyped providing an overall estimate of the genotype error at <1%. Furthermore, the consistent pattern of linkage disequilibrium relationships between the microsatellites across multiple datasets, provide strong reassurance in this regard.

The UK sample set (case-controls) provide ~87% power (at alpha=0.05, 2-sided) to detect a case-control difference between non-P4 containing and P4-containing genotypes for an effect size (specified in terms of odds ratios (OR)) of ~2.0. The equivalent for the UK trios is ~64% power (at alpha=0.05, 2-sided). Power to detect differences in mean testosterone levels between the same two genotype groups in the UK dataset (cases and trio probands combined) reached ~80% for an effect size (between group mean difference) of ~0.56 of a SD (at alpha=0.05, 2-sided). The largest group studied for the relationship between *CYP11A1* variation and testosterone levels (the NFBC1966 controls) provided 80% power to detect (at alpha=0.05) differences between P4-containing and non P4-containing genotypes of~0.28 of a SD. Previous studies of the pentanucleotide (Gharani *et al.*, 1997; Diamanti-Kandarakis *et al.*, 2000) have suggested effect sizes at least twice this.

It is therefore concluded any associations between *CYP11A1* promoter variation and androgen-related phenotypes, have been overestimated in previous studies(Gharani *et al.*, 1997); (Daneshmand *et al.*, 2002); (Diamanti-Kandarakis *et al.*, 2000). This may be

partially due to small sample sizes and extensive data exploration without allowance for multiple testing (Bland and Altman, 1995). In addition, any publication bias favouring reporting of positive results may compound these issues. These factors are increasingly being recognised as contributing to the inconsistency which has affected association studies, particularly in the context of multifactorial traits (Ioannidis *et al.*, 2001); (Colhoun *et al.*, 2003).

There have been many solutions suggested in the literature to overcome the multitude of problems associated with replicating association studies (Ioannidis *et al.*, 2001); (Hirschhorn *et al.*, 2002); (Colhoun *et al.*, 2003); (Cardon and Bell, 2001). Several of these have been incorporated in the current study, including: (a) use of well-sized sample-sets; (b) attempts to seek replication of any nominally-significant findings, particularly those arising from data exploration; (c) use of both case-control and family-based association methods; and (d) analyses based on both intermediate phenotypes and discrete disease categories.

One of the reasons false positive results can be obtained in these types of studies is due to population stratification. If one of these groups has a higher disease prevalence than the others, stratification occurs, because that group will be over-represented in the cases and under-represented in the controls. Any polymorphism therefore found at a higher level in the cases will appear to be associated with disease, giving a false positive result. In PCOS cases of non-European origin, marked differences in the pentanucleotide repeat allele frequencies were found. For example, in both Afro-Caribbean and South Asian subjects, the P6, rather than the P4 allele, was the commonest allele at the pentanucleotide repeat, with frequencies of 46.6% and 48.6% respectively (table 4.29). This suggests that studies of these loci may be particularly susceptible to spurious

associations arising from the failure to match cases and control subjects adequately for ethnic composition. However, by excluding the non-UK samples from our cases-control dataset, this problem is minimized. As reported above, there were no significant differences in allelic frequencies with the NFBC1966 symptomatic cases compared with the cohort controls.

Allele Counts Pentanucleotide	UK Cases	Afro-Caribbean Cases	South Asian Cases
	N (%)	N (%)	N (%)
P4 (169bp)	429 (63.1)	18 (31.0)	23 (32.9)
P6 (179bp)	147 (21.6)	27 (46.6)	34 (48.6)
P7 (184bp)	0	3 (5.2)	0
P8 (189bp)	38 (5.6)	8 (13.8)	6 (8.6)
P9 (194bp)	65 (9.6)	2 (3.5)	6 (8.6)
P10 (199bp)	1 (0.2)	0	1 (1.4)

Table 4.29: Allele frequencies at the pentanucleotide repeat in the UK, Afro-Caribbean and South Asian cases. The P6 allele in both Afro-Caribbean and South Asian populations is the most commonest allele rather than the P4 allele, as seen in the UK sample set ($P<0.001$).

Family-based association studies such as the transmission disequilibrium test can eliminate problems such as population stratification. This is because in these studies, the heterozygous parents act as controls. In this thesis, 141 UK trios, from mixed ethnic origin, were analysed. The different ethnic origins included 112 British/Irish, 2 Afro-Caribbean, 2 Middle Eastern, 18 other European, 5 South Asian and 2 other ethnicities. There was no suggestion of excess transmission for the pentanucleotide or the tetranucleotide repeat in the trios analysed here (tables 4.5 and 4.6). Therefore this study represents a well-designed attempt to address the poor performance of association

studies by incorporating adequate sample numbers and the use of both case-control and family-based association methods for association studies.

Various intermediate traits were also examined in the attempt to identify possible associations with either variant and known factors involved in polycystic ovary syndrome. In this study, analysis of traits such as BMI, weight, insulin, and LH failed to identify any such associations with either variant. It is important to point out that the present association study does not allow the conclusion that these variants have absolutely no effect on PCOS status, testosterone levels, or other intermediate traits analysed here. However, the size of the study means that any such effect must be modest, and certainly much smaller than the effect size implied by the previous positive findings (Gharani *et al.*, 1997); (Daneshmand *et al.*, 2002); (Diamanti-Kandarakis *et al.*, 2000). Functional studies of these microsatellites to determine whether they have any impact on *CYP11A1* expression have been completed. The results of these studies will be discussed in detail in chapter 5.

The current study also does not exclude the possibility that *CYP11A1* variants other than those studied contribute to PCOS susceptibility and population variation in testosterone levels. This will be discussed in chapter 6. However, it seems unlikely that the previously-reported associations at the pentanucleotide repeat were a reflection of linkage disequilibrium with aetiological variants elsewhere in the gene. Otherwise, given the apparent similarities in haplotype structure in these two quite distinct European populations between the pentanucleotide and tetranucleotide repeats, the same effects should have been detected in this study. However, until such time as the whole gene has been surveyed comprehensively to document and test all sources of genomic variation, no conclusive statements are possible.

5.0 Promoter Function

5.1 Introduction

The transcriptional regulation of *CYP11A1* has been well defined in several tissues including adrenal and placenta (Ben Zimra *et al.*, 2002); (Chung *et al.*, 1989); (Moore *et al.*, 1990). There are known tissue- and species-specific regulatory elements within the promoter region (see section 1.5.2) (Rodriguez *et al.*, 1997); (Morohashi *et al.*, 1993). A schematic diagram outlining the main regulatory elements in the *CYP11A1* promoter region is shown in section 1.5.2, figure 1.3.

There have been no studies published to date that have examined the role of the pentanucleotide repeat (as discussed in section 4.2.2) in the regulation of *CYP11A1*. There are many examples in the literature of similar microsatellites having a functional role in the regulation of genes including a pentanucleotide repeat polymorphism, located 1.3 kb 5' of the first exon of the apolipoprotein(a) gene, *APOA1*. Sequences at, or closely-linked to, the *APOA1* locus have been shown to contribute over 90% of the variability in plasma levels of lipoprotein(a) (Lp(a)). Most of the variability comes from a highly polymorphic tandemly repeated cystein-rich motif called kringle 4 (k4) (Utermann, 1989). However, subsequent studies have shown that approximately 10-14% of the variability in plasma levels of Lp(a) is associated with variation in the number of repeats of the $(TTTTA)_n$ repeat polymorphism (Trommsdorff *et al.*, 1995). Subjects with alleles with 11 repeats have been shown to produce the lowest plasma Lp(a) concentrations (Mooser *et al.*, 1995). In addition, expression studies have shown a 5-fold higher transcriptional activity in constructs containing 8 copies of the apolipoprotein(a) pentanucleotide repeat compared to those with 9 copies (Wade *et al.*, 1993). Furthermore, a polymorphic pentanucleotide repeat found in the 5' flanking

region of the sex hormone binding globulin gene (SHBG), directly influences its transcriptional activity by binding a nuclear factor to the pentanucleotide containing 6 repeats rather than 7-10 repeats. It has been hypothesized that this binding may contribute to differences in plasma serum levels of SHBG between individuals (Hogeveen *et al.*, 2001).

The similarities of the *APOA1* pentanucleotide repeat polymorphism and the *CYP11A1* pentanucleotide repeat in terms of sequence composition and genomic location suggest that the *CYP11A1* polymorphism may function in a similar manner. In addition, the presence and location of a tetranucleotide repeat 840 bp into intron 1 of *CYP11A1*, and upstream of the new alternate exon 1a, leads to the hypothesis that this repeat may also function to regulate *CYP11A1* transcripts discussed in chapter 3 (section 3.2.1). Therefore, the aim of this study was to determine if the *CYP11A1* pentanucleotide repeat and/or the tetranucleotide repeat had any effect on the levels of transcription.

5.2 Results

5.2.1 Transient Transfections

Transient transfection analysis is a conventional method for localizing transcriptional regulatory regions. Constructs containing the region to be studied are cloned into expression vectors and then transfected into the cells of choice.

For cloning of the pentanucleotide repeats, the promoterless luciferase basic expression vector pGL3-Basic (Promega, UK) was used. pGL3-Basic vector contains the firefly luciferase gene for identification of activation of the construct and a gene for ampicillin resistance allowing for vector positive clone selection. The luciferase enzyme is a

mono-oxygenase enzyme that catalyses bioluminescent reactions. Enzyme activity can be measured with a high degree of sensitivity, in terms of light production, and this light production is then indicative of activation or repression of the region to be studied. In addition, pGL3-Basic lacks eukaryotic promoter and enhancer sequences so that sensitive monitoring of cloned putative regulatory elements can be performed. To control for transfection efficiency, the pCMV-*Renilla* vector (a pUC-derived vector from Promega, UK) was co-transfected with pGL3-Basic constructs. This vector contains the CMV enhancer and early promoter elements for high-level and constitutive expression of the *Renilla* luciferase gene.

5.2.2 Nucleotide Repeat Constructs

To test for any functional roles for the pentanucleotide (P) repeat by transient transfection, DNA from subjects homozygous for the most common repeat lengths (4, 6 and 9 repeats as ascertained from studies described in chapter 4) was ligated into the promoterless pGL3-Basic vector. To generate the P4, P6 and P9 constructs, primers PentaF and PentaR (see table 5.1 for sequences) amplified the sequence between the 5' end of the clone C18438 (*PS-1*, see section 3.3.1 and figure 5.1) and the 5' end of *CYP11A1* exon 1, containing the pentanucleotide repeat. In addition, the primers were designed to incorporate, by PCR, a *Kpn*I site at the 5' end of the Penta F primer; and a *Hin*dIII at the 5' end of the Penta R primer. The resulting PCR fragment was purified using the Qiagen PCR purification kit (Qiagen, UK, section 2.4.12) and ligated into the PCR cloning vector, pGEM-T Easy (section 2.4.16). Once the pentanucleotide fragments were successfully cloned, all fragments were completely sequenced using primers listed in table 5.2 (see section 2.4.17) to ensure the correct sequence and selection of repeats. Those clones containing the verified sequences were digested with

*Kpn*I and *Hin*dIII for insertion into the transfection vector pGL3-Basic (which was also digested with *Kpn*I and *Hin*dIII, see section 2.4.10). Figure 5.2 shows the pentanucleotide-containing sequence cloned into pGL3-Basic. The sequence was the same for all constructs, except for the differences in repeat length.

Primer name	5'→3' sequence	$MgCl_2$ mM	Annealing °C
PentaF	cgcGGTACCcctcagtgctgagaccctt	1.5	60
PentaR	tatAAGCTTtgtacctgctccacttcagc	1.5	60
Tetra2F	cgcGGTACCccagaagtatggcccgatt	2.5	58
TetraR	tatAAGCTTacttgaccacgctgcgacc	2.5	58
0RepeatR	ttatGAATTCagcctaaggatgcatttc	1.5	60
0RepeatF	ttatGAATTCaagagagtttcattcttg	1.5	60

Table 5.1 PCR primers for the pentanucleotide and tetranucleotide repeat constructs. Primer sequences, with the associated concentration of $MgCl_2$ and annealing temperatures, are as listed. *Kpn*I (GGTACC), *Hin*dIII (AAGCTT) and *Eco*RI (GATTC) as shown. See text for details.

Primer Name	5'→3' Sequence
Penta1F	ggttcaacatgattggagt
Penta2F	ggcttttagtgtatcagaga
Penta3F	gcagcacagatcaagtgtgt
Penta4F	ccaagtgggctgtaattt

Table 5.2 Primers for sequencing of the pentanucleotide repeat constructs.

PentaF

62821 ggaatgtggg gctgcgtaga acaatgggat tgactttaag tcagaaagtt ataaatgtca cctcagtgct gagacccttg gaggaaaaac
62911 tagtccttgg aagactgctt ttctttgtgg aagctcatca ccctgccgct gctcgtgaga cactgccttc cttggctgat gtcattccag
63001 gctcaaggtc atcagtgagg caaaacaggc tttctccata ctctctttat cagaaggttc atgactgatg aggtagtggt cactccagcg
63091 ggaagagcaa caaccactct tgataagtac ttttttttt tcttctaaaa actgttgctc taaatttgtt gaaagtggtt caacagtatt
63181 ggagtctggg gtcaagtggc tgtgtaagga caacttttgc cattgtgggt gatctatggg ctgacacata caacagaaga ggccaggagg
63271 atgtcactcg tgtgtgtgtg tgtgtgtgtg tgtttggtaa cagctctatt gagatataat tcacacacca tacaattcac ccatttaaag
63361 tatgcaattc aatggctttt agtgtatcag agagttgttc aataattacc acaatatact ttagaatatt ttcattttca tcatcctaag
63451 aatccctaca cactttatgt cccatcctct aattcctcta tccccctagc tctaagcaac caccagtcta ttttctgtct cttagatgtg
63541 tgttttaaat gccttatttg ccagaaaata agattttggg gaaaaaaagc aagaagttgt accttccaat gataagtata aggtatagtg
63631 tagattgctg tcatgaaatg ctctacatgg cacatgtatg ttttctgacc tacacatttt taccaagtgg gctgtaattt tgttgaaatg
63721 acttgaaaac tctacaggtg actgaccctt gttgtttggg aaagtggtag gtgcagggtg atgggggtg ggtgtggggg ctgacctggg
63811 ctggaaggat gggtctgggg atatgatatg atgggagggg gtatggaagc agctctgagg ggaaagcagc acagatcaag tgtgtggaca
63901 gggaagagct gacatcctga gttccatggc aaccagattt gccaaggtct tagagtgtgt ccagagtgga gcctgaccac agacctcagc
63991 tcaagggacc cagagcccct ctgagtcagc tgtactgaat tacagcccca aatctgggtc aactggggag agacgacgag gattagggtt

0 Repeat F

64081 ccaaggtgaa actgtgccat tgcgctccag cctgggcaac aagaatgaaa ctctcttAAA ATAAAATAAA ATAAAATAAA ATAAAATAAA

0 Repeat R

64171 ATAAAATAAA Tagcctaagg atgcatttct cagaacttat ccctgttgtt caatgatgtg tgtctataca gtggggccat aactaagacg
64261 tatgttgccc aagctggcaa gatagctctg accttctctt gggcccctca tttcccccaa acacaggttg tctgcagtct tgaccaatgg
64351 ctgccagggc atggactccg ctgcaggggc cagtgggagg ccccagctca ggcaaaagca caggcagata tttcaggagt ctgctagggc
64441 tggcactgag ggcagagaca gaggggtctc cctgtccttt ggagaacctc acgctgcaga aattccagac tgaaccttca taccgagtag
64531 gggaggagct gtctgcgggt ttgagcctgc agcaggagga aggacgtgaa cattttatca gcttctggta tggccttgag ctggtagtta

Penta R

64621 taatcttggc cctggtggcc cagggctaca gtcatcctag CAGTCCCCGC TGAAGTGGAG CAGGTACAGT CACAGCTGTG GGGACAGCAT
64711 GCTGGCCAAG GGTCTTCCCC CACGCTCAGT CCTG

Figure 5.1: Sequence of pentanucleotide repeat and zero repeat promoter construct. Pentanucleotide repeat forward and reverse primer sequences in blue; zero repeat primer sequences in green. Pentanucleotide repeat in uppercase and in purple (P9 is shown). CYP11A1 exon 1 sequence underlined in black and uppercase. Nucleotides are numbered from 5' end of the genomic clone.

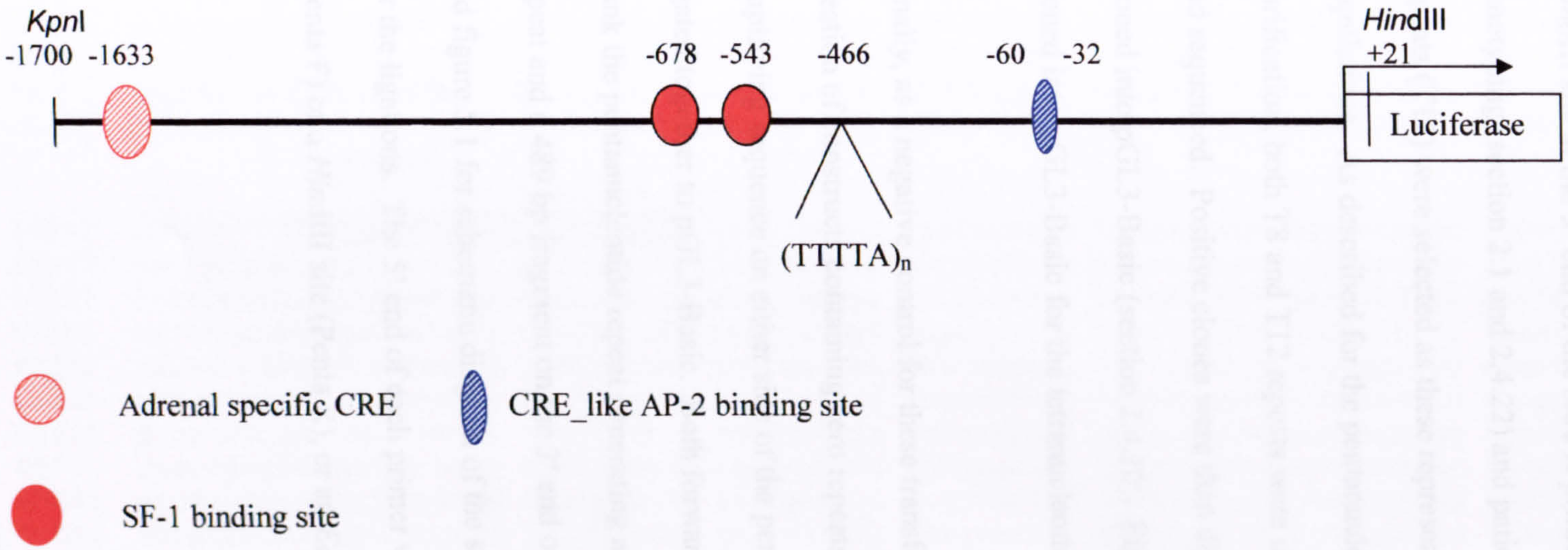


Figure 5.2: Schematic diagram of pentanucleotide repeat constructs ligated into the expression vector, pGL3-Basic. N= 4,6 or 9 repeats. Pentanucleotide repeat $(TTTTA)_n$ as shown. Direction of transcription is indicated by the arrow. Major regulatory elements are shown.

Primers were also designed to amplify a fragment surrounding the tetranucleotide repeat as with the pentanucleotide repeat. Forward (tetra2F) and reverse (tetraR) primers were designed to amplify sequence containing the repeat from the 3' end of *CYP11A1* exon 1 to the 5' end of *CYP11A1* exon 1a (section 3.1). Primer sequences, as listed in table 5.1, were also designed to incorporate a *Kpn*I site at the 5' end of the tetra 2F primer and a *Hind*III site at the 5' end of the tetra R primer. Patients were identified through genotyping (section 2.1 and 2.4.22) and patients homozygous for 8 repeats (T8) and 12 repeats (T12) were selected as these represent the most frequent repeat lengths in our populations. As described for the pentanucleotide repeat above, after PCR and purification, both T8 and T12 repeats were initially cloned into pGEM-T Easy vector and sequenced. Positive clones were then digested with both *Kpn*I and *Hind*III and cloned into pGL3-Basic (section 2.4.10). Figure 5.3 shows a diagram of the sequence cloned into pGL3-Basic for the tetranucleotide repeats T8 and T12.

Finally, as a negative control for these transfection experiments, a strategy for the creation of constructs containing zero repeats was produced. The strategy involved amplifying sequence on either side of the pentanucleotide repeat, which were then ligated together to pGL3-Basic. Both forward and reverse primers were designed to flank the pentanucleotide repeat generating a fragment of 1256 bp on the 5' side of the repeat and a 489 bp fragment on the 3' end of the repeat (table 5.1 for primer sequences and figure 5.1 for schematic diagram of the sequence). Figure 5.4 outlines the strategy for the ligations. The 5' end of each primer was designed to incorporate either *Kpn*I site (Penta F) or a *Hind*III site (Penta R), or an *Eco*RI site (0Repeat F and 0Repeat R).

64621 cttggccctg gtggcccagg gctacagtca tcctagcagt cccCGCTGAA GTGGAGCAGG TACAGTCACA GCTGTGGGGA
64701 CAGCATGCTG GCCAAGGGTC TTCCCCCACG CTCAGTCCTG GTCAAAGGCT GCCAGACCTT TCTGAGTGCC CCCAGGGAGG
64781 GGCTGGGGCG TCTCAGGGTG CCCACTGGCG AGGGAGCTGG CATCTCCACC CGCAGTCCTC GCCCCTTCAA TGAGATCCCC
64861 TCTCCTGGTG ACAATGGCTG GCTAAACCTG TACCATTTCT GGAGGGAGAC GGGCACACAC AAAGTCCACC TTCACCATGT

Tetra 2F

64941 CCAGAATTTC CAGAAGTATG GCCCGATTTA CAGgtaagcc tggcagaggg tgggagccga agggacaggg aggaggaggg
65021 gcctgggtag ccctgctgta gtccccactc ccccaacagg ctgttctctg gttccaaact cctgtctaac ttctttttt
65101 ttttttttt tggtaataca atacagaaat atatcaggga tatttcaaaa ttcaacattt gaatctgtga atgtcttaag
65181 tactctaaac cagacatcta cttttgaaac aattggtggc aagaccgcca gccatctgta aggaagctct ttaatgggtc
65261 caaaccccag gtctggccag gcatggtggc tcacacccat aatcccagca ctttgggagg ccgaggtggg cggatcactt
65341 gagatcaaga gttcgagacc atcctggcca acatgatgaa accccgtctc tactaaaaat acaaaaatta gctgggtgtg
65421 gtggcaggcg cctgtaatcc cagctactcg ggaggctgaa gcaggagaat ggcttgaacc ggagaggcag aggttgcagt
65501 gagccgagat cgcaccattg caatccagcc tgggcgacag agtgagattc tgtctcaaac aaaacaaaac aacctcctcg
65581 gtctgaggag gggagactgg tgaggctaag ttttgaattc ctggccctgt agctgtcagc ccagccagcg agttttaaaa
65661 attgggagaa ggccgggtgc ggtggctcac gcttgtaatc ccagcacttt gggaggctga ggcaggcgga tcacttgagg
65741 acaggagttc gagatcaacc cggccaacat ggtgaaaccc tgtctctact acaaaaatAA ATAAATAAAT AAATAAATAA
65821 ATAAATAAAA Taaaaaaaat tagccgggca tggtggcggg ggtctataat cccagctact cgggaggctg aggcaggaca
65901 ctcgcccgaa cccaggaggt ggaggttgca gtgagccgag atcatgtcac tgcactcccg ccctgggcaa caaaagcgaa

TetraR

65981 actctgtctc aaaaaaaaaa aaaaaaaaaa aatcggggga gggagatttg aaggtcagag cacagagtga aaaggtcgca

66061 gcgtggtcaa gtttgGCGCT AGACAGAC CCAGCATTCC TCCCAGCGCA CGCCGCCGCC

Figure 5.3:Sequence of the tetranucleotide repeat constructs. Associated tetra 2F (uppercase) and tetra R (lowercase) primer sequences in blue. Tetranucleotide repeat in uppercase and coloured red (T8 is shown). Exon 1 sequence inuppercase, exon 1a sequence is in uppercase and boxed. Nucleotides are numbered from 5' end of the genomic clone.

PCR was then performed using the patient DNA described in chapter 2 (section 2.1). Resultant PCR fragments were purified and digested with the appropriate restriction enzyme, as was pGL3-Basic (section 2.4.10). However, after numerous attempts at ligation, no cloned products were obtained and no studies using a construct with zero repeats were possible. Further work on the creation of zero repeat constructs for the tetranucleotide repeat were not completed due to the difficulties with the pentanucleotide zero repeat construct.

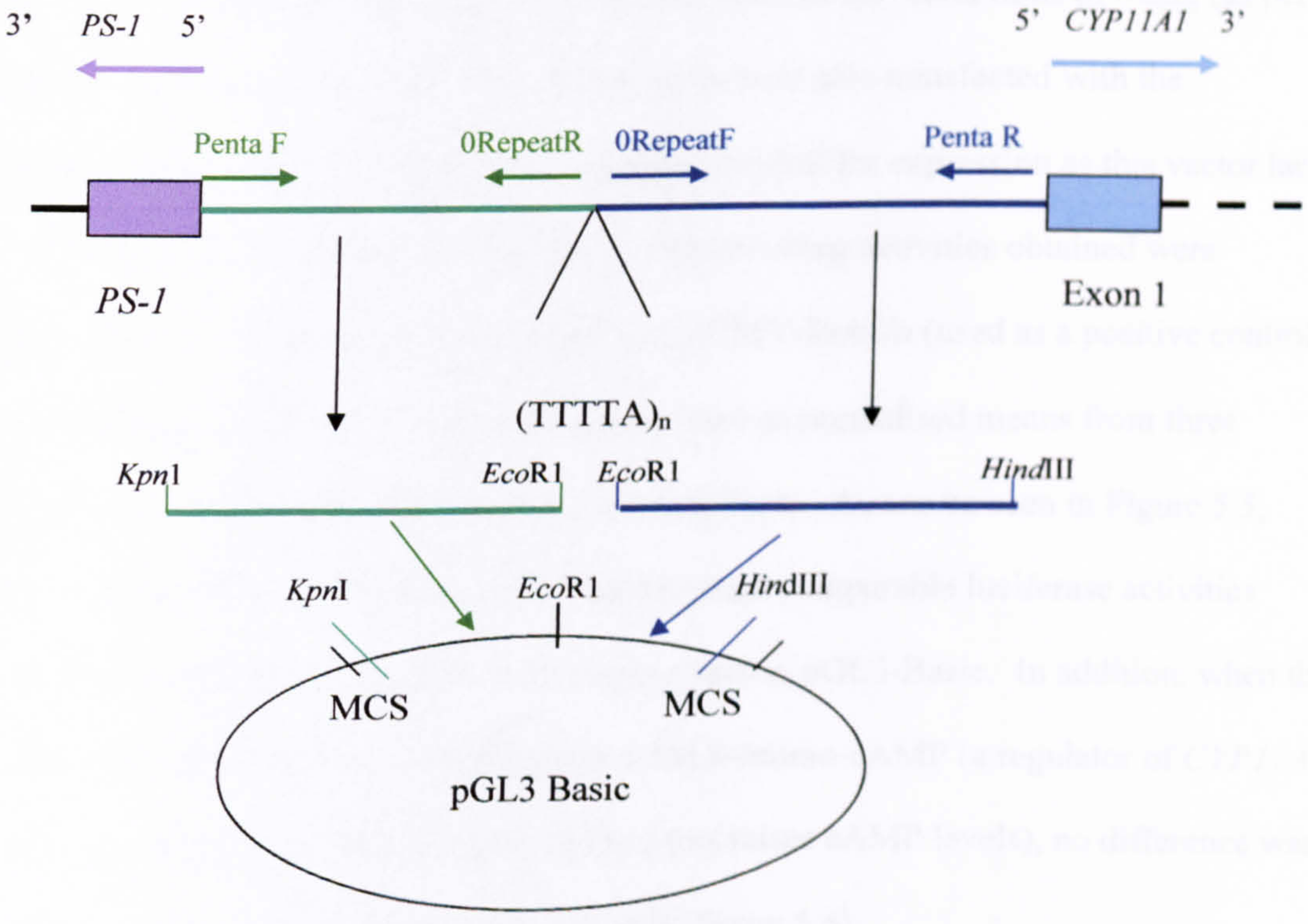


Figure 5.4. Schematic representation of the strategy for producing the zero repeat construct: Primers are as shown. Three ligation reactions take place into the expression vector, pGL3-Basic into the *Kpn*I, *Eco*R1 and *Hin*dIII sites. MCS- multiple cloning site. Pentanucleotide as shown $(TTTTA)_n$.

5.2.3 Transient Transfections of Pentanucleotide Repeats

5.2.3.1 Human Embryonic Kidney 293T cell line

To determine whether the number of pentanucleotide repeats has a direct effect on expression and to test methodology, constructs containing 4, 6 and 9 pentanucleotide repeats (as described in section 5.2.2) were transiently transfected into the human embryonic kidney cell line, 293T. This cell line does not normally possess the potential to produce steroids and is also easily transfectable. These cells were subcultured to approximately 70% confluency in 24-well plates. The constructs containing varying number of repeats were transfected at optimal transfection conditions of 0.2μg (as per manufacturers' instructions). In addition, cells were also transfected with the promoterless vector pGL3-Basic as a negative control for expression as this vector lacks a functional promoter (see section 5.2.1). The resulting activities obtained were normalized to luciferase units obtained from pCMV-*Renilla* (used as a positive control for transfection efficiency). Data are represented as normalised means from three independent experiments performed in quadruplicate. As can be seen in Figure 5.5, transient transfection of P4, P6 or P9 demonstrated comparable luciferase activities when compared with the promoterless empty vector, pGL3-Basic. In addition, when the transfected cells were treated with either 0.5M 8-bromo-cAMP (a regulator of *CYP11A1* gene expression) or 25μM forskolin (a drug that raises cAMP levels), no difference was detected when compared with untreated cells (figure 5.6).

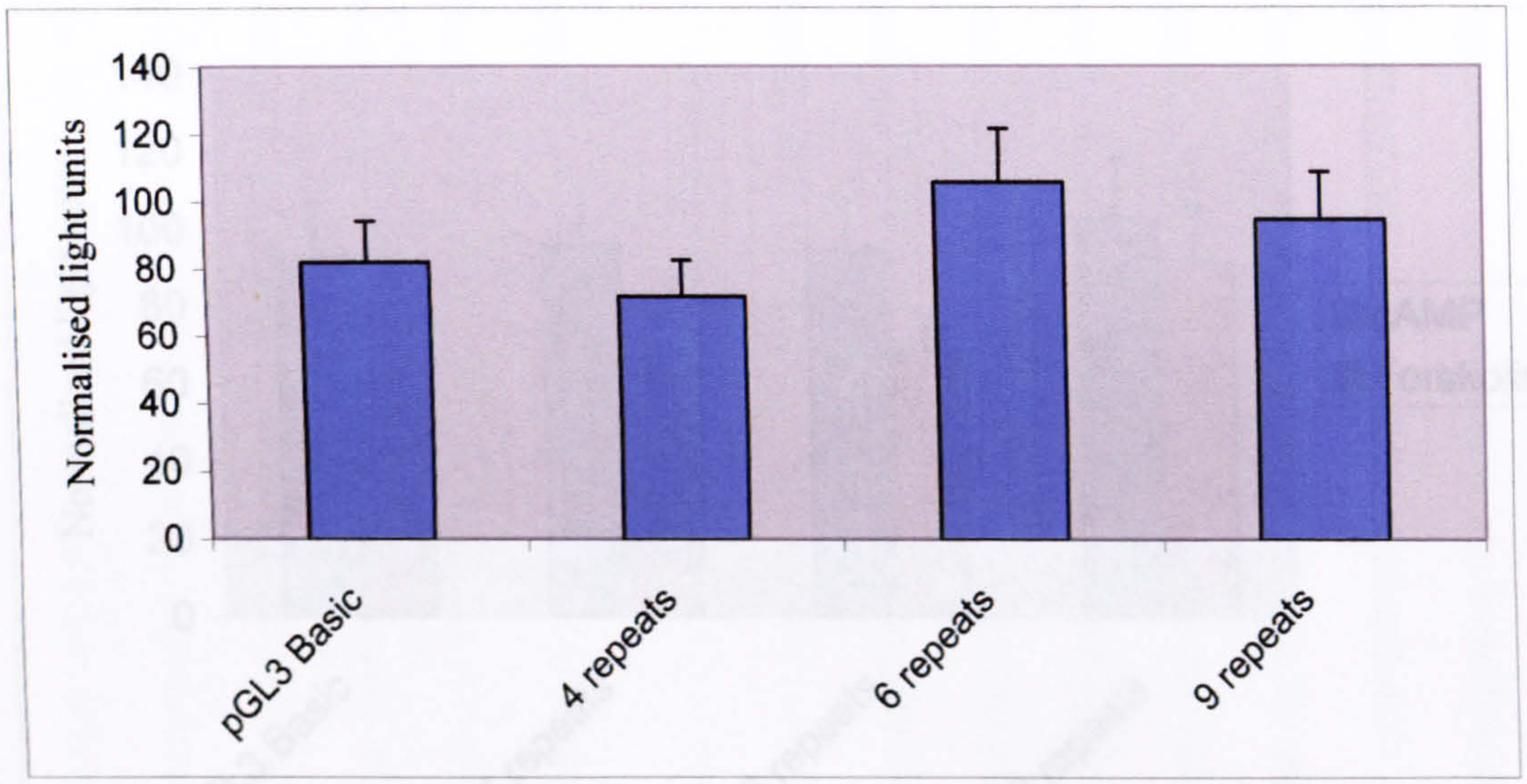


Figure 5.5: Transient transfection of pentanucleotide repeat constructs into human embryonic kidney 293T cells. Human embryonic kidney cells were transiently transfected with luciferase reporter constructs containing P4, P6 and P9 pentanucleotide repeat sequences, the promoterless pGL3-Basic vector and the control expression vector pCMV- *Renilla*. These values are presented as normalised light units from three independent transfections done in quadruplicate.

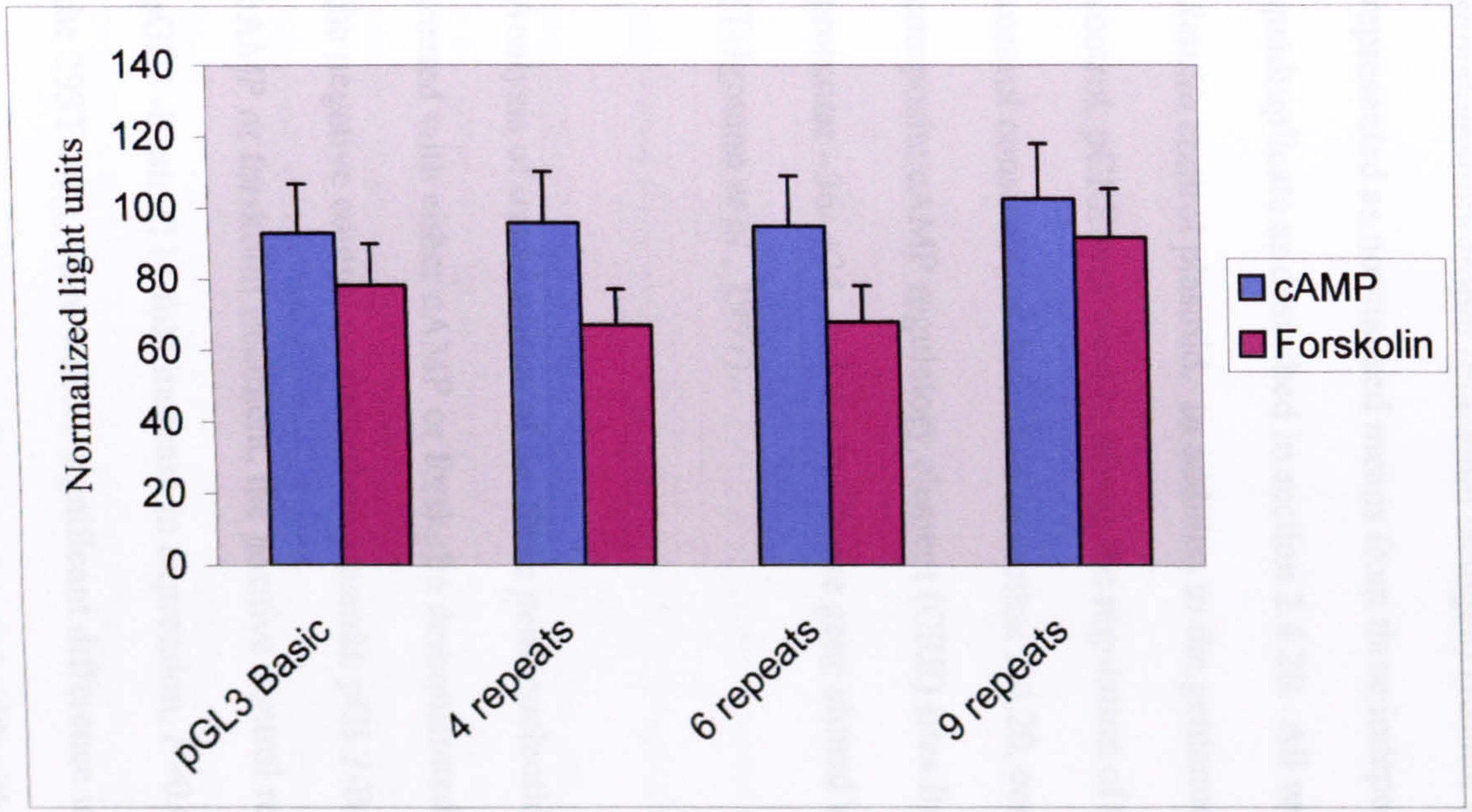


Figure 5.6:Transient transfection of pentanucleotide repeat constructs into 293T cells treated with cAMP and forskolin. Human embryonic kidney cells were transiently transfected luciferase reporter constructs containing P4, P6 and P9 pentanucleotide repeat sequences, the promoterless pGL3-Basic vector and the control expression vector pCMV- *Renilla* treated with 0.5 M 8-Br-cAMP and 25 μM forskolin. These values are presented as normalised light units from three independent transfections performed in quadruplicate.

5.2.3.2 Transfection of JEG-3 Cells

As *CYP11A1* is expressed in the placenta, the human choriocarcinoma cell line, JEG-3 was also used to ascertain possible functional differences between the P4, P6 or P9 repeats. This cell line has a functional steroid production pathway, and has been shown to produce progesterone as well as human chorionic gonadotropin and human somatomammotropin (placental lactogen) (Kohler and Bridson, 1971). Data are represented as normalised means from three independent experiments performed in quadruplicate as described in section 2.4.20. All wells were co-transfected with pCMV-*Renilla* control plasmid. In addition to the pentanucleotide constructs, a positive control, pCRE was used to assess the regulation of transcription in these cells. The control construct, as described in section 2.4.20, consisted of sequences containing two composite cAMP regulatory element (CRE) sites linked to the minimal rat (r) PRL promoter –36/ +34, and the luciferase gene cloned into the pGL3-Basic vector (Telgmann *et al.*, 1997).

Analysis of transfections of the three pentanucleotide constructs in JEG-3 cells not treated with either cAMP or forskolin demonstrated comparable luciferase values with the negative control promoterless plasmid, pGL3-Basic (figure 5.7). In the absence of cAMP or forskolin treatment, the positive control remains active over the promoterless pGL3 -Basic (114 fold increase in expression, P=0.002, figure 5.7). As seen above with the 293T cells, there was no significant difference in pentanucleotide repeat luciferase activity when the JEG-3 cells were treated with either 0.5 M 8-Br-cAMP or 25 μM forskolin and transiently transfected with any of the pentanucleotide repeat constructs. The positive control transfected cells treated with 8-Br-cAMP registered approximately 2-fold increase over untreated (figure 5.7, P= 0.01). This is similar to the results

obtained elsewhere (3-fold induction in ES cells stimulated with 0.5M 8-Br-cAMP) (Telgmann *et al.*, 1997). When compared with the promoterless pGL3 -Basic vector, the pCRE positive control remained active when treated with both cAMP and forskolin (figure 5.7; P=0.02 and P=0.001 respectively). Significance values were calculated using the Wilcoxon Mann-Whitney test in the statistical analysis program STATXACT (UK, see section 2.2.10).

As described above, pCMV-*Renilla* is a common control used in transfection experiments to ascertain the efficiency of the transfection of DNA into the cells. Interestingly however, when the raw luciferase light units for the pCMV-*Renilla* plasmid were studied, *Renilla* expression was found to be significantly higher when treated with either 8-Br-cAMP or forskolin as compared with the untreated values (exact P<0.0001 for both treatments, table 5.3). This was not seen in the transfections of the embryonic kidney 293T cells, possibly indicating a cell-specific response. This implies use of the pCMV-*Renilla* luciferase vector as a positive control may have introduced an experimental artefact. This is because as the values obtained from the transfection of the control plasmid are used to normalise the construct reporter activity to be tested, the use of this control plasmid could result in the misinterpretation of the data due to an under- or over-estimation of reporter activity. Studies examining the influences of the tetranucleotide repeat were not completed due to the above response of the pCMV-*Renilla* control vector. Furthermore, preparation of the zero repeat was not completed due to difficulties in the preparation of this construct.

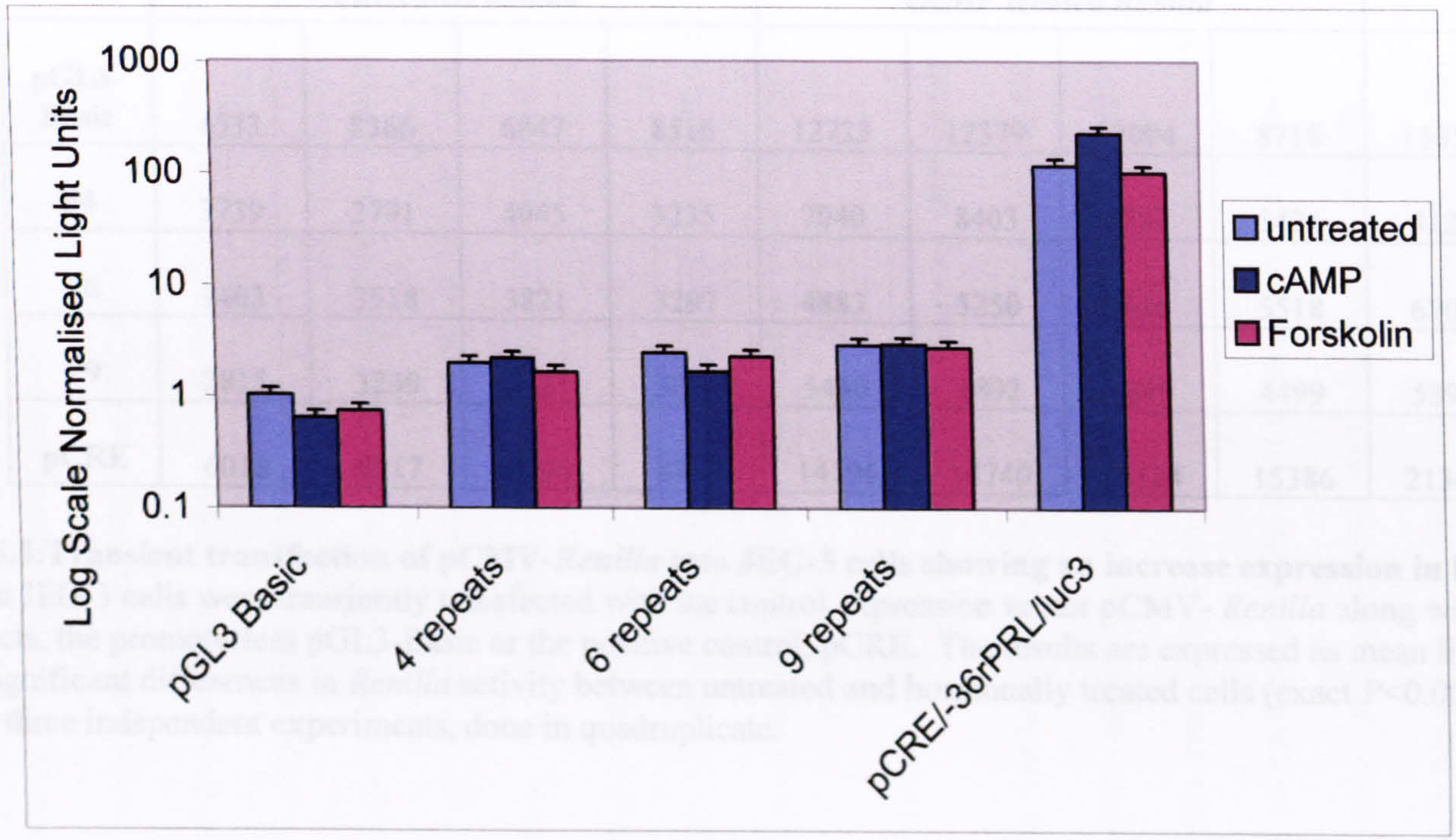


Figure 5.7:Transient transfection of pentanucleotide repeat constructs into JEG-3 cells. Human placenta JEG-3 cells were transiently transfected with luciferase reporter constructs containing P4, P6 and P9 pentanucleotide repeat sequences, the promoterless pGL3-Basic vector, the positive control vector, pCRE/-36rPRL/luc3, and the control expression vector pCMV- *Renilla* untreated or treated with 0.5 M 8-Br-cAMP and 25 μM forskolin. These values are presented as unlogged normalised light units from three independent transfections done in quadruplicate.

	Untreated *Renilla*				cAMP treated *Renilla*				Forskolin treated *Renilla*			
pGL3-Basic	4533	8366	6847	8516	12723	12379	12004	8718	15054	12545	9283	12077
P4	3739	2791	4045	3235	7040	8403	8282	5425	4123	4872	11945	23946
P6	3403	3518	3821	3297	4883	5250	6616	5518	6707	9121	11198	20212
P9	2815	3239	3227	3081	5460	4892	3848	4499	5398	6115	6731	20411
pCRE	6018	9017	8406	8865	14196	14740	16124	15386	21349	17428	8041	19554

Table 5.3:Transient transfection of pCMV-*Renilla* into JEG-3 cells showing an increase expression in luciferase light units. Human placenta JEG-3 cells were transiently transfected with the control expression vector pCMV- *Renilla* along with the pentanucleotide repeat constructs, the promoterless pGL3-Basic or the positive control, pCRE. The results are expressed as mean light units of *Renilla* activity and show significant differences in *Renilla* activity between untreated and hormonally treated cells (exact $P<0.0001$). The values are the mean of at least three independent experiments, done in quadruplicate.

5.3 Discussion

Transient transfection analysis allows the identification of potential regulatory elements controlling gene expression. The results presented here represent the first study of the function of pentanucleotide repeat length of *CYP11A1* with respect to the transcriptional regulation of this gene. Constructs containing approximately 2300 bp of promoter have been found to have increased activity over background (1.5-2 fold increase) (Moore *et al.*, 1992). The data presented here use constructs containing approximately 1700 bp of the promoter region with 4,6 or 9 pentanucleotide repeats. In the human embryonic kidney 293T control cell line, there was no increase in luciferase activity over basal level with any of the three pentanucleotide repeat constructs, either untreated or treated with 8-Br-cAMP or forskolin (figures 5.5 and 5.6). As the embryonic kidney cell line does not normally express *CYP11A1*, these results may suggest that this region does not contain regulatory sites for ubiquitous factors.

Transient transfection experiments were also completed in the choriocarcinoma cell line, JEG-3, which normally expresses *CYP11A1* (section 5.2.3.2). When the pentanucleotide repeat constructs were transiently transfected into the untreated placental cell line, JEG-3, the results did not mirror what has been seen in the literature. Using a *CYP11A1* deletion construct containing 2327 bp of *CYP11A1* promoter sequence relative to the transcriptional start site, Moore and colleagues reported an increase in basal expression 1.5-2 times over background (Moore *et al.*, 1992). When treating the transfected cells with 8-Br-cAMP or forskolin, an increase greater than five-fold in luciferase activity of the 2327 bp construct over untreated cells was found (Moore *et al.*, 1992). Sequence analysis revealed that constructs used in these studies contained the P4 variant. However, these results are not consistent with those presented

in this thesis. As shown in figure 5.7, there was no increase in luciferase activity with P4 or any of the constructs when treated with 8-Br-cAMP or forskolin compared to the untreated values. There are a number of possibilities for this discrepancy between these results and those reported in the literature. Firstly, the constructs reported in the literature contain over 2300 bp of 5' *CYP11A1* promoter sequence, whereas those used in this thesis contain only 1700 bp of 5' promoter sequence (Moore *et al.*, 1992). While there is a known adrenal specific CRE (cAMP responsive element) consensus sequence located between –1697 and –1558 which confers cAMP responsiveness, the same sequence region is not functional in placental cells (Moore *et al.*, 1992). It may be possible that another regulatory CRE element is present between sequences –1700 and –2300bp, which allows for induction in placental cells; however, none have been detected by sequence analysis using the program Gentomatrix.

Secondly, it is possible that during the cloning process, base changes could have occurred, rendering responsive sites inactive. Sequence analysis did reveal several base changes from each construct sequence when compared to the sequence present on the NCBI database (Genbank accession number AC0908426) but these changes were outside any known regulatory regions.

Thirdly, a powerful element located between –155 and –131, which increased transcription *in vitro* by approximately 500% as shown by Huang and Millar, has been described as containing sequences for transcription factors belonging to the LBP family of transcription factors (Huang and Miller, 2000). It is thought that these transcription factors function to modulate the degree of expression as opposed to controlling basal expression. It is possible that the JEG-3 cells used in these experiments no longer

contained these transcription factors. RT-PCR and Northern Blot analysis would need to be completed to answer these questions.

Finally, constructs described in the literature of the *CYP11A1* promoter region more commonly contain sequences up to +49 from the transcriptional start site (Moore *et al.*, 1992); (Ben Zimra *et al.*, 2002). However, for the constructs used in this study, sequences were used up to +21 from the transcriptional start site. This extra sequence could be necessary for complete transcription to occur. Ascertainment of proper transcription initiation with these constructs in JEG-3 cells (and human embryonic kidney 293T cells) will need to be completed, by RNase protection assays.

However, all of these results should be interpreted with caution due to the effect seen on the luciferase activity of the control reporter plasmid, pCMV-*Renilla*, when treated with either 8-Br-cAMP or forskolin (table 5.3). As the values obtained from the transfection of the control plasmid are used to normalise the construct reporter activity to be tested, the use of this control plasmid could result in the misinterpretation of the data due to an under- or over-estimation of reporter activity. Similar results have been reported in the literature. For example, in African Green Monkey kidney (CV-1) cells, the activity of the internal control used to correct for transfection efficiency (pRL-TK luciferase vector, containing the *P.reniformis* luciferase gene), was seen to increase in response to dihydrotestosterone (DHT) and decrease in response to dexamethasone (DEX) (Ibrahim *et al.*, 2000). Furthermore, when the control vector was changed to pRL-TS (a thromboxane synthase promoter *Renilla* luciferase plasmid), activity remained constant when CV-1 cells were treated with DHT but decreased when treated with DEX (Ibrahim *et al.*, 2000). In another study, co-transfection of the luciferase reporter plasmid, pRSV-luc (containing the *P.pyralis* luciferase gene), was shown to be repressed when co-

transfected with transcription factors PPARalpha and ARP-1, but was activated by COUP-TF1 (Everett and Crabb, 1999). These results suggest that some pCMV-*Renilla* vectors used in transient transfection experiments as internal controls may contain sequences that respond specifically to hormonal stimulation, or to co-transfection of other factors.

Cell type may also contribute to the responsiveness of these controls. For example, retinoic acid X receptors, when co-transfected into CV-1 cells with pRL-TK did not have an effect on luciferase values; however, an increase in luciferase activity was seen in H4IIEC3 hepatoma cells (Everett and Crabb, 1999). As the same activation of pCMV-*Renilla* was not seen in the embryonic kidney 293T cells, it is possible that the JEG-3 cells may contain factors that specifically bind to sequences in the *Renilla* luciferase gene. Therefore, vectors will need to be tested for responses to exogenous stimuli to avoid possibly introducing experimental artefacts.

Deriving the zero repeat constructs has been more troublesome. Constructs were designed to contain the identical section of sequence as the pentanucleotide constructs excluding the pentanucleotide repeat. PCR primers were designed flanking both repeats and the complementary primer for each set located at each end of the construct. The resultant PCR products were ligated into the promoterless pGL3 Basic and sequenced. However, subsequent sequencing revealed inappropriate ligations had occurred. Given the inconclusive results for the P4, P6 and P9 experiments, the cloning of the zero repeat was not pursued. Moreover, the tetranucleotide repeats were also not used in transient transfection assays due to the results seen for the pentanucleotide repeat. Taken together, the above data represent a preliminary study into the putative functional role of the pentanucleotide repeat in the regulation of *CYP11A1* transcription.

6.0 FUTURE WORK

In this thesis, several approaches have been described which were used to investigate the role of *CYP11A1* in PCOS, a common complex trait disorder. Both a case control and family based association study was conducted in two distinct populations to examine the relationship with a previously implicated pentanucleotide repeat in the promoter region of *CYP11A1* and a neighbouring microsatellite and PCO status and/or serum testosterone levels. Furthermore, this pentanucleotide repeat was also examined using transient transfection analysis to study the role of this microsatellite in the regulation of *CYP11A1*. Substantial expression analysis and *in silico* examination of the *CYP11A1* locus was completed.

The results of these studies revealed several key findings:

- One of the largest association studies to date of the pentanucleotide polymorphic repeat located in the promoter region as a modifier of the phenotypic expression of PCOS demonstrated that there was no consistent association (Gaasenbeek *et al.*, 2004).

- Preliminary functional analysis of the pentanucleotide repeat by transient transfection did not reveal any differences in expression with respect to repeat length.

- Examination of the *CYP11A1* locus identified two new alternate transcripts for *CYP11A1* and a new putative gene of unknown function (*PS-1*).

6.1 Association Studies

One of the aims of this work was to seek replication of the association results of the pentanucleotide repeat and serum testosterone levels in PCOS women (Gharani *et al.*, 1997) using both case control and family based sample sets. The results of the study described in this thesis revealed that the effect of this locus on PCOS susceptibility reported in the literature has been overestimated. This pointedly does not imply that *CYP11A1* has no role in the aetiology of PCOS, rather that the role must be much smaller than anticipated based on previous estimates (Gharani *et al.*, 1997); (Diamanti-Kandarakis *et al.*, 2000); (Daneshmand *et al.*, 2002). As discussed in chapter 4 (section 4.6), mutation screening of *CYP11A1* exons 1 to exon 6 did not reveal the existence of any other putative SNPs (Gharani, 2000). Finishing the mutation screening of exons 7-9 should be carried out in the PCOS data sets used here for completeness. Forty-four SNPs have now been identified, including 37 intronic SNPs in *CYP11A1* (with 22 of the 44 SNPs validated in the NCBI SNP database, see section 2.4.22). A subset of these SNPs could be typed in the populations described in this thesis including those that have good associated functional data. All SNPs would not have to be typed due to the likelihood of strong LD between them.

Along with the considerable case control and family based data sets collected by members of the laboratory, data from a set of affected sibling-pairs (sibpairs; i.e. a PCOS proband plus at least one affected sister) were also being compiled, with the addition of parents where possible. These data may be used as a pedigree data set for linkage analysis. Furthermore, once the sibpair data set is large enough, this resource may be used to carry out a genome wide scan to identify the genetic component(s) of PCOS. To date, approximately 110 sibpairs (some of which have more than one

affected sibling) have been collected. Estimates on the size of sibpair data sets required for the study of PCOS has revealed great variation depending on the underlying model tested (McCarthy *et al.*, 1998). However, it is thought that roughly 400 sibpairs should provide 80% power to detect loci with λs (λ- sibling relative risk) of 1.5 at a LOD score of 2.6 (and 95% power at a LOD score of 1.5).

The data sets used in this study (and including the sibpair dataset once collected) could be utilised to investigate other putative candidate genes, which are relevant to androgen synthesis, follicular development and function and insulin action and secretion. Some of these genes, for example follistatin, are discussed in detail in section 1.4.

6.2 Functional Role of Pentanucleotide and Tetranucleotide Repeats

There has been no reported study examining the role of the length of the pentanucleotide repeat (as discussed in section 4.2.2) in the regulation of *CYP11A1*. It has been previously hypothesised that this pentanucleotide repeat may function to regulate the transcription of *CYP11A1* (Gharani *et al.*, 1997). The results of a preliminary attempt at deciphering this role are presented in this thesis.

Sequence analysis of transient transfection studies on *CYP11A1* transcription regulation revealed that 5 out of the 6 studies used constructs containing the 4-repeat allele (Chung *et al.*, 1989); (Moore *et al.*, 1990); (Moore *et al.*, 1992);, (Monte *et al.*, 1998); (Rodriguez *et al.*, 1997). Only one study used a clone containing the 6-repeat allele (Takayama *et al.*, 1994).

It was initially hypothesised that, with the association of those lacking the 4 repeat allele and higher serum testosterone levels in women with PCOS, higher number repeat alleles may function to upregulate the *CYP11A1* gene, thereby increasing androgen production (Gharani *et al.*, 1997). However, results presented in this thesis suggest that the pentanucleotide repeat, at least in the placental JEG-3 cells, might not have a role in the transcriptional regulation of the *CYP11A1* gene. It is possible that the placental cell line JEG-3 no longer possess the appropriate factors for transcription of *CYP11A1*. It would be necessary to transfect the different pentanucleotide repeat constructs into another cell line known to express *CYP11A1* (such as an adrenal or ovarian cell line). There are two known ovarian cell lines (HOTT and HGL5), both of which express *CYP11A1* and have been shown to produce androstenedione and progesterone (Rainey *et al.*, 1994); (Rainey *et al.*, 1996). Transient transfection of these constructs into the various ovarian cell lines available would also provide much needed information on the regulation of *CYP11A1* in the human ovary. This in turn, may provide clues to the dysregulation of steroidogenesis that occurs in the polycystic ovary.

It is not known whether the pentanucleotide repeat functions to bind trans-activating factors. A similar pentanucleotide repeat, found in an *alu* repeat sequence located close to the 5' boundary of the human SHBG promoter, has been shown to have a silencing effect on SHBG promoter activity when constructs contained 6 repeat units (Hogeveen *et al.*, 2001). Furthermore, this silencing effect was dependent upon binding of a 46-kDa nuclear protein and interaction with downstream promoter elements (Hogeveen *et al.*, 2001). It is possible that the *CYP11A1* pentanucleotide repeat requires other trans-activating factors binding to this element to function correctly. DNase I footprinting, which identifies DNA elements able to bind nuclear proteins, and electrophoretic mobility shift assays (EMSA), for the identification of DNA-protein complexes could

be completed to answer this question. DNase I footprinting experiments have been performed in placenta and adrenal tissues, and in other species such as bovine and mouse (Chou *et al.*, 1996); (Guo *et al.*, 1994); (Pestell *et al.*, 1993); (Rice *et al.*, 1990). However, such experiments have not been performed in the human ovary. It is possible that ovarian specific factors, when bound to the pentanucleotide repeat, function to control *CYP11A1* expression.

It is also possible that the tetranucleotide repeat, located 840 bp downstream of exon 1 may regulate *CYP11A1* expression. A similar repeat, located in the first intron of the tyrosine hydroxylase (*TH*) gene, has been shown to act as a transcription regulatory element *in vitro* (Meloni *et al.*, 1998). Transient transfection experiments of the tetranucleotide repeats into human ovarian cells would need to be completed to ascertain function in these cells.

Results on the functional role of the pentanucleotide repeat, as shown in table 5.3, have to be viewed with caution. This is due to the increase in luciferase activity seen in the JEG-3 cells transfected with the control plasmid, pCMV-*Renilla,* when treated with cAMP or the drug, forskolin. This may alter the transfection results obtained in this study by providing an under-estimation of luciferase activity of the transfected constructs. As discussed in chapter 5 (section 5.3), there is precedence for this in the literature with several studies reporting similar findings (Everett and Crabb, 1999); (Ibrahim *et al.*, 2000). It is possible that another positive control, not containing the *Renilla* luciferase gene (such as pCMV-LacZ) will not respond to treatment as seen here. Such experiments will need to be completed.

6.3 Novel *CYP11A1* Transcripts and *PS-1*

Experiments conducted for the purposes of this thesis reveal the first evidence of the existence of alternate transcripts for *CYP11A1* and of a gene of unknown function, *PS-1*. The results presented here show that the *CYP11A1* locus contains several extra levels of complexity than originally reported (Chung *et al.*, 1986;) (Morohashi *et al.*, 1987).

Data described in this thesis imply that the 5' end of transcript II and transcript III is missing. The longest open reading frame (figure 3.7(b)) for transcript II begins in exon 3 and for transcript III begins in exon 2 (figure 3.7 (c)). As the 5' end of these transcripts is unknown, it remains a possibility that a functional methionine lies upstream of the known sequence. Western Blot analysis could be completed to determine if transcript II and transcript III produce detectable proteins. There is no commercially available antibody for *CYP11A1*. However, an antibody made by and generously given to our lab by Professor Ian Mason (Edinburgh, UK) has been made against the *CYP11A1* protein purified from adrenal mitochondria. The epitope for this antibody is unknown and could be anywhere along the full-length protein sequence. However, it is possible that the antibody might recognise Western Blot analysis *CYP11A1* alternate transcript proteins and may identify a difference in size. A limitation to such an experiment is that the antibody may not recognise the different proteins.

It is difficult to ascertain a possible function for *PS-1* as the full-length transcript has yet to be determined. *In silico* translation of the total sequence obtained thus far does not reveal any open reading frames. The sequence obtained to date by 5' and 3' RACE and

in silico analyses is 3000 bp of contiguous DNA. The results of 3' RACE suggest that the 3' end of the gene has been identified and the result of the Northern blot analysis, which detected a 6.5 and 7.5kb transcript, would therefore suggest that it is the 5' end of the gene that is missing. This needs to be isolated before the predicted protein encoded by this gene can be identified. The sequence obtained to date may represent 3' UTR.

The contiguous sequence obtained through 5' and 3' RACE and in silico analyses indicates that the sequence represents the 3' UTR of this gene. This gene is expressed in both the placenta and the ovary, which makes *PS*-1 an interesting gene to study. Once the 5' end of the gene is established, the experiments described above for the alternate transcripts of *CYP11A1* could also be performed. However, antibodies would have to be raised to this protein.

6.4 Conclusions

In conclusion, *CYP11A1*, encoding the enzyme P450scc, the first enzymatic step in steroid biosynthesis, was examined to elucidate the role of *CYP11A1* in the actiology of PCOS. Results presented here show that the role of *CYP11A1* may be modest at most. However, a more complex genomic arrangement on chromosome 15q24 has emerged. It is possible that the alternate transcripts found here or the novel gene (*PS-1*) may play an important role in regulation of steroidogenesis. Ultimately, once the elucidation of the dysregulation of steroid production in the polycystic ovary has been determined, this may lead to better therapy for the symptomatology of PCOS.

7.0 REFERENCES

Abecasis,G.R. and Cookson,W.O.C. (2000). GOLD-Graphical Overview of Linkage Disequilibrium. Bioinfomatics *16*, 182-183.

Adams,J., Franks,S., Polson,D.W., Mason,H.D., Abdulwahid,N., Tucker,M., Morris,D.V., Price,J., and Jacobs,H.S. (1985). Multifollicular ovaries: clinical and endocrine features and response to pulsatile gonadotropin releasing hormone. Lancet *2*, 1375-1379.

Adams,J., Polson,D.W., and Franks,S. (1986). Prevalence of polycystic ovaries in women with anovulation and idiopathic hirsutism. Br. Med. J. (Clin. Res. Ed) *293*, 355-359.

Baird,D.T., Baker,T.G., McNatty,K.P., and Neal,P. (1975). Relationship between the secretion of the corpus luteum and the length of the follicular phase of the ovarian cycle. J. Reprod. Fertil. *45*, 611-619.

Baird,D.T., Corker,C.S., Davidson,D.W., Hunter,W.M., Michie,E.A., and Van Look,P. (1977). Pituitary-Ovarian Relationships in Polycystic Ovary Syndrome. J. Clin. Endocrinol. Metab *45*, 798-809.

Barbieri,R.L., Makris,A., Randall,R.L., Daniels,G., Kistnew,R.W., and Ryan,K.J. (1986). Insulin stimulates androgen accumulation in incubations of ovarian stroma obtained from women with hyperandrogenism. J. Clin. Endocrinol. Metab *62*, 904-910.

Barth,J.H. (1988). Alopecia and hirsuties. Current concepts in pathogenesis and management. Drugs *35*, 83-91.

Bell,G.I., Selby,M.J., and Rutter,W.J. (1982). The highly polymorphic region near the human insulin gene is composed of simple tandemly repeating sequences. Nature *295*, 31-35.

Ben Zimra,M., Koler,M., and Orly,J. (2002). Transcription of cholesterol side-chain cleavage cytochrome P450 in the placenta: activating protein-2 assumes the role of steroidogenic factor-1 by binding to an overlapping promoter element. Mol. Endocrinol. *16*, 1864-1880.

Bennett,S.T. and Todd,J.A. (1996). Human type 1 diabetes and the insulin gene: principles of mapping polygenes. Annu. Rev. Genet. *30*, 343-370.

Benyo,D.F., Ravindranath,N., Bassett,S., Hutchison,J., and Zeleznik,A.J. (1993). Cellular aspects of corpus luteum function in the primate. Hum. Reprod. *8 Suppl 2*, 102-106.

Bland,J.M. and Altman,D.G. (1995). Multiple significance tests: the Bonferroni method. Br. Med. J. *310*, 170.

Broustas,C.G., Gokhale,P.C., Rahman,C., Dritschilo,A., Ahmad,I., and Kasid,U. (2004). BRCC2, a novel BH3-like domain-containing protein, induces apoptosis in a caspase-dependent manner. J. Biol. Chem. *279*, 26780-26788.

Brown,J.B. (1978). Pituitary control of ovarian function- concepts derived from gonadotropin therapy. Aust. N. Z. J. Obstet. Gynaecol. *18*, 46-54.

Bulun,S.E., Takayama,K., Suzuki,T., Sasano,H., Yilmaz,B., and Sebastians,S. (2004). Organisation of the human aromotase p450 (*CYP19*) gene. Semin Reprod Med *22*, 5-9.

Burks,D.J. and White,M.F. (2001). IRS proteins and beta-cell function. Diabetes *50 Suppl 1*, S140-S145.

Cardon,L.R. and Bell,J.I. (2001). Association study designs for complex diseases. Nat. Rev. Genet. *2*, 91-99.

Carey,A.H., Chan,K.L., Short,F., White,D., Williamson,R., and Franks,S. (1993). Evidence for a single gene effect causing polycystic ovaries and male pattern baldness. Clin. Endocrinol. (Oxf) *38*, 653-658.

Carey,A.H., Waterworth,D., Patel,K., White,D., Little,J., Novelli,P., Franks,S., and Williamson,R. (1994). Polycystic ovaries and premature male pattern baldness are associated with one allele of the steroid metabolism gene CYP17. Hum. Mol. Genet. *3*, 1873-1876.

Chou,S.J., Lai,R.N., and Chung,B.C. (1996). Characterisation of the upstream sequence of the human *CYP11A1* gene for cell type specific-expression. J. Biol. Chem. *271*, 22125-22129.

Christian,R.C., Dumesic,D.A., Behrenbeck,T., Oberg,A.L., Sheedy,P.F., and Fitzpatrick,L.A. (2003). Prevalence and predictors of coronary artery calcification in women with polycystic ovary syndrome. J. Clin. Endocrinol. Metab *88*, 2562-2568.

Chung,B.C., Matteson,K.J., Voutilainen,R., Mohandas,T.K., and Miller,W.L. (1986). Human cholesterol side-chain cleavage enzyme, P450scc: cDNA cloning, assignment of the gene to chromosome 15, and expression in the placenta. Proc. Natl. Acad. Sci. U. S. A *83*, 8962-8966.

Chung,B.C., Hu,M.C., Lai,C.C., and Lin,C.H. (1989). The 5'-region of the P450XIA1 (P450scc) gene contains a basal promoter and an adrenal-specific activating domain. Biochem. Biophys. Res. Commun. *160*, 276-281.

Ciaraldi,T.P., Morales,A.J., Hickman,M.G., Odom-Ford,R., Yen,S.S., and Olefsky,J.M. (1998). Lack of insulin resistance in fibroblasts from subjects with polycystic ovary syndrome. Metabolism *47*, 940-946.

Clayton,R. (1999). A generalization fo the transmission/disequalibrium test for uncertain haplotype transmission. Am. J. Hum. Genet. *65*, 1170-1177.

Clemens,J.W., Lala,D.S., Parker,K.L., and Richards,J.S. (1994). Steroidogenic factor-1 binding and transcriptional activity of the cholesterol side-chain cleavage promoter in rat granulosa cells. Endocrinology *134*, 1499-1508.

Colhoun,H.M., McKeigue,P.M., and Davey,S.G. (2003). Problems of reporting genetic associations with complex outcomes. Lancet *361*, 865-872.

Conn,P.M. and Crowley,W.F., Jr. (1991). Gonadotropin-releasing hormone and its analogues. N. Engl. J. Med. *324*, 93-103.

Conway,G.S., Honour,J.W., and Jacobs,H.S. (1989). Heterogeneity of the polycystic ovary syndrome: clinical, endocrine and ultrasound features in 556 patients. Clin. Endocrinol. (Oxf) *30*, 459-470.

Cusi,K., Maezono,K., Osman,A., Pendergrass,M., Patti,M.E., Pratipanawatr,T., DeFronzo,R.A., Kahn,C.R., and Mandarino,L.J. (2000). Insulin resistance differentially affects the PI 3-kinase- and MAP kinase-mediated signaling in human muscle. J. Clin. Invest *105*, 311-320.

Daneshmand,S., Weitsman,S.R., Navab,A., Jakimiuk,A.J., and Magoffin,D.A. (2002). Overexpression of theca-cell messenger RNA in polycystic ovary syndrome does not correlate with polymorphisms in the cholesterol side-chain cleavage and 17alpha-hydroxylase/C(17-20) lyase promoters. Fertil. Steril. *77*, 274-280.

Diamanti-Kandarakis,E., Kouli,C.R., Tsianateli,T.C., and Bergiele,A.T. (1998). Theraputic effects of metformin on insulin resistance and hyperandrogenism in polycystic ovary syndrome. Eur J Endocrinol *138*, 269-274.

Diamanti-Kandarakis,E., Bartzis,M.I., Bergiele,A.T., Tsianateli,T.C., and Kouli,C.R. (2000). Microsatellite polymorphism (tttta)(n) at -528 base pairs of gene CYP11alpha influences hyperandrogenemia in patients with polycystic ovary syndrome. Fertil. Steril. *73*, 735-741.

Dong, J., Albertini, D.F., Nishimori, K., Kumar, T.R., Lu, N., and Matzuk, M.M. (1996). Growth differentiation factor-9 is required during early ovarian folliculogenesis. Nature *383*, 531-535.

Dunaif,A., Segal,K.R., Futterweit,W., and Dobrjansky,A. (1989). Profound peripheral insulin resistance, independent of obesity, in polycystic ovary syndrome. Diabetes *38*, 1165-1174.

Dunaif,A., Scott,D., Finegood,D., Quitana,B., and Whitcomb,R. (1996). The insulin-sensitizing agent troglitazone improves metabolic and reproductive abnormalities in the polycystic ovary syndrome. J. Clin. Endocrinol. Metab *81*, 3299-3306.

Dunaif,A. (1997). Insulin resistance and the polycystic ovary syndrome: mechanism and implications for pathogenesis. Endocr. Rev. *18*, 774-800.

Durlinger, A.L., Kramer, P., Karels, B., de Jong, F.H., Uilenbroek, J.T., Grootegeod, J.A., and Themmen, A.P. (1999). Control of primordial follicle recruitment by anti-Mullerian hormone in the mouse ovary. Endocrinology. *140*, 4891-4899.

Ehrmann,D.A., Cavaghan,M.K., Imperial,J., Sturis,J., Rosenfield,R.L., and Polonsky,K.S. (1997a). Effects of metformin on insulin secretion, insulin action, and ovarian steroidogenesis in women with polycystic ovary syndrome. J. Clin. Endocrinol. Metab *82*, 524-530.

Ehrmann,D.A., Schneider,D.J., Sobel,B.E., Cavaghan,M.K., Imperial,J., Rosenfield,R.L., and Polonsky,K.S. (1997b). Troglitazone improves defects in insulin action, insulin secretion, ovarian steroidogenesis, and fibrinolysis in women with polycystic ovary syndrome. J. Clin. Endocrinol. Metab *82*, 2108-2116.

Ehrmann,D.A., Schwarz,P.E.H., Hara,M., Tang,X., Horikawa,Y., Imperial,J., Bell,G.I., and Cox,N.J. (2002). Relationship of Calpain-10 Genotype to Phenotypic Features of Polycystic Ovary Syndrome. J. Clin. Endocrinol. Metab *87*, 1669-1673.

Ek,I., Arner,P., Bergqvist,A., Carlstrom,K., and Wahrenberg,H. (1997). Impaired adipocyte lipolysis in nonobese women with the polycystic ovary syndrome: a possible link to insulin resistance? J. Clin. Endocrinol. Metab *82*, 1147-1153.

Ek,I., Arner,P., Ryden,M., Holm,C., Thorne,A., Hoffstedt,J., and Wahrenberg,H. (2002). A unique defect in the regulation of visceral fat cell lipolysis in the polycystic ovary syndrome as an early link to insulin resistance. Diabetes *51*, 484-492.

Everett,L.M. and Crabb,D.W. (1999). Sensitivity of virally-driven luciferase reporter plasmids to members of the steroid/thyroid/retinoid family of nuclear receptors. J. Steroid Biochem. Mol. Biol. *70*, 197-201.

Fauser,B.C., Pache,T.D., Hop,W.C., de Jong,F.H., and Dahl,K.D. (1992). The significance of a single serum LH measurement in women with cyclic disturbances: discrepencies with between immunoreactive and bioactive hormone estimates. Clin. Endocrinol. (Oxf) *11*, 445-452.

Feinberg,A.P. and Vogelstein,B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. *132*, 6-13.

Ferrannini,E., Natali,A., Capaldo,B., Lehtovirta,M., Jacob,S., and Yki-Jarvinen,H. (1997). Insulin resistance, hyperinsulinemia, and blood pressure: role of age and obesity. European Group for the Study of Insulin Resistance (EGIR). Hypertension *30*, 1144-1149.

Filicori,M., Cognigni,G.E., Samara,A., Melappioni,S., Perri,T., Cantelli,B., Parmegiani,L., Pelusi,G., and DeAloysio,D. (2002). The use of LH activity to drive folliculogenesis: exploring uncharted territories in ovulation induction. Hum. Reprod. Update. *8*, 543-557.

Findlay,J.K. (1986). Angiogenesis in reproductive tissues. J. Endocrinol. *111*, 357-366.

Franks,S., Adams,J., Mason,H., and Polson,D. (1985). Ovulatory disorders in women with polycystic ovary syndrome. Clin. Obstet. Gynaecol. *12*, 605-632.

Franks,S., Kiddy,D.S., Hamilton-Fairley,D., Bush,A., Sharp,P.S., and Reed,M.J. (1991). The role of nutrition and insulin in the regulation of sex hormone binding globulin. J. Steroid Biochem. Mol. Biol. *39*, 835-838.

Franks,S. (1995). Polycystic ovary syndrome. N. Engl. J. Med. *333*, 853-861.

Gaasenbeek,M., Powell,B.L., Sovio,U., Haddad,L., Gharani,N., Bennett,A., Groves,C.J., Rush,K., Goh,M.J., Conway,G., Ruokonen,A., Martikainen,H., Pouta,A., Taponen,S., Hartikainen,A.L., Halford,S., Jarvelin,M.R., Franks,S., and McCarthy,M.I. (2004). Large-Scale Analysis of the Relationship between *CYP11a* Promoter Variation, Polycystic Ovarian Syndrome, and Serum Testosterone. J. Clin. Endocrinol. Metab *89*, 2408-2413.

Gauderman, W.J. Sample size requirements for association studies of gene-gene interaction. Am. J. Epidemiol. *155*, 478-484.

Gharani,N., Waterworth,D.M., Williamson,R., and Franks,S. (1996). 5' polymorphism of the CYP17 gene is not associated with serum testosterone levels in women with polycystic ovaries. J. Clin. Endocrinol. Metab *81*, 4174.

Gharani,N., Waterworth,D.M., Batty,S., White,D., Gilling-Smith,C., Conway,G.S., McCarthy,M., Franks,S., and Williamson,R. (1997). Association of the steroid synthesis gene CYP11a with polycystic ovary syndrome and hyperandrogenism. Hum. Mol. Genet. *6*, 397-402.

Gharani,N. (2000). The molecular genetics of polycystic ovary syndrome. PhD Thesis.

Gilling-Smith,C., Willis,D.S., Beard,R.W., and Franks,S. (1994). Hypersecretion of androstenedione by isolated thecal cells from polycystic ovaries. J. Clin. Endocrinol. Metab *79*, 1158-1165.

Gilling-Smith,C., Story,H., Rogers,V., and Franks,S. (1997). Evidence for a primary abnormality of thecal cell steroidogenesis in the polycystic ovary syndrome. Clin. Endocrinol. (Oxf) *47*, 93-99.

Gillio-Meina,C., Hui,Y.Y., and LaVoie,H.A. (2003). GATA-4 and GATA-6 Transcription Factors: Expression, Immunohistochemical Localization, and Possible Function in the Porcine Ovary. Biol. Reprod. *68*, 412-422.

Givens,J.R. (1988). Familial polycystic ovarian disease. Endocrinol. Metab Clin. North Am. *17*, 771-783.

Gonzalez,A., Abril,E., Roca,A., Aragon,M., Figueroa,M., Velarde,P., Ruiz,R., Fayez,O., Galan,J., Herreros,J., Real,L., and Ruiz,A. (2003). Specific *CAPN10* Gene Haplotypes Influence the Clinical Profile of Polycystic Ovary Patients. J. Clin. Endocrinol. Metab *88*, 5529-5536.

Gougeon,A. (1996). Regulation of ovarian follicular development in primates: facts and hypotheses. Endocr. Rev. *17*, 121-155.

Guo,I.C., Tsai,H.M., and Chung,B.C. (1994). Actions of two different cAMP-responsive sequences and an enhancer of the human *CYP11A1* (P450scc) gene in adrenal Y-1 and placental JEG-3 cells. J. Biol. Chem. *269*, 6362-6369.

Guo Q., Kumar T.R, Woodruff T, Hadsell L.A, DeMayo F.J, and Matzuk M.M (1998). Overexpression of mouse follistatin causes reproductive defects in transgenic mice. Mol Endocrinol *12*, 96-106.

Haddad,L., Evans,J.C., Gharani,N., Robertson,C., Rush,K., Wiltshire,S., Frayling,T.M., Wilkin,T.J., Demaine,A., Millward,A., Hattersley,A.T., Conway,G., Cox,N.J., Bell,G.I., Franks,S., and McCarthy,M.I. (2002). Variation within the type 2 diabetes susceptibility gene calpain-10 and polycystic ovary syndrome. J. Clin. Endocrinol. Metab *87*, 2606-2610.

Halford,S., Freedman,M.S., Bellingham,J., Inglis,S.L., Poopalasundaram,S., Soni,B.G., Foster,R.G., and Hunt,D.M. (2001). Characterisation of a novel human Opsin gene with wide tissue expression and identification of embedded and flanking genes on chromosome 1q43. Genomics *72*, 203-208.

Hanis,C.L., Boerwinkle,E., Chakraborty,R., Ellsworth,D.L., Concannon,P., Stirling,B., Morrison,V.A., Wapelhorst,B., Spielman,R.S., Gogolin-Ewens,K.J., Shepard,J.M., Williams,S.R., Risch,N., Hinds,D., Iwasaki,N., Ogata,M., Omori,Y., Petzold,C., Rietzch,H., Schroder,H.E., Schulze,J., Cox,N.J., Menzel,S., Boriraj,V.V., Chen,X., and et al (1996). A genome-wide search for human non-insulin-dependent (type2) diabetes genes reveals a major susceptibility locus on chromosome 2. Nature Genetics *13*, 161-166.

Harada,N., Ogawa,E., Shozu M, and Yamada K (1992). Genetic studies to characterize the origin of the mutation in placental aromatase deficiency. Am. J. Hum. Genet. *51*, 666-667.

Hill,W.G. and Weir,B.S. (1994). Maximum-likelihood estimation of gene location by linkage disequilibrium. Am. J. Hum. Genet. *54*, 705-714.

Hillier, S.G., Tetsuka, M and Fraser, H.M. (1997). Location and developmental regulation of androgen receptor in the primate ovary. Hum Reprod. *12*, 107-111.

Hirschhorn,J.N., Lohmueller,K., Byrne,E., and Hirschhorn,K. (2002). A comprehensive review of genetic association studies. Genet. Med. *4*, 45-61.

Hogeveen,K.N., Talikka,M., and Hammond,G.L. (2001). Human sex hormone-binding globulin promoter activity is influenced by a (TAAAA)n repeat element within an Alu sequence. J. Biol. Chem. *276*, 36383-36390.

Horikawa,Y., Oda,N., Cox,N.J., Li,X., Orho-Melander,M., Hara,M., Hinokio,Y., Linder,T.H., Mashima,H., Schwarz,P.E.H., del Bosque-Plata,L., Horikawa,Y., Oda,Y., Yoshiuchi,I., Colilla,S., Polonsky,K.S., Wei,S., Concannon,P., Iwasaki,N., Schulze,J., Baier,L.J., Bogardus,C., Groop,L., Boerwinkle,E., Hanis,C.L., and Bell,G.I. (2000). Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nature Genetics *26*, 163-175.

Huang,N. and Miller,W.L. (2000). Cloning of factors related to HIV-inducible LBP proteins that regulate steroidogenic factor-1- independent of human placental transcription of the cholesterol side chain cleavage enzyme, P450scc. J. Biol. Chem. *275*, 2852-2858.

Hughesdon,P.E. (1982). Morphology and morphogenesis of the Stein-Leventhal ovary and of so-called 'hyperthecosis'. Obstet Gynecol Surv. *37*, 59-77.

Hull,M.G. (1992). The causes of infertility and and relative effectiveness of treatment. In Infertility, A.A.Templeton and J.O.Drife, eds. (London: Springer-Verlag), pp. 33-62.

Huxtable,S.J., Saker,P.J., Haddad,L., Walker,M., Frayling,T.M., Levy,J.C., Hitman,G.A., O'Rahilly,S., Hattersley,A.T., and McCarthy,M.I. (2000). Analysis of Parent-Offspring Trios Provides Evidence for Linkage and Association Between the Insulin Gene and Type 2 Diabetes Mediated Exclusively Through Paternally Transmitted Class III Variable Number Tandem Repeat Alleles. Diabetes *49*, 126-130.

Ibrahim,N.M., Marinovic,A.C., Price,S.R., Young,L.G., and Frohlich,O. (2000). Pitfall of an internal control plasmid: response of Renilla luciferase (pRL-Tk) plasmid to dihydrotestosterone and dexamethasone. Biotechniques *29*, 782-784.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

International Mouse Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Ioannidis,J.P.A., Ntzani,E.E., Trikalinos,T.A., and Contopoulos-Ioannidis,J.S. (2001). Replication Validity of Genetic Association. Nature Genetics *29*, 306-309.

Ito,Y., Fisher,C.R., Conte,F.A., Grumbach,M.M., and Simpson,E.R. (1993). Molecular basis of aromatase deficiency in an adult female with sexual infantilism and polycystic ovaries. Proc Natl Acad Sci USA *90*, 11677.

Jacobs,H.S. (1994). Pulsatile GnRH in multifollicular and polycystic ovary patients. In Ovulation induction: basic science and clinical advances., M.Filicori and C.Flamigni, eds. (Amsterdam: Excerpta Medica), pp. 11673-11677.

Jakimiuk,A.J., Weitsman,S.R., Navab,A., and Magoffin,D.A. (2001). Luteinizing hormone receptor, steroidogenesis acute regulatory protein, and steroidogenic enzyme messenger ribonucleic acids are overexpressed in thecal and granulosa cells from polycystic ovaries. J. Clin. Endocrinol. Metab *86*, 1318-1323.

Kiddy,D.S., Hamilton-Fairley,D., Seppala,M., Koistinen,R., James,V.H., Reed,M.J., and Franks,S. (1989). Diet-induced changes in sex hormone binding globulin and free testosterone in women with normal or polycystic ovaries: correlation with serum insulin and insulin-like growth factor-I. Clin. Endocrinol. (Oxf) *31*, 757-763.

Knochenhauer,E.S., Key,T.J., Kahsar-Miller,M., Waggoner,W., Boots,L.R., and Azziz,R. (1998). Prevalence of the polycystic ovary syndrome in unselected black and white women of the southeastern United States: a prospective study. J. Clin. Endocrinol. Metab *83*, 3078-3082.

Kohler,P.O. and Bridson,W.E. (1971). Isolation of hormone- producing clonal cell lines of human choriocarcinoma. J. Clin. Endocrinol. Metab *32*, 683-687.

Kozak,M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell *44*, 283-292.

Le Roith,D. and Zick,Y. (2001). Recent advances in our understanding of insulin action and insulin resistance. Diabetes Care *24*, 588-597.

Le Stunff C, Fallin D, Schork,N.J., and Bougneres P (2000). The insulin gene VNTR is associated with fasting insulin levels and development of juvenile obesity. Nature Genetics *26*, 444-446.

Le Stunff C, Fallin D, and Bougneres P (2001). Paternal transmission of the very common class I INS VNTR allelespredisposes to childhood obesity. Nature Genetics *29*, 96-99.

Legro,R.S., Driscoll,D., Strauss,J.F., III, Fox,J., and Dunaif,A. (1998). Evidence for genetic basis for hyperandrogenaemia in polycystic ovary syndrome. Proc Natl Acad Sci USA *95*, 14956-14960.

Legro,R.S., Kunselman,A.R., Dodson,W.C., and Dunaif,A. (1999). Prevalence and predictors of risk for type 2 diabetes mellitus and impaired glucose tolerance in polycystic ovary syndrome: a prospective, controlled study in 254 affected women. J. Clin. Endocrinol. Metab *84*, 165-169.

Legro,R.S., Bentley-Lewis,R., Driscoll,D., Wang,S.C., and Dunaif,A. (2002). Insulin Resistance in the Sisters of Women with Polycystic Ovary Syndrome: Association with Hyperandrogenaemia Rather than Menstrual Irregularity. J. Clin. Endocrinol. Metab *87*, 2128-2133.

Legro,R.S. (2003). Diagnostic criteria in polycystic ovary syndrome. Semin Reprod Med *21*, 267-275.

Lohmueller,K.E., Pearce,C.L., Pike,M., Lander,E.S., and Hirschhorn,J.N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common diseases. Nature Genetics *33*, 177-182.

Lui,Z. and Simpson,E.R. (1997). Steroidogenic factor-1(SF-1) and SP1 are required for regulation of bovine CYP11a gene expression in bovine luteal cells and adrenal Y1 cells. Mol Endocrinol *11*, 127-137.

Lunenfeld,B. and Insler,V. (1978). Diagnosis and treatment of functional infertility. (Berlin: Gross Verlag).

Mahendroo,M.S., Means,G.D., Mendelson,C.R., and Simpson,E.R. (1999). Tissue specific expression of human P450 AROM. The promoter responsible for expression in adipose tissue is different from that utilized in placenta. J. Biol. Chem. *266*, 11276-11281.

Marshall,J.C. and Eagleson,C.A. (1999). Neuroendocrine aspects of polycystic ovary syndrome. Endocrinol. Metab Clin. North Am. *28*, 295-324.

Mason,H.D., Willis,D.S., Beard,R.W., Winston,R.M., Margara,R., and Franks,S. (1994). Estradiol production by granulosa cells of normal and polycystic ovaries: relationship to menstrual cycle history and concentrations of gonadotropins and sex steroids in follicular fluid. J. Clin. Endocrinol. Metab *79*, 1355-1360.

Mather J.P, Moore A, and Li R.H (1997). Activins, inhibins, and follistatins: further thoughts on a growing family ofregulators. Proc Soc Exp Biol Med *215*, 209-222.

McArthur,J.W., Ingersoll,F.M., and Worcerster,J. (1958). The urinary excretion of interstitial-cell and follicle-stimulating hormone activity by women with diseases of the reproductive system. J. Clin. Endocrinol. Metab *18*, 1202-1215.

McCarthy,M.I., Kruglyak,L., and Lander,E.S. (1998). Sib-pair collection strategies for complex diseases. Genetic Epidemiology *15*, 317-340.

McNatty,K.P., Hillier,S.G., van den Boogaard,A.M., Trimbos-Kemper,T.C., Reichert,L.E., Jr., and van Hall,E.V. (1983). Follicular development during the luteal phase of the human menstrual cycle. J. Clin. Endocrinol. Metab *56*, 1022-1031.

Meloni,R., Albanese,V., Ravassard,P., Treilhou,F., and Mallet,J. (1998). A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element *in vitro*. Hum. Mol. Genet. *7*, 423-428.

Miller,W.L. (1988). Molecular biology of steroid hormone synthesis. Endocr. Rev. *9*, 295-318.

Monte,D., DeWitte,F., and Hum,D.W. (1998). Regulation of the human P450scc gene by steroidogenic factor 1 is mediated by CBP/p300. J. Biol. Chem. *273*, 4585-4591.

Moore,C.C., Brentano,S.T., and Miller,W.L. (1990). Human P450scc gene transcription is induced by cyclic AMP and repressed by 12-O-tetradecanoylphorbol-13-acetate and A23187 through independent cis elements. Mol. Cell Biol. *10*, 6013-6023.

Moore,C.C., Hum,D.W., and Miller,W.L. (1992). Identification of positive and negative placenta-specific basal elements and a cyclic adenosine 3',5'-monophosphate response element in the human gene for P450scc. Mol. Endocrinol. *6*, 2045-2058.

Mooser,V., Francesco,P.M., Bopp,S., Petho-Schramm,A., Guerra,R., Boerwinkle,E., Muller,H.H., and Hobbs,H.H. (1995). Sequence polymorphisms in the Apo(a) gene is associated with specific levels of Lp(a) in plasma. Hum. Mol. Genet. *4*, 173-181.

Morin-Papunen,L.C., Vauhkonen,I., Koivunen,R.M., Ruokonen,A., Martikainen,H.K., and Tapanainen,J.S. (2000). Endocrine and metabolic effects of metformin versus ethinyl estradiol-cyproterone acetate in obese women with polycystic ovary syndrome: a randomized study. J. Clin. Endocrinol. Metab *85*, 3161-3168.

Morohashi,K., Sogawa,K., Omura,T., and Fujii-Kuriyama,Y. (1987). Gene structure of human cytochrome P-450(SCC), cholesterol desmolase. J. Biochem. (Tokyo) *101*, 879-887.

Morohashi,K., Zanger,U.M., Honda,S., Hara,M., Waterman,M.R., and Omura,T. (1993). Activation of CYP11A and CYP11B gene promoters by the steroidogenic cell-specific transcription factor, Ad4BP. Mol. Endocrinol. *7*, 1196-1204.

Mount,S.M. (1982). A catalogue of splice junction sequences. Nucleic Acids Res. *10*, 459-472.

Nelson,V.L., Legro,R.S., Strauss,J.F., III, and McAllister,J.M. (1999). Augmented androgen production is a stable steroidogenic phenotype of propagated theca cells from polycystic ovaries. Mol. Endocrinol. *13*, 946-957.

Nestler,J.E., Barlascini,C.O., Matt,D.W., Steingold,K.A., Plymate,S.R., Clore,J.N., and Blackard,W.G. (1989). Suppression of serum insulin by diazoxide reduces serum testosterone levels in obese women with polycystic ovary syndrome. J. Clin. Endocrinol. Metab *68*, 1027-1032.

Nilsson E., Parrot, J.A., and Skinner, M.K. (2001). Basic fibroblast growth factor induces primordial follicle development and initiates follicle development. Mol Cell Endocrinol *175*, 123-130.

Onishi,M., Yasunaga,T., Tanaka,H., Nishimune,Y., and Nozaki,M. (2004). Gene structure and evolution of testicular haploid germ cell-specific genes, Oxct2a and Oxct2b. Genomics *83*, 647-657.

Parrot J.A and Skinner, M.K (1999). Kit-ligand/stem-cell factor induces primordial follicle development and initiates folliculogenesis. Endocrinology. *140*, 4262-4271.

Peiris,A.N., Aiman,E.J., Drucker,W.D., and Kissebah,A.H. (1989). The relative contributions of hepatic and peripheral tissues to insulin resistance in hyperandrogenic women. J. Clin. Endocrinol. Metab *68*, 715-720.

Pestell,R.G., Hammond,V.E., and Crawford,R.J. (1993). Molecular cloning and characterisation of the cyclic AMP-responsive ovine CYP11A1 (cholesterol side-chain cleavage) gene promoter: DNase 1 protection of conserved consensus elements. J. Mol. Endocrinol. *10*, 297-311.

Plymate,S.R., Matej,L.A., Jones,R.E., and Friedl,K.E. (1988). Inhibition of sex hormone binding globulin production in the human hepatoma (HepG2) cell line by insulin and prolactin. J. Clin. Endocrinol. Metab *67*, 460-464.

Polson,D.W., Mason,H.D., Saldahna,M.B., and Franks,S. (1987). Ovulation of a single dominant follicle during treatment with low-dose pulsatile follicle stimulating hormone in women with polycystic ovary syndrome. Clin. Endocrinol. (Oxf) *26*, 205-212.

Powell,B.L., Haddad,L., Bennett,A., Gharani,N., Sovio,U., Groves,C.J., Rush,K., Goh,M.J., Conway,G., Ruokonen,A., Martikainen,H., Pouta,A., Taponen,S., Hartikainen,A.L., Halford,S., Zeggini,E., Jarvelin,M.R., Franks,S., and McCarthy,M.I. (2005). Analysis of multiple data sets reveals no association between the insulin gene variable number tandem repeat element and polycystic ovary syndrome or related traits. J. Clin. Endocrinol. Metab *in press*.

Prelevic,G.M. (1997). Insulin resistance in polycystic ovarian syndrome. Curr. Opin. Obstet. Gynecol. *9*, 193-201.

Proudfoot,N. and O'Sullivan,J. (2002). Polyadenylation: A Tail of Two Complexes. Current Biology *12*, R955-R957.

Rainey,W.E., Sawetawan,C., Shay,J.W., Michael,M.D., Mathis,J.M., Kutteh,W., Byrd,W., and Carr,B.R. (1994). Transformation of Human Granulosa Cells with the E6 and E7 Regions of Human Papillomavirus. J. Clin. Endocrinol. Metab *78*, 705-710.

Rainey,W.E., Sawetawan,C., McCarthy,J.L., McGee,E.A., Bird,I.M., Word,A., and Carr,B.R. (1996). Human ovarian tumour cells; A potential model for thecal cell steroidogenesis. J. Clin. Endocrinol. Metab *81*, 257-263.

Rebar,R., Judd,H.L., Yen,S.S., Rakoff,J., Vandenberg,G., and Naftolin,F. (1976). Characterization of the inappropriate gonadotropin secretion in polycystic ovary syndrome. J. Clin. Invest *57*, 1320-1329.

Rezniczek,G.A., Abrahamsberg,C., Fuchs,P., Spazierer,D., and Wiche,G. (2003). Plectin 5' transcript diversity: short alternate sequences determine stability of gene products, initiation of translation and subcellular localization of isoforms. Hum. Mol. Genet. *12*, 3181-3194.

Rice,D.A., Kirkman,M.S., Aitken,L.D., Mouw,A.R., Schimmer,B.P., and Parker,K.L. (1990). Analysis of the promoter region of the gene encoding mouse cholesterol side-chain cleavage enzyme. J. Biol. Chem. *265*, 11713-11720.

Rodin,A., Thakkar,H., Taylor,N., and Clayton,R. (1994). Hyperandrogenism in polycystic ovary syndrome. Evidence of dysregulation of 11 ß-hydroxysteroid dehydrogenase. N. Engl. J. Med. *330*, 460-465.

Rodriguez,H., Hum,D.W., Staels,B., and Miller,W.L. (1997). Transcription of the human genes for cytochrome P450scc and P450c17 is regulated differently in human adrenal NCI-H295 cells than in mouse adrenal Y1 cells. J. Clin. Endocrinol. Metab *82*, 365-371.

Saltiel,A.R. (1996). Diverse signaling pathways in the cellular actions of insulin. Am. J. Physiol *270*, E375-E385.

San Millan,J.L., Sancho,J., Calvo,R.M., and Escobar-Morreale,H.F. (2001). Role of the pentanucleotide (tttta)(n) polymorphism in the promoter of the CYP11a gene in the pathogenesis of hirsutism. Fertil. Steril. *75*, 797-802.

Sparkes,R.S., Klisak,I., and Miller,W.L. (1991). Regional mapping of genes encoding human steroidogenic enzymes: P450scc to 15q23-q24, adrenodoxin to 11q22; adrenodoxin reductase to 17q24-q25; and P450c17 to 10q24-q25. DNA Cell Biol. *10*, 359-365.

Stein,I.F. and Leventhal,M.L. (1935). Amenorrhea associated with bilateral polycystic ovaries. Am. J. Obstet. Gynecol. *29*, 181-191.

Steppan,C.M., Brown,E.J., Wright,C.M., Bhat,S., Banerjee,R.R., Dai,C.Y., Enders,G.H., Silberg,D.G., Wen,X., Wu,G.D., and Lazar,M.A. (2001a). The hormone resistin links obesity to diabetes. Nature *409*, 307-312.

Steppan,C.M., Brown,E.J., Wright,C.M., Bhat,S., Banerjee,R.R., Banerjee, R.R., Dai,C.Y., Enders,G.H., Silberg,D.G., Wen,X., Wu,G.D., and Lazar,M.A. (2001b).A family of tissue-specific resistin- like molecules. Proc. Natl. Acad. Sci. *98*, 502-506.

Takayama,K., Morohashi,K., Honda,S., Hara,N., and Omura,T. (1994). Contribution of Ad4BP, a steroidogenic cell-specific transcription factor, to regulation of the human CYP11A and bovine CYP11B genes through their distal promoters. J. Biochem. (Tokyo) *116*, 193-203.

Taponen,S., Martikainen,H., Jarvelin,M.R., Laitinen,J., Pouta,A., Hartikainen,A.L., Sovio,U., McCarthy,M.I., Franks,S., and Ruokonen,A. (2003). Hormonal profile of women with self-reported symptoms of oligomenorrhea and/or hirsutism: Northern Finland birth cohort 1966 study. J. Clin. Endocrinol. Metab *88*, 141-147.

Telgmann,R., Maronde,E., Tasken,K., and Gellersen,B. (1997). Activated Protein kinase A is required for differentiation-dependant transcription of the decidual prolactin gene in human endometrial stromal cells. Endocrinology *138*, 929-937.

The Rotterdam ESHRE/ASRM Sponsored PCOS Consensus Workshop Group. (2004). Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. Fertil. Steril. *81(1)*, 19-25.

Trommsdorff,M., Kochl,S., Lingenhel,A., Kronenberg,F., Delport,R., Vermaak,H., Lemming,L., Klausen,I.C., Faergeman,O., and Kraft,H.G. (1995). A pentanucleotide repeat polymorphism in the 5' control region of the apolipoprotein(a) gene is associated with lipoprotein(a) plasma concentrations in Caucasians. J. Clin. Invest *96*, 150-157.

Urbanek,M., Legro,R.S., Driscoll,D.A., Azziz,R., Ehrmann,D.A., Norman,R.J., Strauss,J.F., III, Spielman,R.S., and Dunaif,A. (1999). Thirty-seven candidate genes for polycystic ovary syndrome: strongest evidence for linkage is with follistatin. Proc. Natl. Acad. Sci. U. S. A *96*, 8573-8578.

Urbanek,M., Wu,X., Vickery,K.R., Kao,L.C., Christenson,L.K., Schneyer,A., Legro,R.S., Driscoll,D.A., Strauss,J.F., III, Dunaif,A., and Spielman,R.S. (2000). Allelic variants of the follistatin gene in polycystic ovary syndrome. J. Clin. Endocrinol. Metab *85*, 4455-4461.

Urbanek,M., Du,Y., Silander,K., Collins,F.S., Steppan,C.M., Strauss,J.F., III, Dunaif,A., Spielman,R.S., and Legro,R.S. (2003). Variation in resistin gene promoter not associated with polycystic ovary syndrome. Diabetes *52*, 214-217.

Utermann,G. (1989). The mysteries of lipoprotein(a). Science *246*, 904-910.

Vafiadis,P., Bennett,S.T., Todd,J.A., Nadeau,J., Grabs,R., Goodyer,C.G., Wickramasinghe,S., Colle,E., and Polychronakos,C. (1997). The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the INS-VNTR-IDDM2 susceptibility locus for type 1 diabetes. Nature Genetics *15*, 293-297.

Vendola, K.A., Zhou, J., Adesayna, O.O., Weill, S.J., and Bondy, C.A. (1998). Androgens stimulate early stages of follicular growth in the primate ovary. J. Clin Invest. *101*, 2622-2629.

Wade,D.P., Clarke,J.G., Lindahl,G.E., Lui,A.C., Zysow,B.R., Meer,K., Shwartz,K.S., and Lawn,R.M. (1993). 5' control regions of the apolipoprotein(a) gene and members of the related plasminogen gene family. Proc Natl Acad Sci USA *90*, 1369-1373.

Wang,C.F. and Gemzell,C. (1980). The use of human gonadotropins for induction of ovulation in women with polycystic ovarian disease. Fertil. Steril. *33*, 479-486.

Waterworth,D.M., Bennett,S.T., Gharani,N., McCarthy,M.I., Hague,S., Batty,S., Conway,G.S., White,D., Todd,J.A., Franks,S., and Williamson,R. (1997). Linkage and association of insulin gene VNTR regulatory polymorphism with polycystic ovary syndrome. Lancet *349*, 986-990.

Webber,L.J., Stubbs,S., Stark,J., Trew,G.H., Margara,R., Hardy,K., and Franks,S. (2003). Formation and early development of follicles in the polycystic ovary. Lancet *362*, 1017-1021.

Weill, S.J., Vendola, K., Zhou, J., Adesanya, O.O., Wang, J., Okafor, J and Bondy, C.A. (1998). Androgen receptor gene expression in the primate ovary: cellular localisation, regulation, and functional correlations. J. Clin. Endocrinol. Metab. *83*, 2479-2485.

White,D.M., Polson,D.W., Kiddy,D., Sagle,P., Watson,H., Gilling-Smith,C., Hamilton-Fairley,D., and Franks,S. (1996). Induction of ovulation with low-dose gonadotropins in polycystic ovary syndrome: an analysis of 109 pregnancies in 225 women. J. Clin. Endocrinol. Metab *81*, 3821-3824.

White,M.F. and Yenush,L. (1998). The IRS-signaling system: a network of docking proteins that mediate insulin and cytokine action. Curr. Top. Microbiol. Immunol. *228*, 179-208.

Wood,J.R., Nelson,V.L., Ho,C., Jansen,E., Wang,C.Y., Urbanek,M., McAllister,J.M., Mosselman,S., and Strauss,J.F., III (2003). The molecular phenotype of polycystic ovary syndrome (PCOS) theca cells and new candidate PCOS genes defined by microarray analysis. J. Biol. Chem. *278*, 26380-26390.

Yen,S.S., Vela,P., and Rankin,J. (1970). Inappropriate secretion of follicle-stimulating hormone and luteinizing hormone in polycystic ovarian disease. J. Clin. Endocrinol. Metab *30*, 435-442.

Yen,S.S. (1980). The polycystic ovary syndrome. Clin. Endocrinol. (Oxf) *12*, 177-207.

Zawadski,J.K. and Dunaif,A. (1992). Diagnostic criteria for polycystic ovary syndrome; towards a rational approach. In PCOS, A.Dunaif, J.R.Givens, and F.Haseltine, eds. (Boston: Blackwell Scientific), pp. 377-384.

0021-972X/04/$15.00/0
Printed in U.S.A.

The Journal of Clinical Endocrinology & Metabolism 89(5):2408–2413
Copyright © 2004 by The Endocrine Society
doi: 10.1210/jc.2003-031640

Large-Scale Analysis of the Relationship between *CYP11A* Promoter Variation, Polycystic Ovarian Syndrome, and Serum Testosterone

MICHELLE GAASENBEEK, BRENDA L. POWELL, ULLA SOVIO, LEMA HADDAD, NEDA GHARANI, AMANDA BENNETT, CHRISTOPHER J. GROVES, KAREN RUSH, MICAELA J. GOH, GERARD S. CONWAY, AIMO RUOKONEN, HANNU MARTIKAINEN, ANNELI POUTA, SAARA TAPONEN, ANNA-LIISA HARTIKAINEN, STEPHANIE HALFORD, MARJO-RIITTA JÄRVELIN, STEVE FRANKS, AND MARK I. McCARTHY

Genomic Medicine, Faculty of Medicine (M.G., L.H., N.G., A.B., S.H., M.I.M.) and Institute of Reproductive and Developmental Biology (M.G., S.F.), Imperial College, Hammersmith Campus, London W12 0NN, United Kingdom; Departments of Epidemiology and Public Health (U.S., M.-R.J.) and Reproductive Endocrinology (K.R., M.J.G., S.F.), Imperial College, St. Mary's Campus, London W2 1PG, United Kingdom; Wellcome Trust Centre for Human Genetics (B.L.P., C.J.G., M.I.M.) and Oxford Centre for Diabetes, Endocrinology and Metabolism (B.L.P., A.B., C.J.G., M.I.M.), Churchill Hospital, Oxford OX3 7LJ, United Kingdom; Department of Obstetrics and Gynaecology (G.S.C.), Bloomsbury Site, University College, London WC1E 6BT, United Kingdom; and Departments of Clinical Chemistry (A.R., S.T.), Obstetrics and Gynecology (S.T., H.M., A.-L.H.), and Public Health Science and General Practice (M.-R.J., S.T., A.P.), University of Oulu, Oulu, Finland 90014

CYP11A, the gene encoding P450scc, a key enzyme in steroid biosynthesis, is a strong biological candidate for polycystic ovary syndrome (PCOS) susceptibility. Four of the five published studies that have examined _CYP11A_ for evidence of linkage and/or association have reported significant relationships with polycystic ovary (PCO) status and/or serum testosterone levels. However, study sizes have been modest, and the current study aimed to reevaluate these findings using significantly larger clinical resources. A pair of _CYP11A_ promoter microsatellites, including the pentanucleotide (D15S520) previously implicated in trait susceptibility, were genotyped in 371 PCOS patients of United Kingdom origin, using both case-control and family-based association methods, and in 1589 women from a population-based birth cohort from Finland characterized for PCO symptomatology and testosterone levels. Although nominally significant differences in allele and genotype frequencies at both loci were observed in the United Kingdom case-control study (for example, an excess of the pentanucleotide four-repeat allele in cases, $P = 0.005$), these findings were not substantiated in the other analyses, and no discernable relationship was seen between variation at these loci and serum testosterone levels. These studies indicate that the strength of, and indeed the existence of, associations between _CYP11A_ promoter variation and androgen-related phenotypes has been substantially overestimated in previous studies. (_J Clin Endocrinol Metab_ 89: 2408–2413, 2004)

POLYCYSTIC OVARY SYNDROME (PCOS) is a heterogeneous endocrine disorder of premenopausal women and the commonest cause of anovulatory infertility and hirsutism (1, 2). A cardinal feature of this condition is a disturbance of ovarian steroidogenesis, with excessive androgen production responsible for many of the principal symptomatic manifestations. In addition, individuals with PCOS display metabolic abnormalities that include prominent defects in both insulin action (3) and β-cell function (4).

PCOS displays strong familial aggregation (5), and the etiology is likely to be multifactorial, individual susceptibility being determined by the action of multiple genetic and environmental risk factors. A wide range of potential candidate genes have been studied in the search for PCOS-susceptibility variants, but few of these have been studied in multiple populations, and even fewer have shown convincing replication (6, 7).

Some of the strongest evidence presented to date has related to a pentanucleotide repeat 5′ to the *CYP11A* gene. The *CYP11A* gene is a credible biological candidate because it encodes P450scc (also known as cytochrome P450scc or cholesterol side-chain cleavage), the enzyme catalyzing the initial and rate-limiting step in adrenal and ovarian androgen biosynthesis. Thecal cells cultured from PCOS subjects show increased androgen production (8) and, in most studies, increased *CYP11A* mRNA expression (9–11) compared with samples from control women. Maintenance of this cellular phenotype on long-term culture (12) strengthens the hypothesis that *CYP11A* overexpression is a stable property of PCOS theca cells that contributes to the excess androgen production that characterizes this condition.

In the search for genetic variants that might be associated with *CYP11A* overexpression, attention has focused on a pentanucleotide repeat approximately 500 bp 5′ to the translation start site (D15S520). Gharani *et al.* (13) reported evidence for linkage to this region in 20 multiplex pedigrees

Abbreviations: PCO, Polycystic ovary; PCOS, PCO syndrome; USS, ultrasound scan.

JCEM is published monthly by The Endocrine Society (http://www.endo-society.org), the foremost professional society serving the endocrine community.

Downloaded from jcem.endojournals.org on November 8, 2005

segregating PCOS (LOD, ~1.74), and in an analysis of 148 PCOS subjects and 59 controls, associations between pentanucleotide length (specifically absence of the 216-bp, four-repeat allele) and both symptomatic PCOS and raised testosterone levels were found. Subsequent studies have, to varying extents, corroborated these findings. Urbanek *et al.* (14) found nominal evidence for excess allele sharing in the region in an analysis of 28 multiplex sibships, as well as a nonsignificant reduction in transmission (47%) of the four-repeat allele to PCOS offspring in 163 parent-offspring trios. Diamanti-Kandarakis *et al.* (15) studied 170 Greek subjects (80 with PCOS) and found a significant increase in the frequency of four-repeat-lacking genotypes in PCOS subjects (26% *vs.* 13% in controls) and reproduced the previously observed association with testosterone. Most recently, Daneshmand *et al.* (10) also reported a reduction in the four-repeat allele frequency in PCOS (14% in 51 PCOS women *vs.* 44% in 280 controls), which was associated with a striking compensatory difference in the frequency of the nine-repeat allele (56% in PCOS women *vs.* 7% in controls). The only published study that has failed to produce any evidence for a relationship between pentanucleotide genotype and phenotype, a Spanish study of 92 hirsute women (only 34 of whom had PCOS) and 33 controls, was also the smallest and, consequently, likely to have been underpowered (16).

On the basis of these data, *CYP11A* is generally considered one of the few well-supported examples of a PCOS susceptibility effect (6, 7). However, the sample sizes included in the studies described have been rather modest. Increasing concern over the poor performance of association studies in the identification and characterization of complex trait susceptibility effects (17–19) prompted us to reevaluate the relationship between *CYP11A* promoter variation and PCOS disease status, polycystic ovary (PCO)-related symptoms, and serum testosterone levels.

In this study, we have extended genotyping of the *CYP11A* pentanucleotide, together with a neighboring tetranucleotide repeat element not previously studied, to almost 2500 individuals, which is around twice the combined size of the previous studies. Our analyses of these adequately sized and well-characterized populations indicate that these variants do not contribute significantly to observed phenotypic variation in either serum testosterone levels or PCOS disease status.

Subjects and Methods

United Kingdom (UK) subjects

Women with PCOS (n = 583) were ascertained from infertility and endocrine clinics, predominantly at St. Mary's and the Middlesex Hospitals in London. The criteria for diagnosis of PCOS were as previously described (5, 20) and, in line with the revised (2003) consensus on the diagnostic criteria for this syndrome (21), rely on the combination of clinical symptoms, ultrasonographic examination, and biochemical data. The essential criterion for the diagnosis of PCOS in this study was the presence of polycystic ovaries on ultrasound in a patient presenting with anovulation (oligomenorrhea or amenorrhea) and/or hyperandrogenism. A diagnosis of hyperandrogenism required clinical (presence of hirsutism or acne) and/or biochemical [serum total testosterone > 0.78 ng/ml (2.7 nmol/liter)] features. Other potential endocrine and neoplastic causes of hyperandrogenemia were excluded by appropriate tests (21). From this wider patient cohort of diverse ethnic origin, all of whom were typed for the variants of interest, we restrict the analyses reported here to the 371 women with exclusively British/Irish origin (based on three generations of reported ethnicity and birthplace). For 141 of these cases (designated probands), DNA was available from both parents, such that parent-offspring trios could be generated; the remaining 230 patients without parental DNA are designated PCOS cases. Family relationships in the trios were confirmed by genotyping with a panel of highly polymorphic microsatellite markers (combined heterozygosity, >99%). Quantitative trait analyses were performed in all PCOS subjects (*i.e.* probands and cases combined) but excluded any woman taking oral contraceptive treatment, pregnant at the time of sampling, or with previous oophorectomy. A small number of women (4%, 14 of 371 women) who had received metformin were included in these analyses; their exclusion did not alter the findings.

As population control subjects for case-control analyses, we used DNA (n = 350) from a random UK population of blood donors available from the Centre for Applied Microbiology and Research (Salisbury, UK), used extensively for such purposes. All control subjects were of British UK origin. Metabolic and hormonal status was not known for these individuals. Use of population controls (rather than subjects known not to have PCOS) is generally associated with only a modest reduction in power, which can readily be compensated by increasing the number of control subjects typed (19).

Finnish subjects

For the cohort study, samples were available from the Northern Finland Birth Cohort of 1966, which is based on 12,058 live-born babies born in the northernmost two provinces of Finland arising from pregnancies with expected dates of delivery during calendar year 1966. Of 5926 cohort members (3088 women) who had undergone clinical review and DNA sampling at age 31, we genotyped 1589 for this study (22). These included 527 women who had reported symptoms of hirsutism and/or oligo/amenorrhea on the 31-year review questionnaire (symptomatic women) and 1062 randomly selected women (cohort controls) from the same cohort without such symptoms (generating an approximate 1:2 ratio of cases to controls, all matched for age given the birth cohort design). No woman who was receiving oral contraceptives at the time of metabolic sampling was included in either group. The cohort cases have been shown as a group to have biochemical features consistent with PCOS (*e.g.* hyperandrogenemia and insulin resistance) (22), but only a proportion have proven PCOS. Ovarian ultrasound scan (USS) examination of these symptomatic women is ongoing, with 67 of 178 so far examined having PCO status confirmed on USS. These 67 individuals are referred to as USS-confirmed cohort cases. For the quantitative trait analyses, which were the main emphasis of analysis in the Finnish subjects, the group size was reduced to 1435 (475 cases and 960 controls) by exclusion of 138 women who were pregnant at the time of the blood sampling and 16 in whom no testosterone measures were available. The single subject who was taking metformin at the time of recruitment is included.

Clinical features of these sample sets are provided in Table 1. All clinical investigations were conducted in accordance with the guidelines in the Declaration of Helsinki, and the study was approved by the relevant ethics committees in UK and Finland. All subjects provided fully informed consent.

Laboratory assays

In the UK women with PCOS and all the Finnish subjects, fasting samples were used for the analysis of testosterone levels. For the UK subjects, testosterone was measured with an in-house RIA using ether extraction and dextran-coated charcoal separation. Genomic DNA was extracted from whole blood using the PUREGENE kit (Gentra Systems, Minneapolis, MN) as described previously (23). For the Finnish subjects, serum testosterone was determined by automated chemiluminescence system (Ciba-Corning ACS-180; Diamond Diagnostics, Holliston, MA).

Genotyping

The pentanucleotide D15S520 was genotyped by PCR amplification in a 15-μl volume containing 30 ng genomic DNA, 10 pmol of deoxynucleotide triphosphates, 1.0 mM $MgCl_2$, 1U *Taq* polymerase (Bioline, London, UK), and 8 μM of each of the forward (FAM-labeled

Downloaded from jcem.endojournals.org on November 8, 2005

TABLE 1. Clinical characteristics of study subjects

	UK subjects			Finnish subjects	
	PCOS cases	PCOS probands	Population controls	Symptomatic women	Cohort controls
n	230	141	331	527	1062
% female	100	100	55.5	100	100
Age (yr)	33.8 (6.9)[a]	31.4 (6.1)[a]	38.6 (8.1)[b]	31	31
BMI (kg/m^2)	27.4 (7.8)	27.0 (7.4)		25.2 (5.4)	24.2 (4.6)
Waist to hip ratio	0.80 (0.07)	0.80 (0.09)		0.82 (0.08)	0.81 (0.08)
Testosterone (ng/ml)[c]	0.64 (0.26)	0.68 (0.34)		0.67 (0.29)	0.59 (0.25)
Glucose (mg/dl)[d]	86 (10)	85 (18)		90 (18)	88 (12)
Symptoms					
Hirsutes	105 (46%)	87 (62%)		295 (56%)	0
Oligomenorrhea	145 (63%)	115 (82%)		312 (60%)	0
Ovarian status	USS confirmed PCOS	USS confirmed PCOS		USS underway, 67/178 confirmed	Not known but asymptomatic

Quantitative data are presented as mean (SD).

[a] Ranges for subject ages were as follows: UK PCOS cases, 19.3–57.6 yr; UK PCOS probands, 16.5–47.7 yr. All Finnish women were aged 31 yr.

[b] Age information is available from only 38% of these control subjects.

[c] To convert to SI units (nanomoles per liter) multiply by 3.467.

[d] To convert to SI units (millimoles per liter) multiply by 0.05551.

5′-CCTGGGCAACAAGAATGAAA for the UK subjects and FAM-labeled 5′-GGTGAAACTGTGCCATTGC for the Finns) and reverse (5′-GCCCAAGAGAAGGTCAGAGC for the UK subjects and 5′-GTTT-GGGGGAAATGAGGGGC for the Finns) primers. Details of PCR conditions are available from the authors. Sequence analysis of the *CYP11A* gene revealed a second (tetranucleotide) microsatellite lying in intron 1 approximately 18.4 kb upstream of exon 2. This tetranucleotide lies immediately upstream of a novel exon that is expressed as the 5′ end of an alternate *CYP11A* transcript (Gaasenbeek, M., M. I. McCarthy, S. Halford, and S. Franks, unpublished observations). We elected to type this microsatellite, which lies approximately 1500 bp from D15S520, because sequence context suggested a potential functional role analogous to that of the pentanucleotide with respect to the usual promoter. This microsatellite was amplified in a separate reaction (primers: forward, HEX-labeled 5′-AGTAGCTGGGATTATAGACCC; reverse, 5′-GGAGACTGGTGAGGCTAAGT; 1.5 mM $MgCl_2$). Resulting PCR products were diluted 1:4 in water, and 0.5 μl of each PCR reaction was mixed with formamide and 450 bp ROX size standards (Applied Biosystems, Warrington, UK), followed by separation and sizing on an ABI PRISM 3700 DNA sequencer (Applied Biosystems). Genotypes were derived using Genotyper software (Applied Biosystems). When possible, Mendel-checking methods were used to confirm genotyping accuracy. In addition, approximately 10% of samples were re-genotyped. From these data, we estimate that the overall genotyping error is less than 1%. For convenience, alleles are designated by prefix P or T, reflecting pentanucleotide or tetranucleotide locus, respectively, and a number reflecting the number of repeat units (*e.g.* P4 includes four repeats at the pentanucleotide). We did not genotype either D15S169 or D15S519 [markers typed in some of the previous studies (13, 14)] either because of distance from the gene (D15S169) or uncertain location (D15S519 does not map to the current genome assembly).

Statistical analysis

Linkage disequilibrium relationships between the two microsatellites were determined in UK and Finnish control subjects using the LDMAX option within the GOLD (Graphical Overview of Linkage Disequilibrium) package (24), which implements the expectation-maximization (EM) algorithm. Between-group comparisons of genotype and allele frequencies used standard contingency table methods implemented in STATXACT (Cytel, Cambridge, MA), which allows exact probabilities (or Monte-Carlo estimates thereof) to be generated. Haplotype frequencies were derived by standard maximum likelihood methods implemented in TRANSMIT (25), with group comparisons assessed by likelihood ratio testing and permutation methods (at least 1000 permutations). In the UK trios, family-based association tests (25, 26) were undertaken using TRANSMIT; alleles and haplotypes with a frequency less than 2% were pooled for their respective analyses. Quantitative trait analyses of serum testosterone levels were performed by ANOVA on log-transformed values, using SPSS (version 10; SPSS Inc., Chicago, IL) and SAS (version 8.2; SAS Institute, Inc., Cary, NC), with and without adjustment for age. For analyses of testosterone levels, all women taking oral contraceptive medication at the time of sampling were excluded, as were women who were pregnant or had undergone oophorectomy.

Power calculations

The UK samples (case control + trios combined) provide approximately 85% power (at $P = 0.05$) to detect effect sizes (odds ratios for PCOS, ~1.4–1.5) significantly smaller than those implied by some earlier reports (odds ratio, >2) (13, 15). The largest group studied for the relationship between *CYP11A* variation and testosterone levels (the Finnish cohort controls) provided more than 90% power to detect (at $P = 0.05$) differences between P4-containing and non-P4-containing genotypes of approximately 0.33 of a SD. Previous studies of the pentanucleotide (13, 15) have suggested effect sizes at least twice this.

Results

In the UK samples, both microsatellites were confirmed as polymorphic. Variation at the pentanucleotide generated seven alleles ranging from 169 bp (P4) to 199 bp (P10), with the P4 allele being most common (frequency, ~58% in controls). The 169-bp (P4) allele is equivalent to the 216-bp allele designated in some previous papers, the difference reflecting use of differing PCR primers (13). Seven alleles were observed at the tetranucleotide ranging from 276 bp (T6) to 304 bp (T13), with the 284-bp (T8) modal (~82% in controls). In the Finnish controls, allele distributions were similar, with the modal alleles again being P4 (69%) and T8 (81%).

In the UK population controls, the two microsatellites were in strong linkage disequilibrium (D′, ~0.86; $P < 0.001$); the most marked departures from equilibrium were positive associations between the P4 and T8 alleles (56.5% observed *vs.* 47.6% expected) and between P6 and T12 (11.3% *vs.* 3.1%) and the complementary negative associations between P6 and T8 (10.8% *vs.* 21.7%) and P4 and T12 (0.4% *vs.* 6.8%). In the Finnish controls, the extent and direction of linkage disequilibrium relationships between alleles at the two loci were similar (D′, ~0.87; $P < 0.0001$).

At the pentanucleotide, comparisons between UK PCOS

Downloaded from jcem.endojournals.org on November 8, 2005

cases and population controls revealed significant differences in genotype frequencies (Monte-Carlo estimate, $P = 0.029$; 99% confidence interval, 0.027–0.030), which were largely attributable to an excess of P4 homozygotes in the cases (47.2% *vs.* 34.8% in controls; Table 2). The same was true of allele frequencies (exact $P = 0.005$, reflecting P4 allele frequencies of 67.6% in cases and 58.2% in controls; Table 2). At the tetranucleotide, there were no significant differences in genotype frequencies ($P = 0.139$), but allele frequency comparisons indicated an excess of T8 alleles in cases ($P = 0.0097$; 88.4% *vs.* 82.4% in controls). However, global comparison of haplotype frequencies (by maximum likelihood testing) revealed no significant differences between cases and controls (exact $P = 0.15$), although, as expected given the allele frequencies, the P4-T8 haplotype was more frequent in cases (65.5% *vs.* 56.5% in controls, $P = 0.005$ uncorrected).

We sought to replicate these associations in the trios (the probands from which had not been included in the previous analyses) by seeking evidence for transmission disequilibrium. No suggestion of excess transmission was seen for the P4 or any other allele at the pentanucleotide (global analysis, $\chi^2 = 4.15$, df = 4, $P = 0.39$; P4 transmissions, $\chi^2 = 0.25$, df = 1, $P = 0.62$), for the T8 or any other allele at the tetranucleotide (global analysis, $\chi^2 = 5.26$, df = 5, $P = 0.39$; T8 transmissions, $\chi^2 = 0.007$, df = 1, $P = 0.93$), or for estimates of haplotype transmission (global analysis, $\chi^2 = 4.41$, df = 5, $P = 0.49$; P4-T8 haplotype, $\chi^2 = 0.017$, df = 1, $P = 0.90$).

We also analyzed the relationship between microsatellite genotypes and testosterone levels. In the UK dataset, intermediate trait information was unavailable for the population controls, so these analyses were performed in all PCOS subjects (*i.e.* a combination of the cases and the trio probands) not receiving oral contraception. There was no association between genotype at either locus and testosterone levels (pentanucleotide, $P = 0.98$, Table 3; tetranucleotide, $P = 0.87$). In particular, given previous findings with regard to the P4 allele (13), we compared testosterone levels in subjects with one or more P4 allele [P4/−, equivalent to 216+ in previous papers (13, 15)] with all other subjects and found no appreciable difference [P4/−: geometric mean, 0.61 ng/ml (2.10 nmol/liter); SD range, 0.38–0.96 ng/ml (1.33–3.32 nmol/liter); non-P4: geometric mean, 0.55 ng/ml (1.92 nmol/liter); SD range, 0.37–0.82 ng/ml (1.29–2.85 nmol/liter); $P = 0.29$). Reanalysis after adjusting for subject age had no significant impact on these findings.

In the Finnish cohort, the distribution of pentanucleotide genotypes did not differ significantly between the symptomatic women and controls, whether genotypes were considered globally (exact $P = 0.075$) or grouped according to presence/absence of the P4 allele ($P = 0.55$). There were no differences between the groups in terms of tetranucleotide genotypes ($P = 0.18$), allele counts (pentanucleotide, $P = 0.27$; tetranucleotide, $P = 0.29$) or estimated haplotype frequencies ($P = 0.27$). These findings were unchanged when the case sample was limited to women with USS-confirmed PCO (data not shown). In neither the group of symptomatic women nor the cohort controls was any significant relationship seen between pentanucleotide (controls, $P = 0.86$; symptomatic women, $P = 0.89$) or tetranucleotide genotype (controls, $P = 0.56$; symptomatic women, $P = 0.27$) and testosterone levels (Table 3). Specifically, testosterone levels in P4/− individuals were no different from those in individuals with other genotypes [controls: P4/−, geometric mean, 0.53 ng/ml (1.83 nmol/liter); SD range, 0.35–0.81 ng/ml (1.20–2.79 nmol/liter) *vs.* non-P4, geometric mean, 0.55 ng/ml (1.89 nmol/liter); SD range, 0.38–0.80 ng/ml (1.30–2.77 nmol/liter), $P = 0.39$; symptomatic women: P4/−, geometric mean, 0.59 ng/ml (2.04 nmol/liter), SD range, 0.39–0.89 ng/ml (1.35–3.10 nmol/liter) *vs.* non-P4, geometric mean, 0.61 ng/ml (2.13 nmol/liter); SD range, 0.40–0.94 ng/ml (1.39–3.25 nmol/liter), $P = 0.49$].

Discussion

The main conclusion of our analysis is that we have failed to detect any consistent association between genomic variation at the two *CYP11A* promoter microsatellites and either PCOS status or serum testosterone despite analyses of well-sized data sets. It is worth pointing out that the nominally significant association that we found at the pentanucleotide

TABLE 2. Allele and genotype frequencies in the UK subjects (PCOS cases and population controls only)

Locus	Pentanucleotide	No. of alleles or genotypes (%)		Tetranucleotide	No. of alleles or genotypes	
		PCOS cases	Population controls		PCOS cases	Population controls
Allele counts	P4 (169 bp)[a]	261 (67.6)	391 (58.2)	T6 (276 bp)	1 (0.3)	0
	P6 (179 bp)	68 (17.6)	176 (26.2)	T8 (284 bp)	352 (88.4)	567 (82.4)
	P7 (184 bp)	0	2 (0.3)	T9 (288 bp)	0	9 (1.3)
	P8 (189 bp)	16 (4.2)	43 (6.4)	T10 (292 bp)	3 (0.8)	4 (0.6)
	P9 (194 bp)	40 (10.4)	59 (8.8)	T11 (296 bp)	3 (0.8)	12 (17.4)
	P10 (199 bp)	1 (0.3)	1 (0.2)	T12 (300 bp)	35 (8.8)	82 (11.9)
				T13 (304 bp)	4 (1.0)	14 (2.0)
Genotype counts	P4/P4 (169/169)	91 (47.2)	117 (34.8)	T8/T8 (284/284)	155 (77.9)	233 (67.7)
	P4/P6 (169/179)	47 (24.4)	97 (28.9)	T8/T9 (284/288)	0	6 (1.7)
	P4/P8 (169/189)	9 (4.7)	23 (6.8)	T8/T11 (284/296)	3 (1.5)	9 (2.6)
	P4/P9 (169/194)	23 (11.9)	34 (10.1)	T8/T12 (284/300)	33 (16.6)	72 (20.9)
	P6/P6 (179/179)	4 (2.1)	22 (6.5)	T8/T13 (284/304)	4 (2.0)	12 (3.5)
	P6/P8 (179/189)	4 (2.1)	16 (4.8)	Other genotypes[b]	4 (2.0)	12 (3.5)
	P6/P9 (179/194)	8 (4.1)	19 (5.7)			
	Other genotypes[b]	7 (3.6)	8 (2.4)			

[a] The P5 allele was not seen in the UK cases and controls.
[b] All genotypes less than 1% frequency are pooled for the purposes of this table; contingency table analyses were performed on unpooled counts.

Downloaded from jcem.endojournals.org on November 8, 2005

TABLE 3. Pentanucleotide genotypes *vs.* testosterone in the Finnish birth cohort and UK cases

Genotype	Finnish subjects				UK subjects	
	Cohort controls		Symptomatic women		PCOS subjects[a]	
	n[b]	Testosterone (ng/ml)[c]	n[b]	Testosterone (ng/ml)[c]	n[b]	Testosterone (ng/ml)[c]
P4/P4	426	0.52 (0.34–0.80)	195	0.60 (0.40–0.92)	112	0.59 (0.39–0.91)
P4/P6	299	0.54 (0.35–0.83)	150	0.57 (0.37–0.88)	60	0.63 (0.40–1.01)
P4/P8	64	0.53 (0.36–0.76)	50	0.59 (0.40–0.86)	13	0.60 (0.44–0.83)
P4/P9	38	0.52 (0.34–0.81)	13	0.59 (0.45–0.79)	30	0.61 (0.35–1.08)
P6/P6	48	0.53 (0.36–0.80)	29	0.61 (0.40–0.92)	8	0.59 (0.42–0.82)
P6/P8	39	0.54 (0.38–0.76)	16	0.66 (0.44–0.96)	6	0.59 (0.40–0.87)
The rest	30	0.59 (0.40–0.85)	14	0.59 (0.36–0.97)	16	0.53 (0.34–0.81)
Significance tests						
Global	944	$P = 0.86$	467	$P = 0.89$	245	$P = 0.98$
P4/– *vs.* rest	944	$P = 0.39$	467	$P = 0.49$	245	$P = 0.29$

Testosterone values are expressed as geometric mean (SD range).
[a] PCOS subjects includes both case and trio probands.
[b] These analyses only include women not receiving oral contraceptive therapy at the time of sampling.
[c] To convert to SI units (nanomoles per liter) multiply by 3.467.

in UK cases and control subjects in this study (see *Results*) lies in the opposite direction to the previous findings and was, in any event, not substantiated in the analyses of the UK trios or the Finnish subjects.

These findings are clearly in conflict with several previous studies (10, 14, 15) that have appeared to corroborate the initial findings (13) that variation at the *CYP11A* pentanucleotide is both linked to and (with specific reference to the absence of the four-repeat allele) associated with both symptomatic PCOS and raised testosterone levels.

We can discount several possible explanations for these discrepant findings. First, we are confident that our failure to detect association was not due to genotyping error; the low rate of discrepancies observed on cryptic duplicate genotyping and the consistent pattern of linkage disequilibrium relationships between the microsatellites across multiple datasets provide strong reassurance in this regard. Second, we can be confident, given the size of the resources typed, that the failure to detect associations with effect sizes similar to those previously reported was not due to lack of power. Third, the explanation is unlikely to lie in between-cohort differences in linkage disequilibrium relationships (for example, between the pentanucleotide and some as yet unidentified etiological variant); the close similarities in haplotype structure observed between the Finnish and UK data sets suggest that there are no profound differences in linkage disequilibrium structure, at least among populations of European origin (in whom the majority of the previous studies were conducted).

Our data suggest that the strength of, indeed the existence of, any association between *CYP11A* promoter variation and androgen-related phenotypes has been overestimated in previous studies through a combination of small sample size and data exploration conducted without due allowance for the associated inflation of the nominal type 1 error rate. This may have been compounded by the effects of any publication bias that has favored studies generating positive results. These factors are increasingly being recognized as contributing to the inconsistency that has plagued association studies, particularly in the context of multifactorial traits (17–19).

A further explanation for the discrepant findings may lie in the possibility of ethnic stratification (26). In this article, we have restricted the reported analyses to individuals of European origin. However, in PCOS cases of non-European origin, we have found marked differences in microsatellite allele frequencies. For example, in both Afro-Caribbean and South Asian subjects, the P6, rather than the P4, allele was the commonest allele at the pentanucleotide repeat, with frequencies of 46.6% and 48.6%, respectively. This suggests that studies of this locus may be particularly susceptible to spurious associations arising from the failure to match cases and control subjects adequately for ethnic composition.

Concerns over the rather poor performance of association studies have led to promulgation of recommendations designed to improve study design (17–19, 27, 28). Several of these have been incorporated in the current study, including: use of well-sized data sets; attempts to seek replication of any nominally significant findings, particularly those arising from data exploration; use of both case-control and family-based association methods; and analyses based on both intermediate phenotypes and discrete disease categories.

It is important to point out, of course, that the present study does not allow us to conclude that these variants have absolutely no effect on PCOS status or on testosterone levels. However, the size of the study means that any such effect must be extremely small and certainly much smaller than the effect size implied by the previous positive findings. We are currently undertaking functional studies of these microsatellites to determine whether they have any impact on *CYP11A* expression. Equally, the present study does not allow us to exclude the possibility that *CYP11A* variants other than those studied contribute to PCOS susceptibility and population variation in testosterone levels. Extensive resequencing and genotyping will be needed to test this hypothesis. However, as discussed above, it seems unlikely that the previously reported associations at the pentanucleotide repeat were a reflection of linkage disequilibrium with etiological variants elsewhere in the gene. Otherwise, given the apparent similarities in haplotype structure, at least as judged from the relationship between the microsatellites in these two quite distinct European populations, we should have expected to detect the same effects in this study.

The aim of this study was to test whether the previously reported associations with the *CYP11A* pentanucleotide held

Downloaded from jcem.endojournals.org on November 8, 2005

up to reanalysis in larger datasets. We conclude that they do not and that the major sites of genomic variation underlying susceptibility to PCOS and population variation in testosterone levels remain to be identified.

Acknowledgments

We acknowledge the many patients, relatives, nurses, and physicians who have contributed to the ascertainment of the various clinical samples used in this study.

Received September 18, 2003. Accepted February 1, 2004.

Address all correspondence and requests for reprints to: Prof. Mark McCarthy, Robert Turner Professor of Diabetes, Oxford Centre for Diabetes, Endocrinology & Metabolism, Churchill Hospital Site, Old Road, Headington, Oxford OX3 7LJ, United Kingdom. E-mail: Mark.mccarthy@drl.ox.ac.uk.

This work was supported by the United Kingdom Medical Research Council (Program Grant G9700120), the Academy of Finland, the European Commission (Framework 5 Award QLG1-CT-2000-01643), and the Wellcome Trust (Project Grant GR069224MA).

References

1. Adams J, Polson DW, Franks S 1986 Prevalence of polycystic ovaries in women with anovulation and idiopathic hirsutism. Br Med J 293:355–359
2. Hull MG 1987 Epidemiology of infertility and polycystic ovarian disease: endocrinological and demographic studies. Gynecol Endocrinol 1:235–245
3. Book CB, Dunaif A 1999 Selective insulin resistance in the polycystic ovary syndrome. J Clin Endocrinol Metab 84:3110–3116
4. Holte J 1996 Disturbances in insulin secretion and sensitivity in women with the polycystic ovary syndrome. Baillieres Clin Endocrinol Metab 10:221–247
5. Franks S 1989 Polycystic ovary syndrome: a changing perspective. Clin Endocrinol (Oxf) 31:87–120
6. Franks S, Gharani N, McCarthy M 2001 Candidate genes in polycystic ovary syndrome. Hum Reprod Update 7:405–410
7. Dunaif A, Thomas A 2001 Current concepts in the polycystic ovary syndrome. Annu Rev Med 52:401–419
8. Gilling-Smith C, Story H, Rogers V, Franks S 1997 Evidence for a primary abnormality of thecal cell steroidogenesis in the polycystic ovary syndrome. Clin Endocrinol (Oxf) 47:93–99
9. Jakimiuk AJ, Weitsman SR, Navab A, Magoffin DA 2001 Luteinizing hormone receptor, steroidogenesis acute regulatory protein, and steroidogenic enzyme messenger ribonucleic acids are overexpressed in thecal and granulosa cells from polycystic ovaries. J Clin Endocrinol Metab 86:1318–1323
10. Daneshmand S, Weitsman SR, Navab A, Jakimiuk AJ, Magoffin DA 2002 Overexpression of theca-cell messenger RNA in polycystic ovary syndrome does not correlate with polymorphisms in the cholesterol side-chain cleavage and 17α-hydroxylase/C(17–20) lyase promoters. Fertil Steril 77:274–280
11. Nelson VL, Legro RS, Strauss III JF, McAllister JM 1999 Augmented androgen production is a stable steroidogenic phenotype of propagated theca cells from polycystic ovaries. Mol Endocrinol 13:946–957
12. Wood JR, Nelson VL, Ho C, Jansen E, Wang CY, Urbanek M, McAllister JM, Mosselman S, Strauss JF 2003 The molecular phenotype of polycystic ovary syndrome (PCOS) theca cells and new candidate PCOS genes defined by microarray analysis. J Biol Chem 278:26380–26390
13. Gharani N, Waterworth DM, Batty S, White D, Gilling-Smith C, Conway GS, McCarthy M, Franks S, Williamson R 1997 Association of the steroid synthesis gene CYP11a with polycystic ovary syndrome and hyperandrogenism. Hum Mol Genet 6:397–402
14. Urbanek M, Legro RS, Driscoll DA, Azziz R, Ehrmann DA, Norman RJ, Strauss JF, Spielman RS, Dunaif A 1999 Thirty-seven candidate genes for polycystic ovary syndrome: strongest evidence for linkage is with follistatin. Proc Natl Acad Sci USA 96:8573–8578
15. Diamanti-Kandarakis E, Batzis MI, Bergiele AT, Tsianateli TC, Kouli CR 2000 Microsatellite polymorphism (tttta) at −528 base pairs of gene CYP11a influences hyperandrogenemia in patients with polycystic ovary syndrome. Fertil Steril 73:735–741
16. San Millan JL, Sancho J, Calvo RM, Escobar-Morreale HF 2001 Role of the pentanucleotide (tttta)(n) polymorphism in the promoter of the CYP11a gene in the pathogenesis of hirsutism. Fertil Steril 75:797–802
17. Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis JG 2001 Replication validity of genetic association. Nat Genet 29:306–309
18. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN 2003 Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177–182
19. Colhoun HM, McKeigue PM, Davey Smith G 2003 Problems of reporting genetic associations with complex outcomes. Lancet 361:865–872
20. Haddad L, Evans JC, Gharani N, Robertson C, Rush K, Wiltshire S, Frayling TM, Wilkin TJ, Demaine A, Millward A, Hattersley AT, Conway G, Cox NJ, Bell GI, Franks S, McCarthy MI 2002 Variation within the type 2 diabetes susceptibility gene calpain-10 (CAPN10) and polycystic ovary syndrome. J Clin Endocrinol Metab 87:2606–2610
21. The Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group 2004 Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS). Hum Reprod 19:41–47
22. Taponen S, Martikainen H, Jarvelin M-R, Laitinen J, Pouta A, Hartikainen A-L, Sovio U, McCarthy MI, Franks S, Ruokonen A 2003 The hormonal profile of women with self-reported symptoms of oligomenorrhea and/or hirsutism: Northern Finland Birth Cohort 1966 Study. J Clin Endocrinol Metab 88:141–147
23. Fahle GA, Fischer SH 2000 Comparison of six commercial DNA extraction kits for recovery of cytomegalovirus DNA from spiked human specimens. J Clin Microbiol 38:3860–3863
24. Abecasis GR, Cookson WO 2000 GOLD—graphical overview of linkage disequilibrium. Bioinformatics 16:182–183
25. Clayton D 1999 A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65:1170–1177
26. Spielman RS, Ewens WJ 1996 The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983–989
27. 1999 Freely associating. Nat Genet 22:1–2
28. Cardon LR, Bell JI 2001 Association study designs for complex disease. Nat Rev Genet 2:91–99

JCEM is published monthly by The Endocrine Society (http://www.endo-society.org), the foremost professional society serving the endocrine community.

Downloaded from jcem.endojournals.org on November 8, 2005