The characterisation of genes localised on chromosome 2p23.3

Patrick John Wightman

(University of Edinburgh)

Thesis presented to the University of Edinburgh for examination for the degree of Ph.D.

August 2000

Declaration

I declare that

a) this thesis has been composed by myself
b) that the work is my own, except where otherwise stated

Patrick John Wightman
August 2000

Acknowledgements

I would like to thank many people for the help, support and guidance I have received during the production of this thesis. Firstly, I would like to thank Professor David Bonthron for his interest, advice and patience that has made this project possible. I would also like to thank Professor David Brock, who gave me the opportunity of doing this PhD in the Human Genetics Unit (now called the Molecular Genetics Section) at the University of Edinburgh. Thanks also to Dr Bruce Hayward for his interest and advice during the course of this thesis. Many thanks also to my colleagues at the University of Leeds, especially Dr Jack Leek who did the fluorescence *in situ* hybridisation experiments and Dr Liz Telford who was involved in the cloning of *KIF3C*. I thank Dr Cathy Abbot for advice on mouse mapping using the Jackson backcross DNA, Dr Lucy Rowe at the Jackson Laboratory for analysing the mouse mapping data, and Dr John Warner for his involvement in cosmid library production. Thanks to the HGMP resource centre for supplying IMAGE clones, PAC pools and PAC filters. I would also like to thank Professor David Porteous and the "West Wingers", who recently joined the Molecular Genetics Section and who have made my stay most enjoyable.

A very big thank you to my flatmates present and past, especially Claire, Jo, and Fiona, whom have kept me sane throughout my PhD. I must also thank my friends and lab-mates, in particular Dawn, Katie, Nicola, Dors, Veronica, Hannah, and Irene, who have always offered encouragement and have made the last few years so much fun. Finally, the biggest thank you goes to my family, who have provided me with encouragement and support throughout my education without which I could not have completed this PhD.

Abstract

The chromosomal region 2p23.3 contains the candidate type 2 diabetes genes encoding glucokinase regulator protein (GKRP) and ketohexokinase (KHK) and is also the location of the non-syndromic recessive sensorineural deafness gene *DFNB9*. Both GKRP and KHK have previously been shown to be metabolically connected and with the co-localisation of *GCKR* and *KHK* to a 500 kb region of human chromosome 2p23.3, the possibility of co-ordinate regulation by common *cis*-acting regulatory elements has been raised. Several genes on 2p, such as *PPP1CB* and *KIF3C*, map to mouse chromosome 12, but mapping of both *GCKR* and *KHK* has shown them to co-localise to mouse Chr 5 (gene order *PPP1CB-KHK-GCKR-KIF3C*). A detailed investigation of transcripts within this genomic region was carried out with the aim of identifying and characterising other transcripts that may encode proteins involved in carbohydrate metabolism and provide evidence for the involvement of this genomic region in the pathogenesis of type 2 diabetes. In addition, the localisation of the DFNB9 gene to the same region, 2p22-23, provided additional interest in transcripts that were identified in this region and also expressed in the inner ear.

A 2 Mb physical contig that spanned both the *KHK-GCKR* genomic region and the DFNB9 interval was constructed using YACs, BACs, PACs and cosmids and assembled by a combination of STS content analysis and cosmid fingerprinting. This physical contig was used as the basis for transcript mapping by a combination of PCR screening of genomic clones for cDNA markers mapping to the 2p23 region and direct sequencing of genomic clones with computer analysis of sequences to search for similarity to ESTs. A total of 14 known genes and 15 ESTs were mapped to the physical contig. Several ESTs were chosen for further investigation based upon the involvement of their predicted encoded proteins in various biochemical pathways with potential roles in the pathogenesis of type 2 diabetes or deafness. Genes that were cloned included *eIF2B4*, the gene that encodes the delta subunit of the guanine nucleotide exchange factor eIF2B, which plays a key role in protein synthesis initiation and has been shown to be stimulated by both glucose and sugar phosphates. The genomic structure was characterised, two different isoforms identified and tissue specific splice forms identified. A gene called *KIAA0064* was also characterised due to its intimate location upstream to *eIF2B4*. An EST showing homology to ribokinase was investigated as it belonged to the same family of kinase proteins as KHK. Three genes were identified as candidate DFNB9 genes: *MPV17*, *KIF3C*, and *KCNK3*. The *KIF3C* gene was cloned, genomic structure characterised and the mouse homologue mapped.

Abbreviations

Abbreviations commonly used in this thesis are listed below. Additional abbreviations are explained where appropriate in the text.

µCi	microcuries
µg	microgram
µl	microlitre
µM	micromolar
A	adenosine residue in a DNA sequence
ATP	adenosine triphosphate
BAC	bacterial artificial chromosome
bp	base-pairs
BSA	bovine serum albumin
C	cytidine residue in a DNA sequence
°C	degrees centigrade
Ca^{2+}	calcium ion
cDNA	complementary deoxyribonucleic acid
CEPH	Centre d'Etude du Polymorphisme humain
Ci	Curie
CIP	calf intestinal phosphatase
cM	centiMorgan
dATP	2′-deoxyadenosine 5′-triphosphate
ddATP	dideoxyadenosine 5′-triphosphate
dCTP	2′-deoxycytidine 5′-triphosphate
ddCTP	dideoxycytidine 5′-triphosphate
ddH_2O	deionized and distilled water
dGTP	2′-deoxyguanosine 5′-triphosphate
ddGTP	dideoxyguanosine 5′-triphosphate
dH_2O	distilled water
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
ddNTP	dideoxynucleoside 5′-triphosphate
DFNB	recessive loci for prelingual non-syndromic hearing impairment
DTT	dithiothreitol

dTTP	2′-deoxythymidine 5′-triphosphate
ddTTP	dideoxythymidine 5′-triphosphate
E. coli	*Escherischia coli*
EDTA	ethylenediaminetetra-acetic acid disodium salt
eIF2	eukaryotic initiation factor 2 (substrate of eIF2B)
eIF2B	eukaryotic initiation factor 2B (guanine nucleotide exchange factor eIF2B)
eIF2B4	the gene encoding the delta subunit of eIF2B
EMBL	European Molecular Biology Laboratory
EST	expressed sequence tag
F1	first generation offspring
FISH	fluorescent *in situ* hybridisation
g	grams or relative centrifugal force
G	guanosine residue in a DNA sequence
GCK	glucokinase
GCKR	glucokinase regulator gene
GKRP	glucokinase regulatory protein
GTP	guanosine triphosphate
HGMP	human genome mapping project
ICRF	Imperial Cancer Research Fund
IDDM	insulin-dependent diabetes mellitus
IMAGE	Integrated Molecular Analysis of Genomes and their Expression Consortium
K^+	potassium ion
kb	kilobase pairs
K_m	Michaelis constant
KIF3C	kinesin-related protein, member of the KIF3 family
KHK	ketohexokinase
l	litres
LOD	logarithm of odds
kDa	kilo Dalton
M	molar
Mb	megabase pairs
mg	milligram
ml	millilitre
MODY	maturity onset diabetes of the young
mM	millimolar

mRNA	messenger ribonucleic acid
N2	backcross mice
NIDDM	non-insulin-dependent diabetes mellitus
NSHI	non-syndromic hearing impairment
nt	nucleotides
p	human chromosome short arm
PAC	P1 artificial chromosome
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PFGE	pulsed field gel electrophoresis
q	human chromosome long arm
RBSK	ribokinase
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
RNase	ribonuclease
rpm	revolutions per minute
RT-PCR	reverse transcriptase polymerase chain reaction
SDS	sodium dodecyl (lauryl) sulphate
SSC	saline sodium citrate
SSCP	single strand conformation polymorphism
STS	sequence tagged site
T	thymidine residue in a DNA sequence
T_A	annealing temperature
TEMED	NNN′N′ tetramethylethylenediamine
TIGR	The Institute of Genomic Research
T_M	melting temperature
tRNA	transfer ribonucleic acid
$tRNA^{Met}$	methionyl-tRNA
Tris	tris hydroxymethyl aminomethane
U	units
UTR	untranslated region
w/v	weight for volume
YAC	yeast artificial chromosome

List of Tables

List of Figures

Table of Contents

Chapter 1

1 Chromosome 2p23.3 – a region of interest for diabetes and deafness

1.1 Chromosome 2p23.3

The research described in this thesis examines in detail a genomic region of chromosome 2p23.3, within which are located the candidate type 2 diabetes genes encoding the glucokinase regulatory protein (GKRP)and fructokinase (KHK), and also a locus for non-syndromic recessive sensorineural deafness, *DFNB9*. The chromosomal region 2p23.3 first became of interest to us when the genes encoding GKRP and KHK were shown to co-localise to a 500 kb region of 2p23 (Warner *et al.*, 1995). GKRP and KHK had previously been shown to have a close metabolic relationship with each other, since fructose-1-phosphate is both the KHK end-product and an allosteric inhibitor of GKRP. Although the intimate location of their encoding genes *GCKR* and *KHK* could be a coincidence, the possibility of common *cis*-acting regulatory elements has been raised (Hayward *et al.*, 1996). An investigation of the *GCKR-KHK* genomic region might therefore also reveal other transcripts that encode proteins involved in carbohydrate metabolism and provide evidence for the involvement of this genomic region in the pathogenesis of type 2 diabetes. Finally, the localisation of a gene for non-syndromic recessive sensorineural deafness, *DFNB9*, to the same small genomic region in chromosome 2p22-p23 (Chaib *et al.*, 1996), provided additional interest in transcripts found in this region that are expressed in the inner ear.

1.2 Type 2 diabetes

1.2.1 Classification and prevalence

Non-insulin-dependent (type 2) diabetes mellitus (NIDDM) is characterised by hyperglycaemia due to defects in insulin secretion and/or action (Sacks & McDonald, 1996). It is a common disorder especially in developed countries where it affects 10-20% of the population older than 45 years of age (King & Rewers, 1993). In some populations such as the Pima Indians, NIDDM is achieving epidemic proportions (Bennett, 1999). Left untreated, NIDDM is a leading cause of death and morbidity. NIDDM differs from the much less common insulin-dependent (type 1) diabetes mellitus (IDDM) in genetic basis,

age of onset (mainly in adults than juveniles), cellular manifestations (peripheral insulin resistance rather than autoimmune destruction of pancreatic islet β-cells), and treatment (often by diet, not by insulin injections).

1.2.2 Genetic aetiology of type 2 diabetes

The genetic aetiology of type 2 diabetes was first investigated by examining type 2 diabetes in older identical (monozygotic) twins. This revealed concordance rates approaching 100% in monozygotic twins but much lower in dizygotic (non-identical) twins (with estimates of concordance ranging from 3% to 37%)(Barnett *et al.*, 1981a; Barnett *et al.*, 1981b). As twins usually share the same environment early on in life, concordance could be the result of genetic or environmental similarity, but the fact that they usually live apart in later life does suggest, for a later-onset disorder such as type 2 diabetes, that this is a genetic disease. Furthermore, a recent study investigating concordance rates for type 2 diabetes in monozygotic twin pairs, initially ascertained as discordant for diabetes, reveals a concordance rate of 76% 15 years of the initial type 2 diabetes in one member of the twin pair (Medici *et al.*, 1999). In addition, the concordance rate for any abnormality of glucose metabolism (either type 2 diabetes or impaired glucose tolerance) at 15 years' follow-up was 96%. However, in another study of type 2 diabetes in twin pairs, concordance was only 26% for type 2 diabetes but 61% for abnormal glucose tolerance (Poulsen *et al.*, 1999). Although different twin populations and experimental methods make it difficult to compare these studies, generally they agree that genetic predisposition is important for the development of abnormal glucose tolerance and that although environmental factors play an important part in the pathogenesis of type 2 diabetes, genetic factors also play a crucial role.

1.3 The search for type 2 diabetes loci

1.3.1 Linkage studies

Although genetic factors have been shown to play a crucial role in determining susceptibility to type 2 diabetes, the search for late onset NIDDM susceptibility loci has proven largely unsuccessful. This can mainly be attributed to the multifactorial nature of NIDDM pathogenesis in which this highly genetic heterogeneous disorder is also further modulated by environmental factors. Linkage studies have to overcome several problems, such as the fact that the mean age at diagnosis is approximately 50 years, making it difficult to identify complete nuclear families in which segregation events can be observed. Often one or both

parents are not available, due both to the high mortality rate of the disorder and to the fact that children of affected individuals may not yet have developed the condition. Without a genetic model that reflects the clear familial aggregation of the disorder, new methods such as genome-wide linkage screening for NIDDM susceptibility genes using affected siblings without parents have had to be employed.

So far, three loci have been identified in two different populations using such approaches (see Table 1.1). The NIDDM1 locus near to D2S125 was identified using non- and quasi-parametric linkage analysis in a genome-wide search for NIDDM genes in Mexican Americans (Hanis *et al.*, 1996). This data suggests that in this population studied, late onset NIDDM results from the action of at least one relatively major susceptibility gene. However, in two other populations (non-Hispanic whites and Japanese), no evidence for linkage of D2S125 with NIDDM was found. Similarily, in the Finnish population two other NIDDM loci, on chromosome 12 (Mahtani *et al.*, 1996) and chromosome 20 (Ghosh *et al.*, 1999) have been identified, but when other populations with late onset NIDDM have been analysed, no linkage was found to these chromosomal regions. These results suggest that different racial and ethnic populations may have different NIDDM susceptibility loci or perhaps susceptibility loci within different populations play stronger or weaker roles in the multistep process leading to the onset of NIDDM.

NIDDM locus	Gene	Chromosomal location	Population
NIDDM 1	unknown	2q37	Mexican Americans
NIDDM 2	*HNF1A*	12q24.2	Finnish
NIDDM3	unknown	20	Finnish

Table 1.1 NIDDM loci.

1.3.2 The candidate gene approach

Genome-wide screening of populations with NIDDM has identified only a limited number of candidate loci, and even these have not been shown to be of central importance in disease pathogenesis in different populations. Therefore, to identify other genes involved in the pathogenesis of type 2 diabetes, the candidate gene approach has been frequently adopted. For the successful identification of NIDDM candidate genes, an understanding of the biochemical pathways underlying both insulin secretion and action is required. One of the first successes was the detection of mutations in the insulin-receptor gene in patients with NIDDM (Orahilly *et al.*, 1991). However, mutations in the insulin-receptor gene account for

a very small proportion of NIDDM cases. The insulin-receptor gene was identified by its role in insulin action, but other candidate genes were identified by their role in insulin secretion.

1.4 Maturity onset diabetes of the young (MODY)

1.4.1 Identification of MODY genes

The main successes in identifying NIDDM susceptibility loci have arisen using linkage studies in families with maturity onset diabetes of the young (MODY), a monogenic form of NIDDM characterised by onset usually before 25 years of age and autosomal dominant inheritance. The genetic analysis of MODY families has identified 5 different genes involved in the pathogenesis of MODY (see Table 1.2). Mutational screening studies in MODY families has also shown that several do not contain mutations in any of these 5 genes, suggesting that at least one other MODY gene exists (Chevre *et al.*, 1998; Froguel & Velho, 1999).

	Gene name	Chromosomal location	Reference
MODY1	Hepatocyte nuclear factor 4 alpha (*HNF4A*)	20q12-13	(Yamagata *et al.*, 1996a)
MODY2	Glucokinase (*GCK* - see Section 1.5)	7p13-15	(Vionnet *et al.*, 1992)
MODY3	Hepatocyte nuclear factor 1 alpha (*HNF1A* – also known as *TCF1*)	12q24.2	(Yamagata *et al.*, 1996b)
MODY4	Insulin promoter factor 1 (*IPF1*)	13q12	(Stoffers *et al.*, 1997)
MODY5	Hepatic transcription factor 2 (*TCF2*)	17q11.2-q21.3	(Horikawa *et al.*, 1997)

Table 1.2 MODY genes and chromosomal location.

1.4.2 Function of MODY genes

Although the exact function of the proteins encoded by MODY genes is unknown, four out of the five MODY genes encode transcription factors (the exception being glucokinase – see Section 1.5). Clinical physiological studies indicate that mutations in these genes are associated with abnormal patterns of glucose-stimulated insulin secretion by pancreatic β-cells. HNF4A (encoded by *HNF4A*, the MODY1 gene) is a key regulator of hepatic gene expression and is a major activator of *HNF1A* (the MODY3 gene) expression. HNF1, in turn activates the expression of a large number of genes specifically in the liver, for example

albumin, alpha-1-antitrypsin, and pyruvate kinase (Courtois *et al.*, 1987). Both *HNF4A* and *HNF1A* are also expressed in other tissues including the pancreas. The interaction between *HNF4A* and *HNF1A* demonstrates the existence of a complex transcriptional regulatory mechanism for the expression of genes in both liver and the pancreas.

Studies investigating the cellular consequences of mutations in MODY genes will reveal important information concerning the pathogenesis of type 2 diabetes. The *HNF1A* rat homologue encodes a transcription factor that binds a sequence required for hepatocyte-specific transcription of several genes (Courtois *et al.*, 1987). It is also suggested, though, that mutations in human *HNF1A* might lead to diabetes by causing abnormal pancreatic islet development during fetal life or by impairing transcriptional regulation of genes that play a key role in normal β-cell function (Vaxillaire *et al.*, 1997). The gene underlying MODY5 encodes another hepatic transcription factor (TCF2). Mutations in either *TCF2* or *HNF1A* can cause MODY by affecting β-cell function and interestingly, deficiency of either TCF2 or HNF1A also affects renal function; individuals with *TCF2* mutations show susceptibility to severe non-diabetic renal disease (Horikawa et al., 1997; Nishigori *et al.*, 1998), while *HNF1A* mutations are associated with reduced renal thresholds for glucose (Menzel *et al.*, 1998).

The MODY4 gene (*IPF1*), is a homeodomain-containing transcription factor and is critically required for the embryonic development of the pancreas and for the transcriptional regulation of endocrine pancreas-specific genes in adults, such as insulin, glucose transporter-2 (GLUT2) and glucokinase in β-cells, and somatostatin in δ-cells (reviewed in Habener & Stoffers, 1998)). The finding that four out of five MODY genes encode transcription factors suggests that the proper transcriptional regulation of glucose homeostasis-specific genes is crucial for maintaining normal blood glucose levels. Understanding the complex transcriptional regulatory mechanisms in both liver and pancreas may reveal novel candidate type 2 diabetes genes.

1.4.3 The role of MODY genes in late onset type 2 diabetes

Interestingly, the MODY3 and MODY4 genes have now also been linked to late onset type 2 diabetes. The NIDDM2 and MODY3 loci co-localise, and NIDDM2 and MODY3 families are both characterised by a reduced insulin secretory response that subsequently progresses to diabetes. It is suggested that both forms of NIDDM are due to allelic mutations in the *HNF1A* gene (Lehto *et al.*, 1997). Mutations in the MODY4 gene, *IPF1*, have now also

been demonstrated in late onset type 2 diabetes (Hani *et al.*, 1999). *IPF1* mutations that result in a profound alteration of IPF1 function cause MODY 4 but less severe mutations cause a late-onset form of type 2 diabetes. This demonstrates that the identification of genes underlying MODY may also lead to the discovery of genes involved in late-onset type 2 diabetes.

1.5 Glucokinase (GCK)

1.5.1 *GCK*- a good candidate type 2 diabetes gene

The glucokinase (*GCK*) gene encodes the major enzyme that phosphorylates glucose on entry into the liver and pancreatic islet β-cells. The gene encoding GCK was considered a good candidate as a type 2 diabetes gene when it was discovered that GCK played an important role in determining the rate of glucose uptake and thereby acting as a key component in "glucose-sensing" in pancreatic islet β-cells, controlling insulin release.

1.5.2 Evidence for glucokinase as the "glucose-sensor"

Glucose metabolism is initiated by phosphorylation of the sugar to glucose-6-phosphate by hexokinases. In all cells except hepatocytes and insulin-secreting pancreatic cells (β-cells), glucose is phosphorylated by hexokinases I, II or III, all of which have a Michaelis constant (K_m) for glucose in the 0.01 to 0.1 mM range. Glucose metabolism rates depend on the concentration of glucose within the cell, and in all cells except hepatocytes and pancreatic β-cells, this is controlled by the activity of plasma membrane glucose transporters (GLUT) 1, 3, and 4 (Burant *et al.*, 1991).

However, investigation of glucose transport into hepatocytes revealed that these cells appear freely permeable to glucose (Cahill, 1958a; Cahill, 1958b) and that their rate of glucose phosphorylation is dependent on glucose concentration across the physiological range (4-9 mM). Further investigation revealed that glucose transport into hepatocytes and pancreatic β-cells is facilitated by glucose transporter 2 (GLUT2). Therefore, unlike the other glucose transporters, GLUT2 is non rate-limiting for glucose metabolism (Johnson *et al.*, 1990), it facilitates rapid equilibrium between extracellular and intracellular glucose so that the concentration of glucose in hepatocytes and pancreatic β-cells is within the physiological range (4-9 mM).

In all tissues except liver and pancreas, hexokinases I, II, and III phosphorylate glucose for entry into the glycolytic pathway. However, in hepatocytes and pancreatic β-cells, a different hexokinase was identified with different kinetic properties to that of hexokinases I, II, and III (DiPietro, 1962). This hexokinase was named glucokinase (hexokinase IV or GCK) and its properties included a much lower affinity for glucose than the other hexokinases, a sigmoid glucose concentration-reaction velocity relationship, and no inhibition by the reaction product glucose-6-phosphate (Vinuela, 1963).

Investigation of insulin secretion by β-cells shows that only sugars which are substrates for hexokinases and for glycolysis elicit insulin release, suggesting a link between sugar metabolism and insulin release (German, 1993). It was also found that the insulin output of cultured β-cells is regulated by glucose according to a sigmoidal curve (Liang *et al.*, 1992), with a threshold for insulin production around the physiological glucose concentration of 4mM (Figure 1.1). The resemblance of this sigmoidal curve to the glucose concentration-reaction velocity curve of GCK suggested that GCK might play a key role in this homeostatic process of sensing glucose and modulation of insulin release (Matschinsky, 1990; Randle, 1993). The functional characteristics of GCK allow it to increase glucose phosphorylation in response to hyperglycaemia and it is this variable flux into glycolysis that is thought to determine insulin output. Thus, in effect, the extracellular glucose concentrations are "sensed" by glucokinase.

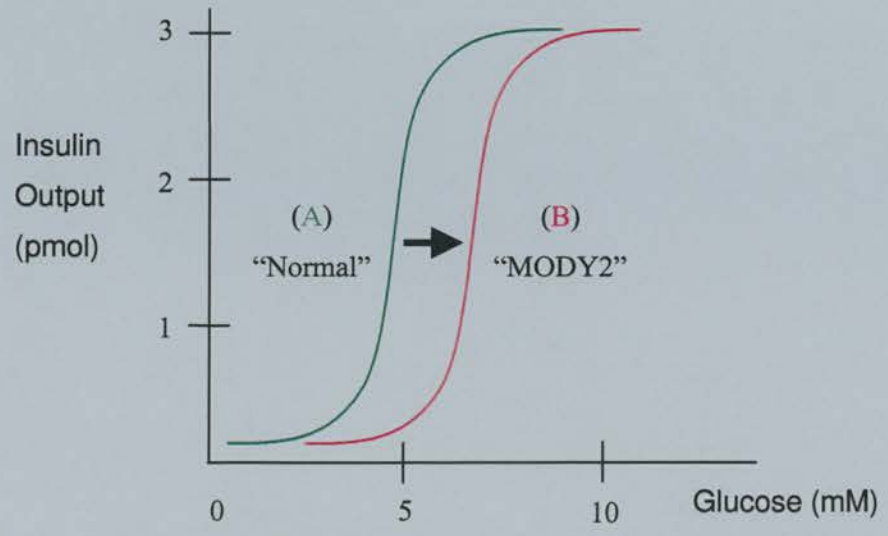


Figure 1.1 Sigmoidal relationship between glucose concentration and insulin output in the pancreatic β-cell. (Chen *et al.*, 1994). The threshold for insulin production is around physiological glucose concentration of 4mM. A) "normal" glucose homeostasis (green line); B) postulated effect of mutations in *GCK* that "reset glucose sensing"(red line), causing insufficient insulin release in response to high levels of glucose, as seen in mild type 2 diabetes (MODY2).

The downstream mechanism between "glucose sensing" by GCK and insulin release involves a complex series of steps that presently are still poorly defined. Until recently, the level of ATP in the pancreatic β-cell, which affects cell membrane potential through ATP-dependent K^+ channels, was thought to be the main mechanism for insulin release. Briefly, enhanced glucose metabolism in the β-cell leads to an increased ATP/ADP concentration ratio and closure of ATP-dependent K^+ channels. This causes membrane depolarisation and opening of voltage-dependant Ca^{2+} channels. The influx of β-cell Ca^{2+} triggers the release of insulin into the blood stream through exocytosis of insulin secretory granules in the β-cell (Ashcroft & Gribble, 1999). This is an example of a glucose-induced ionic event.

It is now thought that "non-ionic" glucose actions in the β-cell also play an important role in insulin secretion. This hypothesis is based upon evidence showing β-cell insulin secretion that is independent of potassium ion channel function (Aizawa *et al.*, 1994) and insulin release even during Ca^{2+} deprivation (Macfarlane *et al.*, 2000). It should also be noted that in pancreatic β-cells, while glucose is the major insulin secretagogue, there are a host of other physiological regulators that stimulate insulin release from pancreatic β-cells. These include neurotransmitters, neuropeptides, circulating hormones, and amino acids (Dunne, 2000). Whereas glucose must be metabolised in the β-cell for insulin secretion, neuromodulators influence the insulin secretory process following their interaction with specific cell-surface receptors.

It is clear that the sensing mechanisms for insulin secretagogues and corresponding signal transduction pathways required for insulin release are highly complex. Much research is now being focused on "non-ionic" glucose action in the β-cell and also how neuromodulators influence the insulin secretory process. The elucidation and understanding of signal transduction pathways involved in insulin release in the β-cell will almost certainly identify novel type 2 diabetes candidate genes.

1.5.3 Mutations in *GCK* can cause MODY

Linkage of MODY to the *GCK* locus in chromosome 7 (MODY2) and subsequent demonstration of mutations in *GCK* (Froguel *et al.*, 1992), confirmed that in this case, the candidate gene approach had been successful. The cDNA encoding GCK was first cloned from liver (Andreone *et al.*, 1989) and later an islet-specific glucokinase was isolated from a rat insulinoma cDNA library. This, differs from the liver cDNA at its 5′ end (Magnuson &

Shelton, 1989). Analysis of the *GCK* genomic structure reveals that both isoforms originate from the same gene. Both are, encoded by 10 exons, and differ in their first exons only, the liver and islet isoforms both sharing exons 2-10. In addition, a further, less abundant liver splice form exists, that is produced as result of inclusion of a "cassette" exon between exons 1 and 2 (reviewed in Iynedjian, 1993). The mutations in *GCK* that cause MODY2 have been shown to reduce glucokinase activity (to approximately 50% of normal, since only one allele is mutated) (Gidhjain *et al.*, 1993). This reduction in GCK activity then appears to "reset the glucose homeostat", resulting in the threshold of circulating glucose levels which induce insulin secretion to be raised (Figure 1.1).

1.5.4 Regulation of GCK activity

1.5.4.1 Effectors of GCK activity

As mutations in *GCK* underlie MODY2, genes that encode modulators of GCK activity such as regulatory proteins, could also be considered as good candidate genes for type 2 diabetes. Possible mechanisms that could alter the rate of glucose phosphorylation by GCK would be changes in cellular content of GCK and/or GCK catalytic activity (Chen et al., 1994). Studies of cellular GCK protein levels reveal increased levels of GCK only after a long time period of sustained high glucose concentration (Chen et al., 1994). However, in studies of possible GCK allosteric effectors, GCK activity was shown to increase fivefold in cultured rat islets under high glucose concentrations despite little change in mRNA levels. This suggests that regulation of glucokinase is mostly post-transcriptional (Liang et al., 1992).

The metabolite fructose has also been shown to increase uptake and phosphorylation of glucose and its conversion to glycogen, lactate, CO_2 and amino acids (reviewed in Vanschaftingen & Vandercammen, 1989). Further investigation into the effect of fructose revealed that hepatic glucokinase is inhibited by fructose-6-phosphate (F-6-P), that this inhibition is relieved by fructose-1-phosphate (F-1-P) and that these effects of fructose phosphates require a specific regulatory protein called glucokinase regulatory protein (GKRP) – see Figure 1.2 (Van Schaftingen, 1989).

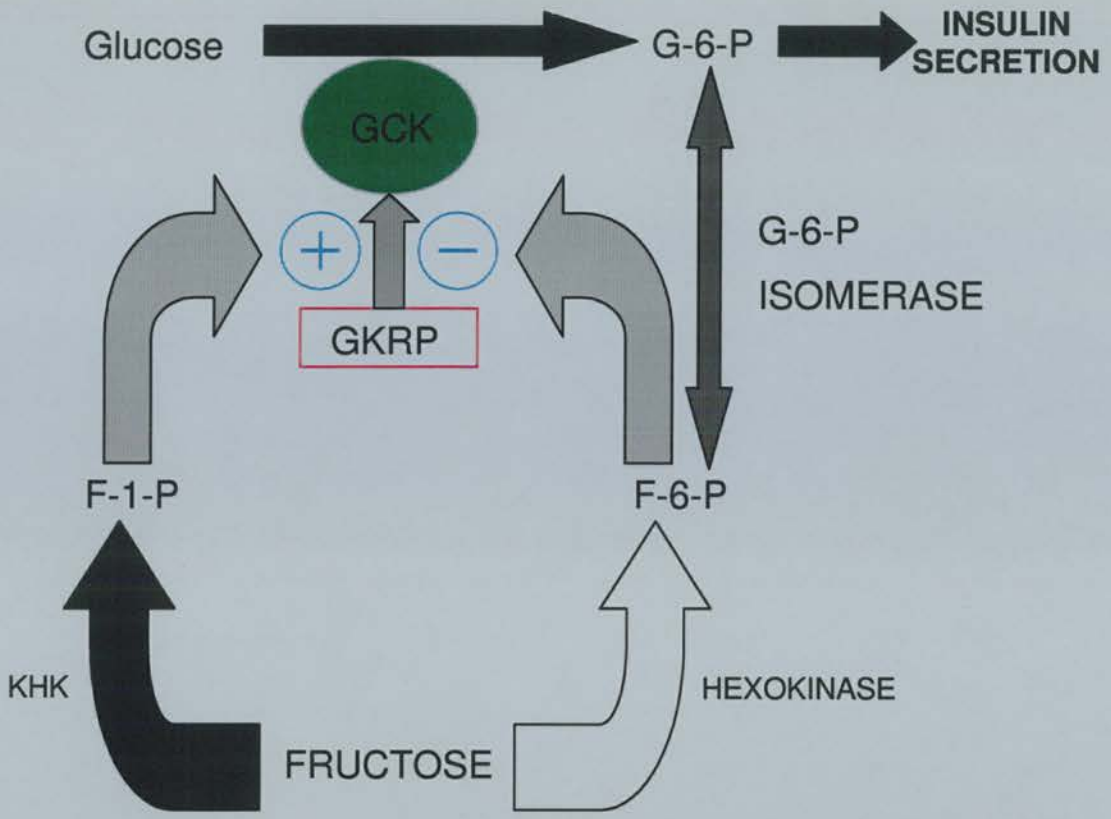


Figure 1.2 The effect of fructose phosphates on glucokinase regulatory protein (GKRP) inhibition of glucokinase (GCK). Abbreviations: KHK, Ketohexokinase; F-1-P, fructose-1-phosphate; F-6-P, fructose-6-phosphate; G-6-P, glucose-6-phosphate. In the presence of F-6-P, GKRP interacts non-covalently with GCK causing its inhibition. F-1-P (the product of fructose phosphorylation by KHK), relieves inhibition of GCK by GKRP. F-6-P concentrations can passively reflect those of G-6-P via G-6-P isomerase. As insulin secretion by the pancreatic β-cell depends on the rate of glycolytic flux, effectors of GCK activity such as the action of fructose phosphates through GKRP may play an important role in the pathogenesis of type 2 diabetes.

1.5.4.2 Glucokinase regulatory protein (GKRP)

The investigation of the 65 kDa glucokinase regulatory protein (GKRP) in rat liver reveals that in the presence of F-6-P, GKRP interacts non-covalently with GCK causing its inhibition (Van Schaftingen, 1989). However, F-1-P prevents the association of GCK and GKRP, thus antagonising the inhibition exerted by F-6-P and upregulating GCK activity (Malaisse *et al.*, 1990). The fructose effect (Figure 1.2) has also been shown to exist in the pancreatic islet which also contains both GKRP and KHK (Malaisse et al., 1990). The F-1-P/F-6-P ratio is probably mainly alterable through variation in F-1-P concentration, because F-6-P will passively follow G-6-P (via G-6-P isomerase), suggesting F-6-P as a mediator of indirect end-product inhibition of GCK (Randle, 1993). This important regulatory role of GKRP and fructose phosphates on GCK was confirmed by the restoration

of fructose phosphate responsiveness of purified GCK by recombinant rat GKRP (Detheux & Vanschaftingen, 1994). Further investigation of the inhibitory function of GKRP on GCK has also subsequently revealed the importance of subcellular localisation of both GKRP and GCK in controlling the activity of this system.

1.5.4.3 Subcellular localisation of GCK and GKRP

Studies into the mechanism by which glucose stimulates GCK activity reveal the importance of the subcellular location of GCK and the essential role that GKRP plays in controlling this subcellular localisation of GCK. One study shows that in livers of fasted rats, GCK is found predominantly in the hepatocyte nuclei bound to GKRP, forming an inactive complex. However in livers of re-fed rats, GCK is translocated to the cytoplasm where it is unbound and active (Fernandez-Novell *et al.*, 1999). A diagrammatic model of GCK nuclear-cytoplasmic translocation and the role of GKRP is shown in Figure 1.3. Apart from high glucose, both fructose-1-phosphate and sorbitol can also stimulate translocation of GCK from nuclei to cytoplasm (Mukhtar *et al.*, 1999). The important role of GKRP in GCK subcellular localisation is shown in studies using mutant GCK that are unable to bind GKRP. In this case, GCK is unable to accumulate in the hepatocyte nuclei even at low glucose concentration (de la Iglesia *et al.*, 1999).

Although all these studies show that GKRP sequesters GCK in the hepatocyte nuclei, there are conflicting reports concerning the subcellular localisation of GKRP during levels of high glucose. This conflict may be due to the different techniques used to observe subcellular localisation GKRP during high and low levels of glucose. Research using specific antibodies against GKRP and laser confocal fluorescence microscopy suggest no translocation of GKRP from nuclei to the cytoplasm at any level of cellular glucose (Brown *et al.*, 1997). However, other confocal microscopic studies using quantitative imaging show that GKRP is translocated with GCK from the nuclei during high levels of glucose (Mukhtar et al., 1999; Toyoda *et al.*, 1995). The latter study indicated that GKRP translocation out of the hepatocyte nuclei with GCK results in only a fractional decrease in total nuclei GKRP. As this results in only small changes in nuclear/cytoplasmic ratios of GKRP compared to that for GCK, changes in GKRP can only be seen using quantitative imaging. This suggests that in addition to the role in nuclear retention of GCK, GKRP may also be involved in nuclear export or import of glucokinase (Mukhtar et al., 1999). The involvement of GKRP in the subcellular localisation of GCK, coupled to the sigmoidal kinetics of GCK, confers a markedly extended responsiveness and sensitivity to changes in glucose concentration in the

hepatocyte. The full implications of this regulatory mechanism for both glycogen synthesis in the liver and insulin release by the pancreatic β-cell have still to be understood.

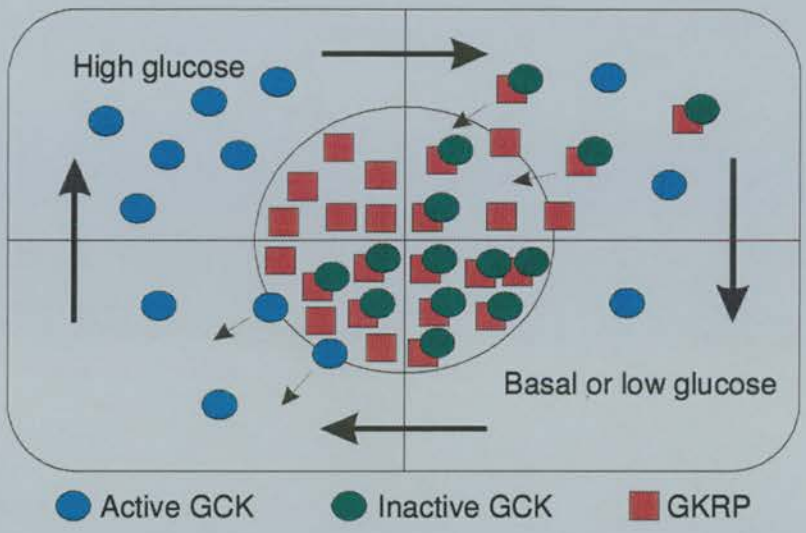


Figure 1.3 Model showing a nuclear-cytoplasmic translocation cycle for hepatic glucokinase (GCK), and the role that the glucokinase regulatory protein (GKRP) plays in this cycle. During low levels of cellular glucose, GCK is bound in an inactive complex with GKRP in the hepatocyte nuclei. High levels of glucose stimulate GCK translocation from the hepatocyte nuclei to the cytoplasm where it is unbound and active. There are conflicting reports as to whether GKRP is or is not translocated with GCK from the hepatocyte nuclei. This diagram is adapted from (Shiota *et al.*, 1999).

1.5.4.4 Mice mutant for glucokinase regulatory protein

Further evidence for an important role for GKRP in glucose homeostasis has recently been obtained by the creation of a mouse knockout of *Gckr*, the gene encoding gkrp (Farrelly *et al.*, 1999; Grimsby *et al.*, 2000). Before the creation of this mouse knockout, it was thought that gain of function mutations in *GCKR* would most likely lead to enhanced glucokinase inhibition and the development of MODY phenotype. Loss of function mutations in *GCKR* would be predicted to result in higher glucokinase activity. However, the mouse *Gckr* knockout revealed that loss of Gkrp causes a secondary loss of glucokinase protein and activity in mutant mouse liver. This loss is shown to be primarily because of post-transcriptional regulation of glucokinase, indicating a positive role for GKRP in maintaining glucokinase levels and activity.

As discussed previously (see Section 1.5.4.3), GKRP plays an important role in the subcellular localisation of GCK by sequestering GCK in hepatocyte nuclei until high levels of glucose stimulate GCK translocation to the cytoplasm where it is unbound and active (see Figure 1.3). In the hepatocytes of *Gckr* -/- mice, Gck is not found in the nucleus under any conditions. This result is similar to that seen in experiments using mutant Gck that cannot bind Gkrp, showing that mutant Gck is unable to accumulate in hepatocyte nuclei (de la Iglesia et al., 1999). Therefore, the *Gckr* -/- mice confirm the function of Gkrp as an anchor to sequester and inhibit glucokinase in the hepatocyte nucleus.

In the *Gckr*-mutant mice, the disruption of the GKRP sequestration mechanism on GCK in hepatocyte nuclei results in a subsequent decrease in GCK activity (Farrelly et al., 1999; Grimsby et al., 2000). Interestingly, although glucose and insulin levels in *Gckr* -/- mice (both fasting and fed) were relatively unchanged compared to *Gckr* +/+ mice, total liver glycogen concentration was decreased (33% lower in -/- mice compared to +/+ mice). This suggests impaired glucose metabolism and is consistent with impaired liver glycogen synthesis observed in MODY2 patients with defective GCK activity (Velho *et al.*, 1996). Other physiological changes observed in the *Gckr* -/- mice were elevated levels of the gluconeogenic enzyme phosphoenolpyruvate carboxykinase (expression negatively regulated by insulin) and higher glucose levels after a glucose challenge. These additional effects suggest that the *Gckr* -/- mice develop resistance to insulin.

To summarise (see Figure 1.3), during the fasting state, GKRP maintains a reserve pool of GCK in the hepatocyte nucleus. After feeding, the GKRP/GCK association is disrupted, and GCK is rapidly released and mobilised into the cytoplasm to provide phosphorylation activity. After glucose levels lower, GCK moves back into the nucleus where it is sequestered until required. This process ensures that glucose phosphorylation is minimal when the liver is in the fasting, glucose-producing phase. It also enables hepatocytes to rapidly mobilise glucokinase into the cytoplasm to phosphorylate and store or metabolise glucose after the ingestion of dietary glucose. Before this regulatory process was discovered, the level of hepatic glucose output was thought to be regulated only by the relative rates of glucose phosphorylation and dephosphorylation by GCK and glucose-6-phosphatase, respectively. The additional level of regulation by GKRP on GCK allows the rapid response to glucose or fructose in hepatocytes resulting in a rapid on/off switch for glucose sensing by GCK in the liver.

Although the exact mechanism by which GKRP regulates GCK activity has still to be fully characterised, the creation of the *Gckr* -/- mouse reveals that GKRP plays a pivotal role in determining GCK localisation in hepatocyte nuclei and controlling GCK translocation to the cytoplasm in response to high levels of glucose and fructose. Interestingly, the mouse *Gckr* knockout also shows subtle biochemical changes such as decreased total liver glycogen concentration. This is consistent with impaired liver glycogen synthesis found in individuals with MODY2 and confirms *GCKR* as an excellent candidate type 2 diabetes gene.

1.6 Characterisation of *GCKR*

To investigate a possible role of GKRP in the pathogenesis of type 2 diabetes, the human gene encoding GKRP (called *GCKR*) was cloned so that a search for diabetogenic mutations in *GCKR* could be performed. The full length human *GCKR* cDNA (2194 bp) was cloned (Warner et al., 1995) by first designing degenerate oligonucleotide pools from regions of amino acid sequence conserved between rat and *Xenopus* GKRP for use in PCR screening a hepatoblastoma cDNA library (HepG2). Two HepG2 PCR products of expected size were amplified showing close homology to rat *GCKR* and these were used to re-screen the HepG2 library by hybridisation. One full length cDNA clone was sequenced and this revealed an open reading frame encoding 625 residues that is 88% identical to the rat GKRP (627 residues) but only 59% identical to the fructose phosphate-insensitive *Xenopus* GKRP (619 residues). To obtain genomic clones, a primer pair was designed that amplified a 258 bp PCR product that included a 187bp intron at position 558-559 of the cDNA sequence (Genbank accession number Z48475) and this was used to PCR screen the ICI YAC library (Anand *et al.*, 1990). Two non-chimeric YACs, 26BA11 and 29IH8 were obtained (sizes 300 kb and 500 kb respectively) and these were each used to perform chromosomal localisation by *in situ* hybridisation. Both *GCKR*-containing YACs localised to chromosome 2p23 (Warner et al., 1995) confirming a previous localisation to 2p22.3-p23 (Vaxillaire *et al.*, 1994).

The analysis of the *GCKR* genomic structure (see Chapter 2, Figure 2.2) reveals the gene to consist of 19 exons and 18 introns, spanning 27 kb, and investigation of the *GCKR* mRNA in the liver and pancreatic islet reveals no tissue specific alternative splice forms (Hayward & Bonthron, 1998). Initial screening of *GCKR* by SSCP analysis revealed a common polymorphism in the Scottish population within exon 15, which alters residue 446 from proline, conserved in rat and *Xenopus*, to leucine. Identification of other polymorphisms and

introduction of these mutations into recombinant human GKRP may indicate whether they possess any pathophysiological significance. Further investigation of the possible role of *GCKR* in the pathogenesis of type 2 diabetes by amplification of individual exons and mutational analysis by SSCP (Orita *et al.*, 1989) or sequencing has still to be carried out.

1.7 Fructokinase (KHK)

1.7.1 Characterisation of *KHK*

The stimulation of glucose phosphorylation in the liver by fructose can be explained by the presence of fructokinase (KHK). This enzyme catalyses the first step of metabolism of dietary fructose by phosphorylating fructose to F-1-P (Figure 1.2). The fructose effect has also been shown to exist in the pancreatic islets, where in mouse and rat islets in the presence of glucose (5.5 mM), a concentration of 20 μM fructose doubles the rate of insulin release (Ashcroft *et al.*, 1972).

The phosphorylation of fructose to F-1-P by KHK and the role of F-1-P/F-6-P ratio in the GKRP regulation of GCK activity in both the liver and pancreatic islet, suggests that KHK could play a role in the pathogenesis of type 2 diabetes. To investigate this possible role, the gene encoding KHK was characterised (Bonthron *et al.*, 1994). The *KHK* cDNA was isolated by screening a HepG2 cDNA library by low-stringency hybridisation with the entire rat *KHK* coding region. The longest cDNA clone (*pHKHK3a*; Genbank accession number X78677) contains an open reading frame encoding 298 residues but two other shorter cDNA clones were also identified (*pHKHK3d* and *pHKHK1-2*). These two shorter clones were derived from mRNA that had utilised an upstream polyadenylation site resulting in a shorter 3'UTR. It was also noticed that *pHKHK1-2* is an alternative splice form that differs from the other two clones by containing an alternative exon (Genbank accession number X78678). Inspection of the two alternative exons reveals that they show enough similarity to suggest that they may have arisen from an intragenic duplication event within *KHK*.

To characterise the *KHK* genomic structure, a P1 clone (J0788) was isolated from the ICRF P1 reference library 700 (Francis *et al.*, 1994). *KHK* was shown to consist of 9 exons spanning 14 kb (see Chapter 2, Figure 2.4), with two alternative splice forms arising from a pair of homologous exons (3a and 3c). The functional significance of these two different *KHK* splice forms is unknown but there is some evidence that they may perform different functions. The two *KHK* isoforms are evolutionary conserved between human, mouse and

rat, suggesting distinct conserved functions, and the alternative splicing has been shown to be tissue specific (Hayward & Bonthron, 1998). In both human and rat, the 3c isoform is exclusively expressed in tissues expressing high levels of KHK (liver, kidney, and duodenum), while other tissues use only the 3a isoform. There is also a developmental splicing shift from the 3a isoform to the 3c form when comparing foetal and adult tissues. As KHK is thought to act as a dimer, the tissue specificity of KHK isoforms suggests that the dimer consists of A-A or C-C isoforms but not a mixture of both.

1.7.2 Molecular basis of essential fructosuria

An inherited deficiency of KHK was first shown to cause the benign metabolic disorder essential fructosuria using a biochemical assay for fructokinase (Schapira, 1961-1962). To genetically confirm this, the gene encoding KHK was screened for mutations in a well characterised family, in which three out of eight siblings were known to have essential fructosuria (Bonthron et al., 1994). The finding of *KHK* mRNA in lymphoblastoid cells allowed the analysis by RT-PCR of RNA from a lymphoblastoid cell line from one essential fructosuria patient. Cloning and sequencing of RT-PCR products (the coding region was amplified in two overlapping segments), revealed several single base substitutions. Using a PCR/restriction digest assay, it was shown that all the affected individuals were compound heterozygotes for two mutations Gly40Arg and Ala43Thr. An additional conservative amino acid change (Val49Ile) was present on the *KHK* allele bearing Ala43Thr but this is found to be a common polymorphism in normal Europeans. Both Gly40Arg and Ala43Thr lie in a conserved region of the protein (in both alternative splice forms), produce non-conservative amino acid changes, and were not present in >100 control alleles. Therefore it was concluded that these amino acid changes were responsible for the fructosuric phenotype.

1.8 Co-localisation of *GCKR* and *KHK*

The isolation from the ICRF P1 reference library 700 of a P1 clone (J0788) that contains the whole of *KHK*, and then the isolation of YACs (26BA11 and 29IH8) and a PAC (J16101) which contain *GCKR* (Warner et al., 1995), allowed refinement of the genomic localisations of *GCKR* and *KHK* by fluorescent *in situ* hybridisation (FISH). This showed that both genes map to chromosome 2p23.2-23.3 and furthermore, two-colour interphase FISH suggested that the two genes lie within 500 kb of each other (Hayward et al., 1996). This is an intriguing finding and although this physical proximity could be a coincidence, it is noteworthy because of the intimate metabolic links between the two genes' products.

Although the clustering of non-homologous genes by function is rare in vertebrates, exceptions do exist such as the immunoglubulin recombination activating genes *RAG1* and *RAG2* (Oettinger *et al.*, 1992). It is also conceivable that the intergenic region between *GCKR* and *KHK* could contain regulatory elements that are common to both genes.

1.9 Other known candidate type 2 diabetes genes at chromosome 2p23.3

Another candidate type 2 diabetes gene, the protein serine/threonine phosphatase 1 beta subunit (*PPP1CB*), also maps to chromosome 2p23 (Saadat *et al.*, 1994). Insulin resistance is associated with decreased rates of insulin-mediated glycogen synthesis in skeletal muscle (DeFronzo *et al.*, 1992). Type 1 protein phosphatase (PP1), which activates insulin stimulation of glycogen synthase, is decreased in insulin-resistant subjects (Kida *et al.*, 1990). Because the PP1 catalytic β subunit is part of the PP1 major isoform in the glycogen bound PP1 complex, its gene (*PPP1CB*) is also a good candidate for type 2 diabetes (Prochazka *et al.*, 1995).

In a study involving 19 candidate genes (including *GCKR* and *PPP1CB*), whose products are implicated in insulin secretion or action, no evidence for linkage of these candidate genes to NIDDM was found by non-parametric methods in affected sib pairs from the French population (Vionnet *et al.*, 1997). This suggests that in these French families, none of the genes investigated are major contributors to the pathogenesis of NIDDM. However, these negative linkage results do not exclude the possibility that mutations in these genes may play smaller roles in the polygenic background of NIDDM or a major role in other populations.

1.10 The DFNB9 interval

It became apparent during the course of this work that the *GCKR-KHK* interval on chromosome 2p23 also coincided very closely with the location of a gene for non-syndromic recessive sensorineaural deafness (DFNB9) (Chaib et al., 1996). Consideration was therefore given during our characterisation of novel genes in this region, to whether any of these genes would be good candidate *DFNB9* genes. The identification of candidate deafness genes was based primarily on the location of the gene within the DFNB9 interval and putative function of the encoded function.

1.11 Aim of this investigation

This thesis describes a detailed examination and characterisation of a genomic region of 2p23.3, the location of the candidate type 2 diabetes genes *GCKR* and *KHK*, and the gene for non-syndromic recessive sensorineural deafness, *DFNB9*. Chapter 2 describes the mapping of *Gckr* and *Khk* in the mouse genome that reveals information about the genetic linkage of these two genes through evolution. The creation of a detailed transcript map of the *GCKR-KHK* intergenic region and DFNB9 interval is described in Chapter 3. This was aided by the construction of a YAC, BAC, PAC and cosmid physical contig spanning 2 Mb across this genomic region. Chapter 4 describes the investigation and characterisation of transcripts with possible roles in biochemical pathways relating to carbohydrate metabolism. The search, identification and characterisation of candidate *DFNB9* genes on chromosome 2p23.3 is described in Chapter 5. Chapter 6 describes future work that could be carried out to further investigate the function of the genes identified by the research described in this thesis. Materials and methods are described in Chapter 7.

Chapter 2

2 Mapping of the genes encoding glucokinase regulatory protein and ketohexokinase in the mouse

2.1 Introduction

2.1.1 Co-localisation of the glucokinase regulatory protein and ketohexokinase to human chromosome 2p23

The mapping of *GCKR* and *KHK* genes by fluorescent *in situ* hybridisation using the *KHK* P1 clone J0788 and the *GCKR* P1 clone J16101 reveals that both genes co-localise to human chromosome 2p23.2-23.3 (Hayward *et al.*, 1996). The intimate metabolic connection between GKRP and KHK, added to the close proximity of the *GCKR* and *KHK* genes may suggest a genetic regulatory factor (see Chapter 1). Mapping of the *Gckr* and *Khk* genes in the mouse would provide an interesting insight into their genetic relationship and may add further circumstantial support to the possibility of their co-ordinate regulation. Also, if these two genes were found to reside in a region of conserved synteny between human and mouse chromosomes, this may facilitate the identification of other human transcripts located in chromosome 2p23.2-23.3 that have mouse orthologues near *Gckr* and *Khk*.

2.1.2 Comparative genomics

The divergence of the various mammalian lineages approximately 70 million years ago led to chromosomal rearrangements resulting in alteration of the original order of genes and, in some cases, the exchange of segments among chromosomes (Ohno, 1973). However, the number of rearrangements has been sufficiently few to allow the comparison of mammalian genome organisation and detailed mapping has enabled the creation of comparative maps. Although these comparative maps provide an interesting insight of genome organisation and evolution, it is their use in disease gene identification, for example by the accurate cross referencing of model organism genes with mapped mammalian phenotypes that can help facilitate the identification of genes mutated in human disease states via the positional candidate approach. By identifying

homologous human disease genes in other organisms such as mouse, puffer fish, fruit fly, nematode, yeast, and bacteria, the investigators can take advantage of the experimental systems available for each organism permitting the rapid elucidation of molecular mechanisms involved in the human disease process. The existence of cross-reference databases like XREFdb (http://www.ncbi.nlm.nih.gov/XREFdb/), designed to establish cross-references between model organism genes and mammalian phenotypes will also increase the rate at which these cross-species connections will be established (Bassett *et al.*, 1997).

While studies in organisms such as fruit fly and nematodes have advantages such as ease of handling and genetic manipulation coupled with a short life cycle and large numbers of progeny, for the study of human disease phenotypes the mouse has advantages because it is evolutionarily closer to humans and as both organisms are mammalian, the pathogenic mechanisms of diseases may be similar in both organisms. Therefore, mutations in orthologues of human genes in the mouse is more likely to produce phenotypes similar to that seen in the mutation of the human gene. The mouse is also relatively cheap to maintain, easy to handle, and has large numbers of progeny.

The mouse is a primary model organism for the Human Genome Project and considerable emphasis has been placed on the genetic and physical mapping of the mouse genome world-wide. Both the mouse and human genomes have now been characterised sufficiently to identify all the large chromosomal regions showing conserved synteny (DeBry & Seldin, 1996). However, as the number of markers mapped in both the human and mouse genomes increases, many smaller regions of conserved synteny may yet be identified. The identification of mouse genetic loci which may be homologous to human mutations that cause genetic disease often provides powerful animal models for the disease process. In addition, the ability in the mouse to target new mutations to genes that may be involved in the human genetic disease has opened up vital avenues for the exploration of the mammalian gene function. Many mouse models of human disease phenotypes using gene "knock-outs" and "knock-ins"* have now been produced

*Mouse gene "knock-out" models involve the investigation of gene function by disruption of a gene and the examination of the resultant mouse phenotype, for example by gene replacement or insertion of a selectable marker. Mouse gene "knock-in" experiments involve trying to reverse a mouse disease phenotype by the insertion of a functional gene sequence (often to correct a gene disrupted during a "knock-out" experiment).

and this can reveal much information involving the pathogenic mechanisms of genetic diseases which could not be revealed by the study of the disease in either human patients or other model organisms. More information on mouse models can be found at the "Mouse knockout mutation database" web-site (http://www.biomednet.com/db/mkmd).

2.1.3 Genetic mapping in the mouse

The first genetic linkages in the mouse were established by analysing the genetic linkage relationship between loci that could be analysed by visually scoring the segregation of alleles based on pigmentation, morphological or behavioural phenotypes, for example the pink-eye dilution (*p*) and albino locus (*c*) (Haldane, 1915). With the progress of research into biochemistry and molecular biology, many more markers have now become available for genetic mapping and these include protein and enzyme polymorphisms, immunological responses, antigenic determinants, endogenous retroviruses and DNA sequences (reviewed in Rowe *et al.*, 1994). It is the recent rapid expansion of known DNA sequences that has allowed many new loci to be mapped in the mouse. Molecular biology techniques such as cDNA cloning, isolation of mini- and microsatellite markers, and anonymous DNA sequences derived from genomic and chromosome specific libraries now provide a substantial pool of useful markers for the creation of genetic maps.

2.1.4 Recombinant inbred (RI) strains

The production of homozygous inbred strains by the systematic inbreeding of the progeny of a cross of two progenitor founder strains (Bailey, 1971; Taylor, 1978) has been of major importance in the genetic mapping of the mouse genome. Once genetically homogeneous lines are produced, each line is typed with a variety of biochemical and molecular markers to produce haplotype information for each line. New markers that contain a DNA sequence polymorphism between the 2 progenitor founder strains are typed in the RI lines as either one type or the other. By comparing haplotype information for a new marker with existing haplotype information, the marker in question can be placed on the genetic map. Although RI strains provide an unlimited DNA resource, and all haplotype information is additive, this method of mapping can be limited by the lack of sequence divergence between the 2 progenitor strains used to produce the RI strains.

2.1.5 Interspecific backcrosses

The difficulty encountered in identifying allelic differences among laboratory mouse strains has been largely overcome by using interspecific backcrosses, which exploit the genetic diversity inherent among wild mouse species (Avner *et al.*, 1988). The crosses involve 2 mouse species whose evolutionary distance has allowed accumulation of differences at the DNA sequence level and yet interbreed under laboratory conditions. An example of a breeding scheme used to generate an interspecific backcross between the laboratory inbred strain C57BL/6J and *Mus spretus* (Copeland & Jenkins, 1991) is shown in Figure 2.1.

Although this cross produces fertile female animals and can be used to establish the backcross generation, the F1 males are sterile. This means that only recombinational data from the female F1 mice can be obtained. To overcome the problem of no male genetic map for this interspecific cross, other wild mouse species that are more closely related to laboratory strains and produce fertile F1 hybrids of both sexes are used, for example *Mus musculus castaneus* or *molossinus*. One disadvantage of interspecific crosses is the limited amount of DNA that can be obtained from each backcross animal. However, the increasing use of PCR to type the backcross panel instead of methods such as Southern blotting, should allow an unlimited number of loci to be mapped from each backcross panel. Immortal cell lines from the backcross animals is another method to ensure enough DNA is available for typing but as these lines can be unstable, other methods are preferable. As the number of backcross animals and number of loci typed increase, the mapping of markers and determination of gene order will become more accurate. Although combining mapping data from different interspecific backcrosses cannot be done directly, mapping of a common set of anchor loci on each panel allows the mapping data to be combined with respect to the anchor loci.

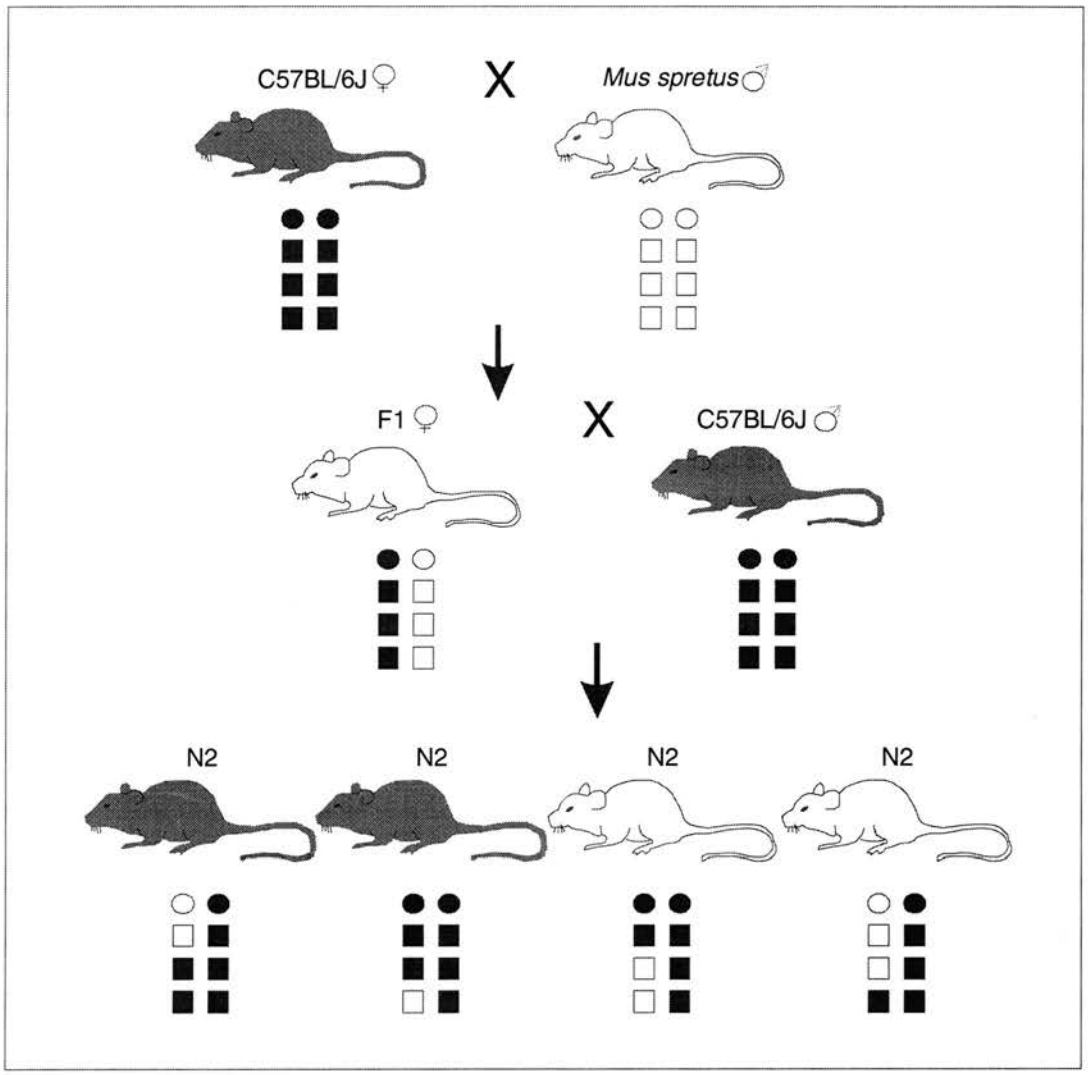


Figure 2.1 Breeding scheme used to generate an interspecific mouse backcross mapping panel (taken from Copeland and Jenkins, 1991) NB. Only one pair of chromosomes is shown for each parent. Only recombination events occurring in the F1 parent can be scored in the N2 backcross mice.

2.1.6 Mapping using the Jackson interspecific backcross DNA panel

The Jackson laboratory established two mouse interspecific backcross DNA panels to be used as a resource for the generation of a high-resolution (<1 Centimorgan) genetic map of the mouse genome (Rowe *et al.*, 1994). The two backcross DNA panels consist of (C57BL/6JEi x SPRET/Ei)F_1 x C57BL/6JEi and the reciprocal backcross (C57BL/6JEi x SPRET/Ei)F_1 x

SPRET/Ei, each containing 94 animals. Initial characterisation of the genetic maps was carried out by mapping MIT simple sequence length polymorpism (SSLP) markers (Dietrich *et al.*, 1992), proviral loci (Stoye & Coffin, 1988), and several other sequence-defined genes (Ko *et al.*, 1994) so that the Jackson map could be anchored to other published maps such as the European Collaborative Interspecific Backcross (EUCIB) map (Breen, 1994). Many restriction fragment length polymorphisms (RFLPs) between inbred mouse strains have been typed using molecular probes and Southern blot analysis. However, the use of PCR as a mapping tool has accelerated the assignment of new markers to the mouse genetic map.

Various types of loci can be mapped by PCR. Oligonucleotide primers can be designed within variant sequence from cDNA sequences, or from sequence flanking VNTRs (variable numbers of tandemly repeated mini-satellites) –including polymorphic dinucleotide repeats; (Weber & May, 1989) (Love *et al.*, 1990). To produce a high density of markers located on the Jackson mouse genetic maps, both arbitrarily-primed PCR (AP-PCR) and motif-primed PCR (MP-PCR) have been employed to "fingerprint" the panel DNAs. In AP-PCR , PCR primers are designed from arbitrary nucleotide sequence irrespective of coding potential (Serikawa *et al.*, 1992). In MP-PCR, PCR primers are derived from conserved promoter elements and protein motifs (Birkenmeier *et al.*, 1992). One advantage of using MP-PCR over AP-PCR is that the loci mapped relate to actual gene sequences.

2.1.7 Other mapping methods

Although *Mus domesticus* and *Mus spretus* have a high degree of DNA sequence difference, some regions of the genome may not contain polymorphisms useful for genetic mapping. In such cases, mapping can be carried out using radiation hybrid mapping, for example the T31 Mouse Radiation Hybrid Panel of 100 cell lines supplied from Research Genetics (McCarthy *et al.*, 1997). Radiation hybrid mapping does not require sequence difference between mouse strains, but only differences between mouse and hamster (the host cell line). One draw back may occur if mapping coding sequences, as the evolutionary closeness of mouse and hamster may result in the co-migration of hamster and mouse PCR fragments in a given assay.

2.1.8 Aims

This chapter describes the mapping of the mouse *Gckr* and *Khk* genes using the Jackson Laboratory interspecific backcross mapping panel ((C57BL/6JEi x SPRET/Ei) x SPRET/Ei). To map *Gckr* and *Khk*, I identified novel sequence variants between the two mouse species C57BL/6JEi and SPRET/Ei for both genes. These sequence variants were used to design PCR/restriction digest assays that could distinguish between the presence or absence of the C57BL/6JEi alelle and SPRET/Ei allele for *Gckr* and *Khk* in each interspecific backross animal. The haplotype information obtained for *Gckr* and *Khk* was compared to the haplotypes of known chromosomal markers and this used to place each gene onto the Jackson (C57BL/6JEi x SPRET/Ei) x SPRET/Ei interspecific backcross map.

2.2 Methods

2.2.1 Method of mapping

The Jackson Laboratory Backcross DNA panel Mapping Resource which consists of 94 backcross animals from the interspecific cross (C57BL/6JEi x SPRET/Ei) x SPRET/Ei, was typed using a PCR/restriction digest assay to reveal the variants.

2.2.2 Mapping of *Gckr*

2.2.2.1 Molecular reagents

Comparison of the mouse and human *GCKR* cDNA sequences (Genbank accession numbers X68497 and Z48475) reveals a similarity of 83.9%. It was also noticed that that there are regions of exact identity corresponding to sequences that have previously been used to design primers for the cloning of the human *GCKR* cDNA (Hayward *et al.*, 1997). Therefore, primer pairs could be chosen that would be predicted to amplify *GCKR* PCR products from both human and mouse DNA. Two primers were chosen (Gre2f (5′-dGATATTCCAGGAGGAGGGGCA-3′) and Gre7r (5′-dGCTCACTGGATTGAAGCCAACC-3′)) and used in a long range hot start PCR program: 5 min 94°; 2 x (94°C,20 s; 63°C,30 s; 68°C, 4 min); 8 x (94°C,20 s; 65°C,30 s; 68°C, 4 min); 20 x (94°C,20 s; 65°C,30 s; 68°C, 4 min + 10 s per cycle). The long range PCR was carried out in long range buffer #1 (Materials and Methods, Chapter 6), a final concentration of 350 μM dNTPs, thin wall PCR tubes and a final reaction volume of 30 μl.

Gre2f and Gre7r were found to amplify a ~4 kb PCR product corresponding to *GCKR* exons 2-8 from both C57BL/6Ei and SPRET/Ei mouse strains (predicted size from human *GCKR* genomic structure: ~3.7 kb – see Figure 2.2). The PCR product ends were sequenced using the Amersham Thermosequenase cycle sequencing kit.

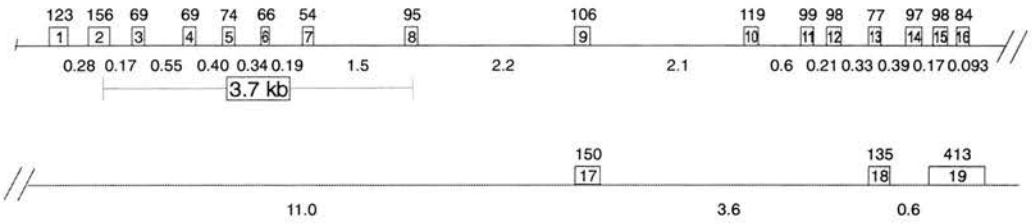


Figure 2.2 Genomic structure of human *GCKR*. Boxes 1-19 represent exons. Exon sizes (bp) are indicated above the exon boxes and intron sizes (kb) are also shown. The 3.7 kb region between exons 2-8 corresponding to the region amplified in mouse *Gckr* using primer Gre2f and Gre7r is indicated. This figure was adapted from (Hayward *et al.*, 1998).

Although no sequence variants were found in *GCKR* IVS2 between C57BL/6Ei and SPRET/Ei, the C57BL/6Ei and SPRET/Ei sequences were found to diverge 145 bp upstream of the IVS7-exon 8 splice junction, where a poly(A) sequence (on the sense strand) was found in C57BL/6Ei but the start of a B1 repetitive element-like sequence was present in SPRET/Ei (Figure 2.3). It was also noticed that the C57BL/6JEi PCR product was overall 200-300 bp larger than that of SPRET/Ei. A single nucleotide variant 55 bp upstream of the IVS7-exon 8 splice junction was shown to alter a *Xcm*I cutting site (present in SPRET/Ei but absent in C57BL/6JEi – see Figure 2.3). To type this polymorphism, a new primer GreSB (5′-dCTTGGTTGAGGAATCTATTTCTAG-3′) within intron 7 was used with the exon 8 primer Gre7r to generate a PCR product from both C57BL/6JEi and SPRET/Ei. Standard buffers (1.5 mM $MgCl_2$) and a hot start PCR program: 5 min 94°C; 30 x (94°C,45 s; 55°C,45 s; 72°C, 1 min) were used.

2.2.2.2 *Gckr* allele detection

The primers GreSB and Gre7R were used to amplify a ~200 bp *Gckr* fragment that, if derived from a SPRET/Ei allele, cuts with *Xcm*I to yield fragments of size ~150 bp and ~50 bp. Restriction digest products were visualised by electrophoresis through a 3.5 % agarose gel.

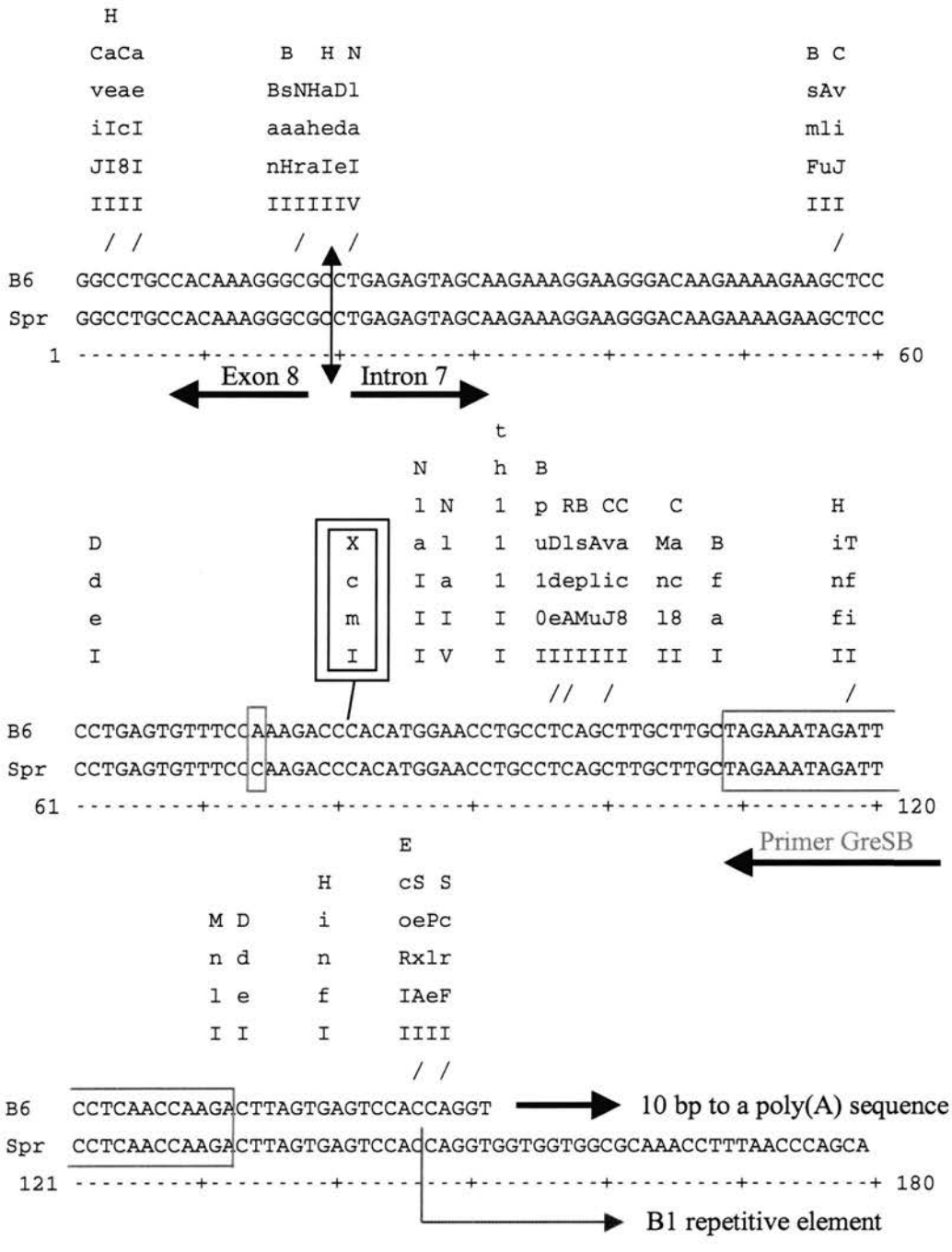


Figure 2.3 Comparison of C57BL/6JEi (B6) and SPRET/Ei (Spr) *GCKR* intron 7 sequences (anti-sense strand). Restriction enzyme sites are indicated above the sequences; a C/A variant is found at nucleotide 74, creating a *Xcm*I site in *Mus spretus*.

2.2.3 Mapping of *Khk*

2.2.3.1 Molecular reagents

As only partial mouse *Khk* cDNA sequence was available at the time of this study, the human and rat *KHK* cDNA sequences (Bonthron *et al.*, 1994) (Donaldson *et al.*, 1993) were compared to identify conserved regions of sequence which may also be conserved in the mouse. This revealed some highly conserved regions (the overall similarity between human and rat *KHK* was 87%) from which the primer KhkM4 (5′-dTGAGGGGCTTGTACAGTCGTCGAG-3′) was designed. Another primer KhkR9 (5′-dCCACCTGGCACCCGAATCTC-3′), that was designed from the partial mouse *Khk* cDNA sequence, was used with KhkM4 to PCR amplify a ~400 bp genomic fragment corresponding to exons 6-8 (predicted size from human *KHK* genomic structure was 451 bp – see Figure 2.4). Standard PCR reaction conditions were used with a hot start PCR program: 5 min 94°; 30 x (94°C,45 s; 64°C,45 s; 72°C, 1 min).

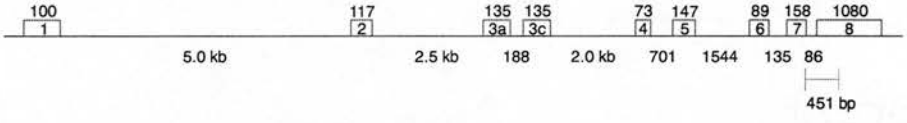


Figure 2.4 Genomic structure of human *KHK*. Boxes 1-8 represent exons. Exon sizes (bp) are indicated above the exon boxes and intron sizes are also shown (intron sizes are bp unless indicated as kb). Corresponding 451 bp region amplified in mouse *Khk* using primers KhkM4 and KhkR9 is indicated.

2.2.3.2 *Khk* allele detection

As the amplified mouse *Khk* PCR product was small (~400 bp), various restriction enzymes were used on the C57BL/6JEi and SPRET/Ei PCR products, to look for differences in cutting sites between the two mouse strains which could subsequently be used to reveal the variants and type the interspecific backcross. The restriction enzymes tried were:- *Eco*RI, *Hin*dIII, *Mbo*I, *Sac*I, *Xba*I. Out of the restriction enzymes tried, only *Mbo*I was found to reveal a variant sequence between the C57BL/6JEi and SPRET/Ei mouse strains. It was found that *Mbo*I restriction enzyme digestion of the ~400 bp genomic fragment amplified using the primers KhkM4 and KhkR9 produces a fragment of 300 bp from the C57BL/6JEi *Khk* allele and 220 bp from SPRET/Ei. PCR products were visualised by electrophoresis through a 2 % agarose gel.

2.3 Results

2.3.1 Allele typing

The *Gckr* and *Khk* alleles were typed as described in the methods section and the products from the PCR/restriction digest assay visualised by electrophoresis through agarose gel (Figures 2.5 and 2.6).

2.3.2 Map position of *Gckr* and *Khk*

Analysis of the typing data for both *Gckr* and *Khk* reveals complete concordance between genotypes, with both genes mapping to the proximal part of mouse chromosome 5 and co-segregating with a number of other loci including *D5Mit149*. The relative positions of *Gckr* and *Khk* compared to other genes and markers located on the same region of the mouse chromosome 5 genetic map were estimated by statistical analysis of the typing data (summarised in Figure 2.7). This places *Gckr* and *Khk* at the following genetic location (marker-map distance between marker and next closest marker (Centimorgans) ± standard error(Centimorgans): *D5Mit1*-3.19 ± 1.81-*D5Bir5*-1.06 ± 1.06-*Nos3/Dpp6/Fgl2*-1.06 ± 1.06-*Htr5a/Gbx1/En2/Nkx1-1/Plk-ps/D5Mit149/D5Xrf391/D5Bir6/Znt3/Gckr/Khk*-1.06 ± 1.06-*D5Mit351/D5Xrf47*-1.06 ± 1.06-*Crmp1/D5Bir7/Msx1*-2.13 ± 1.49-*Bapx1*. *D5Mit* and *D5Bir* are anonymous mouse chromosome 5 markers, *D5Xrf47* is a marker designed from an anonymous cDNA clone, all the other markers are genes that co-localise with or map near to *Gckr* and *Khk*. The typing data was deposited under accession number MGD-JNUM-37599 at http://www.jax.org/resources/documents/cmdata.

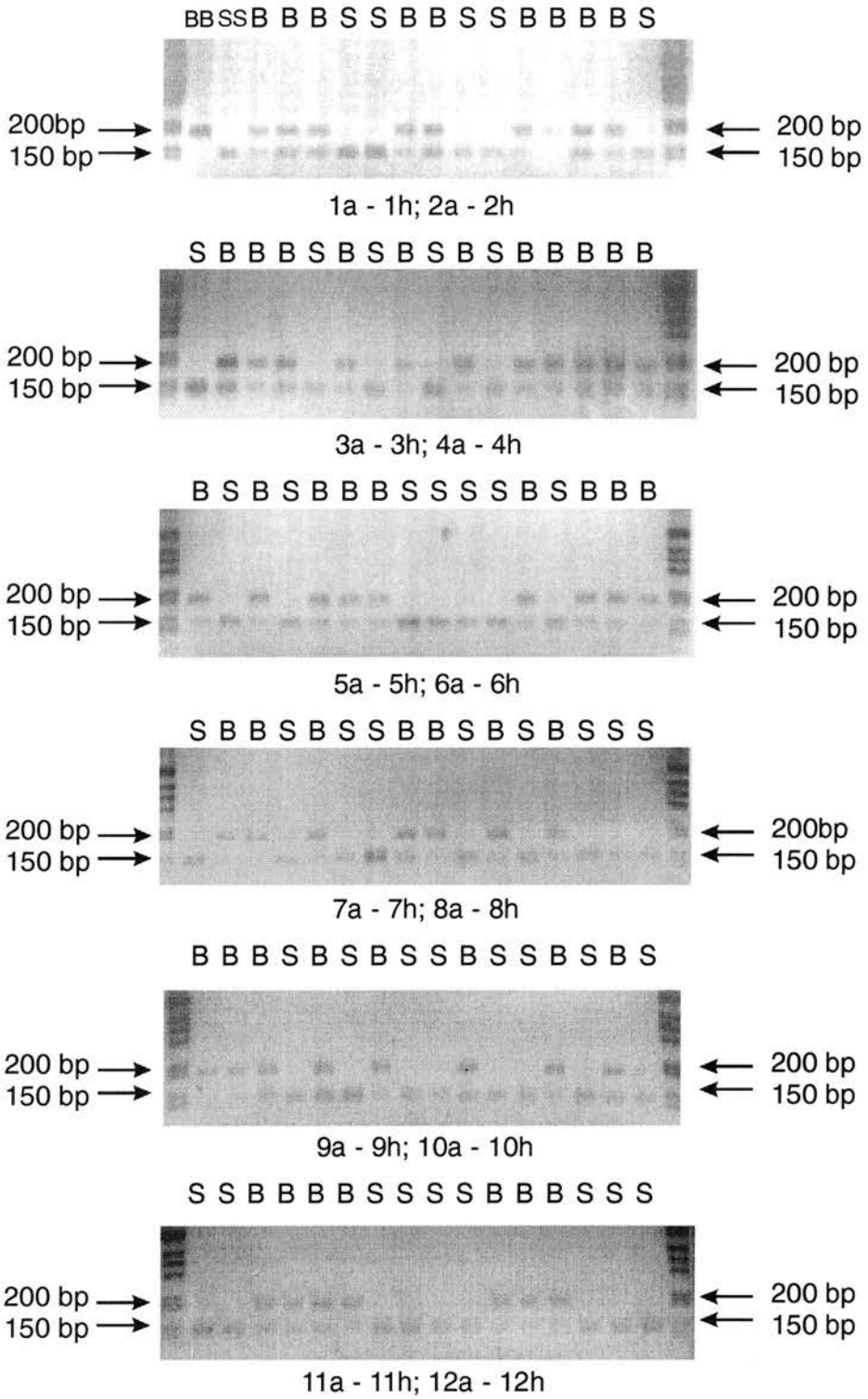


Figure 2.5 *Gckr* allele typing. The typing of each interspecific backcross animal is indicated above each gel image (B:heterozygote C57BL/6JEi / SPRET/Ei type; S: homozygous SPRET/Ei type). Lane 1A contains parental C57BL/6JEi homozygous alleles (BB) and lane 1B contains parental SPRET/Ei homozygous alleles (SS). The size of each allele is shown (C57BL/6JEi:200 bp and SPRET/Ei: 150 bp). The end lanes of each gel contain 1 kb ladder (fragment sizes (bp) shown are 506, 396, 344, 298, 220, 201, 154, and 134).

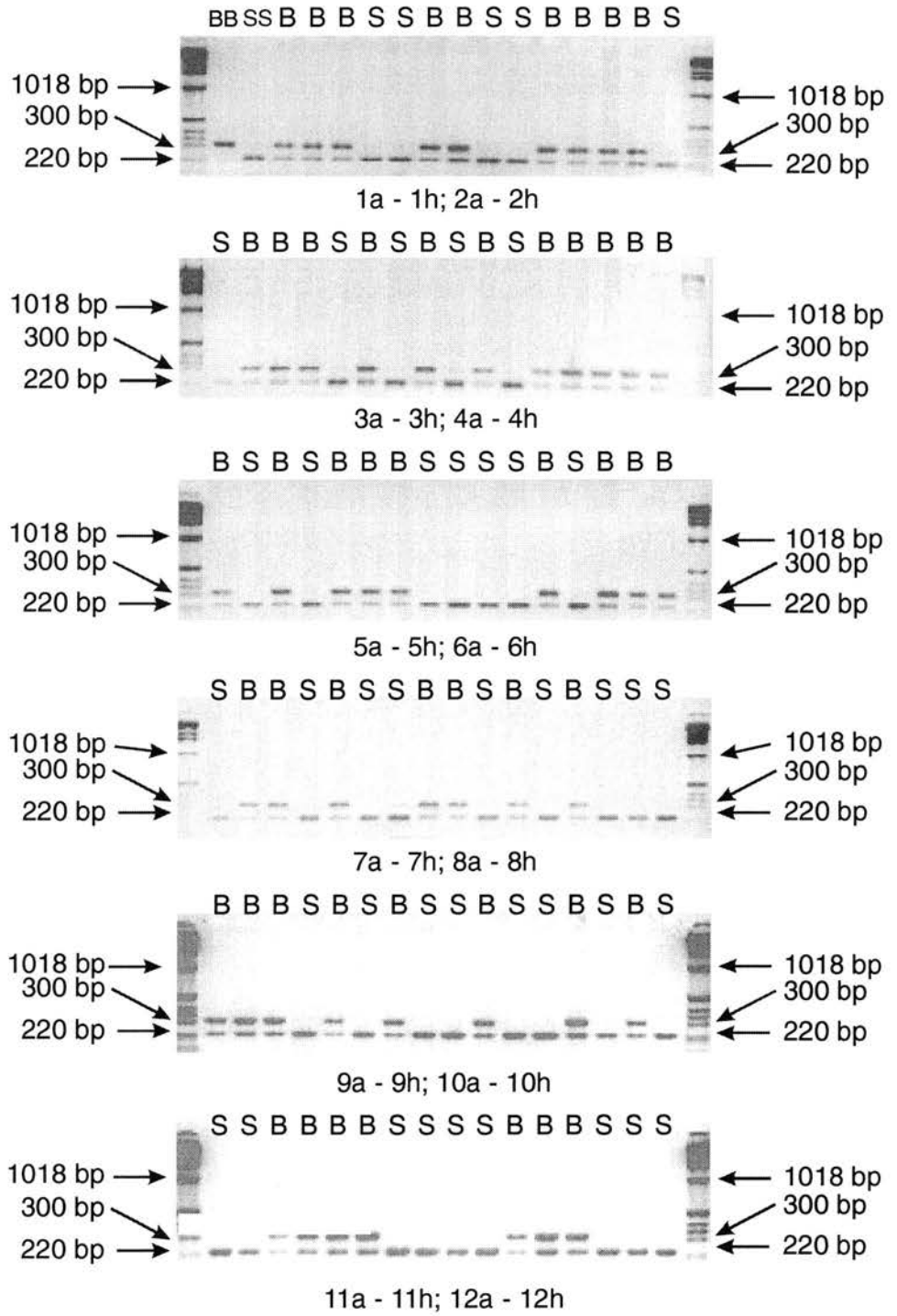


Figure 2.6 *Khk* allele typing. The typing of each interspecific backcross animal is indicated above each gel image (B:heterozygote C57BL/6JEi / SPRET/Ei type; S: homozygous SPRET/Ei type. Lane 1A contains parental C57BL/6JEi homozygous alleles (BB) and lane 1B contains parental SPRET/Ei homozygous alleles (SS). The size of each allele is shown (C57BL/6JEi: 300 bp and SPRET/Ei: 220 bp). The end lanes of each gel contain 1 kb ladder (fragment sizes (bp) shown are 1018, 506, 396, 344, 298, 220, 201, 154, and 134). The 1018 bp DNA fragment is indicated.

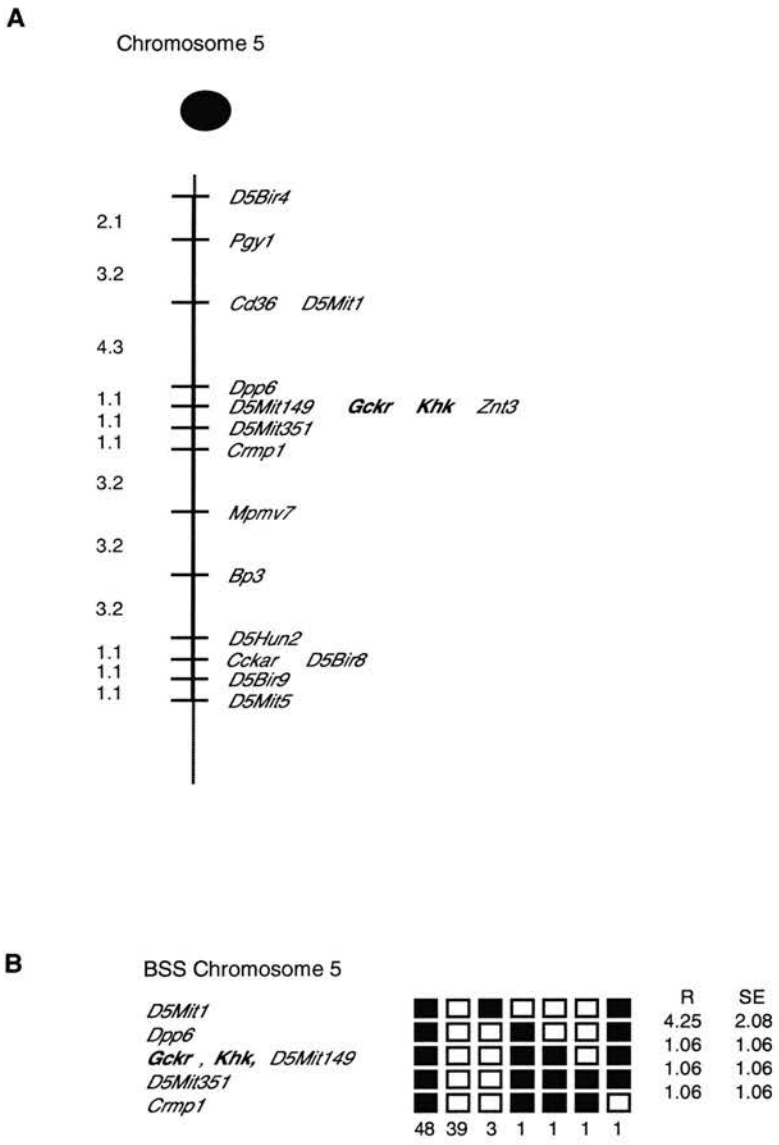


Figure 2.7 Co-localisation of *Gckr* and *Khk* on mouse chromosome 5.

A: map figure showing the proximal part of chromosome 5, with markers linked to *Gckr* and *Khk*. The map is depicted with the centromere (black circle) toward the top. The gene order and relative positions of markers used in this study are shown. Map distances in centimorgans are shown on the left.

B: Haplotype figure showing loci linked to *Gckr* and *Khk* on chromosome 5. Loci are listed in order with the most proximal on top. The black boxes represent the C57BL/6J allele, and the white boxes the SPRET/Ei allele. The number of animals with each haplotype is given at the bottom of each column of boxes. The percentage recombination (R) between adjacent loci is given to the right of the figure, with the standard error (SE) for each R.

2.4 Discussion

2.4.1 Co-localisation of *Gckr* and *Khk*

The mapping of the *Gckr* and *Khk* genes reveals that they co-localise to the proximal region of mouse chromosome 5 (see Figures 2.7 and 2.8). In addition, the mapping of *GCKR* and *KHK* in both human and mouse genomes reveals a new region of conserved synteny between human chromosome 2 and mouse chromosome 5. As both rat *GCKR* (Detheux *et al.*, 1993)and *KHK* (Hayward & Bonthron, 1998) map to rat chromosome 6, this suggests that *GCKR* and *KHK* are highly closely linked genes that have remained linked during evolution. *GCKR* encodes the regulatory protein of glucokinase, which binds to and inhibits glucokinase in liver and probably pancreatic islet (Malaisse *et al.*, 1990; Van Schaftingen, 1989). This inhibitory interaction is promoted by fructose-6-phosphate and relieved by fructose-1-phosphate, the product of ketohexokinase (KHK). The postulated metabolic link between KHK and GKRP therefore makes the co-localisation of their genes in human, rat, and as shown here also in mouse, noteworthy.

The present localisation defines a new region of conserved synteny (Figure 2.8), which has since been supported by the mapping of other genes located on human chromosome 2p to mouse chromosome 5 (Table 2.1). To date, there is not a physical and transcript map for this proximal region of mouse chromosome 5, therefore it is not known whether the gene order is the same within the regions of conserved synteny on human chromosome 2p23.3 and the proximal region of mouse chromosome 5. In the next chapter (Chapter 3), a detailed physical and transcript map of human chromosome 2p23.3 is described (Figures 3.5 and 3.8) that shows relative gene order at the *GCKR-KHK* genomic region. The analysis of two mouse genomic clones reveals that the two pairs of genes *Ucn-Mpv17* (Zhao *et al.*, 1998), and *EIF2B4-KIAA0064* (see Chapter 4, Figure 4.18 in Sections 4.2), are in the same intimate genomic arrangement in both the human and mouse genome. Although this is not proof that the relative gene order is the same at the region of conserved synteny between human chromosome 2p23.3 and mouse chromosome 5, it does emphasise the that these two genomic regions are highly evolutionary conserved.

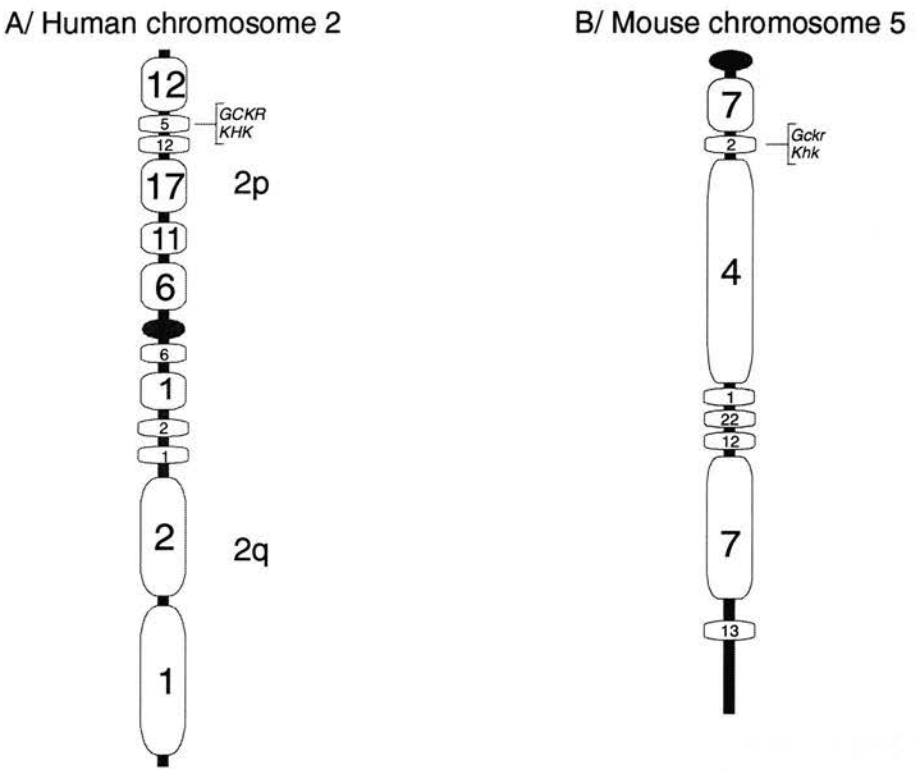


Figure 2.8 Comparative map of A/ human chromosome 2 showing homologous mouse chromosomes, and B/ mouse chromosome 5 showing homologous human chromosomes. Localisation of *GCKR* and *KHK* is indicated. This information shown in this figure was gathered from the Davis human/mouse homology map at the NCBI web-site (http://www.ncbi.nlm.nih.gov/Homology).

The region of conserved synteny between human chromosome 2 and mouse chromosome 5 must be relatively small, as other genes near *GCKR* and *KHK*, like the mouse homologue of the human phosphatase gene *PPP1CB,* map to the distal region of mouse chromosome 12 (Saadat *et al.*, 1994). The human *KIF3C* gene, a member of the kinesin family (see Chapter 5, Section 5.3), which resides on the other side of *KHK* and *GCKR* compared to *PPP1CB* (see Chapter 3, Figures 3.5 and 3.8), also maps to mouse chromosome 12 (Figure 2.9 shows the relative gene order on chromosome 2p23.3 and mapping of mouse gene orthologues). Located near *KIF3C* is the *KCNK3* gene whose mouse homologue also maps to chromosome 5 (Fujita *et al.*, 1998).

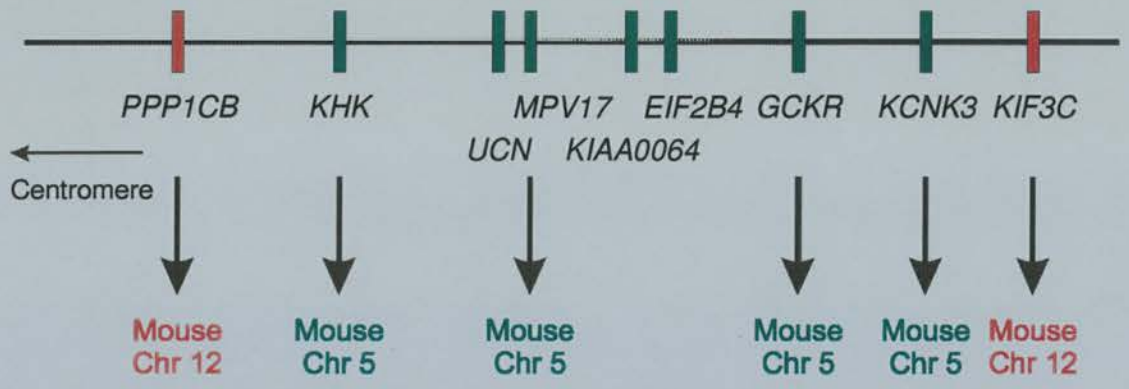


Figure 2.9 Relative gene order on human chromosome 2p23.3. Mapping of mouse gene orthologues is indicated if known. The estimated physical distance between *PPP1CB* and *KIF3C* according to YAC contigs is <2 Mb. This diagram is not drawn to scale.

It is only possible at present to make an imprecise estimate of the physical size of the region of conserved synteny between human chromosome 2p23.3 and mouse chromosome 5. YAC contig mapping indicates *PPP1CB* and *KHK* to be <400 kb apart (Hayward *et al.*, 1996; Hayward *et al.*, 1997). The physical size of the region between *KIF3C* and *PPP1CB* indicated by YAC, BAC, and PAC contig mapping is <2.0 Mb. Therefore the region of conserved synteny between human chromosome 2p23.3 and mouse chromosome 5 must also be <2.0 Mb in size. The co-localisation of *Gckr* and *Khk* in the mouse despite their non-syntenic relationship with *Ppp1cb* and *Kif3c*, which are adjacent in man, adds circumstantial support to the possibility of co-ordinate regulation of *GCKR* and *KHK*.

2.4.2 Other genes that co-localise with *Gckr* and *Khk*

The *Gckr* and *Khk* genes co-localise to the proximal part of the Jackson chromosome 5 BSS map, co-segregating with a number of other loci (Figure 2.7). Several of these markers were designed from ESTs or gene sequences and therefore it is possible that the human homologues are located in the region of synteny on human chromosome 2p23.3. The mapping of markers flanking the mouse chromosome 5 *Gckr/Khk* locus show that towards the centromere, for example *dpp6* maps to human chromosome 7 while towards the telomere, for example *Crmp1*, maps to human chromosome 4. The mapping of flanking markers to mouse *Gckr* and *Khk* to human chromosomes 7 and 4 confirms that the region of conserved synteny between human chromosome 2p23 and mouse chromosome is quite small

(< 2.2 cM from the mouse genetic maps and <2.0 Mb from human YAC contigs). Examination of the loci co-segregating with *Gckr* and *Khk*, reveals that several genes map either to human chromosome 7 or chromosome 4 and several ESTs including one showing homology to a zinc transporter gene (*Znt3*), show homology to a human ESTs that map to human chromosome 2p23.

The Jackson interspecific backcross BSS panel provides a useful tool in the search for novel genes that map close to *GCKR* and *KHK* on human chromosome 2p23. Any genes/ESTs that co-localise with *Gckr* and *Khk* in the mouse are worth further investigation as there is a good chance that the human homologue maps to human chromosome 2p23. Good candidate mouse genes that could map to human chromosome 2p23 include *Znt3* and several mouse cDNA clones that have not been mapped in the human genome. These include: *D5Ertd260e*, a mouse cDNA clone showing similarity to human KIAA0064 cDNA sequence (see Chapter 4, Section 4.3); *D5Xrf391*, an EST (Genbank F13207) showing homology to *S.cerevisiae* YJR072C gene; and several anonymous cDNA clones *D5Ertd422e* (J0214A03), *D5Ertd477e* (J0239B08), *D5Wsu178e* (C0034C05). (For further information, see the Jackson web-site at http://jax.org/resources/documents/cmdata/bkmap/BSS5refs).

A more recent search of the literature reveals that there are now several other genes that have been mapped to human chromosome 2p23 and mouse chromosome 5 – see Table 2.1 for a full list. As the glucokinase regulatory protein (GKRP) and ketohexokinase (KHK) are metabolically connected and their genes co-localise in both human and mouse genomes, an investigation of transcripts near *GCKR* and *KHK* might reveal other genes that encode proteins that are either functionally related to or interact with GKRP and KHK. Examination of the proteins encoded by the genes co-localising with *GCKR* and *KHK* (see Table 2.1), does not immediately reveal any such related genes as there seems to be a wide variety of protein functions and tissue expression patterns. However, it is probable that this is not an exhaustive list of the genes that reside at human chromosome 2p23 and mouse chromosome 5. Indeed, studies described elsewhere in this thesis show that this region of chromosome 2p23 is extremely gene-dense, so that only a detailed transcript map of chromosome 2p23.3 might reveal other genes encoding proteins with related functions to GKRP and KHK.

Mouse gene	Gene name/function	Expression profile based upon cDNA clones
Gckr	Glucokinase regulator protein	Liver, Pancreas, Testis.
Khk	Ketohexokinase	Brain, Colon, Kidney, Liver, Spleen, Tonsil, colon, connective tissue
Znt3	Zinc transporter gene	Testis, head and neck.
Mpv17	Peroxisomal protein A potential glomerulosclerosis and deafness gene in the mouse	Aorta, Bone, Brain, CNS, Heart, Kidney, Lung, Ovary, Prostate, Skin, Stomach, Testis, Tonsil, Uterus, Whole embryo, lung.
Ucn	Urocortin A neuropeptide related to urotensin I and corticotrophin-releasing factor	Brain, Colon, Germ Cell, Kidney, Lymph, Muscle, Prostate, Testis, Uterus, Whole embryo, lung, ovary.
Kcnk3	Potassium ion channel	Brain, Foreskin, Kidney, Lung, Neural, Pancreas, Placenta, Pooled, Prostate, Uterus, brain.
Fosl2	Fos-related antigen 2	Breast, Colon, Foreskin, Heart, Liver, Pancreas, Prostate, Skin, Uterus, colon, head_neck, lung, ovary.
Cenpa	Centromeric protein A	Adrenal gland, Brain, Breast, Colon, Germ Cell, Pooled, Testis, Tonsil, Uterus, Whole embryo.

Table 2.1 Genes that map to human chromosome 2p23 and mouse chromosome 5. For function of encoded proteins, see Chapter 3, Table 3.4.

Chapter 3

3 Physical and transcript mapping in chromosome 2p23.3

3.1 Introduction

3.1.1 Aim

To aid the search for transcripts and mapping of chromosomal markers located in the *GCKR-KHK* genomic region and DFNB9 interval on chromosome 2p23.3, a detailed physical contig spanning this genomic region was constructed. At the start of this physical contig construction, BAC sequences produced as part of the human genome project were not available and only YAC and PAC libraries were available for screening. However, the recent availability of BAC clone sequence within this chromosomal region as part of the human genome project, although not complete and consisting of unordered contigs of DNA sequence, has been used to supplement the physical contig that I have constructed.

To search for transcripts within the physical contig spanning the *GCKR-KHK* genomic region and DFNB9 interval, the genomic clones that were used to construct the contig were PCR screened using ESTs that mapped to chromosome 2p23.3 by radiation hybrid mapping according to Genemap'99 (http:/www.ncbi.nlm.nih.gov/genemap). Alternatively, the analysis of DNA sequence obtained by direct sequencing of PAC and cosmid clones also allowed the identification of transcripts located in this genomic region. Although other methods of transcript identification were contemplated such as cDNA selection and exon trapping (see Section 3.1.3), they were found not to be required, largely because of the high gene density that became apparent within this genomic region.

3.1.2 Contig Assembly

3.1.2.1 Identification of YAC and P1 clones

The original physical contig constructed at the *GCKR-KHK* genomic locus by Hayward *et al.*, 1996, is shown in Figure 3.1. The YACs 29IH8 and 26BA11 were identified by the PCR screening of a YAC library (Anand *et al.*, 1990) using primers designed from within *GCKR* (Warner *et al.*, 1995). The YACs 18AG7 and 3AG3 were identified by PCR screening the same YAC library using primers from within *KHK* and STS D5, respectively (STS D5 is a

STS designed from sequence at one of the ends of YAC 26BA11; STS D3, designed from the other end of YAC 26BA11, lies within *GCKR* intron 7). The P1 clone J0788 which contains the whole of *KHK* was identified by hybridisation screening 40 000 clones of the ICRF P1 reference library 700 (Francis *et al.*, 1994) using a 2.0 kb *KHK* cDNA insert (Bonthron *et al.*, 1994). The P1 clone J16101, which has been shown to contain *GCKR* exons 1-7, was identified by screening the same P1 library with a cloned 623 bp RT-PCR product that corresponded to *GCKR* codons 7-214 (Hayward *et al.*, 1996).

The EST content analysis of the physical contig shown in Figure 3.1 reveals none of the YACs to contain both *GCKR* and *KHK*. This meant that STS content analysis of the YAC clones had to be used to establish the *GCKR-KHK* physical linkage. This was carried out by PCR screening of the YAC clones 3AG3, 18AG7, 26BA11 and 29IH8 using STSs created from sequence at the ends of YAC 26BA11 (STSs D3 and D5 - STS D3 is located within *GCKR* intron 7) and one end of the *KHK*-containing PAC J0788 (STS J0788 5.3/5.4).

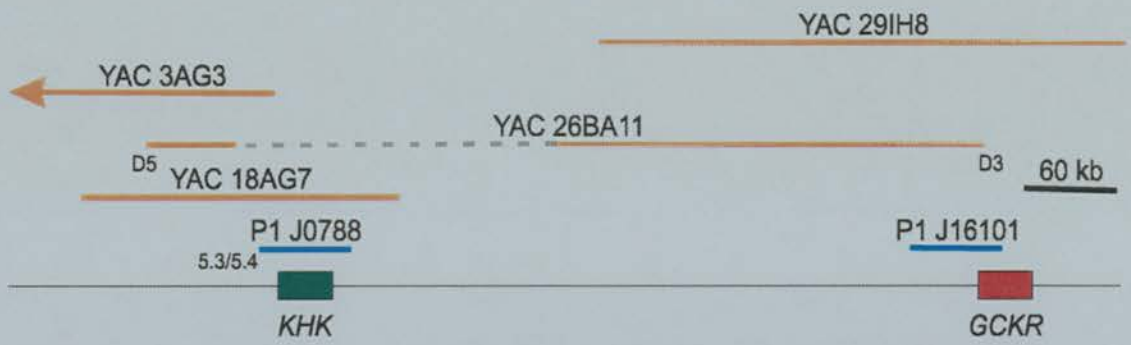


Figure 3.1 The original YAC and P1 clone physical contig at the *KHK-GCKR* genomic locus constructed by Hayward *et al.*, 1996. The distance between *KHK* and *GCKR* was estimated from FISH experiments to be ~500 kb. The examination of the STS markers mapping to these YAC clones reveals that it is not possible to assemble an internally consistent contig. This was later found to be due to a deletion in YAC 26BA11 – indicated by a dashed line. The lack of chromosomal markers that had been mapped to the physical contig meant that at this time the contig orientation with respect to the centromere was not known.

The results of this STS content analysis showed the *KHK*-containing YAC 18AG7, and YAC 3AG3, to both contain the YAC 26BA11 STS D5 (YAC 26BA11 contains part of *GCKR*). Also, the STS JO788 5.3/5.4, an STS from one end of the *KHK*-containing P1 clone J0788, is positive for YAC 26BA11. Although these results established a physical linkage between *GCKR* and *KHK*, examination of the STS markers mapping to these YAC clones reveals that it is not possible to assemble an internally consistent contig . This was later proved to be due to a deletion within YAC 26BA11 (see Discussion section 3.4.1.1).

Although FISH analysis suggests *GCKR* and *KHK* to be less than 500 kb apart (Hayward *et al.*, 1996), the deletion within YAC 26BA11 meant that it was impossible to estimate the distance between *GCKR* and *KHK* using the existing physical contig. The identification of this aberration within YAC 26BA11 also meant that other genomic clones would have to be identified to span the deletion within YAC 26BA11. This would be essential for an accurate estimation of the physical distance between *GCKR* and *KHK*, and also would be required for a detailed investigation of transcripts that resided within the *GCKR-KHK* intergenic region.

As part of this project was to identify candidate deafness genes within the DFNB9 interval (see Chapter 5), a genomic region also located on chromosome 2p23, it was decided to extend the *GCKR-KHK* physical contig to span the DFNB9 interval. This would be performed by identifying YAC clones by PCR screening YAC libraries for chromosomal markers and ESTs that were known to map to chromosome 2p23 by radiation hybrid mapping.

3.1.2.2 Contig construction strategy

To produce a more detailed contig between *KHK* and *GCKR*, YAC clones containing *GCKR* were subcloned into cosmids. For cosmid contig assembly, a high resolution, fluorescence based semi-automated technique for DNA fingerprinting was used (Carrano *et al.*, 1989) – see Methods Section 3.2.2. One of the advantages of subcloning the YAC inserts into cosmid vectors is that the cosmid DNA could be directly sequenced for gene identification without the need for further subcloning into plasmid vectors. Also, a detailed contig would allow the rapid elucidation of the relative order and orientation of transcripts located on the contig.

For the identification of new genomic clones at the *GCKR-KHK* interval, the cosmid STSs (shown in Table 3.1) were used to either PCR screen the RCPI1 PAC library, or screen chromosome 2-specific PAC and cosmid libraries by radioactive hybridisation, – see Figure 3.2 for a general outline of the contig construction strategy. To confirm clone overlap and for further library screening, the insert ends of the PAC clones initially identified were sequenced using the "T7" and "SP6" universal primers. STSs were designed from these sequences only if they were non-repetitive elements. New STSs were used to screen all the genomic clones used in the physical contig to confirm clone overlap. Also, to aid contig assembly, STSs that did not "hit" any of the genomic clones in the physical contig were used to re-screen the PAC library. This strategy was used to extend the physical contig with the aim of spanning the deletion within YAC 26BA11.

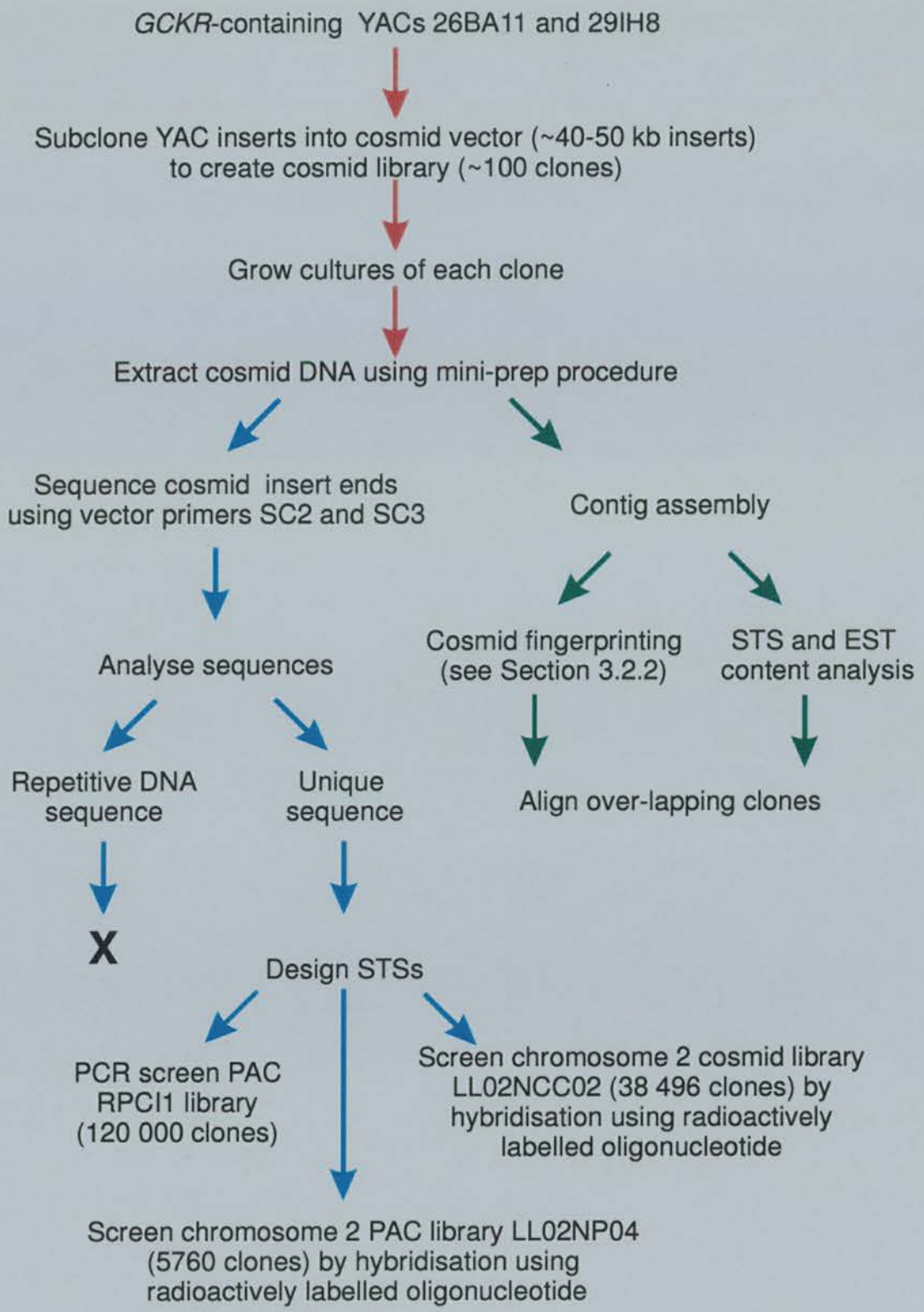


Figure 3.2 General strategy employed for contig construction at the *GCKR-KHK* interval.

3.1.3 Approaches for identifying transcripts

Once a genomic region has been cloned, for example the physical contig spanning the *GCKR-KHK* genomic region and DFNB9 interval, the next step is to identify transcripts for further identification. This section describes five methods for identifying transcripts.

3.1.3.1 cDNA selection

The method of direct selection allows rapid amplification of coding sequences within large genomic regions. It is based upon the hybridisation of cDNAs to an immobilised genomic clone/library and "enrichment" of specific hybrids (Lovett *et al.*, 1991). For example, a partial *Sau*3A genomic clone digest is labelled with biotin linkers and immobilised on streptavidin coated paramagnetic beads. A high quality cDNA library is also partially digested with *Sau*3A and hybridised to the immobilised genomic library (it is important to use a complex cDNA library - if the gene of interest is not represented in the cDNA library, it cannot be cloned by direct selection). Several rounds of washing are performed to remove non-specific hybrids to leave the specific cDNA-genomic DNA hybrids. The specific cDNAs can be eluted and utilised for PCR amplification using the linker oligonucleotide sequences that were attached to the partially digested cDNA before hybridisation. The advantages of this method are that it is a relatively quick procedure once all the conditions have been optimised. However, this technique is susceptible to several problems. Repeat elements within the cDNA genomic sequences must be blocked successfully to prevent non-specific hybridisation. The use of a high quality cDNA library is essential as some libraries may contain contaminants or may not contain a good representation of sequences. Also, the basis of this technique relies upon the hybridisation of cDNA to genomic DNA. This means the homology is "patchy" and may cause hybridisation problems especially if the gene only contains short exons.

3.1.3.2 Exon trapping

Exon trapping facilitates the capturing of exons from genomic DNA by relying on the cellular splicing machinery (Duyk *et al.*, 1990). Various exon trapping strategies have been devised but all rely on the *in vivo* selection for splice sites flanking exon sequences in genomic DNA. One strategy (Heiss *et al.*, 1996) involves the cloning of the genomic DNA of interest for example from a cosmid or YAC, into an exon trap vector containing splice sites, a promoter and polyadenylation signals. The purified recombinant DNA is transfected into COS-7 cells which allows any exons within the genomic DNA to be transcribed and spliced. After the total RNA is isolated, cDNA synthesis with vector specific primers is

performed and the products cloned to generate an exon trap library, or alternatively, are used as complex probes on cDNA libraries. The advantages of using an exon trapping strategy are that unlike in techniques such as cDNA selection, genes whose expression patterns are tissue specific or developmentally regulated will be isolated with the same efficiency as ubiquitously expressed genes. Exon trapping is only limited by its requirement for the presence of functional 3′ and 5′ splice sites flanking a target exon. Therefore intronless or single intron genes would be missed using this approach.

3.1.3.3 CpG island positional cloning

CpG islands are relatively short stretches (~1 kb) of G+C-rich (60-70%) genomic DNA in which the frequency of non-methylated CpG dinucleotides is substantially higher than elsewhere in genomic DNA (Cross & Bird, 1995). A large proportion of genes with a tissue-specific pattern of expression have been found with CpG islands at their 5′ ends, often as part of the promoter sequence (Bird *et al.*, 1987). Because of their location, CpG islands are frequently used as markers for the presence of genes in uncharacterised genomic DNA. Although it is still largely unclear what the exact function of CpG islands is, DNA methylation has been shown to repress transcription and has been implicated in alterations of gene expression. Methylation of CpG islands and subsequent silencing of associated transcription units have been found to occur in genes located on the inactive X chromosome (Pfeifer *et al.*, 1990), genes silenced by genomic imprinting (Razin & Cedar, 1994), and genes silenced in transformed cell lines and tumours (Antequera & Bird, 1993; Bird, 1996). CpG islands have also been shown to often contain multiple binding sites for transcription factors (Pfeifer *et al.*, 1990) and an open chromatin conformation (Antequera *et al.*, 1989). One mechanism by which DNA methylation can cause transcriptional repression is by the binding of methyl-CpG binding proteins to methylated CpG islands, directly interfering with the binding of sequence-specific transcription factors (Hendrich & Bird, 1998). CpG islands have now also been associated with replication origins (Delgado *et al.*, 1998).

The identification within genomic DNA of rare cutting restriction enzyme sites for example *Bss*HII, *Eag*I, and *Sac*II, suggests a high "GC" content and the possible presence of a CpG island. If cloning and sequencing of these restriction fragments reveals a CpG island, there is a high likelihood that the CpG island may be adjacent to a gene sequence. This method of CpG island identification has been used successfully to identify a number of gene sequences.

3.1.3.4 Sequencing of genomic clones

Once genomic clones have been mapped to the chromosomal region of interest, the sequencing of these clones and sequence analysis may reveal transcript sequences. While BACs, PACs and cosmids may be sequenced directly, it is more efficient to create a plasmid subclone library from the larger genomic clones to allow rapid sequencing of a large number of novel DNA sequences. Two types of subclone library can be created: 1/ a restriction digest subclone library produced by cloning restriction fragments, and 2/ a shotgun library produced by cloning DNA fragments created by physical disruption or use of DNAse I. A shotgun library is preferable, as it will provide a much greater variety of subclone inserts and more random sampling of the sequence of the large-clone insert.

The computer analysis of subclone sequences using programs such as BLAST (Altschul *et al.*, 1990) can identify possible gene sequences by looking for similarity to sequences found in cDNA databases. Although this method requires much time and resources, once enough unique sequences have been obtained, they can be used to produce a contig of sequences. This valuable resource can be used for characterisation of genomic structures once a gene sequence has been identified. Gene sequences not found in cDNA libraries can be identified by use of gene prediction programs, for example GRAIL (Roberts, 1991), that identify exons on the basis of common structural features. The use of the nucleotide identification (NIX) computer program interface (based at HGMP) which encompasses a group of sequence similarity search programs and gene prediction programs (commercially available from Genscan and the Sanger Centre) was found to be a highly useful aid to the analysis of genomic sequence in the search for transcripts.

3.1.3.5 PCR screening for ESTs

Many EST primer pairs have been designed from cDNA clones and their position on human chromosomes mapped by radiation hybrid mapping. This EST mapping information has been gathered to produce a transcript map, allowing ESTs and genes to be placed in relation to chromosomal markers -see Genemap'99 at the NCBI web-site (http://www.ncbi.nlm.nih.gov/genemap). Once the chromosomal location of a genomic clone is known (for example a YAC clone), ESTs mapping to the same genomic region can be used to PCR screen the YAC clone. The screening of genomic clones using ESTs is quick and easy to perform but for this method to be successful, the location of the genomic clone on genetic maps must be accurately known.

3.1.4 Identification and analysis of transcripts

During the course of the physical contig construction at the *GCKR-KHK* and DFNB9 intervals, the insert ends of cosmid and PAC clones had been sequenced for the design of novel STSs. Further sequence analysis of the generated sequences turned out to be a highly useful method for searching for transcripts. The mapping of chromosomal markers to the YAC clones in the chromosome 2p23.3 physical contig allowed the selection of ESTs (mapped by radiation hybrid mapping) for PCR screening the clones within the physical contig. This again turned out to be a successful method for transcript identification at the *GCKR-KHK* and DFNB9 intervals. A search for CpG islands by searching for rare restriction enzyme sites within the cosmid genomic clones was also performed. Due to the success of searching for transcripts by direct sequencing and EST PCR screening of genomic clones, the other methods such as cDNA selection and exon trapping were found not to be required. Once transcripts had been identified, they could be accurately placed onto the physical contig and in many cases the relative gene order and orientation could be deduced. Sequence analysis of the transcripts identified was performed to look for clues to the encoded protein's function and this was used to choose transcripts for further investigation (the investigation of individual transcripts is described in Chapters 4 and 5).

3.2 Methods

3.2.1 YAC contig assembly

The chromosomal marker D2S1237 had previously been mapped to the genomic YAC clone 3AG3 (Hayward *et al.*, 1996; Prochazka *et al.*, 1995) near to the *GCKR-KHK* region. The examination of radiation hybrid maps for chromosomal markers mapping near to D2S1237 was carried out to identify other possible chromosomal markers for PCR screening of the YACs 29IH8, 26BA11, 18AG7 and 3AG3. With the identification of new chromosomal markers mapping to the *GCKR-KHK* YAC contig, a search of the Genome Database (GDB) was performed to identify other YACs, for example CEPH mega-YACs and Whitehead YACs that also contained these chromosomal markers.

Once YACs had been identified and obtained (CEPH mega-YACs were obtained from the HGMP resource centre, Hinxton), further STS and EST content analysis was performed by PCR to identify overlapping YAC clones. To extend this contig to include the DFNB9 interval, YACs mapping between *GCKR* and *KIF3C* (a gene known to reside in the DFNB9 interval), were identified by PCR screening YAC libraries using STSs and ESTs that were known to map to the chromosome 2p23.3 by radiation hybrid mapping. Identification of YACs mapping to the DFNB interval was carried out by Dr J. Leek. All primer sequences for chromosome 2 polymorphic markers were obtained from GDB (http://www.hgmp.mrc.ac.uk/gdb/).

3.2.2 Cosmid fingerprinting and cosmid contig assembly

For more detailed mapping of the *GCKR-KHK* interval and gene identification, the YACs 26BA11 and 29IH8 were subcloned by Dr J. Warner into the Supercos1 vector. Once cosmid DNA had been extracted by the alkaline lysis technique (Section 6.2.1.2), the method of cosmid fingerprinting was used to assemble the cosmid clones into a contig (summarised in Figure 3.3).

Briefly, a universal primer labelled with one of three fluorescent dyes (FAM/HEX/ROX-5′-dTCCCAGTCACGACGTTGT-3′) was annealed to a complementary synthetic oligonucleotide to create a double-stranded oligonucleotide linker with a 5′-overhang complementary to one produced by a restriction enzyme. The complementary oligonucleotides were designed with *Hin*dIII, *Eco*RI and *Bam*HI overhangs (5′-dAGCTACAACGTCGTGACTGG-3′, 5′-dAATTACAACGTCGTGACTGG-3′,

5′-dGATCACAACGTCGTGACTGG-3′, respectively). The cosmid DNA was digested with the appropriate restriction enzyme and ligated to the fluorescent dye-labelled oligonucleotide linker. A second digestion step using a different restriction enzyme, for example *Hin*fI, was performed to create smaller restriction fragments and an alternative fingerprint. To size the dye-labelled restriction fragments, the restriction products were electrophoresed through a polyacrylamide gel and run using the Genscan software package on the ABI 377 fluorescent DNA sequencer. The cosmid fragments were sized by comparison to an internal lane size standard (Genescan 500 (TAMRA-labelled)).

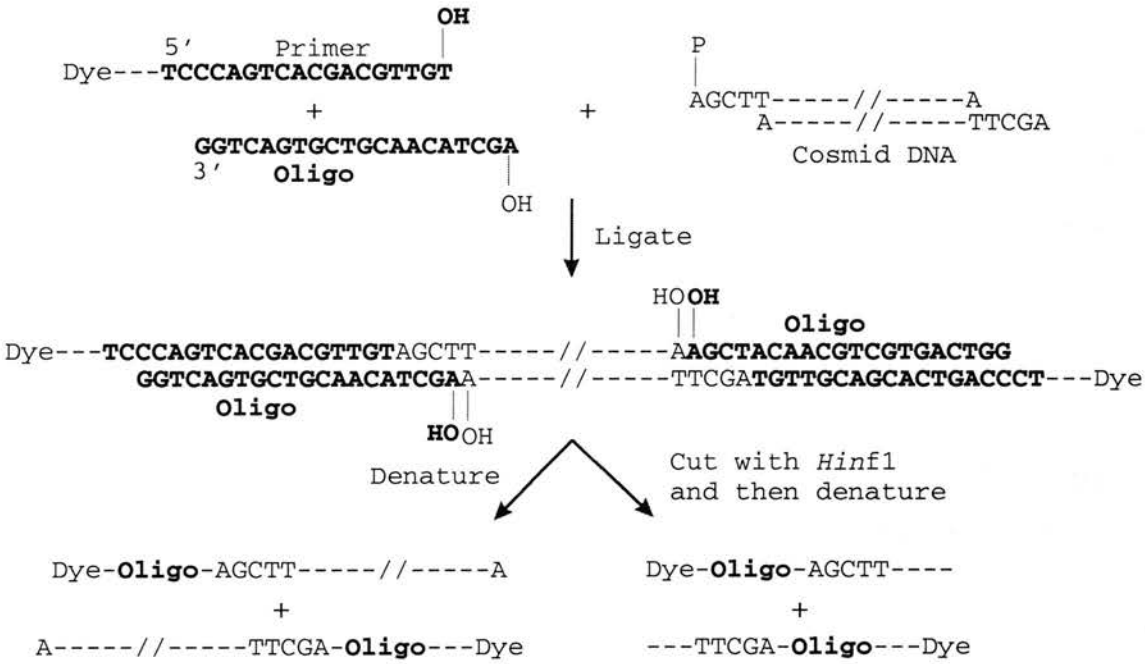


Figure 3.3 Annealing, restriction enzyme, and ligation reactions for the preparation of the fluorochrome-labelled restriction fragments. The cosmid DNA is shown as a *Hin*dIII restriction fragment for illustration purposes only; the fluorochrome-labelled oligonucleotide is shown in bold type. In actuality, the reaction is initiated with uncut cosmid and the restriction enzyme and ligase are added simultaneously. This diagram was adapted from Carrano *et al.*, 1988.

The cosmids were assembled into a contig of overlapping clones, based on the number of shared fragments (of equal size) that each cosmid contained. The benefits of using this technique included a single reaction mixture for the simultaneous restriction digest and ligation reaction (only applicable if the double-stranded oligonucleotide linker destroyed the restriction enzyme recognition site upon ligation to the restriction fragment being labelled).

Advantages of using an automatic fluorescent sequencing machine included a fast through put of samples with four different dyes run per lane (three fingerprinting reactions and one internal size standard) and that an internal size standard in each lane allowed highly accurate sizing of the restriction fragments. To confirm clone overlap inferred from the cosmid fingerprinting, cosmid clones were PCR screened using novel STSs (Table 3.1) designed from sequence at the ends of the genomic DNA cosmid clones's inserts. The primers used to obtain these sequences were designed from the Supercos1™ vector sequence flanking the multiple cloning site ("SC2": 5'-dTGGAAGTCAACAAAAAGCAGAGC-3' and "SC3": 5'-dGAGGCCCTTTCGTCTTCAA-3').

3.2.3 PAC contig assembly

To identify PAC clones located on chromosome 2p23, STSs and ESTs were used to PCR screen the RCPI1 PAC library, constructed by P. de Jong -see Figure 3.4 below for strategy used to screen the RCPI1 PAC library.

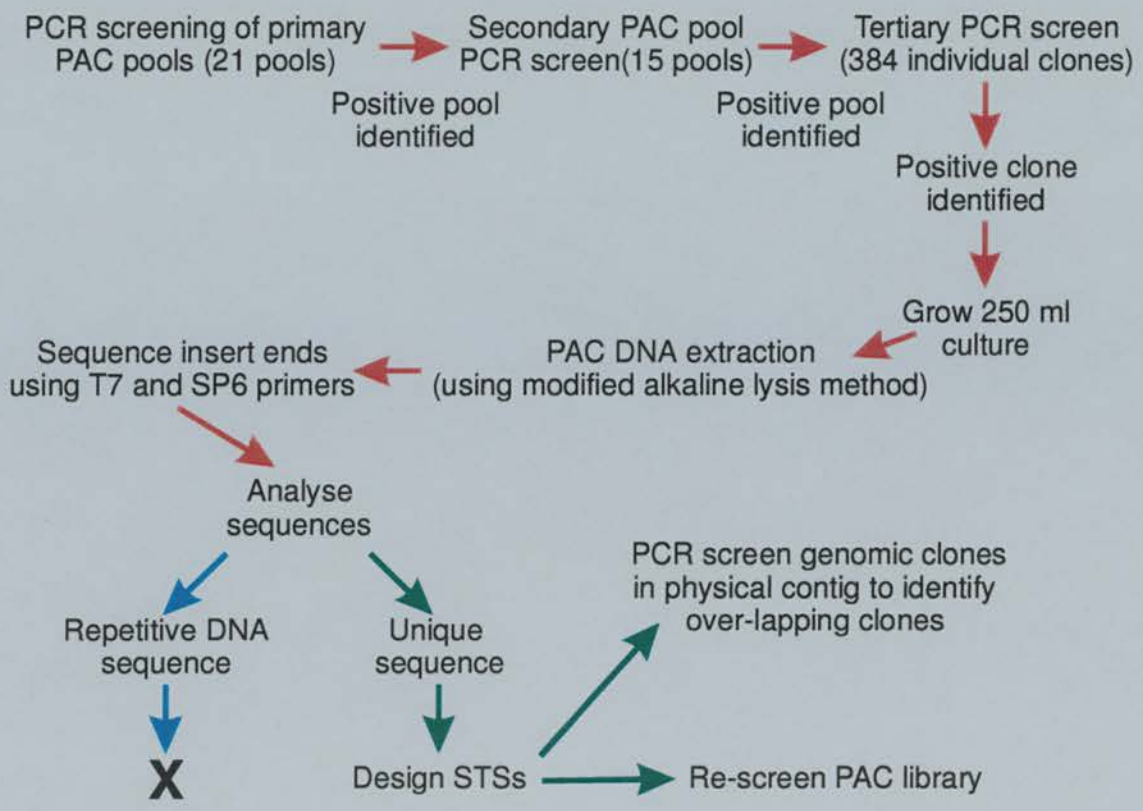


Figure 3.4 Strategy for identifying PAC clones. STS or EST primer pairs were used to PCR screen the PAC library RPCI1 (constructed by P. de Jong, Roswell Park Cancer Institute, Buffalo). The RPCI1 library consists of approximately 120 000 clones.

STS	Primer pair sequences (5′-3′)	Annealing temp. (°C)	Expected size (bp)
A4′2	dAGGATGAAGTAGAAGTAGCACGC dACCCATCTTAGTTATTTCACAGGA	60	250
A4′3	dTAGCTACCAAGCACCTAAAATGT dGTGGGAAAGAGAAGCACTCA	62	189
A12′2	dATTGGTCTGGTAGCTCAGTCACT dAACATCAGCCCTACACTCAGC	62	265
A12′3	dTCTACACACTAAACCGTCCCT dTTCATTTCTTTGGTACCCTTC	58	170
G9′2	dGGAGAATGAGATCTATCAAAGTGAGA dCCATTCATCGTCCTCCTTTTC	60	220
G9′3	dTGGAGGTGGTGAAACTAGA dCTCCTTTACATTCCCATACTC	60	119
J16101 #14	dGAATTGTTCCTTTTTCTCCCCG dGGCCTCAACAAATGTATTGATCAG	60	160
J16101 #28	dTCTTCGAGATGTCCTCTTCAAC dGAGGAGTGCCATTTCCATAGAG	60	160
119A8 SP6	dTCATGCAGTTATGGAGATAGAC dGAACATAAAGGCTGGCTCC	58	308
103I5 T7	dGGTACCCAAAGTCCTGTCTG dTGCTACCCCTTCTCCTACAG	62	288
103I5 SP6	dGTGTTTAATCATCTCCTGCTA dGAGTTAAAAGATATTGTGTGACAG	58	110
13L13 T7	dACGACGACGTTTTGCAGA dACATCCACAAGGTGTCCCAT	58	162
17F24 SP6	dGTAGATCAGCCAAAATGCCC dTTTCTGCCATATGATTCCCC	58	225
17F24 T7	dAGCATTTGGCCAGACGTAGT dATCTTCCCACCTCAAGCCTC	60	220

Table 3.1 Primer sequences for STSs generated from sequencing ends of cosmid and PAC clones. STS nomenclature: ′2 and ′3 refer to STSs designed from cosmid sequence obtained using SC2 and SC3 primers, respectively; T7 and SP6 refer to STSs designed from PAC sequence obtained using T7 (5′-dTAATACGACTCACTATAGGG-3′) and SP6 (5′-dTATTTAGGTGACACTATAG-3′) primers respectively.

3.2.4 BAC clone identification

The recent availability of BAC sequences produced as part of the Human Genome Project facilitated the identification of BACs located within the *GCKR-KHK* region by means of searching the Genbank DNA sequence databases using EST sequences from within the *GCKR-KHK* contig. This identified three partially sequenced BACs in the form of unordered contigs. Further sequence analysis using gene identification programs based at the HGMP resource centre, Hinxton, (submitted using the NIX interface), identified several EST and gene sequences located on these BACs (Table 3.3). The gene sequences that were

identified were used to place the BAC clones on to the *GCKR-KHK* interval physical contig (see Figure 3.8).

3.2.5 Mapping of ESTs to *GCKR-KHK* region

ESTs mapped to the chromosome 2p23.3 region by radiation hybrid mapping according to Genemap'99 were used to PCR screen genomic clones mapping to the *GCKR-KHK* region (for list of ESTs, see http://www.ncbi.nlm.nih.gov/genemap). The sequencing of insert ends from PAC and cosmid clones and comparison to the Genbank EST DNA database using the BLAST program also identified several transcript sequences. A summary of all the genes and ESTs that were identified to map within or near the *GCKR-KHK* genomic region can be found in Table 3.4.

3.2.6 Identification of CpG islands

CpG islands are often associated with transcript sequences and can be identified by searching for rare cutter restriction enzyme sites within genomic DNA sequences. As a preliminary investigation into the presence of CpG in the *GCKR-KHK* interval, a search for rare cutting restriction enzyme sites was performed on genomic clones using the restriction enzyme *Eag*I (cutting site: CGGCCG). The restriction analysis of the *KHK*-containing P1 clone J0788 using the *Eag*I restriction enzyme had previously revealed five internal *Eag*I cutting sites (Hayward & Bonthron, 1998). The largest *Eag*I fragment (~40 kb) was radioactively-labelled with [α^{32}P] dCTP by random priming and used to screen a colony filter of the cosmid library described in Section 3.2.2 (performed by Dr B. Hayward). This initial screening of the cosmid library using the P1 clone J0788 *Eag*I fragment was performed to increase the chances of identifying cosmids containing *Eag*I restriction enzyme sites. Positive cosmids (cosmids: A1, A7, A10, B2, B8, B10, C2, C7, C8, C12, D10, and D11) were chosen for restriction analysis using *Eag*I. The resulting cosmid *Eag*I restriction fragments were ligated into the pGEM vector and the inserts sequenced. Computer analysis of the sequences was performed to ascertain the presence of a CpG island. The methods used are as described in the methods (Chapter 6).

3.3 Results

3.3.1 Physical mapping

3.3.1.1 YAC clone contig assembly

A YAC physical contig was constructed that spanned both the *GCKR-KHK* and DFNB9 intervals on chromosome 2p23.3. This contig is shown in Figure 3.5 and consists of eleven YAC clones identified by screening either the ICI YAC or CEPH mega-YAC libraries. Figure 3.5 also displays the chromosomal markers and genes that were used to identify overlapping YAC clones. The physical size for this genomic region, based on YAC sizes, is estimated at <2 Mb.

3.3.1.2 Cosmid clone contig assembly

Cosmid fingerprinting produced a set of restriction fragment sizes for each cosmid clone (overlapping cosmid clones share identical sized restriction fragments). Therefore, to identify overlapping cosmid clones, the fingerprint data was transferred into a Microsoft Excel spread sheet and cosmid clones put into order based on number of "common" restriction fragments (data not shown). This alignment was carried out manually as there were only 92 cosmid clones.

The analysis of cosmid fingerprinting data identified cosmid clones sharing common sized restriction fragments, indicating that they were likely to be overlapping. To confirm cosmid clone overlap, STSs designed from sequence at the ends of cosmid clones' inserts were used to PCR screen the cosmid clone library (see Table 3.1 for STS details). By combining the data from both the fingerprinting experiment and the STS library screening, a cosmid contig was constructed (Figure 3.6 shows a selected number of cosmids shown to be overlapping by both cosmid fingerprinting and STS screening).

Described later in this results section is the search for and mapping of ESTs/transcripts. One EST that was mapped to cosmid B2 was WI-18702, designed from the human homologue of the mouse *Mpv17* gene (a recessive kidney disease and deafness gene; this gene is discussed in Chapter 5, Section 5.2). The human *MPV17* gene consists of 8 exons (Karasawa *et al.*, 1993). PCR analysis of the YAC 26BA11 and its cosmid subclone B2 revealed that *MPV17*

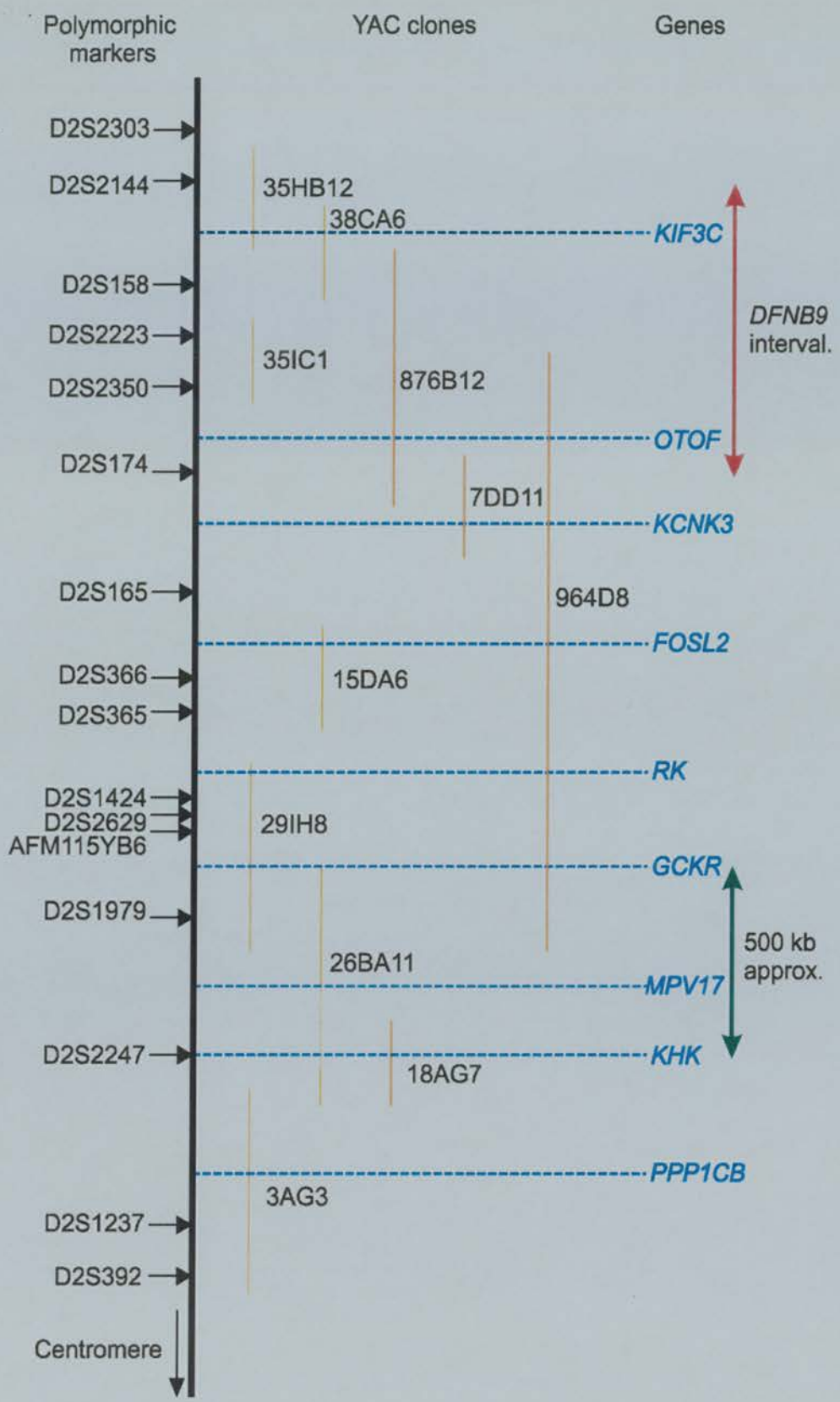


Figure 3.5 YAC contig of chromosome 2p23.3. Polymorphic markers and genes mapping to the YAC clones are indicated. The *GCKR-KHK* and DFNB9 intervals are shown on the right hand side.

exons 1-5 were present, but exons 6-8 were missing from YAC 26BA11. STS and EST analysis of YAC 26BA11 (Hayward et al., 1996), had previously suggested the presence of an internal deletion and that the deletion included the *KHK* gene. The absence of *MPV17* exons 6-8 suggests that the deletion in YAC 26BA11 starts within the *MPV17* gene and includes the *KHK* gene (see Figure 3.6). As no genomic clones had so far been identified that span this deletion, other genomic clones would have to be identified for the construction of a complete *GCKR-KHK* interval physical contig.

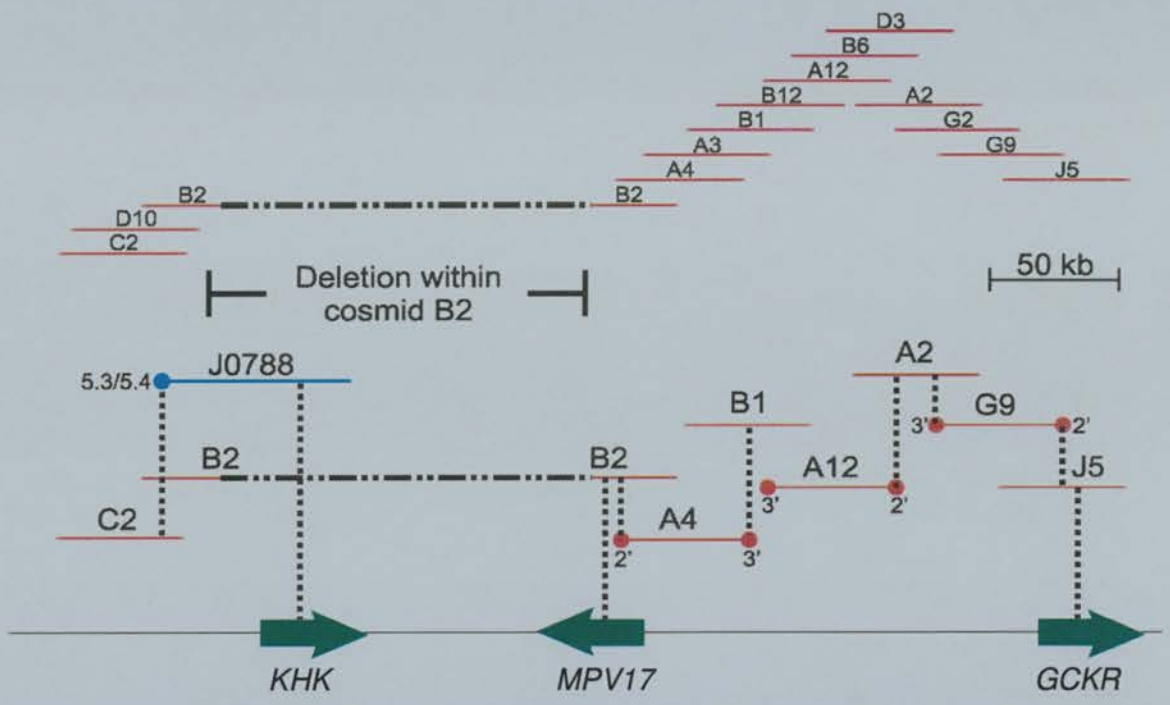


Figure 3.6 Cosmid contig assembled by cosmid fingerprinting and STS/EST content. The upper contig was assembled by analysis of cosmid fingerprinting data; the lower contig was assembled by EST/STS content analysis ('2 and '3 refer to sequence at the ends of a cosmid's insert, used to design the STS). The positions of the three genes *KHK*, *MPV17* and *GCKR* are shown (arrow indicates 5′ to 3′ orientation). All clones are cosmids except the *KHK*-containing P1 clone J0788 (STS 5.3/5.4 was designed from one end of this clone). A deletion was identified within cosmid B2; this cosmid, contains the 5′ but not the 3′ end of the *MPV17* gene (deletion is shown by a dotted/dashed line).

3.3.1.3 PAC contig assembly

For the identification of other genomic clones within the *GCKR-KHK* interval, the RPCI1 PAC library was screened by PCR using STSs designed from the insert ends of cosmid clones (see Table 3.1 for primer sequences and Figure 3.4 for PAC library screening strategy). Once PAC clones were identified, their insert ends were sequenced and novel STSs designed. These STSs were used to re-screen all the genomic clones used to construct

the physical contig and thus allow the PAC clones to be accurately placed onto the YAC/cosmid contig.

In total, 15 PAC clones were identified and placed onto the *GCKR-KHK* and DFNB9 interval physical contig - see Figure 3.7 and Table 3.2 for PAC clone names (Table 3.2 describes PAC clones identified that are not shown in Figure 3.7). STS analysis of the PAC clones revealed that a complete contig spanning from *KHK* to *GCKR* could not be constructed. Figure 3.7 shows the PAC clone contig (at the time of construction, it was not known which genomic clone (J0788 or L1313) was closest to *GCKR*). Re-screening the PAC library with a STS designed from sequence at the end of PAC L1313 (STS L1313-SP6) did not identify any new PAC clones. As one end of the P1 clone J0788 insert is a repetitive DNA sequence (the opposite insert end sequence was used to design STS 5.3/5.4), which could not be used for STS design. Screening of the chromosome 2-specific PAC library LL02NP04 filters by hybridisation of STS oligonucleotides again did not identify any new PAC clones.

STS/EST	PAC clone name
KIF3C (stSG4510)	14H17, 97K3
KCNK3 (see Section 5.5.2.1)	249B24
FOSL2 (STS11971)	17F24
17F24-SP6 (Table 3.1)	190C22, 294C20
17F24-T7 (Table 3.1)	40P7, 159C16, 195F13
RBSK (stSG15128)	22F16, 137K13

Table 3.2 PAC clones mapping to chromosome 2p23.3 (but not shown on physical contig in Figures 3.7). All clones were identified by PCR screening the RPCI1 PAC library using the ESTs/STSs indicated.

3.3.1.4 Completion of contig using BAC clones

Although screening of the RCPI1 PAC library had proved to be highly successful in identifying clones mapping to chromosome 2p23, there was still a discontinuity in the physical contig spanning the *GCKR-KHK* interval. However, the recent BAC clone sequences that have been deposited into the Genbank sequence database by the Washington University School of Medicine as part of the Human Genome project, have allowed the identification of genomic clones that complete the *GCKR-KHK* interval contig.

The searching of the Genbank sequence database using sequence from *KHK*, *MPV17*, and *GCKR* identified three partially sequenced BAC clones (NH0195B17, NH0538J11, and 45_M_3, respectively). A summary of the database sequence analysis for these three BAC clones is described in Table 3.3. The analysis of the *GCKR*-containing BAC clone 45_M_3

confirmed the presence of *GCKR* and an EST already mapped to the physical contig by PCR. However, sequence analysis of the *KHK*-containing BAC clone NH0195B17 and the *MPV17*-containing BAC clone revealed them to contain two genes in common, *SMVT* and *CAD*. Thus, by addition of these two BAC clones to the existing PAC contig, a complete and contig was constructed, spanning from *KHK* to *GCKR* (Figure 3.7).

BAC clone	Transcripts/ESTs revealed by sequence analysis
NH0195B17	*KHK, SMVT, CAD, EMILIN, CGR11*
NH0538J11	*MPV17, SMVT, CAD, ZNT3, UCN, TFIIIC2*
45_M_3	*GCKR*, leucine zipper transcription factor showing similarity to *SREBP-2*

Table 3.3 Sequence analysis of three partially sequenced BAC clones mapping to the *GCKR-KHK* interval. Sequencing was performed by the Washington University School of Medicine as part of the Human Genome Project. For full gene names and encoded protein function (if known), see Table 3.4.

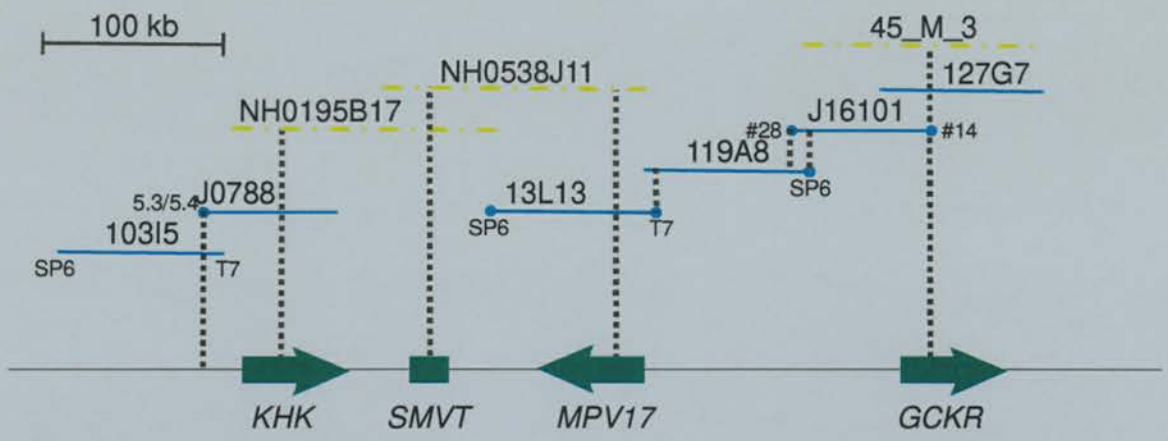


Figure 3.7 PAC and BAC clone contig spanning the *GCKR-KHK* genomic region. The contig was constructed by EST and STS content of the PAC clones, and sequence analysis of the BAC clones. PAC clones are indicated by a solid line; BAC clones are indicated by dashed/dotted line. STSs were designed from sequence at the ends of the PAC clones (named SP6 and T7). The position of the genes *KHK*, *SMVT*, *MPV17* and *GCKR* are indicated (arrow direction indicates 5′-3′ orientation of the gene).

3.3.2 Transcript and EST mapping

All the transcripts and ESTs that were mapped to the *GCKR-KHK* and DFNB9 interval physical contig are summarised in Table 3.4. This Table describes 14 known genes and 15 ESTs (including THC contigs assembled by "The Institute of Genome Research" (TIGR)). Brief descriptions of the encoded protein function are given (if known), including a reference (if the gene/EST is described elsewhere in this Thesis, a chapter/section number is given).

The transcripts and ESTs shown in Table 3.4 were all identified by either sequence analysis or EST PCR screening of the genomic clones used to construct the chromosome 2p23.3 physical contig. The construction of a highly detailed YAC, BAC, PAC and cosmid clone contig of the *GCKR-KHK* interval allowed all transcripts and ESTs to be accurately mapped. This detail allowed the relative order and orientation of many if the genes to ascertained. A combined physical contig and transcript map covering the *GCKR-KHK* interval is shown in Figure 3.8.

Gene/ EST	Protein encoded/ putative identification.	Other information
GCKR	Glucokinase regulatory protein.	Plays a role in sequestering glucokinase in the nuclei of hepatocytes. See Chapter 1,Section 1.6.
KHK	Ketohexokinase.	KHK phosphorylates fructose to fructose-1-phosphate. See Chapter 1, Section 1.7.
PPP1CB	Protein phosphatase 1, catalytic subunit, β isoform. One of three catalytic subunits of protein phosphatase 1 (PP1).	PP1 is one of 4 major serine/threonine-specific protein phosphatases involved in the dephosphorylation of a variety of proteins. See Chapter 1, Section 1.9.
EIF2B4	Guanine nucleotide exchange factor (delta subunit). A subunit of the heteropentameric guanine nucleotide exchange factor.	The eIF2B complex controls a key regulatory step in initiation during protein synthesis. See Chapter 4, Section 4.2.
KIAA0064	Otherwise unknown.	Sequence analysis reveals the putative translation product to contain a serine protease active site motif and PX domain. Intimately adjacent to the *EIF2B4* gene. See Chapter 4, Section 4.3.
RBSK	Ribokinase. Identified by its similarity to bacterial and yeast ribokinases.	Role is to phosphorylate ribose to ribose-5-phosphate. See Chapter 4, Section 4.4.
KIF3C	Kinesin-like gene.	Forms a heterodimer with KIF3A. Probable role in intracellular transport. See Chapter 7, Section 7.3.

Continued on next page.

Table 3.4 List of genes and ESTs mapping to the *GCKR-KHK* and *DFNB9* physical contig (Figures 3.5 and 3.8).

Gene/ EST	Protein encoded/ putative identification.	Other information
OTOF	The *DFNB9* gene. Shows similarity to the *C. elegans* spermatogenesis factor FER-1 and human dysferlin.	Thought to be involved in vesicle membrane fusion. See Chapter 7, Section 7.6.
MPV17	Peroxisomal protein. Recessive kidney and deafness gene in mouse.	Involved in regulation of cellular reactive oxygen species. See Chapter 7, Section 7.2.
UCN	Urocortin. A neuropeptide that is a member of the corticotropin-releasing hormone family.	Plays a role in the mammalian stress response and is related to fish urotensin and corticotropin-releasing factor (CRF). See Chapter 7, Section 7.2.4.2.
KCNK3 (*TASK*)	A member of the mammalian potassium ion channel family.	KCNK3 currents are very sensitive to small changes in extracellular pH, suggesting that TASK has a role in cellular responses to changes in extracellular pH. See Chapter 7, Section 7.5.
EMILIN	*E*lastin *m*icrofibril *i*nterface *l*ocated prote*in*.	An extracellular matrix glycoprotein.
FOSL2	FOS-like antigen 2. Encodes a leucine zipper protein that can dimerise with proteins of the JUN family, thereby forming the transcription factor complex AP-1.	Genbank accession no. X16706. Implicated as regulators of cell proliferation, differentiation, and transformation.
CGR11 (stSG2766)	Cell growth regulator 11. Protein contains an EF-hand domain.	Genbank accession no. U66468. Has been shown to be transcriptionally regulated by p53.
TFIIIC2 (Also known as KIAA0011)	General transcription factor IIIC2 (GTF3C2).	GTF3C family proteins are essential for RNA polymerase III to make a number of small nuclear and cytoplasmic RNAs, including 5S RNA, tRNA, and adenovirus-associated (VA) RNA of both cellular and viral origin.
WI-19785	Xfin like/Zinc finger protein.	Genbank accession no. T16385
stSG3685	Shows similarity to retinitis pigmentosa 3 (RP3).	Genbank accession no. H03977
stSG15582	Weakly similar to sodium iodide symporter.	Genbank accession no. R69136
stSG3204	KIAA0764	Genbank accession no. T15764
SGC32173	Shows similarity to an ATP(GTP) binding protein.	Genbank accession no. G25480

Continued on next page.

Table 3.4 (Continued from previous page). List of genes and ESTs mapping to the *GCKR-KHK* and *DFNB9* physical contig (Figures 3.5 and 3.8).

Gene/ EST	Protein encoded/ putative identification.	Other information
ZNT3	Zinc transporter gene. Solute carrier protein family 30, member 3; SLC30A3.	Genbank accession number NM_003459
CAD	Carbamoyl-phosphatase synthetase 2, aspartate transcarbamylase.	Encodes a trifunctional protein which is associated with the enzymatic activity of the first 3 enzymes in the 6-step pathway of pyrimidine biosynthesis: carbamoylphosphate synthetase, aspartate transcarbamoylase, and dihydroorotase.
SMVT	Sodium multi-vitamin transporter Solute carrier family 5, member 6; SLC5A6.	This protein may be involved in the entry of pantothenate, biotin, and lipoate in all cell types
WI-14636	Anonymous EST	Genbank accession no. G20997
WI-11296	Anonymous EST	Genbank accession no. G24562
WI-18604	Anonymous EST	Genbank accession no. H05599
stSG4391	Anonymous EST	Genbank accession no. R37380
EST1	THC 291225	SREBP-like (sterol regulatory element binding protein-2) leucine zipper transcription factor.
EST2	THC 257039	Anonymous EST

Table 3.4 (Continued from previous page). List of genes and ESTs mapping to the *GCKR-KHK* and *DFNB9* physical contig (Figures 3.5 and 3.8). The putative function of the encoded protein is indicated if known. THC: EST contig constructed by "The Institute for Genomic Research" (TIGR).

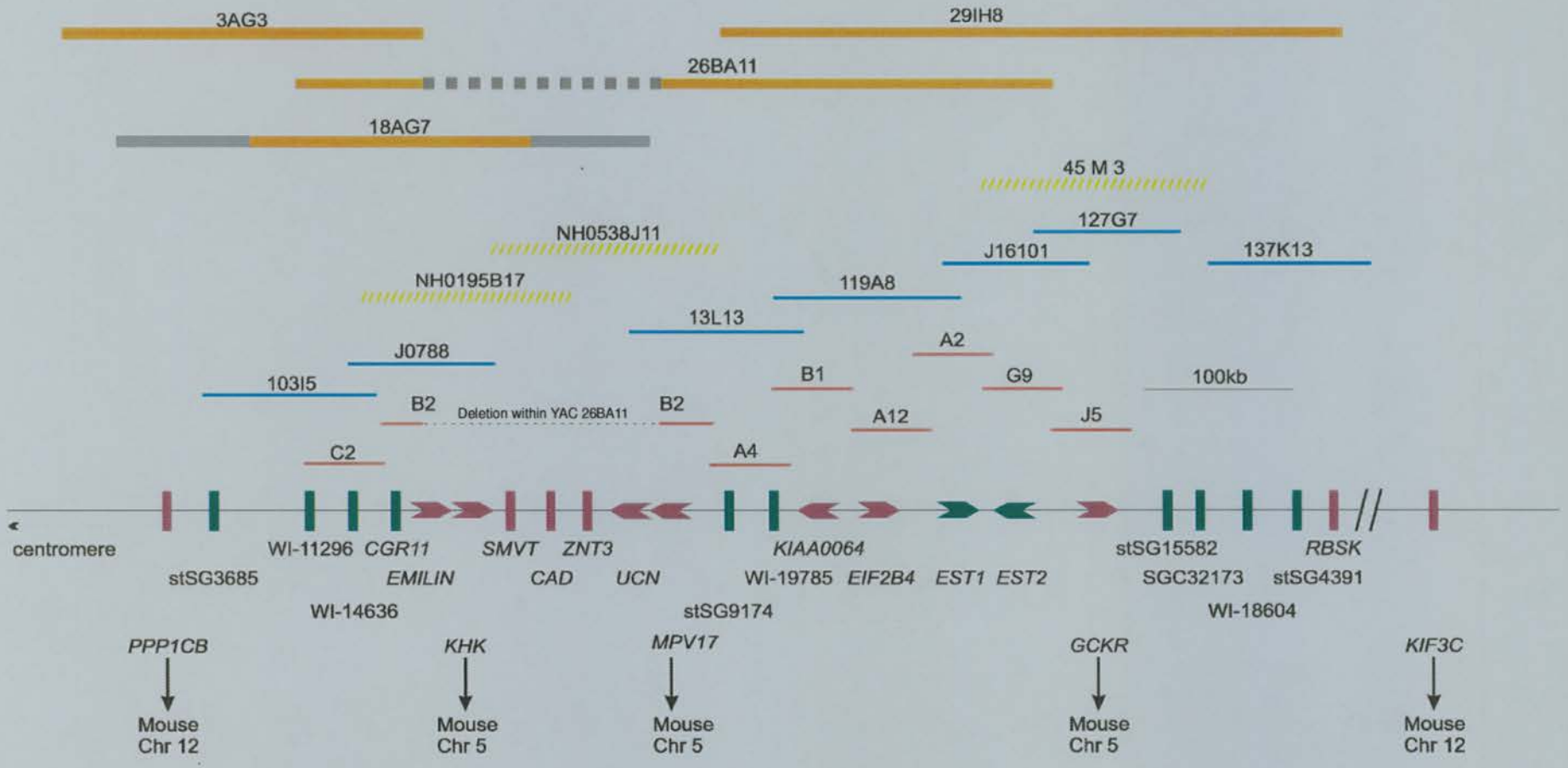


Figure 3.8 Physical contig and transcript map of *GCKR-KHK* genomic region. YACs, BACs, PACs, and cosmids are represented by orange, yellow (hatched), blue, and red lines, respectively. A dotted grey line represents a deletion within YAC 26BA11; a solid grey line represents chimeric regions of YAC 18AG7. Transcripts and ESTs mapping to the genomic clones are shown; characterised genes and ESTs are indicated by pink and green arrows/boxes, respectively (direction of arrow represents 5′-3′ gene orientation. Where known, the chromosomal mapping of the mouse gene homologue is indicated.

3.3.3 Cosmid *Eag*I restriction fragment analysis

As a preliminary search for CpG islands (performed as part of the search for transcripts), 12 cosmids were chosen for *Eag*I restriction analysis (all positive for the P1 clone J0788 large *Eag*I restriction fragment - see methods). All were found to contain identical internal *Eag*I restriction fragments (analysis of other randomly chosen cosmids revealed the presence of 2 out of 3 of these fragments). The *Eag*I restriction fragments produced from cosmid C2 (see Figure 3.9) were cloned and sequenced. Sequence analysis revealed two of these fragments (the smallest and largest) to be vector sequences. The third fragment was found to show significant similarity to a partially sequenced BAC clone and a CpG island clone (Figure 3.10 and Table 3.5).

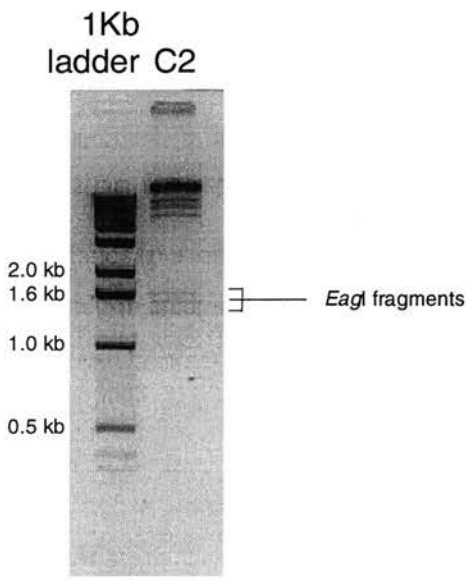


Figure 3.9 *Eag*I restriction digestion of cosmid C2.

```
C2 EagI frag.   4 gccgccgaccatcccaggtctaccgcacccaatcccccaggcttgcgagcgcttc 58
                  ||||||  |||||||||||||||||||  |||||||||||||||||||||||||||
NH0503P10 120972 gccgcccaccatcccaggtctaccgcagggaatcccccaggcttgcgagcgcttc 121026

C2 EagI frag. 59 cgagggcaggaaagcctgtactgcagccccgccgccctctcc 100
                  ||||||||||||||||||||||||||||||||||||||||
NH0503P10 121027 cgagggcaggaaagcctgtactgcagccccgccgccctctcc 121068
                  |||||||||||||||||||||||  |||||||||||||
33H5          216 gggcaggaaagcctgtactacagcnccgccgccctctcc    178
```

Figure 3.10 Sequence alignment of cosmid B8 *Eag*I fragment and sequences from BAC NH0503P10 (nucleotides 120972-121068) and clone 33H5 (nucleotides 216-178). See Table 3.5 for additional clone information.

Clone name	Genbank accession no.	Clone information
NH0503P10	AC013472	*Homo sapiens* chromosome 2 BAC clone RP11 NH0503P10, working draft sequence, 19 unordered pieces. Sequence analysis shows this clone to contain the STS WI-14636 and sequence showing similarity to dihydropyrmidase (Genbank accession no. Q09296).
33H5	Z60767	*Homo sapiens* CpG island DNA genomic *Mse*I fragment (216 bp).

Table 3.5 Clone information for Figure 3.10

Sequence analysis of the database sequence for BAC clone NH0503P10 revealed the presence of an EST sequence WI-14636, an EST already mapped by PCR screening to the physical contig on chromosome 2p23.3. The discovery that one of the cosmid *Eag*I fragments was identical to the CpG island genomic clone 33H5 sequence suggested that a transcript might reside nearby. To investigate whether a transcript could be associated with this CpG island sequence, further sequence analysis of the surrounding genomic region to the *Eag*I fragment is required. PCR screening for ESTs had previously mapped the *CGR11* gene to cosmid C2 therefore the CpG island sequence could be associated with the *CGR11* sequence. Although the method of identifying CpG islands by rare cutter restriction analysis of genomic clones had in this case identified a CpG island sequence, this experiment was not continued due to the success of searching for transcripts by direct sequence analysis and PCR screening using ESTs of genomic clones (described previously in this chapter).

3.4 Discussion

3.4.1 Physical mapping on chromosome 2p23.3

3.4.1.1 Chromosome 2p23.3 physical contig

The construction of a physical contig spanning the *GCKR-KHK* genomic region and DFNB9 interval is a valuable resource for the detailed examination of transcripts within the *GCKR-KHK* intergenic region and the identification of candidate deafness (*DFNB9*) genes. This chapter describes the construction of a YAC contig that spans the *GCKR-KHK* genomic region and DFNB9 interval, covering approximately 2 Mb on chromosome 2p23.3 (Figure 3.5). Also described is a more detailed physical contig that spans the *GCKR-KHK* region, consisting of YAC, BAC, PAC, and cosmid clones. The contigs were assembled using a combination of STS and EST content analysis and cosmid fingerprinting (Figures 3.6 and 3.7).

Although *GCKR* and *KHK* had been estimated by FISH to be <500 kb apart on chromosome 2p23.3 (Hayward et al., 1996), EST content analysis of YACs reveals none to contain both *GCKR* and *KHK*. STS analysis of YACs did allow *GCKR* and *KHK* to be physical linked, but the analysis of STSs revealed this original YAC contig to be internally inconsistent (see introduction, Section 3.1.2.1, Figure 3.1). The most likely explanation for this aberration was a deletion within YAC 26BA11 at the *KHK* locus.

The physical proximity of *GCKR* and *KHK*, estimated to be ~500 kb apart by FISH (Hayward et al., 1996),could not be refined from the YAC physical contig due to the deletion within YAC 26BA11. Initially, the size of the deletion within YAC 26BA11 was unknown. EST analysis of the YAC 26BA11 and the cosmid subclone library revealed that exons 1-5 of the *MPV17* gene were present but exons 6-8 were absent. As *KHK* is also absent from YAC 26BA11, the deletion in YAC 26BA11 is likely to span from within *MPV17* to within the *KHK*-containing P1 clone J0788 (see Figure 3.6).

For the identification of genomic clones that might span the gap within YAC 26BA11, the RPCI1 PAC library was screened. In total, 4 PAC clones and 2 P1 clones were identified to map at the *GCKR-KHK* interval. These included the clones 13L13 and J0788 that contain the genes *MPV17* and *KHK* respectively. Although contig construction suggests these two clones to partly contain some of the genomic DNA deleted in YAC 26BA11, STS and EST

content analysis of clones 13L13 and J0788 reveals these PAC clones to be non-overlapping. Further screening of PAC libraries did not identify any other PAC clones able to span the YAC 26BA11 deletion.

3.4.1.2 Identification of chromosome 2p23.3 BAC clones

As part of the Human Genome Project, partially sequenced BAC clone sequences have been released into the Genbank database. EST and STS screening of the Genbank database using sequences from *MPV17* and *KHK*, identified the BAC clones NH0195B17 and NH0538J11. Computer analysis of the database sequences for BACs NH0195B17 and NH0538J11 reveals them to contain *KHK* and *MPV17* respectively (*KHK* and *MPV17* are located on opposite sides of the YAC 26BA11 deletion). Further sequence analysis of the database sequences for these two BAC clones using the BLAST program reveals both BACs to contain the genes *SMVT* and *CAD*, thus indicating that these two BACs are overlapping and span the YAC 26BA11 deletion (see Figures 3.7 and 3.8). The complete YAC, BAC, PAC and cosmid contig confirms the physical proximity of *GCKR* and *KHK* that was first shown by FISH (Hayward et al., 1996). The physical distance between *GCKR* and *KHK* is estimated by FISH to be ~500 kb and the physical contig confirms this to be a good estimate of the *GCKR-KHK* interval size.

Further database searching using sequence from *GCKR* identified the chromosome 2 BAC clone 45_M_3. Although the genomic region surrounding *GCKR* had already been cloned in the form of YAC, PAC and cosmid clones (Figure 3.8), further analysis of the 45_M_3 BAC database sequence confirms the presence of an EST (similar to a sterol regulatory binding protein transcription factor SREBP-2). This EST had previously been identified by direct sequence analysis of the cosmid clones within the chromosome 2p23.3 physical contig (Table 3.4).

3.4.1.3 Orientation of *GCKR-KHK* contig to DFNB9 interval

The physical linkage between the *GCKR-KHK* locus and the DFNB9 interval was established by YAC 964D8, a CEPH mega-YAC originally known to contain the STSs D2S2350 and D2S1979 (Figure 3.5). However, initially it was unclear as to the correct orientation of *GCKR* and *KHK* with respect to the DFNB9 interval. This was due to a lack of markers mapping to YAC 964D8, and radiation hybrid mapping of both *KHK* and *GCKR* that was inconclusive as to the correct positional orientation of these two genes. Further STS analysis of the YAC 964D8 reveals it to also contain STSs D2S174, D2S165, D2S366,

D2S365, D2S1424, D2S2629, and AFM115YB6 but not D2S2247 (Figure 3.5). EST content analysis of YAC 964D8 reveals it to contain *OTOF*, *CENPA*, *KCNK3*, *FOSL2*, *RBSK*, and *GCKR* but not *ZNT3*, *CGR11* or *KHK* (Figure 3.5). This mapping data strongly suggests that the correct orientation of *KHK* and *GCKR* with respect to the centromere and the DFNB9 interval is centromere/*KHK*/*GCKR*/DFNB9. As a note of caution, examination of other YAC contigs on chromosome 2p23.3 such as the Whitehead 3.2 contig, reveal high frequencies of internal deletions in the chromosome 2p23.3 region. This phenomenon may be responsible for the lack of YACs that contain both *GCKR* and *KHK*. Therefore, as the orientation of the *GCKR* and *KHK* is based upon the STS and EST content of YAC 964D8, the existence of a deletion in this YAC might alter the *GCKR-KHK* orientation with respect to the DFNB9 interval. Although it is has been previously noticed that large scale YAC contig projects contain high frequencies of region specific internal deletions (Collins *et al.*, 1995), an explanation for this phenomenon has yet to be found.

3.4.2 Transcript mapping in chromosome 2p23.3

3.4.2.1 Transcript mapping

The chromosome 2p23.3 physical contig is an ideal tool for transcript searching by PCR screening for ESTs and direct sequencing of clones combined with computer analysis of sequences. In total, 14 known gene sequences and a further 15 ESTs were mapped to chromosome 2p23.3 (see Table 3.4). Within the *GCKR-KHK* interval (see Figure 3.8), 7 genes and 5 ESTs map to the physical contig. The transcript sequences that have been mapped to the cosmid contig are in their correct order relative to *GCKR* and *KHK*. However, the correct positional order of *CAD* and *SMVT* relative to *GCKR* and *KHK* is unknown and will require further sequencing of BAC clones NH0195B17 and NH0538J11 for its elucidation. Gene orientation is indicated on Figure 3.8 if known.

3.4.2.2 CpG Island mapping

To identify other transcripts at the *GCKR-KHK* locus, a preliminary search for CpG islands located within the chromosome 2p23.3 physical contig was performed by searching for rare cutter restriction enzyme sites. This experiment involved the cloning and sequencing of *Eag*I restriction fragments from cosmid clones within the chromosome 2p23.3 physical contig. As a preliminary experiment, cosmids positive for the P1 clone J0788 *Eag*I large fragment were chosen for *Eag*I restriction analysis. This was carried out to increase the

chances of identifying cosmids containing *Eag*I restriction sites (and would also aid cosmid contig construction).

Cosmids positive for the P1 clone J0788 *Eag*I large fragment, were all found to contain three *Eag*I restriction fragments. The cloning and sequencing of these three *Eag*I restriction fragments from the cosmid C2 revealed that two of the fragments were vector sequence. The third fragment showed significant similarity to a genomic CpG island clone 33H5 (Figure 3.10 and Table 3.5). This might indicate the presence of a CpG island on cosmid C2 and the presence of a transcript. Further sequencing of the surrounding region to the cloned *Eag*I fragment is required to investigate whether the CpG island sequence is associated with a transcript sequence. As cosmid C2 is known to contain *CGR11* (Figure 3.8), this CpG island sequence might be associated with this gene.

BLAST analysis of the C2 *Eag*I fragment also revealed significant similarity to sequence from the chromosome 2 BAC clone NH0503P10 (Figure 3.10 and Table 3.5). As the database sequence for BAC NH0503P10 is partial and made up of unordered contigs, it is difficult to accurately place this BAC clone onto the chromosome 2p23.3 physical contigs shown in Figures 3.5 and 3.8. However, further analysis of the BAC clone sequence using BLAST program reveals it to contain the chromosome 2 STS WI-14636, which also maps to YACs 18AG7 and 26BA11, and PAC J0788 (see Figure 3.8). This data places the BAC NH0503P10 close to *KHK* on the physical contig. Further analysis of the BAC NH0503P10 database sequence reveals it to also contain a sequence highly similar to the gene encoding dihydropyrmidase (Table 3.5). Further work is required to place this BAC accurately onto the *GCKR-KHK* physical contig and to identify other transcripts that it may contain. Although the search for CpG island-like sequences had proven successful, this experimental approach was not continued due to the success of transcript identification by direct sequencing and PCR screening of genomic clones.

3.4.2.3 Examination of transcripts

Characterised genes and ESTs that were identified to reside on chromosome 2p23.3 are described in Table 3.4. Initial inspection of the putative protein functions encoded by transcripts on chromosome 2p23.3 reveals a variety of protein functions with no obvious common theme. In addition to the candidate type 2 diabetes genes encoding the glucokinase regulator protein and ketohexokinase (discussed in Chapter 1), examples of other known genes residing on chromosome 2p23.3 include: the *MPV17* gene, of which the mouse homologue is a recessive kidney and deafness gene (see Chapter 5, Section 5.2); the

urocortin gene, which encodes a neuropeptide; *EMILIN*, which encodes a extracellular matrix protein; *TFIIIC2*, which encodes a RNA polymerase III general transcription factor; and *KCNK3*, a potassium ion channel gene.

The aim of the following chapters is to 1) identify and characterise transcripts encoding proteins that might be functionally related to GKRP and KHK or involved in biochemical pathways related to carbohydrate metabolism with a possible involvement in type 2 diabetes (see Chapter 4), and 2) identify candidate non-syndromic recessive sensorineural deafness DFNB9 genes (see Chapter 5). Examination of the encoded protein functions of known genes residing on Chromosome 2p23.3 reveals none to encode proteins that are obviously related to GKRP, KHK, or carbohydrate metabolism. However, many anonymous ESTs also map to this genomic region. Therefore, to search for genes that might have related functions to GKRP, KHK, or carbohydrate metabolism, the ESTs were examined and several chosen for investigation (see Chapter 4).

In the case of candidate *DFNB9* genes, *MPV17* is an obvious candidate deafness gene due to the mouse knock-out phenotype (see Chapter 5, Section 5.2). *KCNK3* is also a potential candidate deafness gene due to the important role of potassium ions within the inner ear (see Chapter 5, Section 5.5). These two genes plus a kinesin-like gene (*KIF3C*, see Chapter 5, section 5.3) that was also identified as a candidate deafness gene, are all discussed in Chapter 5.

The transcript map shown in Figure 3.8 does indicate this genomic region to be gene rich with the transcripts packed closely together. This suggests that promoter and possible enhancer sequences for some of these transcripts may be located within or near other transcripts. Added to the finding that the this relatively small *GCKR-KHK* genomic region on chromosome 2p23.3 shows conserved synteny with a small region of mouse chromosome 5 (see Chapter 2), this may provide circumstantial evidence that genes at the *GCKR-KHK* locus are evolutionary linked and that a common *cis*-acting regulatory element acting on transcripts at the *GCKR-KHK* genomic region might exist.

The physical contig spanning the *GCKR-KHK* genomic region and DFNB9 interval provides an ideal framework for transcript mapping, transcript characterisation, and for the investigation of genomic organisation on chromosome 2p23.3. The location of genes and ESTs is very important when considering their candidacy for involvement in human diseases especially if the disease gene maps to a specific genomic interval. The physical contig shown in Figure 3.5 defines the genomic region containing *GCKR*, *KHK*, and the DFNB9

interval on chromosome 2p23.3. Only genes and ESTs mapping to this genomic region were considered for further investigation in the following chapters (see Chapters 4 and 5).

Apart from EST mapping and gene investigation, the chromosome 2p23.3 physical contig and transcript map gives an interesting insight into genomic organisation and helps to define the size of the region of conserved synteny with mouse chromosome 5 (see Chapter 2). Furthermore, a gene for familial syndromic esophageal atresia has been recently mapped to chromosome 2p23-24 (Celli *et al.*, 2000). The physical contig and transcript map at chromosome 2p23.3 described in this chapter will be a useful tool for the identification of possible candidate genes that could underlie this life-threatening congenital anomaly.

Chapter 4

4 Characterisation of transcripts within the *GCKR-KHK* genomic region

4.1 Introduction

As part of a detailed examination of the *GCKR-KHK* genomic region on chromosome 2p23.3, this chapter describes the identification and investigation of transcripts encoding proteins with 1) possible functions related to the glucokinase regulatory protein (GKRP –is encoded by *GCKR*) and ketohexokinase (KHK), or 2) a role in biochemical pathways relating to carbohydrate metabolism. The investigation of transcripts that map to the *GCKR-KHK* genomic region will use the candidate gene approach to identify transcripts for further scrutiny.

The candidate gene approach is exemplified by both *GCKR* and *KHK* as good candidate type 2 diabetes genes. Their candidacy is based upon biochemical evidence that reveals an important inhibitory interaction of GKRP with glucokinase (GCK), and a modulation of GKRP inhibition on GCK by fructose-1-phosphate, a product of KHK (see Chapter 1, Figure 1.2). The intimate co-localisation of *GCKR* and *KHK* to a 500 kb region of chromosome 2p23.3 raises an intriguing possibility that other transcripts encoding proteins that are metabolically related to GCKR and KHK may also be located on this region of chromosome 2p23.3. To investigate this hypothesis, transcripts were mapped to a physical contig that spans the *GCKR-KHK* intergenic region (see Chapter 3, Figure 3.8), and as described in this chapter, several transcripts were chosen for further investigation.

The mapping of transcripts to the physical contig spanning the *GCKR-KHK* genomic region (described in Chapter 3), reveals this genomic region to be gene dense, with at least 7 genes and 5 ESTs mapping to the ~500 kb genomic region between *GCKR* and *KHK*. These transcripts are used in this chapter as the basis for an investigation into genes located at the *GCKR-KHK* genomic locus that encode proteins with possible functions related to GKRP, KHK, or carbohydrate metabolism. Further examination of the genes that map to the *GCKR-KHK* region shows only one gene, *EIF2B4*, to fit this criterion (see Section 4.2). However, a number of ESTs designed from anonymous cDNA clones also map to this genomic region (Chapter 3, Table 3.4). Therefore, the ESTs mapping to this genomic region were investigated. To prioritise the ESTs for investigation, sequence analysis was performed on the Genbank database cDNA clone sequences from which the ESTs were designed. The

cDNA clone sequence analysis may reveal a putative function of the encoded protein and provide useful information that can be used to decide which ESTs to choose for further investigation.

This chapter describes the investigation of three transcripts at the *GCKR-KHK* locus: 1) *EIF2B4*, encoding the delta subunit of initiation factor eIF2B – a protein complex that plays a key role in the regulation of protein synthesis; 2) *KIAA0064*, a putative gene arranged in an intimate "head to head" arrangement to *EIF2B4*; and 3) *RBSK*, encoding ribokinase - a member of the same kinase family as KHK. Ultimately, the aim of this research was to gather evidence for the possible involvement of the *GCKR-KHK* genomic region on chromosome 2p23.3 in the pathogenesis of type 2 diabetes.

4.2 The guanine nucleotide exchange factor delta subunit (*EIF2B4*)

4.2.1 Introduction

4.2.1.1 Identification of *EIF2B4*

The first putative transcript to be identified for further investigation was a sequence showing significant similarity to the *Saccharomyces cerevisiae* "general control" *GCD2* gene. This was identified by the sequencing of insert ends from cosmid subclones of YACs 29IH8 and 26BA11, two *GCKR*-containing YACs (see Chapter 3, Figure 3.8). Further analysis of the *GCD2*-like sequence reveals it to encode the delta subunit (eIF2Bδ) of the human guanine nucleotide exchange factor eIF2B, an initiation factor complex known to play a key role in protein synthesis initiation.

The *EIF2B4* transcript was chosen for further investigation because of previous research showing eIF2B activity to be stimulated by insulin , glucose, and sugar phosphates (Gilligan *et al.*, 1996; Singh & Wahba, 1995; Welsh *et al.*, 1997b). This suggested that eIF2B plays a major role in regulating protein synthesis in relation to cellular levels of both insulin and metabolites. Therefore, defective eIF2B function might contribute to the pathogenesis of type 2 diabetes. As the initial aim of this investigation was to identify and characterise transcripts on chromosome 2p23.3 that co-localise with the candidate type 2 diabetes genes *GCKR* and *KHK*, and which themselves might encode proteins that play a role in the pathogenesis of type 2 diabetes, *EIF2B4* represented such a candidate gene and warranted further investigation.

4.2.1.2 Eukaryotic protein synthesis

The translation of mRNA into protein is a complex multi-step process mediated by proteins termed translation factors. It can be divided into three phases: initiation, elongation and termination (reveiwed in Hershey, 1991). During initiation, the methionyl-tRNA and several initiation factors associate with the 40S ribosomal unit to form the 43S pre-initiation complex; this complex binds to mRNA and migrates to the correct AUG initiation codon, after which the addition of the 60S ribosomal subunit occurs. This is followed by elongation, during which amino acids from amino acyl-tRNAs are added to the growing peptide in the order dictated by the mRNA bound to the ribosome. The termination phase allows the completed protein to be released from the ribosome.

4.2.1.3 Identification of initiation factor genes

Initiation factor genes were first identified by gene expression studies in starving yeast, although at the time their regulatory function in translation was not realised. Yeast initiation factors are encoded by *GCN* and *GCD* genes, names derived from the role of these proteins in "**g**eneral amino acid **c**ontrol". In both yeast and mammals, starvation and stress result in a reduction in the rate of protein synthesis. Down-regulation of protein synthesis conserves cell resources and limits cell division under adverse growth conditions. The starvation of *Saccharomyces cerevisiae*, for example by growth of the yeast in media lacking an amino acid, results in de-repression of transcription of at least 40 different genes encoding amino acid biosynthetic enzymes in many unrelated metabolic pathways and this enables the cell to alleviate the nutrient starvation conditions. This response (called "general amino acid control") is mediated by an increase in levels of the transcriptional activator protein GCN4. In yeast, "general control" is mediated by both positive (GCN) and negative (GCD) regulatory factors (Greenberg *et al.*, 1986; Harashima & Hinnebusch, 1986; Hinnebusch, 1986).

The genes that encode GCN and GCD factors were identified by examining the effect of mutations in yeast genes on general control. The effect of mutations in yeast *GCN* and *GCD* genes could be measured by using reporter genes such as *GCN4*::*lacZ* fusion or *lacZ* fusions with biosynthetic enzyme genes such as *HIS4C* (*HIS4C* encodes histidinol dehydrogenase, which converts histidinol to histidine in the last step of histidine biosynthesis). Positive regulatory genes were identified by the effect of mutations that block de-repression of the reporter gene (*gcn* mutations e.g. *GCN2*), negative regulatory genes were identified by the effect of mutations which result in constitutive de-repression of the reporter gene (*gcd* mutations e.g. *GCD1* ,*GCD22*, *GCD6* and *GCD7*).

Various combinations of mutations in the *GCN* and *GCD* genes, used to explore interactions among the GCN and GCD factors, provide genetic evidence that some factors interact more closely. For example, deletion of *GCN3* is normally non-lethal and the double mutant *gcd6-1* and *gcd7-201* is non-lethal, but deletion of *GCN3* in the presence of *gcd6-1* and *gcd7-201* is lethal (Bushman *et al.*, 1993). These genetic experiments suggests that subgroups of the *GCN* and *GCD* genes may code for subunits of larger protein complexes. Biochemical proof that these protein complexes exist was provided later by co-purification of protein complexes through biochemical fractionation steps and co-immunoprecipitation using antibodies against the proteins of interest. For example, GCD1, GCD2 and GCN3 were

shown to be integral components of a high-molecular-weight complex of approximately 600 kDa (Cigan *et al.*, 1991).

The role of GCD and GCN factors in translation regulation was realised when research into mammalian protein synthesis using biochemical techniques identified several proteins involved in protein synthesis regulation that showed homology to the GCD and GCN factors. The study of translation in yeast using yeast mutants such as the temperature-sensitive lethal *gcd1-101* mutation (*EIF2B3*) confirmed the role of GCD and GCN factors in translation regulation. Following a shift to the non-permissive temperature, yeast with the *gcd1-101* mutation show reduced amounts of incorporation of radio-labelled amino acids into proteins and reduced amounts of a translation initiation intermediate containing charged initiator tRNA (Bushman *et al.*, 1993).

Research into regulation of eukaryotic protein synthesis using genetic and biochemical techniques has identified many initiation factors with important functions (see next section). Sequence analysis reveals many of the mammalian and yeast initiation factors are similar at both the amino acid and DNA level. Therefore, the cloning of *GCN* and *GCD* genes in yeast can be used to identify mammalian homologues by sequence comparison. The importance of studying translation regulation in yeast is also demonstrated by the many genetic experiments that have been performed to reveal the interaction of GCD and GCN factors. Although the exact interaction between initiation factors and the mechanisms by which translation is regulated have still to be fully elucidated, it is clear that this will require further research using both yeast and mammalian systems.

4.2.1.4 Translation initiation

In eukaryotic initiation, 12 initiation factors have been identified so far that are involved in the translation initiation process (eIF1A, eIF2, eIF2B, eIF2C, eIF3, eIF4A, eIF4B, eIF4E, eIF4F, eIF4G, eIF5 and eIF6). All perform important functions during translation initiation but one factor in particular, eIF2B, plays a key role in recycling of translation initiation factor 2 (eIF2) for translation initiation (Hershey, 1994; Proud, 1992; Samuel, 1993; Webb & Proud, 1997). A simplified overview of the role of eIF2B in translation initiation is shown in Figure 4.1.

The translation initiation process begins with the formation of a ternary complex comprising eIF2, GTP, and charged initiator $tRNA^{Met}$ and subsequent binding of this complex to the 40S ribosomal subunit. Binding of mRNA and recognition of the start codon (AUG) by initiator

$tRNA^{Met}$ leads to GTP hydrolysis and release of the functionally inactive eIF2-GDP binary complex. The exchange of GDP bound to eIF2 for GTP by eIF2B allows eIF2 to be recycled for further rounds of initiation (Figure 4.1). Regulation of translation initiation may occur at this point, so that any mechanisms that affect eIF2B activity could play a significant role in overall protein synthesis.

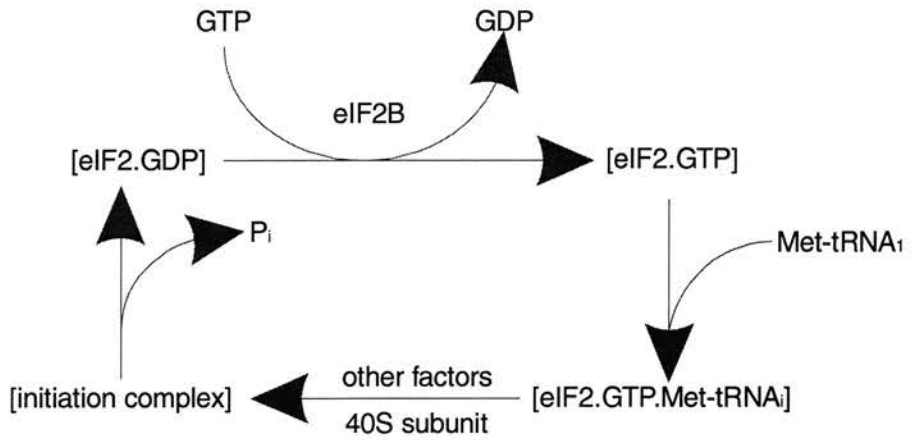


Figure 4.1 Role of eIF2B in translation initiation. eIF2B catalyses the exchange of GDP for GTP which re-activates eIF2, allowing further rounds of initiation.

4.2.1.5 The eIF2B complex

In mammals, analysis of the protein complex eIF2B reveals it to comprise five subunits: α, β, γ, δ, and ε (now known as subunits 1 to 5, respectively), and these subunits are the mammalian homologues of the yeast general control factors GCN3, GCD7, GCD1, GCD2 and GCD6, respectively. The yeast subunits all show considerable amino acid sequence similarity to their mammalian counterparts (α:40-42%, β:35-36%, γ:19%, δ:31%, ε:30%); (Kimball *et al.*, 1996; Price & Proud, 1994). The fact that all five subunits of yeast eIF2B were first identified as translational regulators of GCN4 strongly suggests that regulation of guanine nucleotide exchange on eIF2 is a key control point for translation in yeast cells just as in mammalian cells.

While other guanine nucleotide-exchange factors are generally monomeric, eIF2B is a much more complex protein, comprising five subunits. The reasons for this extra complexity may reflect the key role that eIF2B plays in regulation of general protein synthesis in which eIF2B interacts with its substrate eIF2 - itself a three subunit protein. It is now becoming clear that eIF2B activity is itself regulated by many different cellular factors. Therefore to fully understand eIF2B subunit function, the regulatory mechanisms acting on eIF2B must also be understood. A summary of eIF2 and eIF2B subunit functions is shown in Figure 4.2.

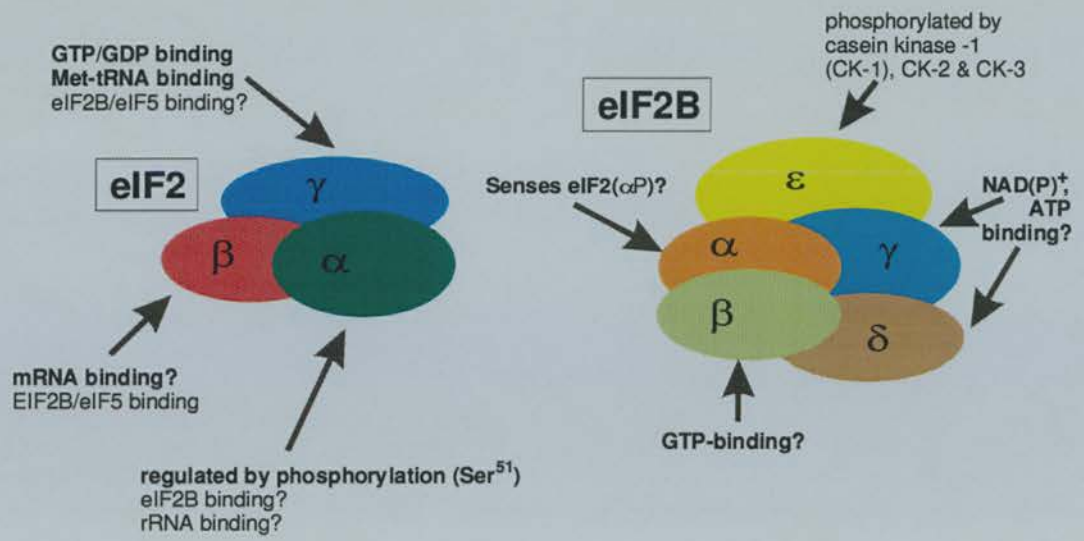


Figure 4.2 eIF2 and eIF2B – possible subunit functions and phosphorylation sites. The sizes of the subunits are roughly proportional to their actual molecular masses. It should be noted that the inter- and intra- subunit interactions are, as yet, not fully understood.

4.2.1.6 Regulatory mechanisms of eIF2B

4.2.1.6.1 Phosphorylation of eIF2α subunit

Regulatory mechanisms known to alter eIF2B activity include phosphorylation of residues in subunits of both eIF2B or eIF2 (see Figure 4.2). The best characterised of these is phosphorylation of the eIF2α subunit on Ser^{51}, blocking eIF2 recycling by competitively inhibiting eIF2B (Dholakia *et al.*, 1989; Gaspar *et al.*, 1994). In yeast, reduction in general protein synthesis during general control (which is coupled to an increase in expression of GCN4) is due to phosphorylation of eukaryotic initiation factor 2 (eIF2) by the eIF2α subunit kinase GCN2 (Cigan *et al.*, 1993). In mammals, the mechanism of eIF2α phosphorylation is used to reduce protein synthesis in response to haem-deficiency, viral infection, amino acid deficiency and certain stress conditions such as heat shock – see Figure 4.3. There are two kinases in mammalian cells (the haem-controlled repressor and the double-stranded RNA-activated inhibitor) that have eIF2α phosphorylation capabilities (Proud *et al.*, 1991). However, inhibition of peptide-chain initiation in skeletal (but not cardiac) muscle of diabetic rats which correlates with a reduction of eIF2B activity, shows no change in the level of eIF2α phosphorylation (Karinch *et al.*, 1993). This suggested the existence of other regulatory mechanisms.

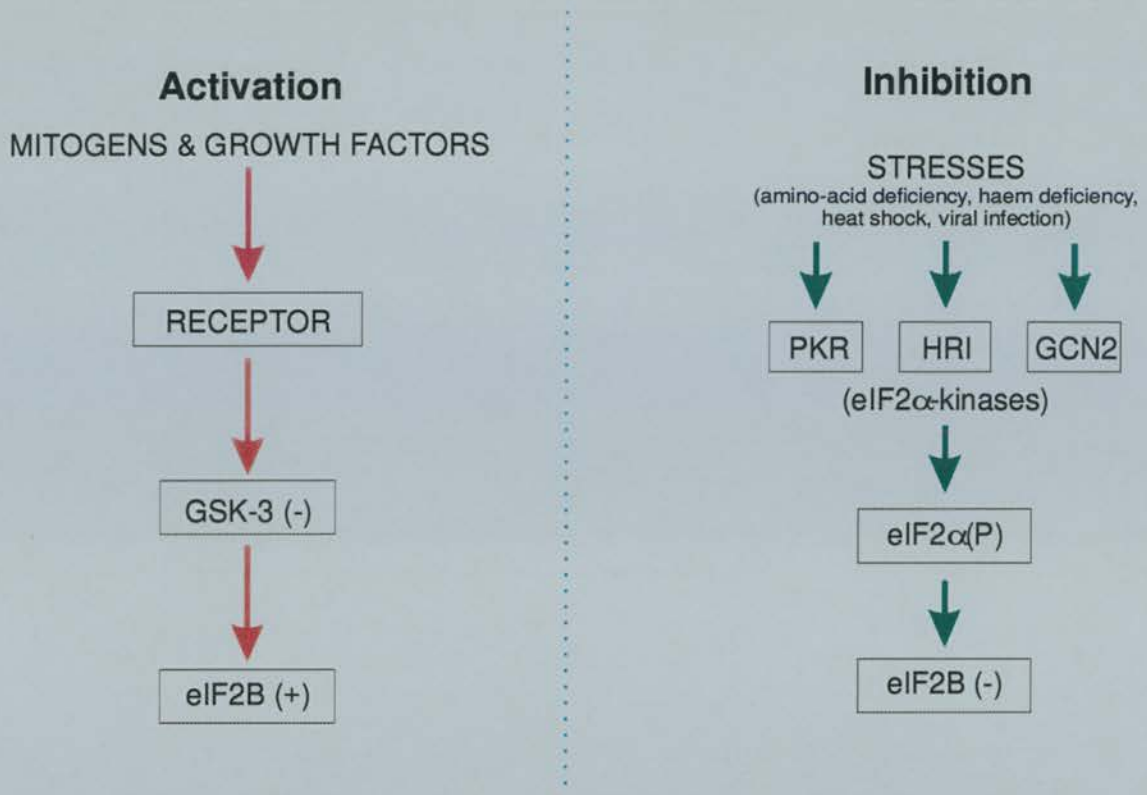


Figure 4.3 Possible mechanisms by which mitogens and growth factors may activate eIF2B, and by which cell stresses may inhibit this protein - (+) activation; (-) inhibition. Arrows only imply connections between the components shown, not the nature of the connection or whether it is direct. Abbreviations: GSK-3, glycogen synthase kinase-3; PKR, RNA-activated kinase; HRI, haemin-regulated inhibitor; GCN2, *Saccharomyces cerevisiae* eIF2α protein kinase ; eIF2α(P) phosphorylation of eIF2α subunit.

4.2.1.6.2 Phosphorylation of eIF2Bε subunit

Research into direct regulation of eIF2B activity reveals that phosphorylation of the eIF2Bε subunit (at the Ser^{540} residue) by glycogen synthase kinase-3 (GSK-3) which is an insulin sensitive protein kinase, results in inhibition of eIF2B GDP/GTP exchange activity, giving a regulatory role for this subunit (Singh *et al.*, 1996b). The likely mechanism by which insulin is thought to stimulate eIF2B activity is summarised in Figure 4.4.

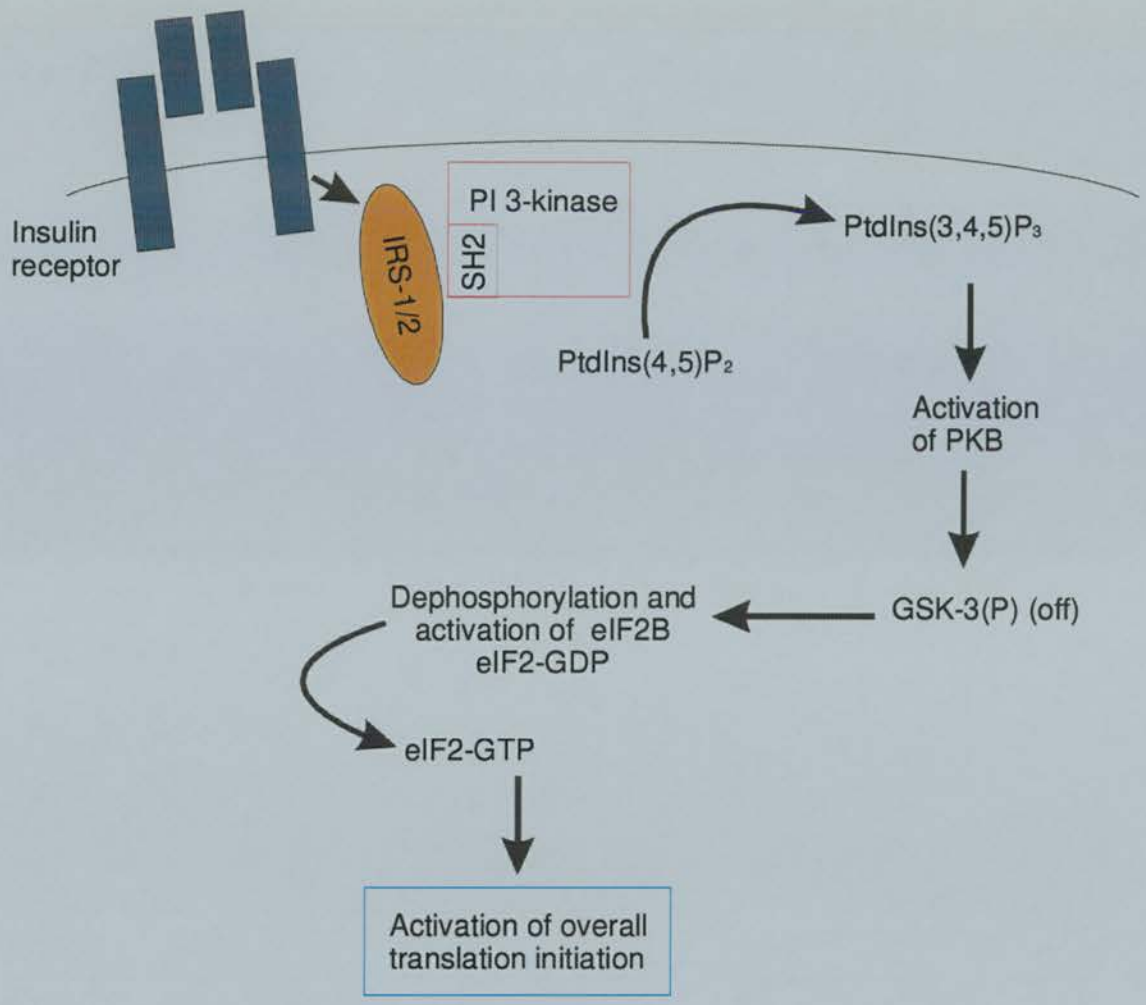


Figure 4.4 Overview of the signalling pathway likely to be involved in the regulation of eIF2B by insulin – adapted from Proud *et al*, 1997. Briefly, the binding of insulin to its receptor activates the intrinsic tyrosine kinase activity of the receptor, leading to binding and subsequent phosphorylation of the insulin receptor substrates (IRS-1/2). The resulting phosphotyrosine residues of the insulin receptor substrate include docking sites for the SH2 domains (regions of Src homology) of phosphatidylinositol 3-kinase (PI 3-kinase) and binding leads to activation of this enzyme. PI-3-kinase phosphorylates PtdIns(4,5)P_2 (phosphotidylinositol(4,5)diphosphate) to PtdIns(3,4,5)P_3 (phosphotidylinositol(3,4,5)triphosphate), and this molecule can either act directly or indirectly (by stimulation of other protein kinases) to activate protein kinase B (PKB). PKB in turn phosphorylates GSK-3 (glycogen synthase kinase-3) leading to its inactivation. Inactivation of GSK-3 leads to dephosphorylation of the eIF2Bε subunit and activation of eIB2B resulting in increased activation of peptide chain initiation by insulin. This cascade of phosphorylation/dephosphorylation of factors suggests a complex mechanism in which insulin regulates eIF2B activity.

4.2.1.6.3 Allosteric effectors

There is evidence for regulation of eIF2B by allosteric effectors. Both sugar phosphates and inositol phosphates have been shown to directly stimulate eIF2B by allosteric activation (Singh & Wahba, 1995). Other studies have shown the activity of purified mammalian eIF2B to be inhibited by NAD^+ and $NADP^+$ while NADH and NADPH stimulate activity (reviewed in Price & Proud, 1994). This might be explained by sequences in the γ and ε eIF2B subunits that show features also found in nucleotide binding enzymes, suggesting possible binding to adenine and guanine nucleotides and NADH/NADPH. In diabetic rats, eIF2B activity is reduced in skeletal muscle but not in cardiac muscle and this correlates with an increase in the ratio of $NADPH/NADP^+$ in heart but not in the skeletal muscle (Karinch *et al.*, 1993). This may indicate that NADPH plays an important role in maintaining eIF2B activity in the heart of diabetic patients and that in this case, phosphorylation of eIF2B subunits and other initiation factors like eIF2 is not involved. Also, photo-affinity labelling studies using 8-azidopurine nucleotides suggests that regions of the δ and γ subunits may constitute an ATP-binding domain (Rowlands *et al.*, 1988) and that the β subunit is involved in GTP binding (Dholakia *et al.*, 1989).

4.2.1.6.4 Other regulatory mechanisms

In isolated rat islets of Langerhans, exposure to elevated glucose concentrations activates eIF2B (Gilligan *et al.*, 1996) with no change in eIF2α phosphorylation (eIF2α phosphorylation inactivates eIF2B activity) and no inactivation of GSK-3 activity (GSK-3 phosphorylates eIF2Bε causing eIF2B inactivation). Whether eIF2B stimulation by glucose is by direct allosteric activation or by a novel mechanism has still to be elucidated.

4.2.1.7 Medical implications

The stimulation of eIF2B activity by glucose and insulin (Karinch *et al.*, 1993; Proud & Denton, 1997; Welsh & Proud, 1993) provides evidence that eIF2B is likely to play an important role in the cellular response to both insulin and glucose. The enhancement of eIF2B activity in response to high glucose levels may contribute to the activation of preproinsulin mRNA translation in the pancreatic islet. Therefore, insufficient eIF2B activity may lead to decreased insulin production by the pancreas and contribute to the pathogenesis of type 2 diabetes. As eIF2B activity is stimulated by insulin, defective eIF2B function may also contribute to the pathogenesis of type 2 diabetes in terms of insulin action, for example if hepatocytes are unable to produce adequate amounts of proteins and enzymes

involved in glucose storage. Furthermore, the signalling pathways that mediate eIF2B activity, for example the regulatory mechanism shown in Figure 4.4 in which insulin stimulates eIF2B activity, represent potential drug targets for disease states like diabetes.

4.2.1.8 Experimental aims

As part of the characterisation of the genomic region on chromosome 2p23.3, this section describes the identification and cloning of the *EIF2B4* gene that is located within the *GCKR-KHK* intergenic region. Sequence analysis of *EIF2B4*, characterisation of genomic structure and investigation into transcript expression is performed. These results have provided new information concerning the generation of different eIF2Bδ isoforms. The potential role of *EIF2B4* in the pathogenesis of type 2 diabetes and the significance of the co-localisation of *EIF2B4* with the candidate type 2 diabetes genes *GCKR* and *KHK* is discussed.

4.2.2 Methods

4.2.2.1 Isolation of genomic and cDNA Clones

The *EIF2B4* transcript was first identified by analysis of sequences produced by the sequencing of cosmid insert ends from the cosmid subclone library created from YAC 29IH8, a YAC known to contain *GCKR* (see Chapter 3). Sequencing using the SC2 primer (5′-dTGGAAGTCAACAAAAAGCAGAGC-3′) on cosmid "G4" reveals a sequence showing significant similarity to the *S. cerevisiae GCD2* gene. Further analysis shows this cosmid sequence to be identical to several human cDNA clones containing parts of the human *EIF2B4* gene, including EST H45644 (see Figure 4.5 for alignment between G4-SC2 cosmid sequence and EST H45644 - sequences are anti-sense). The GAP alignment shown in Figure 4.5 reveals that the G4-SC2 cosmid sequence from nucleotides 1 to 121 is almost identical to the cDNA sequence from H45644. This suggests that the G4-SC2 sequence starts within an exon and extends into intronic sequence. Indeed examination of the boundary between matching and non-matching sequence reveals an exon-intron boundary consensus sequence between nucleotides 119 and 120 of the cosmid G4-SC2 sequence.

```
H45644   101 CTCCGTGATCACCAGATCCACAAGCTCTGGGGGAGTCACATCATAGACTA 150
                             ||||||||||||||||||||||||||||||||
G4-SC2      1 ..................CACAAGCTCTGGGGGAGTCACATCATAGACTA 32

H45644   151 GATTCAACAACCGTAGGAGTGCGTGGTTCTGCCAGTTAGCCAGCGCAACA 200
             |||||||||||||||||  ||||||||||||||| |||||||| |||||
G4-SC2    33 GATTCAACAACCGTAGGGATGCGTGGTTCTGCCACTTAGCCAGCTCAACA 82.

H45644   201 TGTTCTCCCCGCTTACATTGCAGATCATCAGGGTCATCTAGCTCATTAGA 250
             |||||||||||||||||||||||||||||||||||||||         |
G4-SC2    83 TGTTCTCCCCGCTTACATTGCAGATCATCAGGGTCATCT.........GC 123

H45644   251 GACAAAGGCATCAGTCTGCACACGCTCACAGAGNCTTGTATGTTTCACAG 300
              |   | |    |  |   |    ||  |||| :      | |     |
G4-SC2   124 AATGGAAGGCGTACCCATTATGTTCTTTCAGAAAAGAAGTTCTAGTTTAT 173

H45644   301 CAAACCAGACACTGGTACATTATGGGGCTCGAGCCACCAGGGGCTAACTG 350
             |   ||     |       | | |       ||  | | || |    ||
G4-SC2   174 CCGCCCCTCTCCCTCCCTCTCAAGTCTTCACAGGTAGCTGGAGGCTTCTC 223

H45644   351 TGCTGTCCCTACCCGTGACATCACAGACCCATTGGGCCAAGAGTGCATGA 400
             |  || | |    | |  |     |||  ||  ||    ||  ||||
G4-SC2   224 TTTTGCCTCCTTACCTAGCTCATTAGAGACAAAGGCATCAGTCTGCACAC

H45644   401 GCTCCCAATAGNCACCTTNGG 421
             |||
G4-SC2   274 GCT.................. 276
```

Figure 4.5 GAP alignment of EST H45644 (upper sequence) with cosmid sequence G4-SC2 (lower sequence) – sequences are shown as anti-sense. An exon-intron boundary is shown at nucleotides 119-120 (cosmid sequence).

Cloning of the *EIF2B4* transcript (see following sections) reveals it to include the EST WI-12589. To obtain other genomic clones that contain *EIF2B4*, PCR screening using WI-12589 (WI-12589A: 5′-dAGGGAGTATGGCATTTATTAACC-3′); WI-12589B: 5′-dTAGTCTATGATGTGACTCCCCC-3′) was performed on the YACs that map to the *GCKR-KHK* genomic region (3AG3, 18AG7, 26BA11, and 29IH8 – see Chapter 3, Figure 3.8) and the cosmid subclone libraries created from YACs 29IH8 and 26BA11.

4.2.2.2 Sequencing of cDNA clones

To clone *EIF2B4*, overlapping cDNA clones containing parts of the *EIF2B4* sequence were identified by BLAST searching using the cosmid G4-SC2 sequence. The human cDNA clones giving significant similarity scores to the G4-SC2 sequence were obtained from the IMAGE consortium (IMAGE clone numbers 176330, 632738, 357973, and 173249) and sequenced on both DNA strands. Sequence analysis reveals these cDNA clones to all overlap and contain poly(A) tails. The longest cDNA sequence was 573 bp from clone H45644 and this included the poly(A) tail. The poly(A) tail indicates that the sequence obtained is from the 3' end of the gene and that clones containing the 5' end were still to be identified.

To identify cDNA clones containing the 5' end of *EIF2B4*, searching of "The Institute of Genomic Research" (TIGR) human gene index (HGI) database of overlapping cDNA clones (THC contigs) using the cDNA clone 176330 sequence was performed. This reveals the 176330 sequence to be part of a larger cDNA clone contig corresponding to THC180051. This contig spans 1639 bp of DNA and consists of 25 cDNA clones. To complete the *EIF2B4* gene sequence, cDNA clones were chosen from THC180051 that would span the whole 1639 bp (Figure 4.6).

The overlapping *EIF2B4* cDNA clones (IMAGE clone numbers 512237, 365427, 274777, and 380606) were obtained from the IMAGE consortium. The cDNA clones were "mini-prepped" using the alkaline lysis technique and the clone inserts sequenced on both strands using the Thermosequenase cycle sequencing kit (Amersham). Primers used for sequencing were the universal primers M13F (5′-dGTTTTCCCAGTCACGAC-3′) and M13R (5′-dCAGGAAACAGCTATGAC-3′) and internal insert primers designed from sequence obtained from the M13F and M13R sequences.

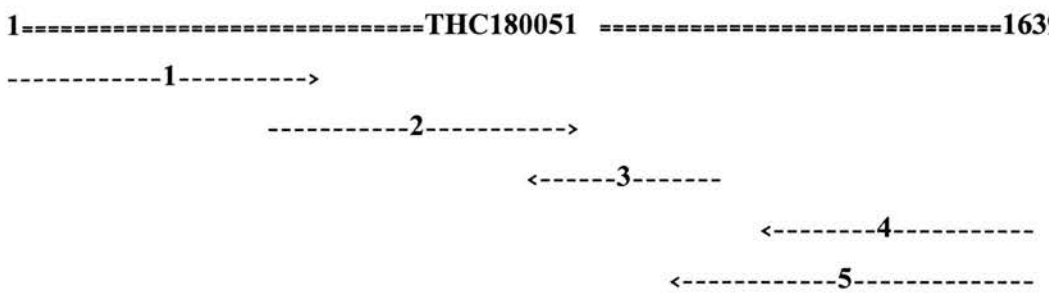


Figure 4.6 cDNA clone contig for *EIF2B4* (see Table 4.1 for clone identification). THC180051 is a group of overlapping cDNA clones constructed by "TIGR".

Clone no.	IMAGE clone number	Genbank database sequence		cDNA library
		Left (nt)	Right (nt)	
1	176330	20	434	Adult brain N2b5HB55Y, Soares
2	512237	394	871	Corneal stroma, Stratagene
3	365427	821	1109	Fetal heart NbHH19W, Soares
4	274777	1214	1639	Retina N2b4HR, Soares
5	380606	1097	1639	Retina N2b4HR, Soares

Table 4.1 cDNA clone IMAGE clone numbers for Figure 4.6

4.2.2.3 Structural analysis

To determine the *EIF2B4* genomic structure, oligonucleotides were designed from the cDNA sequence and used to obtain sequence data across the exon/intron boundaries from the genomic cosmid clones G4 and A4, both of which contained the *EIF2B4* gene. Sizes of introns were determined by amplification across each intron using cosmid genomic DNA as template and primers within adjacent exons. The fragments were analysed on a 1·5 % agarose gel and sized against a Gibco BRL 1 kb ladder. As most introns were found to be small, their exact sizes were determined by sequencing across the whole intron.

4.2.2.4 Alternative *EIF2B4* splice forms

The presence of alternative *EIF2B4* splice forms was investigated by RT-PCR (see Methods in Chapter 6). RT-PCR was performed using human total RNA as template from brain, heart, and muscle. PCR was also performed on a HepG2 cDNA (created from hepatocyte RNA). The primers used were designed from the *EIF2B4* cDNA sequence and are described in the Results section (see Section 4.2.3.5, Table 4.2). PCR products were excised from agarose gel, the DNA purified and sequenced using the Thermosequenase cycle sequencing kit (Amersham).

4.2.3 Results

4.2.3.1 Physical mapping of *EIF2B4*

Sequencing of cosmid G4 using primer SC2 reveals it to contain *EIF2B4*. This cosmid is a subclone produced from the *GCKR*-containing YAC 29IH8. YAC 29IH8 has previously been mapped by FISH to chromosome 2p23.3 (Warner *et al.*, 1995). Therefore, the mapping of *EIF2B4* to YAC 29IH8 indicates *EIF2B4* also to map to chromosome 2p23.3. Analysis of the *EIF2B4* sequence reveals it also to contain the EST WI-12589. PCR screening of YACs mapping to 2p23.3 using WI-12589 reveals that both YACs 29IH8 and 26BA11 contain *EIF2B4*. PCR screening of the cosmid subclone libraries created from both YACs 29IH8 and 26BA11 using WI-12589 reveals several positive cosmid clones (A4, A8, B1, D3, F3, G6, G7 and H6). Cosmid fingerprinting and STS content analysis had previously been used to assemble the cosmid subclone libraries into a cosmid contig and this indicates cosmid A4 to map between *GCKR* and *KHK* approximately 150 kb from *GCKR* (Figure 4.7).

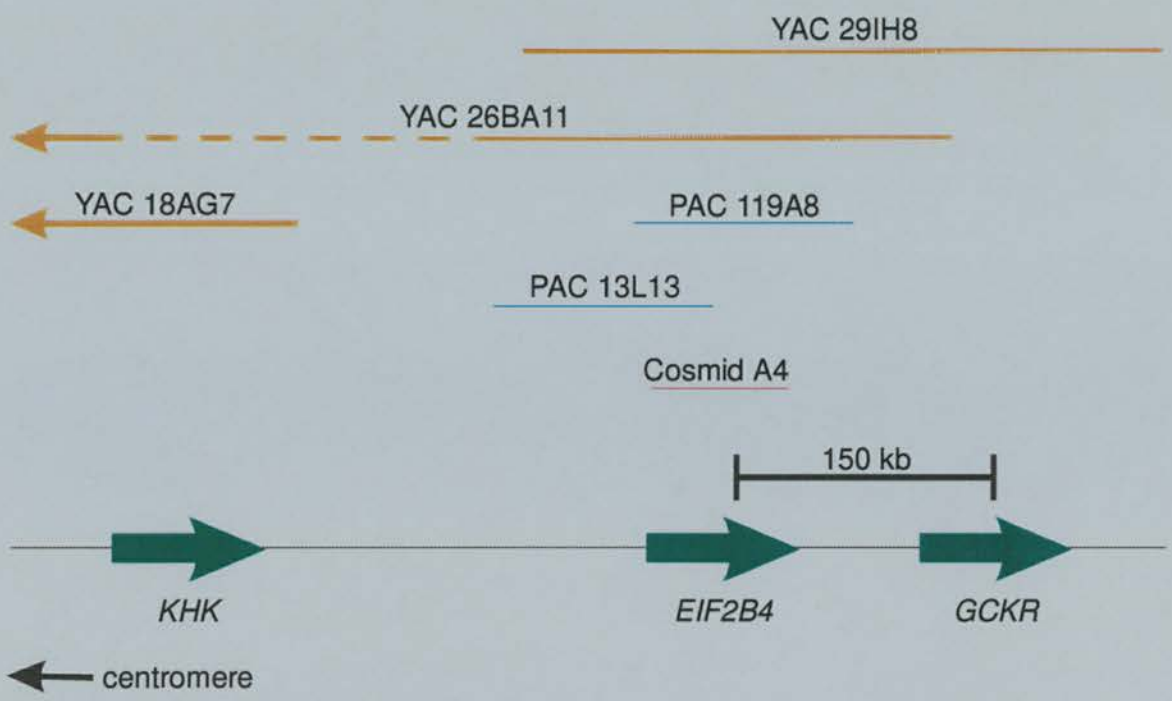


Figure 4.7 Co-localisation of *EIF2B4* with *GCKR* and *KHK*. The arrows represent genes, gene orientation (5′to 3′) is indicated by the arrow direction. Genomic clones containing *EIF2B4* are indicated above the *EIF2B4* arrow. This diagram is a simplified version of Figure 3.8 shown in Chapter 3.

4.2.3.2 *EIF2B4* cDNA sequence

Once the cDNA clones 176330, 512237, 365427, 274777, and 380606 (see Figure 4.6) had been sequenced, their sequences were aligned and assembled into an overlapping contig. Initially, alignment of the cDNA clone sequences did not reveal an open reading frame. However, comparison of the cDNA sequences from clones 380606 and 274777 with homologous mouse, rat, and rabbit *EIF2B4* cDNAs reveals a 4 bp deletion in the human cDNA (Figure 4.8A; see also Figure 4.9iii, the 4 bp deletion corresponds to nucleotides 45-48), which later proved to be situated at the position of the second splice junction sequence (Figure 4.12, the 4 bp deletion corresponds to nucleotides 1074-1077).

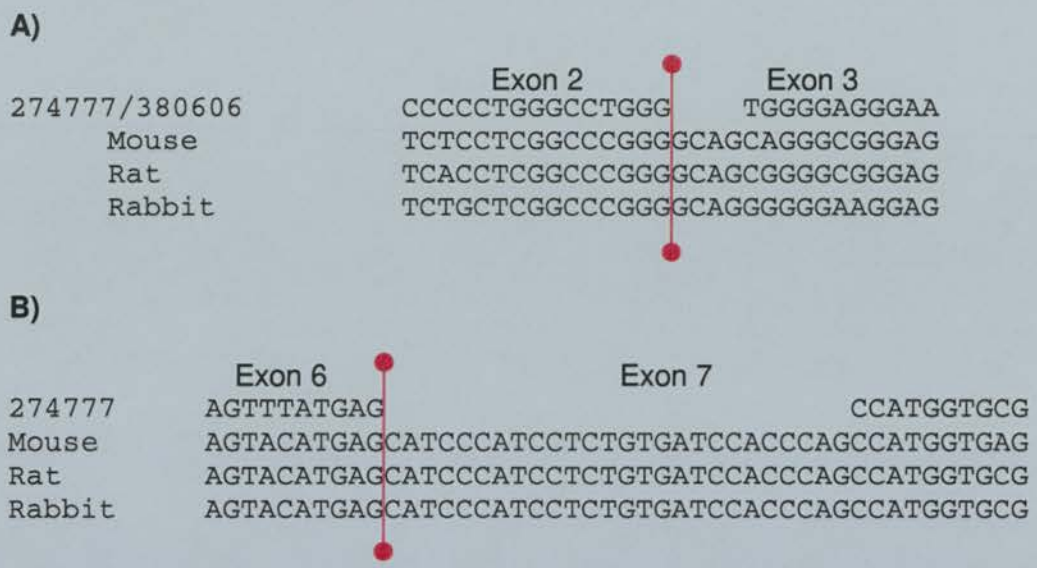


Figure 4.8 Comparison of human cDNA clone sequences 274777 and 380606 with the corresponding regions of *EIF2B4* sequences from mouse, rat and rabbit (Genbank accession numbers M98036, Z48225, and X75451 respectively). A) a 4 bp deletion is found in cDNA clone 274777; B) a 26 bp deletion is found in cDNA clones 274777 and 380606. The red vertical bars represent the positions of splice junctions, as subsequently determined.

To obtain a complete cDNA sequence, an RT-PCR product spanning the region in the cDNA clone sequence containing the 4 bp deletion was generated from human fetal kidney RNA. Its sequence includes the missing 4 bp, thus aligning perfectly with other species' sequences. Examination of the cosmid genomic sequence at the IVS2 acceptor site reveals a cryptic GCAG acceptor site as the first 4 nucleotides of exon 3 (Figure 4.12, nucleotides 1074-1077; see also Figure 4.13).

Similarly, comparison of the cDNA clone 274777 sequence to the corresponding *EIF2B4* cDNA sequences for mouse, rat and rabbit revealed clone 274777 to contain a 26 bp deletion (Figure 4.8B; see also Figure 4.9iii, the 26 bp deletion corresponds to nucleotides 557-582). This deletion later proved to be situated at the IVS6 acceptor site (Figure 4.12, the 26 bp deletion corresponds to nucleotides 2484-2509), and examination of the genomic sequence reveals a cryptic CCAG acceptor site within exon 7 positioned exactly before the deletion splice site (Figure 4.12, nucleotides 2506-2509).

The complete composite cDNA sequence for the cDNA clones described in Figure 4.6 (TIGR cDNA clone contig THC180051) is shown in Figures 4.9i and 4.9iii. It is 1643 bp long, consisting of an open reading frame of 1569 bp, 19 bp of 5' non-coding sequence and 75 bp of 3' non-coding sequence, with a consensus poly(A) addition signal (Figure 4.9iii, nucleotides 1557 to 1562). The 522 amino acid predicted translation product (Figures 4.9i and 4.9iii) has a molecular weight of 57478 Da.

It is known that the mouse eIF2Bδ exists as 2 different isoforms (Henderson *et al.*, 1994) – the "short" isoform (Genbank accession number M98036) and the "long" isoform (Genbank accession number M98035). Both mouse isoforms share 1577 bp of identical sequence toward the 3' end and only differ in their 5' ends. The TIGR contig THC180051 (Figure 4.6) contains the human equivalent to the mouse short isoform and none of the cDNA clones in the contig contain the human equivalent of the long isoform. Searching the human EMBL sequence databases with the mouse long isoform 5' end did not reveal any cDNA clones showing homology to the alternative 5' end.

However, the comparison of the mouse *EIF2B4* long isoform with the human genomic *EIF2B4* sequence reveals that the human alternative 5' exon sequence did indeed exist and is situated in the first intron, adjacent to the second exon (Figure 4.12, nucleotides 572-704). Although none of the human cDNA clones that were sequenced contain this alternative 5' exon, RT-PCR performed on heart and brain mRNA using a primer designed from sequence within the alternative first exon and another primer within exon 12 (primers DL5 and D2 respectively, see Section 4.2.3.5, Table 4.2), produces a RT-PCR product of the correct theoretical size (Figure 4.14). Sequencing of this RT-PCR product proved that the human alternative 5′ exon (the "long" isoform) was transcribed in both heart and brain. The complete human cDNA sequence for the "long" isoform by comparison with the mouse long isoform is 1708 bp long, consisting of an open reading frame of 1632 bp, 39 bp of 5' non-coding sequence with the 3' end (exons 2-13) identical to that of the short isoform. The 543

amino acid predicted translation product has a molecular weight of 59622 Da (Figure 4.9ii and 4.9iii). The short and long isoform cDNA sequences have been submitted to the EMBL sequence database and assigned accession numbers AJ011305 and AJ011306 respectively.

i) Short Isoform 5' end (exon 1a)

```
                                                    gagcctaggactgagggcg 19

 20 ATGGCTGCTGTGGCCGTGGCTGTTCGCGAGG                                      50
  1 M  A  A  V  A  V  A  V  R  E  D
```

ii) Long Isoform 5' end (exon1b)

```
                                tccctggctcgcctcgcgcctgccggcggatccgtggtc 37

 38 ATGCCAACCCAGCAGCCGGCTGCGCCGAGTACTCGTGCCCCCAAACCCTCCCGGAGTCTC 97
  1 M  P  T  Q  Q  P  A  A  P  S  T  R  A  P  K  P  S  R  S  L

 98 TCTGGCTCACTTTGTGCCCTGTTTTCTGATGCAG                            131
 21 S  G  S  L  C  A  L  F  S  D  A  D
```

iii) 3' end (exons 2-13)-continued on next page

```
                                          ACTCGGGATCCGGGATGAAGGCGGAGCTT 29
  1                                         S  G  S  G  M  K  A  E  L

 30 CCCCCTGGGCCTGGGGCAGTGGGGAGGGAAATGACCAAAGAAGAAAAGCTGCAGCTTCGG 89
 10 P  P  G  P  G  A  V  G  R  E  M  T  K  E  E  K  L  Q  L  R

 90 AAGGAAAAGAAACAGCAGAAGAAGAAACGGAAGGAAGAAAAGGGGGCAGAACCAGAGACT 149
 30 K  E  K  K  Q  Q  K  K  K  R  K  E  E  K  G  A  E  P  E  T

150 GGCTCTGCTGTATCTGCAGCCCAATGTCAAGGCCCAACCAGAGAACTGCCAGAATCGGGC 209
 50 G  S  A  V  S  A  A  Q  C  Q  G  P  T  R  E  L  P  E  S  G

210 ATTCAGTTGGGCACTCCTCGGGAGAAAGTTCCAGCTGGTCGGAGTAAGGCCGAACTTCGG 269
 70 I  Q  L  G  T  P  R  E  K  V  P  A  G  R  S  K  A  E  L  R

270 GCTGAGCGTCGAGCCAAGCAGGAGGCCGAGCGGGCCCTGAAACAGGCAAGAAAAGGGGAA 329
 90 A  E  R  R  A  K  Q  E  A  E  R  A  L  K  Q  A  R  K  G  E

330 CAAGGAGGACCACCTCCTAAGGCCAGCCCCAGCACAGCTGGAGAAACCCCCTCAGGAGTG 379
110 Q  G  G  P  P  P  K  A  S  P  S  T  A  G  E  T  P  S  G  V

380 AAGCGTCTCCCTGAGTACCCTCAGGTTGATGACCTACTTCTGAGAAGGCTTGTTAAAAAA 449
130 K  R  L  P  E  Y  P  Q  V  D  D  L  L  L  R  R  L  V  K  K

450 CCAGAGCGTCAACAGGTTCCTACACGAAAGGATTATGGATCCAAAGTCAGTCTCTTCTCT 509
150 P  E  R  Q  Q  V  P  T  R  K  D  Y  G  S  K  V  S  L  F  S

510 CACCTACCCCAGTACAGCAGACAAAACTCTCTGACCCAGTTTATGAGCATCCCATCCTCT 569
170 H  L  P  Q  Y  S  R  Q  N  S  L  T  Q  F  M  S  I  P  S  S

570 GTGATCCACCCAGCCATGGTGCGACTCGGCCTGCAGTACTCCCAGGGCCTGGTCAGTGGC 629
190 V  I  H  P  A  M  V  R  L  G  L  Q  Y  S  Q  G  L  V  S  G
```

Continued on next page.

Figure 4.9 Human *EIF2B4* cDNA sequence and putative translation product. **i)** short 5′ exon 1A, start codon located at nt 20-22, **ii)** long 5′ exon 1B, start codon located at nt 38 to 40 and **iii)** 3′ end - the 4 bp deletion in cDNA clone 274777 corresponds to nucleotides 45-48. The 26 bp deletion in cDNA clones 274777 and 380606 corresponds to nucleotides 557-582.

iii) 3' end (exons 2-13) - continued from previous page

```
 630 TCCAATGCCCGGTGTATTGCCCTGCTTCGTGCCTTGCAGCAGGTGATTCAGGATTACACA 689
 210 S  N  A  R  C  I  A  L  L  R  A  L  Q  Q  V  I  Q  D  Y  T

 690 ACACCGCCTAATGAAGAACTCTCCAGGGATCTAGTGAATAAACTAAAACCCTACATGAGC 749
 230 T  P  P  N  E  E  L  S  R  D  L  V  N  K  L  K  P  Y  M  S

 750 TTCCTGACTCAGTGCCGTCCCCTGTCAGCGAGCATGCACAACGCCATCAAGTTCCTTAAC 809
 250 F  L  T  Q  C  R  P  L  S  A  S  M  H  N  A  I  K  F  L  N

 810 AAGGAAATCACCAGTGTGGGCAGTTCCAAGCGGGAAGAGGAGGCCAAGTCAGAACTTCGA 869
 270 K  E  I  T  S  V  G  S  S  K  R  E  E  E  A  K  S  E  L  R

 870 GCAGCCATTGATCGGTATGTGCAAGAGAAGATTGTGCTAGCAGCTCAGGCAATTTCACGC 929
 290 A  A  I  D  R  Y  V  Q  E  K  I  V  L  A  A  Q  A  I  S  R

 930 TTTGCTTACCAGAAGATCAGTAATGGAGATGTGATCCTGGTATATGGATGCTCATCTCTG 989
 310 F  A  Y  Q  K  I  S  N  G  D  V  I  L  V  Y  G  C  S  S  L

 990 GTATCACGAATTCTTCAGGAGGCTTGGACAGAGGGCCGGCGGTTTCGGGTGGTAGTGGTG 1049
 330 V  S  R  I  L  Q  E  A  W  T  E  G  R  R  F  R  V  V  V  V

1050 GACAGCCGGCCATGGCTGGAAGGAAGGCACACACTACGTTCTCTAGTCCATGCTGGTGTC 1109
 350 D  S  R  P  W  L  E  G  R  H  T  L  R  S  L  V  H  A  G  V

1110 CCAGCCTCCTACCTGCTGATTCCTGCAGCCTCCTATGTGCTCCCAGAGGTTTCCAAGGTG 1169
 370 P  A  S  Y  L  L  I  P  A  A  S  Y  V  L  P  E  V  S  K  V

1170 CTATTGGGAGCTCATGCACTCTTGGCCAATGGGTCTGTGATGTCACGGGTAGGGACAGCA 1229
 390 L  L  G  A  H  A  L  L  A  N  G  S  V  M  S  R  V  G  T  A

1230 CAGTTAGCCCTGGTGGCTCGAGCCCATAATGTACCAGTGCTGGTTTGCTGTGAAACATAC 1289
 410 Q  L  A  L  V  A  R  A  H  N  V  P  V  L  V  C  C  E  T  Y

1290 AAGTTCTGTGAGCGTGTGCAGACTGATGCCTTTGTCTCTAATGAGCTAGATGACCCTGAT 1349
 430 K  F  C  E  R  V  Q  T  D  A  F  V  S  N  E  L  D  D  P  D

1350 GATCTGCAATGTAAGCGGGGAGAACATGTTGCGCTGGCTAACTGGCAGAACCACGCATCC 1409
 450 D  L  Q  C  K  R  G  E  H  V  A  L  A  N  W  Q  N  H  A  S

1410 CTACGGTTGTTGAATCTAGTCTATGATGTGACTCCCCCAGAGCTTGTGGATCTGGTGATC 1469
 470 L  R  L  L  N  L  V  Y  D  V  T  P  P  E  L  V  D  L  V  I

1470 ACGGAGCTGGGGATGATCCCTTGCAGTTCTGTACCTGTTGTTCTACGAGTCAAGAGCAGT 1529
 490 T  E  L  G  M  I  P  C  S  S  V  P  V  V  L  R  V  K  S  S

1530 GACCAGTGAcgggggaaacacagggttaataaatgccatactccctaaaaaaaaaaaaaa 1589
 510 D  Q  *

1590 aaaa
```

Figure 4.9 (Continued from previous page). 3′ end of *EIF2B4* cDNA sequence and putative translation product. The stop codon is located at nucleotides 1536-1538, poly(A) addition signal sequence is underlined (nucleotides 1557-1562) with the poly(A) at nucleotides 1576-1594.

4.2.3.3 Comparison of human eIF2Bδ to other mammalian eIF2Bδ subunits

The human eIF2Bδ amino acid sequence (short isoform) was compared to the sequence of other mammalian eIF2Bδ subunits from which a consensus sequence was produced (Figure 4.10). This reveals the eIF2Bδ subunit to be highly conserved in mammals, with 82% of residues being conserved between human, mouse, rat. and rabbit.

```
Human                d g g  a   ppg   v r  m k                      aepet g       cq g t  lpes  i l.g pr   v    s
Mouse                e r e  t   spr   a r  l q                      adqei g       rq d i  lpgp  s lgg ag   l    s
Rat                  e r e  t   spr   a r  l q                      adqei g       rq d v  lqgt  s lgg tg   l    s
Rabbit               d g g  a   sar   g k  m q                      tdvdt a       cp g a  apgp  s s.s pg   v    t
Consensus  MAAVAVAVRE -S-S-MK-EL ---PGA-G-E -T-EEKLQLR KEKKQQKKKR KEEKG----- -SAVSAAQ-- -P-RE----G -Q---T--EK -PAGR-KAEL

Human                           pp k s       tps  v  lp ypqv   .l     v k pe                              n    f
Mouse                           vp q c       tts  v  vp htpa   pt     l r pd                              s    y
Rat                             ps q c       ats  v  vp htqa   pt     l r pd                              s    y
Rabbit                          pp q s       apa  g  lt htqa   pt     v r se                              n    y
Consensus  RAERRAKQEA ERALKQARKG EQGG--P-A- PSTAGE---G -KR--E---- DD--LLRRL- -K--RQQVPT RKDYGSKVSL FSHLPQYSRQ -SLTQ-MSIP

Human                       v            r             n              ms       l     h      ln  i svg  s       s
Mouse                       i            h             s              is       m     c      lt  v gms  s       s
Rat                         v            h             n              is       m     c      fn  v gms  s       s
Rabbit                      i            r             n              ic       l     y      ln  i gvs  t       a
Consensus  SSVIHPAMVR LGLQYSQGL- SGSNARCIAL L-ALQQVIQD YTTPP-EELS RDLVNKLKPY --FLTQCRP- SASM-NAIKF --KE-T---S -KREEEAK-E

Human       ra i           a    s   yq   n                      wt   r          w      t rs  h     a
Mouse       ke l           a    s   st   d                      rv   r          r      m hs  r     t
Rat         ke i           s    s   sk   d                      wv   r          r      m hc  r     t
Rabbit      qa a           a    l   sk   n                      ws   k          r      m rf  r     a
Consensus  L--A-DRYVQ EKIVLA-QAI -RFA--KIS- GDVILVYGCS SLVSRILQEA --EGR-FRVV VVDSRP-LEG RH-L--LV-A GVP-SYLLIP AASYVLPEVS

Human                                                                        k  e h a     nh a
Mouse                                                                        k  d q a     sh p
Rat                                                                          k  d q t     nn s
Rabbit                                                                       e  d h a     sh p
Consensus  KVLLGAHALL ANGSVMSRVG TAQLALVARA HNVPVLVCCE TYKFCERVQT DAFVSNELDD PDDLQC-RG- -V-LANWQ-- -SLRLLNLVY DVTPPELVDL

Human
Mouse
Rat
Rabbit
Consensus  VITELGMIPC SSVPVVLRVK SSDQ
```

Figure 4.10 Comparison of human eIF2Bδ amino acid sequence with other mammalian eIF2Bδ proteins (mouse, rat, and rabbit, Genbank accession numbers M98036, Z48225, and X75451, respectively).

4.2.3.4 Structural organisation of the human *EIF2B4*

The genomic organisation of *EIF2B4* was determined by sequencing across exon-intron boundaries using cosmids G4 and A4 as template and primers designed from the *EIF2B4* cDNA sequence. The gene was completely sequenced except for intron 11 and the results are summarised in Figures 4.10, 4.11 and 4.12. The human *EIF2B4* gene spans approximately 5.9 kb of genomic DNA and consists of 14 exons (1A, 1B, 2–13) and 12 introns (there are 2 alternative first exons, with exon 1B located directly in front of exon 2) – see Figure 4.11. The exons range in size from 44 bp to 246 bp and introns from 118 bp to 1.8 kb (Figures 4.10 and 4.11). The first methionine codon of the open reading frame for the short isoform is located in exon 1a and for the long isoform in exon 1b. Both long and short isoforms share the stop codon and poly(A) addition signal (Gil *et al.*, 1987) which are located in the last exon, exon 13. An *Alu* repeat region is found in intron 11, located 187 bp upstream from exon 12 (Figure 4.12). All introns show the consensus sequence (C/T/A)AG-exon-GT(G/A) at their boundaries (Figure 4.13). The present gene structure is the first described for any *EIF2B* subunit. The *EIF2B4* genomic sequence was submitted to the EMBL sequence database and assigned accession numbers AJ011307 (exons 1-11) and AJ011308 (exons 12-13).

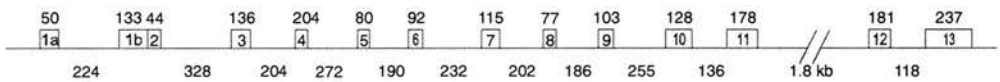


Figure 4.11 Genomic organisation of *EIF2B4*. Exons are shown as boxes. The diagram is not drawn to scale; exact exon and intron sizes (bp) are indicated (except intron 11 – approx. 1.8 kb).

EIF2B4 genomic sequence

```
ggagccctgcgatccgatgtgggaaggcggtgggacgtcggttctgggacgcaatcgcctgggcatgctg 70
ggaactgtagtctaaagggcaagagccacctgtccggatagggagcccaaagtggagttggccctcattg 140
caacccgtagtgcggggcggcctcggggcccgtcggaaattgtagtccgattggggctgccggttcgctt 210
ctgcggctcagcctccagcccaggtggcgctgtggaccctgcagccgcggagcgaggcagcggggcggct 280
ccgtgaccacggcggcgGAGCCTAGGACTGAGGGCGATGGCTGCTGTGGCCGTGGCTGTTCGCGAGGgtg 350
                                  M  A  A  V  A  V  A  V  R  E  D    11
agtgaagagccgggcctgcctactggtacgcggacgcgacgagctccggacagctagtgccgggccttga 420
gcgcctttgggccccgccgctccgttccagcgccacgctggctccggtctacacagtctcgagcgcagtc 490
ccgcgcaccggctggttgctgtgaggagggtcggtgaccctccattcctttgatccgccgcgcagtgcgt 560
ccccacgcggaTCCCTGGCTCGCCTCGCGCCTGCCGGCGGATCCGTGGTCATGCCAACCCAGCAGCCGGC 630
                                                  M  P  T  Q  Q  P  A
TGCGCCGAGTACTCGTGCCCCCAAACCCTCCCGGAGTCTCTCTGGCTCACTTTGTGCCCTGTTTTCTGAT 700
 A  P  S  T  R  A  P  K  P  S  R  S  L  S  G  S  L  C  A  L  F  S  D
GCAGACTCGGGATCCGGGATGAAGGCGGAGCTTCCCCCTGGGCCTGGGgtaagtgaggcttccctcccaa 770
A  D  S  G  S  G  M  K  A  E  L  P  P  G  P  G                       25
gccgcctgccaagcgctacgcaggaggctgtgcagctttattccgccccagctcccatcccgtgtcggtt 840
tatttctcagtaaagccaggagattttacagaggtcttgcaaccctagggatatatgattctgtccctgg 910
aggagaaacgtcagcctctgtggatgtaagcttgcaacaggtagcttgtggaagttaggaagaactaggg 980
atttaaccatgtggaaatcttgtaagcatccaatagtatttgttccccttggaattgttggggaagagc 1050
agtatcctatgtctgtatctcagGCAGTGGGGAGGGAAATGACCAAAGAAGAAAAGCTGCAGCTTCGGAA 1120
                          A  V  G  R  E  M  T  K  E  E  K  L  Q  L  R  K  41
GGAAAAGAAACAGCAGAAGAAGAAACGGAAGGAAGAAAAGGGGGCAGAACCAGAGACTGGCTCTGCTGTA 1190
 E  K  K  Q  Q  K  K  K  R  K  E  E  K  G  A  E  P  E  T  G  S  A  V   64
TCTGCAGCCCAATGTCAAGgtgagtgaggggtctcttttaaagggctggggtgtggagctggtagaggga 1260
S  A  A  Q  C  Q  G                                                    71
tgtggtgggagagataagggagagacaggacaaagttgatttgagcaagtgtgtaaagcaaagaggttta 1330
caaggagagaagctagttttagctggggcacggatatcacctgcctgcagcttgggcatttttttttta 1400
atgcttttagtagGCCCAACCAGAGAACTGCCAGAATCGGGCATTCAGTTGGGCACTCCTCGGGAGAAAG 1470
                  P  T  R  E  L  P  E  S  G  I  Q  L  G  T  P  R  E  K  V  90
TTCCAGCTGGTCGGAGTAAGGCCGAACTTCGGGCTGAGCGTCGAGCCAAGCAGGAGGCCGAGCGGGCCCT 1540
  P  A  G  R  S  K  A  E  L  R  A  E  R  R  A  K  Q  E  A  E  R  A  L  113
GAAACAGGCAAGAAAAGGGGAACAAGGAGGACCACCTCCTAAGGCCAGCCCCAGCACAGCTGGAGAAACC 1610
 K  Q  A  R  K  G  E  Q  G  G  P  P  P  K  A  S  P  S  T  A  G  E  T   136
CCCTCAGgtatcttcccttcattttaagacctcccttactcctaattataacgccagctcaggctgccca 1680
P  S  G                                                                139
atagtgaaacagctctcccctctcattcttgggacccagagttctgaattgttcctgcaaccccattatc 1750
ctatccctatttccttttctaccctattctcccgttttctccacctcatcctgtttttcttgatgcccct 1820
tataccctaagagcaaattacacatttcaagacctacagtgtcctgaaaacatgatgattatctttcagG 1890
AGTGAAGCGTCTCCCTGAGTACCCTCAGGTTGATGACCTACTTCTGAGAAGGCTTGTTAAAAAACCAGAG 1960
 V  K  R  L  P  E  Y  P  Q  V  D  D  L  L  L  R  R  L  V  K  K  P  E   162
CGTCAACAGgtaggaagtgggtttggtgcctggctagaaagagaatagggaaatggactggagggaggaa 2030
R  Q  Q                                                                165
aggtgatgggtgcaaagtgagaaggataggtgcttgggtcagagtagtcttactgttcacttcctttcca 2100
gctttatcccatccttttaaactctgggccattggtcctgttcccttgttttgtcaagGTTCCTACACG 2170
                                                        V  P  T  R   169
AAAGGATTATGGATCCAAAGTCAGTCTCTTCTCTCACCTACCCCAGTACAGCAGACAAAACTCTCTGACC 2240
 K  D  Y  G  S  K  V  S  L  F  S  H  L  P  Q  Y  S  R  Q  N  S  L  T   192
CAGTTTATGAGgtaggatcctatgaaattgtcataacttttgagtctgaggagttttagtaagtcagtat 2310
Q  F  M  S                                                             196
tgaaagtttctagtaatctgagatggactctgtaggaaaaagtatgaacccaaggcagaactgtagagga 2380
gtggggaaatctgctagtgaagggagttccttttcatcaggggcaggtgggcagtagatgctcaagctcc 2450
ctttcaaagtatgtgacaccgccatcccatcagCATCCCATCCTCTGTGATCCACCCAGCCATGGTGCGA 2520
                                  I  P  S  S  V  I  H  P  A  M  V  R   208
CTCGGCCTGCAGTACTCCCAGGGCCTGGTCAGTGGCTCCAATGCCCGGTGTATTGCCCTGCTTCGTGCCT 2590
L  G  L  Q  Y  S  Q  G  L  V  S  G  S  N  A  R  C  I  A  L  L  R  A  L  232
TGCAGCAGgtatgtcccatcctgttctctcttatgatccaaccccacctccccaatactaccccagctt 2660
  Q  Q  234                                                  Continued on next page.
```

Figure 4.12 The complete *EIF2B4* genomic sequence and its putative translation product. Exon 1b (the "long" alternative first exon) is immediately adjacent to exon 2, corresponding to nucleotides 572-704 (shaded yellow). The 4 bp deletion in cDNA clone 274777 corresponds to nucleotides 1074-1077. The 26 bp deletion in cDNA clones 274777 and 380606 corresponds to nucleotides 2484-2509. The cDNA deletions are shaded green.

EIF2B4 genomic sequence (continued from previous page)

```
atgtcacccaaggtcatctctgggggacagagggaatactcctaccctcagaacattttcctcagtgaag 2730
cttacccccagcccccgttcagagatctttattaaggggtattttgcagatcaccttgtgtaccctatag 2800
GTGATTCAGGATTACACAACACCGCCTAATGAAGAACTCTCCAGGGATCTAGTGAATAAACTAAAACCCT 2870
V  I  Q  D  Y  T  T  P  P  N  E  E  L  S  R  D  L  V  N  K  L  K  P  Y 258
ACATGAGgtagggacaacactataagtccactcccagctcgcccaatctttgtcttgtactctttttcct 2940
  M  S                                                                 260
gcttttctatctctcttttcattttgtgttctatttttcttttaaagacttttttttaaatgtgaaggc 3010
tttgtaagttgtgaagagttgcgcagatataaggtattgctgttgtattatagCTTCCTGACTCAGTGCC 3080
                                                        F  L  T  Q  C  R 266
GTCCCCTGTCAGCGAGCATGCACAACGCCATCAAGTTCCTTAACAAGGAAATCACCAGTGTGGGCAGTTC 3150
  P  L  S  A  S  M  H  N  A  I  K  F  L  N  K  E  I  T  S  V  G  S  S 289
CAAGCGGGAAGAGGAGgtgatgagtatgagaaataaggaaggcatgcacctgtgggtaaaattaagaaac 3220
 K  R  E  E  E                                                        294
atgttaaaaactcccgggaaggccaacgtaagccttttttgccctagtcaagggttaaggaatggtttg 3290
gccattctgacttgtgccatctctgagccctctataggcttgtcatcatacacatattcaagcacctggc 3360
atgatcatcagagggataaaagtgtccatttcttgttcactaaggattttactccatttagGCCAAGTCA 3430
                                                          A  K  S  297
GAACTTCGAGCAGCCATTGATCGGTATGTGCAAGAGAAGATTGTGCTAGCAGCTCAGGCAATTTCACGCT 3500
E  L  R  A  A  I  D  R  Y  V  Q  E  K  I  V  L  A  A  Q  A  I  S  R  F 321
TTGCTTACCAGAAGATCAGTAATGGAGATGTGATCCTGGTATATGGATGgtatggtccagaccttgtgac 3570
  A  Y  Q  K  I  S  N  G  D  V  I  L  V  Y  G  C                     337
tgagcagattgggagttggaatgatctgaagaagggactacccacttgcctttgggaataagttagtcac 3640
agttactgacatttataactgaatcctcctcttcattcgctctagCTCATCTCTGGTATCACGAATTCTT 3710
                                              S  S  L  V  S  R  I  L  345
CAGGAGGCTTGGACAGAGGGCCGGCGGTTTCGGGTGGTAGTGGTGGACAGCCGGCCATGGCTGGAAGGAA 3780
Q  E  A  W  T  E  G  R  R  F  R  V  V  V  V  D  S  R  P  W  L  E  G  R 369
GGCACACACTACGTTCTCTAGTCCATGCTGGTGTCCCAGCCTCCTACCTGCTGATTCCTGCAGCCTCCTA 3850
  H  T  L  R  S  L  V  H  A  G  V  P  A  S  Y  L  L  I  P  A  A  S  Y 392
TGTGCTCCCAGAGgtaagtacagaggaaaaggactccaaagttggcggtgaaaaggtgtagtagtataga 3920
 V  L  P  E                                                           396
cataatctcaccttcccaaagttacagttttgtttttgttttgagataaggatgg               3975
intron 11 ( approx. 1.8 kb )
aaaaatacaacaaatcagctgggagtggtggtgcatgcctgtaatcccagctactcgggaggctgaggca 4045
ggagaatcacttgaacccaggaggcagaggttgtggtgagccgagatcgcgccattgcactccagcctgg 4115
gcaacaagagcagaagttcatctcaaagaaaaaaaaaaaaagatttattacttgttgtctgtttttgttt 4185
ctttcccttctttttcacaggtgttgagatttattacttgtcttaatgttgctgtcctgtaagtaggggac 4255
cttattttgcattgatacaacaatgacatttttttcttcctcttctcctgtgaaacagGTTTCCAAGGTG 4325
                                                        V  S  K  V  400
CTATTGGGAGCTCATGCACTCTTGGCCAATGGGTCTGTGATGTCACGGGTAGGGACAGCACAGTTAGCCC 4395
L  L  G  A  H  A  L  L  A  N  G  S  V  M  S  R  V  G  T  A  Q  L  A  L 424
TGGTGGCTCGAGCCCATAATGTACCAGTGCTGGTTTGCTGTGAAACATACAAGTTCTGTGAGCGTGTGCA 4465
  V  A  R  A  H  N  V  P  V  L  V  C  C  E  T  Y  K  F  C  E  R  V  Q 447
GACTGATGCCTTTGTCTCTAATGAGCTAGgtaaggaggcaaaagagaagcctccagctacctgtgaagac 4535
 T  D  A  F  V  S  N  E  L  D                                         457
ttgagagggagggagaggggcggataaactagaacttcttttctgaaagaacataatgggtacgccttcc 4605
attgcagATGACCCTGATGATCTGCAATGTAAGCGGGGAGAACATGTTGCGCTGGCTAACTGGCAGAACC 4675
        D  P  D  D  L  Q  C  K  R  G  E  H  V  A  L  A  N  W  Q  N  H 478
ACGCATCCCTACGGTTGTTGAATCTAGTCTATGATGTGACTCCCCCAGAGCTTGTGGATCTGGTGATCAC 4745
  A  S  L  R  L  L  N  L  V  Y  D  V  T  P  P  E  L  V  D  L  V  I  T 501
GGAGCTGGGGATGATCCCTTGCAGTTCTGTACCTGTTGTTCTACGAGTCAAGAGCAGTGACCAGTGACGG 4815
 E  L  G  M  I  P  C  S  S  V  P  V  V  L  R  V  K  S  S  D  Q  *     522
GGGAAACACAGGGTTAATAAATGCCATACTCCCTaccctcagcaactctgcctttgtttctcttttagca 4885
tctccaccacttaagttaggagtccagacttcacaaccccttttatcacctgctactcagacctttgtga 4955
agacctgctcaagtaactagctccatgccagtacattgggactaacctgaagaccc               5011
```

Figure 4.12 (Continued from previous page). The complete *EIF2B4* genomic sequence and its putative translation product. The translation termination signal is marked with an asterisk and the poly(A) addition signal is underlined. An *Alu* repeat sequence in intron 11 is also underlined.

```
Exon1-GGCCGTGGCTGTTCGCGAGGgtgagtgaagagccgggcct---357 bp---tgccctgttttctgatgcagACTCGGGATCCGGGATGAAG-Exon2
Exon2-AGCTTCCCCCTGGGCCTGGGgtaagtgaggcttccctccc---324 bp---atcctatgtctgtatctcagGCAGTGGGGAGGGAAATGAC-Exon3
Exon3-ATCTGCAGCCCAATGTCAAGgtgagtgaggggtctctttt---204 bp---ttttttaatgctttagtagGCCCAACCAGAGAACTGCCA-Exon4
Exon4-AGCTGGAGAAACCCCCTCAGgtatcttcccttcattttaa---272 bp---acatgatgattatctttcagGAGTGAAGCGTCTCCCTGAG-Exon5
Exon5-AAAAACCAGAGCGTCAACAGgtaggaagtgggtttggtgc---190 bp---tgttcccttgttttgtcaagGTTCCTACACGAAAGGATTA-Exon6
Exon6-TCTCTGACCCAGTTTATGAGgtaggatcctatgaaattgt---232 bp---tgacaccgccatcccatcagCATCCCATCCTCTGTGATCC-Exon7
Exon7-TGCTTCGTGCCTTGCAGCAGgtatgtcccatcctgttctc---202 bp---tcaccttgtgtaccctatagGTGATTCAGGATTACACAAC-Exon8
Exon8-AAACTAAAACCCTACATGAGgtagggacaacactataagt---186 bp---gtattgctgttgtattatagCTTCCTGACTCAGTGCCGTC-Exon9
Exon9-GTTCCAAGCGGGAAGAGGAGgtgatgagtatgagaaataa---255 bp---aaggattttactccatttagGCCAAGTCAGAACTTCGAGC-Exon10
Exon10-GTGATCCTGGTATATGGATGgtatggtccagaccttgtga---136 bp---ctcctcttcattcgctctagCTCATCTCTGGTATCACGAA-Exon11
Exon11-CCTCCTATGTGCTCCCAGAGgtaagtacagaggaaaagga---1.8 kb---cctcttctcctgtgaaacagGTTTCCAAGGTGCTATTGGG-Exon12
Exon12-CTTTGTCTCTAATGAGCTAGgtaaggaggcaaaagagaag---118 bp---gggtacgccttccattgcagATGACCCTGATGATCTGCAA-Exon13
```

Figure 4.13 Exon-intron splice junction sequences. Intron sequences at the 5'- and 3'-end of the exons are in lower case letters. Exon sequences are in capital letters.

4.2.3.5 Alternative *EIF2B4* splice forms

As described previously, the human *EIF2B4* gene is expressed as both long and short isoforms that only differ by their 5′ first exon. To investigate the tissue expression of these two isoforms, RT-PCR was performed using primers designed from within either of the two *EIF2B4* alternative first exons (primer G7R from exon 1A and primer DL5 from exon 1B), in conjunction with a primer in exon 12 (primer G2) - see Table 4.2.

Primer name	Oligonucleotide sequence	Annealing temp. (°C)
G7R	5'-TAGGACTTGAGGGCGATGGCTG-3'	62
G2	5'-AGTGCATGAGCTCCCAATAGC-3'	62
DL5	5'-AAACCCTCCCGGAGTCTCTC-3'	62

Table 4.2 Oligonucleotide sequences for RT-PCR. The primers G7R and DL5 were designed from the *EIF2B4* alternative short and long first exon, respectively. Primer G2 was designed from *EIF2B4* exon12 sequence.

Therefore, RT-PCR primer pair G7R/G2 amplifies *EIF2B4* exons 1A-12 (theoretical size of PCR product = 1230 bp) and primer pair DL5/G2 amplifies *EIF2B4* long exons 1B-12 (theoretical size of PCR product = 1237 bp). RT-PCR was performed using total human RNA from brain, heart, and muscle. Also, PCR was performed on a human HepG2 cDNA library.

The results reveal that both long and short isoforms are expressed in brain, heart, muscle and liver. However, while heart, muscle and liver yielded one single PCR product of the correct theoretical size, RT-PCR from brain RNA yielded multiple products (see Figure 4.14 for comparison of RT-PCR products from heart and brain). This suggests that multiple *EIF2B4* splice forms may exist in the brain.

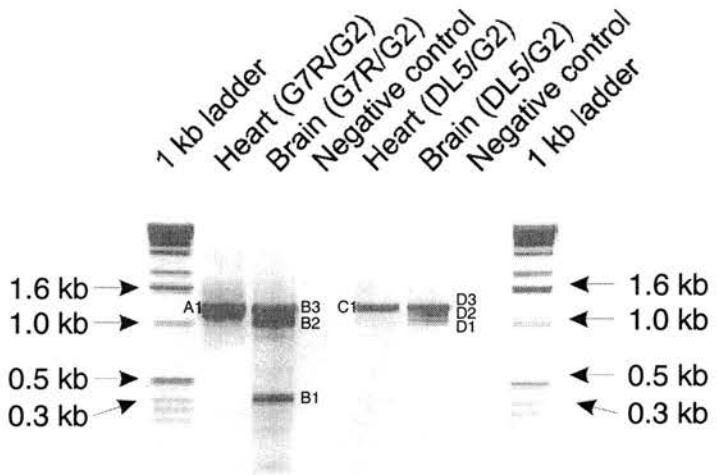


Figure 4.14 RT-PCR products for both short and long *EIF2B4* isoforms amplified from heart and brain RNA. PCR products were run on a 1% agarose gel and stained with ethidium bromide.

To investigate the RT-PCR products from heart and brain (shown in Figure 4.14), the individual PCR products were excised from the agarose gel, purified and sequenced. The purified RT-PCR fragments are shown in Figure 4.15 and the results of sequencing are summarised in Table 4.3.

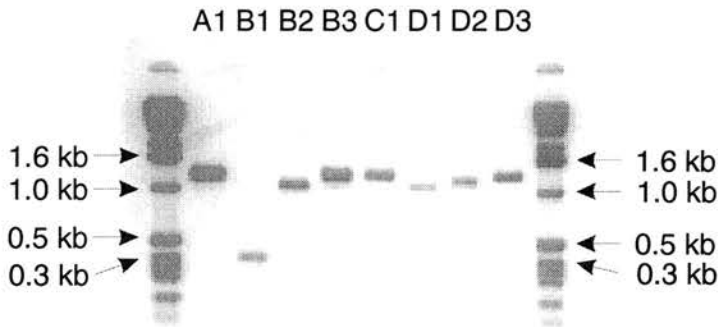


Figure 4.15 Purified RT-PCR products. Lane labels refer to RT-PCR shown in Figure 4.14 and described in Table 4.3. Outer lanes contain 1 kb ladder marker.

RT-PCR product	mRNA template	5′- exon	Exons Missing	Comment
A1	Heart	1a	None	Full length
B1	Brain	1a	-	Artefact RT-PCR product
B2	Brain	1a	Exons 2 and 3	Removes 180 nucleotides
B3	Brain	1a	None	Full length
C1	Heart	1b	None	Full length
D1	Brain	1b	Exons 7 and 8	Removes 192 nucleotides
D2	Brain	1b	Exon 10	Removes 128 nucleotides
D3	Brain	1b	None	Full length

Table 4.3 Results of sequencing RT-PCR products of *EIF2B4* from heart and brain mRNA. RT-PCR products relate to that shown in Figures 4.14 and 4.15.

The sequencing of the RT-PCR products from heart and brain (summarised in table 4.3) shows that in the heart, the long and short isoforms are both expressed in their full length forms (Table 4.3, RT-PCR products "A1" and "C1"). However, sequencing of the multiple *EIF2B4* brain RT-PCR products reveals that apart from the full length long and short isoforms (Table 4.3, RT-PCR products "B3" and "D3"), other shorter splice forms exist in the brain (Table 4.3, RT-PCR products B2, D1 and D2).

Sequencing of these shorter RT-PCR products reveals they are due to the removal of one or two exons. The exon 1A isoform "B2" (see Table 4.3), contains exon 1A but not exons 2 and 3, resulting in the loss of 180 nucleotides (60 amino acids). The removal of exons 2 and 3 however, does preserve the original open reading frame and therefore would produce a protein of 462 amino acids (full length short isoform is 522 amino acids long).

The exon 1B isoform "D1"(see Table 4.3), contains exon 1B but not exons 7 and 8, resulting in the loss of 192 nucleotides (64 amino acids). The removal of exons 7 and 8, like the B2 splice form, preserves the original open reading frame and would therefore produce a protein of 479 amino acids (full length long isoform is 543 amino acids). In contrast, the exon 1B isoform "D2" contains exon 1B but not exon 10, resulting in the loss of 128 nucleotides. This does not preserve the original open reading frame, and if translated would result in a truncated protein of 405 amino acids. It should be noted that the although the removal of exon 10 does not preserve the original open reading frame, a new open reading frame coding for 92 amino acids that spans from exon 11 to within exon 12 is produced.

4.2.3.6 Genomic arrangement of mouse alternative first exons

A search of the EMBL mouse cDNA database using the long alternative 5' end *EIF2B4* sequence revealed several clones with significant similarity including the IMAGE clone 556389 (see Figure 4.16 and Table 4.4). BLAST sequence analysis of the clone 556389 database sequence (Genbank accession number AA103972) reveals that it contains both the short and long alternative first exons in a similar organisation to that found in human genomic DNA (exon 1A (short)-exon1B (long)-exons2-13).

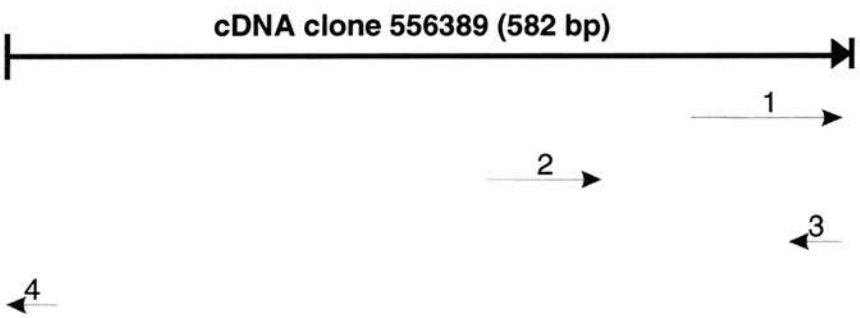


Figure 4.16 Pictorial representation of a BLAST search using mouse cDNA clone 556389. See Table 4.4 below for sequence identification.

Sequence	Genbank acc. no.	Sequence description
1	M98035	*Mus musculus* guanine nucleotide exchange factor delta subunit (long isoform) mRNA, nucleotides 6-144.
2	M98036	*Mus musculus* guanine nucleotide exchange factor delta subunit (short isoform) mRNA, nucleotides 8-100.
3	Z66245	*Homo sapiens* CpG island DNA genomic *Mse*1 fragment, clone 84b2, reverse read cpg84b2.rt1a, nucleotides 274-305.
4	D31764	*Homo sapiens* mRNA for *KIAA0064* gene, nucleotides 1-29.

Table 4.4 Clone identification for Figure 4.16.

Closer inspection of the BLAST analysis of EST AA103972 reveals that there is a 13 bp gap between the regions of identity for sequences M98035 and M98036 (corresponding to nucleotides 427-439, Figure 4.17). The analysis of EST AA103972 also reveals homology to a CpG island clone (Genbank accession number Z66245) and a human putative transcript called *KIAA0064* (Genbank accession number D31764) - see Figure 4.16. As this mouse clone contains both long and short 5' end alternative *Eif2b4* exons plus 13 bp of sequence between them, and a region of sequence showing similarity to the human putative transcript *KIAA0064*, it is assumed that this clone is in fact derived from genomic mouse DNA and not a cDNA clone. To investigate whether the same features existed in the human genome, further sequencing was performed in the 5′ flanking region of the human *EIF2B4* gene. A comparison of the human 5′ region of the *EIF2B4* gene to mouse clone AA103972 is shown in Figure 4.18.

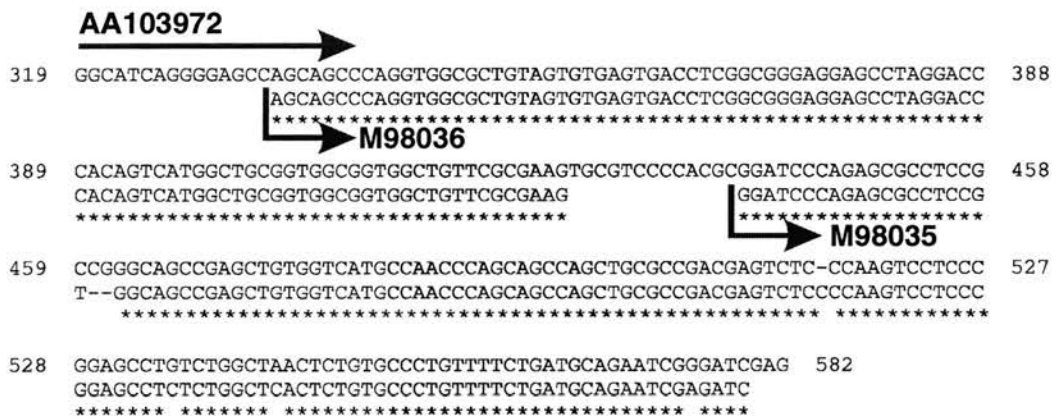


Figure 4.17 Comparison of mouse EST AA103972 sequence (nucleotides 319-582) and the mouse *Eif2b4* short (M98036) and long (M98035) isoforms. The AA103972 sequence is the upper sequence.

The sequence comparison shown in Figure 4.18 reveals that the putative transcript *KIAA0064* is present in both human and mouse genomes in the same "head to head" orientation to the *EIF2B4* gene. Other noticeable features is a high degree of sequence similarity between the human and mouse *KIAA0064-EIF2B4* intergenic regions, even in the non-coding sequences. This can partly be explained by conservation of the human and mouse *EIF2B4* and *KIAA0064* genes' 5′UTRs. The major difference revealed in Figure 4.18 is that while in human, there is 263 bp between the end of *EIF2B4* exon 1a and the start codon of exon 1b, there is only 13 bp between the end of mouse *Eif2b4* exons 1a and the start codon of exon 1b. Further sequence analysis shows that both the human and mouse *EIF2B4-KIAA0064* intergenic regions are CpG islands (Figure 4.18 - human: nucleotides 61-801; mouse: nucleotides 51-535). The human CpG island (740 bp) is larger than in the mouse (484 bp) due the extra sequence found between human *EIF2B4* exons 1a and 1b.

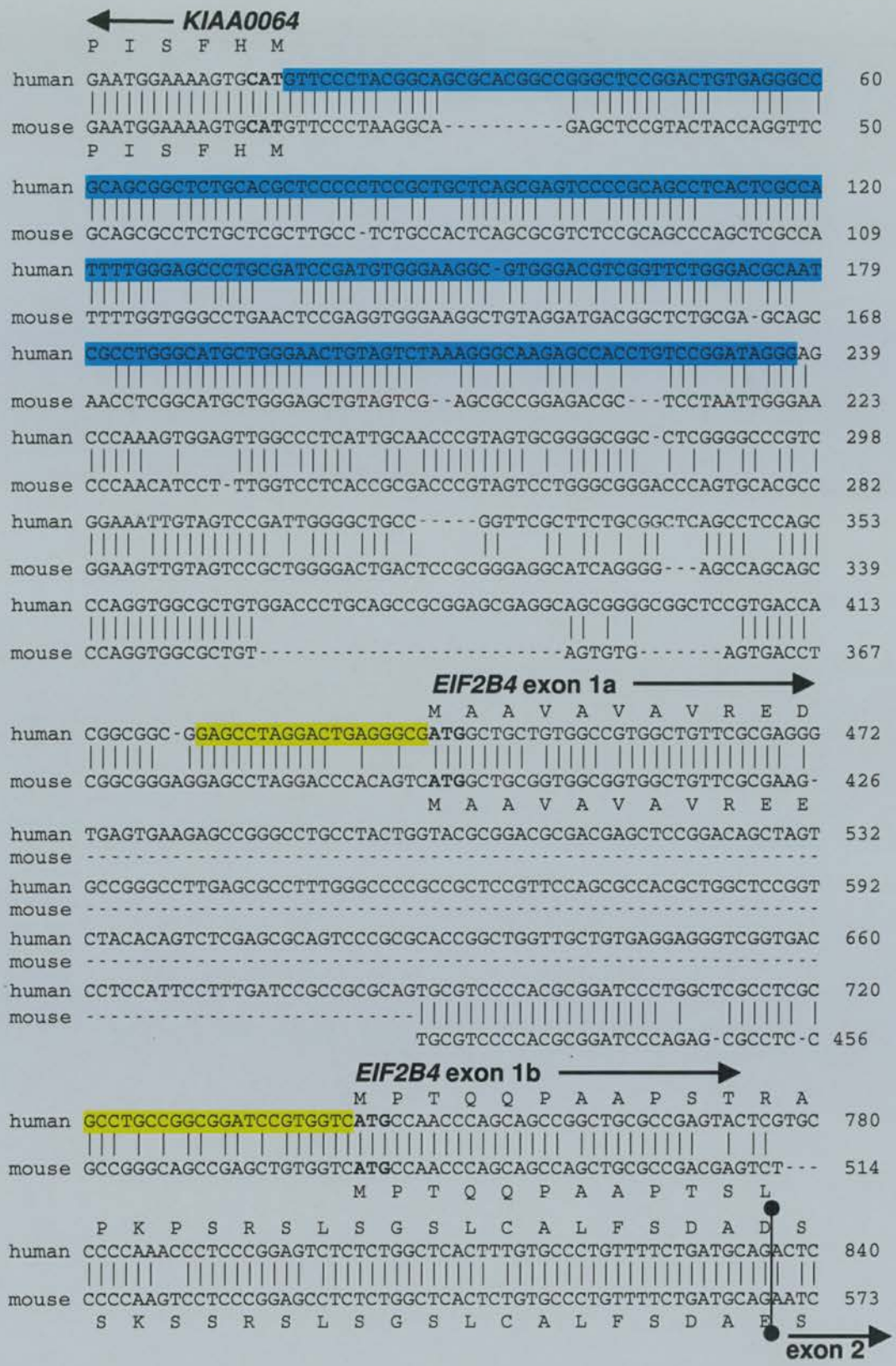


Figure 4.18 Comparison of the 5′end of the human *EIF2B4* gene and the mouse EST AA103972 sequence. The coding regions of *EIF2B4* exons 1a and 1b, and *KIAA0064* are indicated (putative translation products are shown). The blue shaded region indicates the KIAA0064 5′UTR. The yellow shaded regions indicate 5′UTRs for *EIF2B4* short and long isoforms.

4.2.4 Discussion

4.2.4.5 *EIF2B4* chromosomal localisation

Section 4.2 describes the cloning and investigation of the human *EIF2B4* gene. *EIF2B4* encodes the delta subunit of the guanine nucleotide exchange factor eIF2B, a protein complex that plays a key role in the regulation of protein synthesis in yeast and mammals. This transcript was chosen for investigation primarily because it co-localises with the genes glucokinase regulator (*GCKR*) and ketohexokinase (*KHK*) on chromosome 2p23.3, and also because previous studies have shown eIF2B activity to be stimulated by glucose (Gilligan *et al.*, 1996), sugar phosphates (Singh & Wahba, 1995) and insulin (Welsh *et al.*, 1997b). This suggests that like the candidate type 2 diabetes genes *GCKR* and *KHK*, *EIF2B4* can also be considered a candidate type 2 diabetes gene.

The assembly of a physical contig using YAC, BAC, PAC and cosmid clones spanning this genomic region has shown *EIF2B4* to reside between these two genes approximately 150 kb from *GCKR* and 350 kb from *KHK* (Figure 4.7; see also Chapter 3, Figure 3.8). The physical contig also reveals that the *EIF2B4*, *KHK* and *GCKR* genes are all arranged in the same 5′-3′ orientation with respect to the centromere. The mapping of *Gckr* and *Khk* to mouse chromosome 5 (as described in Chapter 2) reveals a new region of conserved synteny between human chromosome 2p23 and mouse chromosome 5 (Wightman *et al.*, 1997). The mapping of *EIF2B4* to between *GCKR* and *KHK*, suggests that the mouse *Eif2b4* gene also probably maps to mouse chromosome 5.

4.2.4.6 *EIF2B4* cDNA sequence and genomic structure

As in the mouse, the human *EIF2B4* gene is expressed as two different isoforms. The "short" isoform mRNA is 1643 nucleotides in length and consists of an open reading frame of 1569 bp, 19 bp of 5' non-coding sequence and 75 bp of 3' non-coding sequence, with a consensus poly(A) addition signal. The predicted short isoform protein is of 522 amino acids (Figures 4.9i and 4.9iii) and a molecular weight of 57478 Da. The "long" isoform mRNA is 1708 nucleotides in length and has an open reading frame of 1632 bp, 39 bp of 5' non-coding sequence and 75 bp of 3' non-coding sequence, with a consensus poly(A) addition signal. The predicted long isoform protein has 543 amino acids (Figure 4.9ii and 4.9iii) and a molecular weight of 59622 Da. Comparison of the short and long isoforms reveals that they only differ at their 5′ ends and the analysis of the *EIF2B4* genomic structure shows that this is due to the existence of two alternative first exons.

The structural analysis of the human *EIF2B4* gene shows it to consist of 14 exons (1a, 1b, and 2-13) that are expressed as two isoforms, both consisting of 13 exons but differing by their 5' first exons. The short isoform is highly conserved between human, mouse (Henderson *et al.*, 1994), rat (Price *et al.*, 1996) and rabbit (Price *et al.*, 1994), and significant homology is found with the yeast *GCD2* gene (overall DNA sequence identity is 87%, 87%, 86%, and 43% respectively). The long isoform which has previously only been described in the mouse, shares 87% identity at the nucleotide level to the human long isoform. Although it is not known why there are two alternative first exons, examination of the encoded amino acids shows that the short 5' exon generally encodes hydrophobic residues and the long 5' exon encodes more hydrophilic residues. This may indicate a difference in protein interactions either between the eIF2Bδ subunit and the other eIF2B subunits and/or the eIF2B substrate, eIF2.

The comparison of mammalian eIF2Bδ at the amino acid level between human, mouse, rat and rabbit reveals a very high degree of sequence conservation, with 82% of residues conserved between all four mammals. Closer inspection shows that sequence similarity increases towards the C-terminus where two long stretches of conserved amino acid sequences can be found between residues 385 to 466 and 482 to 524 (Figure 4.10). It is notable that the peptide sequence encoded by exon 12 is completely conserved between human, mouse, rat and rabbit, and that this region shows very high homology to the yeast *GCD2* peptide (56% identical). It is also very striking that the residues that vary between human, mouse, rat, and rabbit are not randomaly distributed throughout the eIF2Bδ peptide sequence but are grouped together. The sequence conservation between species indicates eIF2Bδ to also have a highly conserved function. The peptide regions that show complete identity between human, mouse, rat, and rabbit are therefore likely to be either the functional regions of eIF2Bδ and/or the regions that interact with the other eIF2B subunits or the eIF2B substrate, eIF2.

4.2.4.7 Cryptic splice sites

During the cloning of *EIF2B4*, two cDNA clones were sequenced that were shown to contain deletions (Figure 4.8), later shown to be caused by the presence of cryptic splice sites within the *EIF2B4* cDNA sequence. The *EIF2B4* cDNA clones 380606 and 274777 were both found to have a 4 bp deletion at the start of exon 3 due to a cryptic GCAG acceptor site (see Figures 4.9 and 4.12). Similarly, the cDNA clone 274777 also contained a 26 bp deletion located at the start of exon 7, due to the presence of a cryptic CCAG acceptor site (see

Figures 4.9 and 4.12). The translation of mRNA containing the 4 bp and 26 bp deletions would create truncated proteins of 29 and 214 amino acids respectively (size of short isoform is 522 amino acids). These truncated proteins would most likely be non-functional.

4.2.4.8 Tissue expression of *EIF2B4*

The investigation of *EIF2B4* tissue expression by RT-PCR shows that both the short and long isoforms are expressed in heart, brain, muscle and liver. Database searching for *EIF2B4* sequences also shows the existence of short isoform *EIF2B4*-containing cDNA clones, produced from a variety of tissues. This evidence suggests *EIF2B4* to be ubiquitously expressed. Interestingly, database searching does not identify any human *EIF2B4* long isoform cDNA clones. Whether this is due to lower expression levels of the *EIF2B4* long isoform is not known. Although RT-PCR shows the eIF2Bδ long isoform to be expressed in various tissues, this method is non-quantitative and therefore RNase protection would be a better method to investigate the comparative expression levels of the *EIF2B4* short and long isoforms.

4.2.4.9 Alternative *EIF2B4* splice forms

Although RT-PCR shows that full length short and long *EIF2B4* isoforms are expressed in heart, brain, muscle and lung, it was discovered that shorter alternative *EIF2B4* splice forms exist in the brain (but not heart, muscle or lung). The brain *EIF2B4* splice forms are summarised below in Figure 4.19.

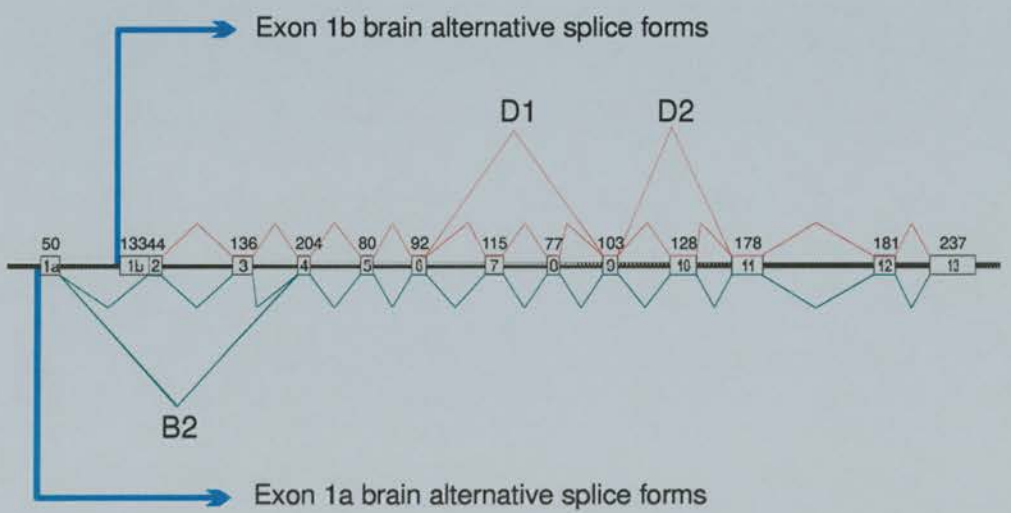


Figure 4.19 Schematic representation of the brain *EIF2B4* alternative splice forms. B2, D1, and D2 refer to alternative splice forms described in Table 4.3.

The *EIF2B4* splice form B2 has exons 2 and 3 missing, splice form D1 has exons 7 and 8 missing and splice form D2 has exon 10 missing. While the missing exons in splice forms B2 and D1 preserves the original open reading frame, the removal of exon 10 in splice form D2 does not preserve the original open reading frame, and if translated would create an alternative 3′ end (but also a truncated eIF2Bδ protein). It is conceivable that as these alternative splice forms encode most of the eIF2Bδ protein, that they do have some functionality. As the alternative splice forms are not present in heart, muscle, or liver, the alternative splice forms might have a brain specific function. This function is as yet unknown.

4.2.4.10 *EIF2B4* promoter region

Analysis of the 5′ flanking region of *EIF2B4* reveals that the putative transcript *KIAA0064* is located upstream (Figure 4.16). The two genes are orientated in opposite directions with 466 bp separating the two start codons of *KIAA0064* and *EIF2B4* (short isoform). This intimate "head to head" orientation suggests that these two genes may share promoter elements. Sequence analysis of the intergenic region in both the human and mouse genome reveals a high degree of sequence conservation, including the presence of a CpG island (Figure 4.18). CpG islands are stretches of non-methylated DNA rich in CpG dinucleotides (in which C occurs immediately 5′ to G) and are found at the promoters of a large proportion of genes (Bird, 1993). It has recently been shown that CpG islands are coincident with DNA replication origins in mammalian chromosomes (Delgado *et al.*, 1998). The intimate location of the *KIAA0064* transcript in a "head to head" orientation to *EIF2B4*, suggests that this CpG island could be associated with the promoter sequences for both *EIF2B4* and *KIAA0064*. This genomic organisation of *EIF2B4* and *KIAA0064* probably reflects the high gene density of chromosome 2p23.3 described in Chapter 3. Due to the intimate location of *KIAA0064* to *EIF2B4*, the *KIAA0064* transcript was further investigated (see Section 4.3).

4.2.4.11 *EIF2B4* mutations and subunit interactions

There is considerable evidence that the phosphorylation of the eIF2B substrate, eIF2, on its alpha subunit (eIF2α), is an important regulatory mechanism used by eukaryotic systems to inhibit guanine nucleotide exchange by eIF2B and thereby decrease protein synthesis in response to certain cellular factors. This protective measure may occur in response to haem-deficiency, viral infection, nutrient deficiency and other stress conditions like heat shock (Scheper *et al.*, 1997; Scheper *et al.*, 1998). Yeast genetic experiments and rat

biochemical data implicate the eIF2Bδ subunit in the "sensing" of eIF2α phosphorylation and have identified mutations in the yeast and rat gene sequence that can cause eIF2B insensitivity to its substrate's phosphorylation (Pavitt *et al.*, 1997).

The double point mutations that change the eIF2Bδ residues Leu^{377}→Lys and Glu^{381}→Gln lead to insensitivity to eIF2αP in both yeast (Vazquez de Aldana & Hinnebusch, 1994) and rat (Kimball *et al.*, 1998). This would indicate that this region of the peptide plays an important role in "sensing" the phosphorylation state of eIF2αP. As the human eIF2Bδ peptide sequence also contains Leu^{377} and Glu^{381}, it is possible that mutations in the *EIF2B4* gene sequence that change these two amino acids could also lead to eIF2αP insensitivity. Therefore, the presence of *EIF2B4* mutations in the natural human population that alters the residues Leu^{377} and Glu^{381}, might make eIF2B unresponsive to eIF2αP. This may have health implications by making a system unable to respond effectively (by decreasing protein synthesis) to certain cellular conditions such as viral infection and nutrient deficiency.

The complex eIF2B structure appears to relate to the many regulatory mechanisms that can affect eIF2B activity. Although the exact function of the eIF2Bδ subunit is unknown, it is intriguing that the eIF2Bα, eIF2Bβ, and eIF2Bδ subunits share mutual sequence similarities at their C-termini (Price *et al.*, 1996) and that these three subunits can actually form a complex in yeast (which however, does not possess guanine nucleotide exchange activity (Yang & Hinnebusch, 1996)). This suggests a close interaction between these subunits and a role for the three subunits (α, β, and δ) in "sensing" eIF2αP. Other mechanisms that alter eIF2B activity include the dephosphorylation of eIF2Bε subunit by GSK-3 (Welsh *et al.*, 1997a) for example in response to insulin (see Figure 4.4). Sugar phosphates are thought to allosterically stimulate eIF2B activity but the mechanism by which glucose stimulates eIF2B activity is unknown. It is clear that understanding of all the regulatory pathways that act on eIF2B will undoubtedly help to understand the eIF2B subunit functions.

4.2.4.12 Further research

The cloning of short and long isoforms of *EIF2B4* that is described in this section adds to the already complex eIF2B structure. Further investigation into the eIF2Bδ subunit function should therefore take into account the existence of the two eIF2Bδ isoforms. Added to eIF2B structure complexity is the expression of alternative *EIF2B4* splice forms (for both short and long isoforms) in the brain. It would be interesting to further investigate these

EIF2B4 splice forms, to see if they are translated and functional, and whether they possess a brain specific function.

The identification of point mutations in the *EIF2B4* subunit by yeast studies yeast (Vazquez de Aldana & Hinnebusch, 1994) and rat biochemical experiments (Kimball *et al.*, 1998) that can make the eIF2B complex insensitive to substrate phosphorylation, thereby preventing inhibition of eIF2B activity and the regulation of protein synthesis in response to viral infection and stress conditions, suggests that eIF2Bδ plays an important role in the eIF2B complex. It also suggests that mutations in the *EIF2B4* gene could play an important role in some disease processes by affecting the cell's ability to respond to viral infection.

The *EIF2B4* gene was originally investigated because of its co-localisation with the candidate type 2 diabetes genes *GCKR* and *KHK* on chromosome 2p23.3, and because of its candidacy as a type 2 diabetes gene. Although defective eIF2B function could be involved in the pathogenesis of type 2 diabetes, more knowledge of eIF2B subunit function is required before the *EIF2B4* gene can be called a legitimate type 2 diabetes gene. However if *EIF2B4* is linked to a disease process in the future, the characterisation of the *EIF2B4* genomic structure that is described in this section will allow mutation screening to be performed.

4.3 The putative *KIAA0064* gene

4.3.1 Introduction

4.3.1.1 Mapping of *KIAA0064*

In Section 4.2, the investigation of the 5′ flanking region to the *EIF2B4* gene revealed the presence of part of the putative *KIAA0064* gene (see Figure 4.18). Sequencing of the *EIF2B4-KIAA0064* intergenic region showed there to be 424 nucleotides between the *KIAA0064* start codon and the *EIF2B4* exon 1a start codon. Due to the intimate "head to head" arrangement of *EIF2B4* and *KIAA0064*, the putative *KIAA0064* gene was further investigated.

KIAA0064 was first mapped to the *GCKR-KHK* genomic region using the EST named stSG149, designed from a cDNA clone sequence (Genbank accession number G43097). This EST had been previously mapped, using radiation hybrid panels, to the genetic interval D2S165-D2S352 (Genemap99: http://www.ncbi.nlm.nih.gov/genemap/). In the present project, as described in Chapter 3, PCR screening of genomic clones located to the *GCKR-KHK* genomic region was used to map stSG149 to the YACs 26BA11 and 29IH8, and to PACs 13L13 and 119A8. Therefore, even before the identification of *KIAA0064* upstream of *EIF2B4* by sequencing (described in Section 4.2), the EST stSG149 (originally designed from *KIAA0064*) was known to map to the *GCKR-KHK* intergenic region.

4.3.1.2 The KIAA0064 transcript

The putative *KIAA0064* cDNA sequence was deduced during a study to predict coding sequences of unidentified human genes by analysis of clones from a cDNA library of the human immature myeloid cell line KG-1 (Nomura *et al.*, 1994). In that study, cDNA clones were only chosen for further characterisation if they fitted the following criteria: 1) sequencing of the 5′-terminal ends of the cDNA clone inserts revealed no significant similarity to known gene sequences, and 2) the cDNA clone inserts were >2 kb in size, and corresponded to at least 90% of the transcript sizes shown by Northern hybridisation. The *KIAA0064*-containing cDNA clone fitted these criteria and was completely sequenced (Genbank accession number D31764). This strategy was utilised in order to increase the chances of sequencing only cDNA clones that contained a complete coding region from a novel gene.

The 2043 bp *KIAA0064* cDNA sequence contains an open reading frame encoding 470 amino acids. Gene expression analysis by Northern hybridisation revealed that *KIAA0064* is ubiquitously expressed (Nomura *et al.*, 1994). At the time of its original description, the *KIAA0064* sequence showed no similarities to any sequence in the Genbank/EMBL databases. However, as described in this chapter, database comparisons now show the protein encoded by *KIAA0064* to be highly similar to three other hypothetical proteins (two from *D. melanogaster* and one from *C. elegans*). Although the function of KIAA0064 remains unknown, amino acid sequence analysis shows KIAA0064 to contain two protein motifs: 1) a motif found at the active site of serine proteases (the trypsin family), and 2) a motif known as the phox homology (PX) domain. This latter motif is a novel domain that has so far been identified in the NADPH oxidase subunits, sorting nexins, and phosphatidylinositol 3-kinases (Ponting, 1996). Although there is no experimental evidence for the function of the PX domain, a short amino acid motif within the PX domain (proline-X-X-proline, where X is any amino acid), may act as a binding partner for Src homology (SH) 3 domains of other proteins (Kay *et al.*, 2000; Mayer & Eck, 1995; Pawson & Gish, 1992). SH3 domains are 50-70 amino acids long and are often present in eukaryotic signal transduction and cytoskeletal proteins. The ligand specificity of SH3 domains was revealed by investigating the peptide binding properties of SH3 domain-containing proteins, for example Src and phosphatidyl inositol 3-kinase regulatory subunit (Rickles *et al.*, 1994). By displaying peptide libraries on beads or phage, the optimal ligand preference for SH3 domains was found to be for a proline rich peptide sequence based around a proline-X-X-proline core (where X is any amino acid).

This section describes the elucidation of the genomic structure and precise location of *KIAA0064*, further examines the protein motifs present in KIAA0064, and assesses any possible functional relationship between *KIAA0064* and the adjacent *EIF2B4* gene.

4.3.2 Methods

4.3.2.1 Isolation of genomic clones

The primer pair for stSG149 (a: 5′-dGGATTGCCCCTTCTCTTTTC-3′ and b: 5′-dAACACAGACCTCTGCCCATC-3′), which amplifies a PCR product of size 180 bp (corresponding to the *KIAA0064* 3′UTR nucleotides 1817 to 1996, Genbank accession number D31764), was used to screen the *KHK-GCKR* YAC and PAC contig.

The cosmid clone 'A4', which had been previously been shown to contain the whole of the *EIF2B4* gene (see Section 4.2), was also found to contain the 5′ end of *KIAA0064* by direct sequencing upstream from *EIF2B4*, using a primer designed in *EIF2B4* intron 1 called G12int (5′-dAGGCCCGGCTCTTCACTC-3′).

4.3.2.2 *KIAA0064* structure analysis

Oligonucleotides were designed from the *KIAA0064* cDNA sequence (Genbank accession number D31764) and used to obtain sequence data across the exon-intron boundaries by direct sequencing on the cosmid clone A4. Intron sizes were determined either by sequencing across the whole intron or by PCR amplification across the intron using cosmid genomic DNA as template and primers within adjacent exons. The PCR products were analysed on a 1·5 % agarose gel and sized against a Gibco BRL 1 kb ladder.

4.3.2.3 Sequence analysis

The cDNA and protein sequence analysis was performed using the BLAST program (Altschul *et al.*, 1990), based at NCBI. Protein alignments were produced using the GCG ClustalW program (Thompson *et al.*, 1994).

4.3.3 Results

4.3.3.1 Chromosomal mapping

Fine STS content mapping of stSG149 to clones in the *GCKR-KHK* physical contig shows *KIAA0064* to map approximately 150 kb from *GCKR* and 350 kb from *KHK* (Figure 4.20 is a simplified version of Figure 3.8 shown in Chapter 3). The genomic clones shown to contain *KIAA0064* by PCR screening were YACs 29IH8 and 26BA11, PACs 13L13 and 119A8, and cosmid subclone A4 (refer to Figure 4.20).

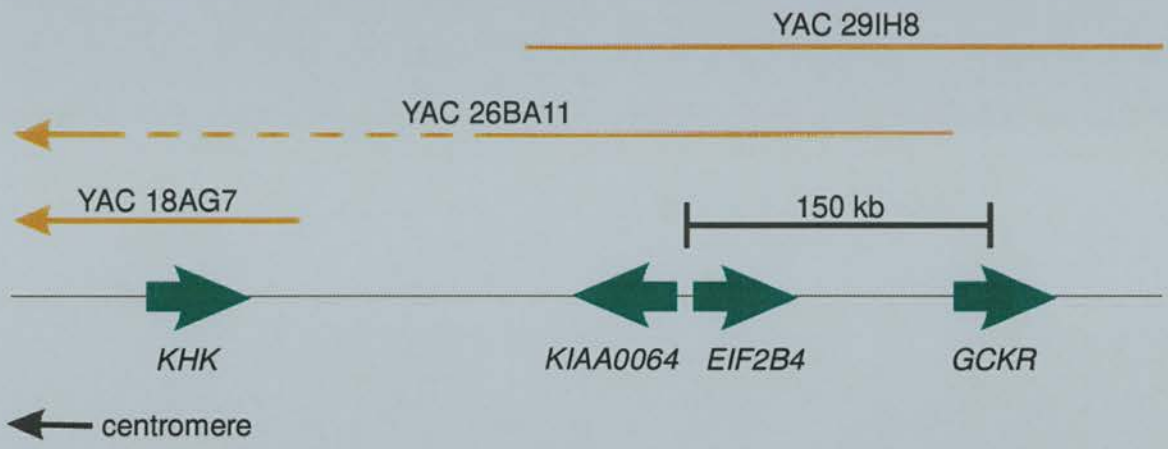


Figure 4.20 Co-localisation of *KIAA0064* with *EIF2B4*, *GCKR* and *KHK*. The arrows represent genes, gene orientation (5'to 3') is indicated by the arrow direction. YAC genomic clones are indicated. This diagram is a simplified version of Figure 3.8 in Chapter 3.

4.3.3.2 *KIAA0064* genomic structure

The DNA sequencing and inter-exon PCR analysis of cosmid A4 reveals the human *KIAA0064* gene to span approximately 6.2 kb of genomic DNA, consisting of 15 exons and 14 introns. The complete *KIAA0064* genomic sequence (except intron 2) was obtained by sequence analysis of cosmid A4 (Figure 4.21). The sequences were deposited in the EMBL sequence database and assigned the accession numbers as follows: *KIAA0064* exons 1-2, AJ404855; and *KIAA0064* exons 3-15, AJ404856. The first methionine codon of the open reading frame is located in exon 1 at nucleotides 457–459, the stop codon at nucleotides 5413–5415 (Figure 4.21). The exons range in size from 42 bp to 522 bp and the introns from 85 bp to 1.1 kb. The *KIAA0064* genomic structure and its intimate positional relationship to *EIF2B4* are shown in Figure 4.22. The genomic sequence of the *KIAA0064* coding regions from cosmid A4 agreed with the published cDNA sequence (accession number D31764). All intron/exon boundaries conform to the consensus sequence (C/T/A)AG-exon-GT(G/A), see Table 4.5.

KIAA0064 genomic sequence

```
cctcgcgaacagccacggccacagcagccatcgccctcagtcctaggctccgccgccgtggtcacggagc 70
cgccccgctgcctcgctccgcggctgcagggtccacagcgccacctgggctggaggctgagccgcagaag 140
cgaaccggcagccccaatcggactacaatttccgacgggccccgaggccgccccgcactacgggttgcaa 210
tgagggccaactccactttgggctCCCTATCCGGACAGGTGGCTCTTGCCCTTTAGACTACAGTTCCCAG 280
CATGCCCAGGCGATTGCGTCCCAGAACCGACGTCCCACCGCCTTCCCACATCGGATCGCAGGGCTCCCAA 350
AATGGCGAGTGAGGCTGCGGGGACTCGCTGAGCAGCGGAGGGGGAGCGTGCAGAGCCGCTGCGGCCCTCA 420
CAGTCCGGAGCCCGGCCGTGCCGTGCCGTAGGGAAC ATGCACTTTTCCATTCCCGAAACCGAGTCCCGCA 490
                                      M  H  F  S  I  P  E  T  E  S  R  S  12
GCGGGGACAGCGGCGGCTCCGCCTACGTGgtgaggagcggccggagccgagccgggccgggcaggggcgg 560
  G  D  S  G  G  S  A  Y  V                                              21
ggataacgggcccgagccatctcggagcgcctctgaggcctgaccgcccatctcgggcctggtcctgcgc 630
cgactcccgacggcctctctgggagcttatctcatgtgtgtgattcgtgcagtggtcgcccgagctcctg 700
tagcccttaggccgggtcgcttccccgagctaggccgcagctgccctggtgggcgtcactggcctcttgg 770
gcagtgagactgcaacacaagcgtccgaatttcccccggactggctcttcagctttgggacttggccctg 840
gcgaaaacaagcgaactatccctcgggagattagcgcgttactccggcgagttgccaggcaacttccggc 910
cagggttctcaggggtgtgcgcagaccgcaaccctaaaatgcttgactacttttgtgtcctcag GCCTAT 980
                                                                  A  Y   23
AACATTCACGTGAATGGAGTCCTGCACTGTCGGGTGCGCTACAGCCAGCTCCTGGGGCTGCACGAGCAGg 1050
N  I  H  V  N  G  V  L  H  C  R  V  R  Y  S  Q  L  L  G  L  H  E  Q     46
tgggactagcacccctgccttgagacagcttcaagcccttccctacacgtggacatcagtgtctcccttt 1120
acccgctcttcttgttcccgtagctgtagttcccctttaatcactgccacagggtagggccctggtttat 1190
cttgtgacatttccgggaactccc                                               1214
intron 2 (approx. 1.1 kb)
ggggaacagggactgagggaagggaggggcacccaagagaacattatgagcatatgtagattgcctcatc 1284
tcattttctgtgtttatgtgaagggttgtatctctttctctaaatag CTTCGGAAGGAGTATGGGGCCAA 1354
                                                  L  R  K  E  Y  G  A  N   54
TGTGCTTCCTGCATTCCCCCCAAAGAAGCTTTTCTCTCTGACTCCTGCTGAGGTAGAACAGAGGAGAGAG 1424
 V  L  P  A  F  P  P  K  K  L  F  S  L  T  P  A  E  V  E  Q  R  R  E   77
CAGTTAGAGAAGTACATGCAAGCTGgtgagtggttgcaggaaactaggttgactatattgaggactatgg 1494
Q  L  E  K  Y  M  Q  A  V                                              86
ggagagacttaaaaacagctggttctgtggaagggcctgatttaatttctagatctaggaaggctcttgt 1564
ttgatctgttatgcagttagcaaagtaaatagtctacgaggatgtgtaactctgtgtactagcaaaacaa 1634
tttgggcactaacatcgatgagaaatactagtataacactagtcttaatatcagtggagagaacttaatt 1704
tgtatctgatgtgctttagattaacatcattgttatttccctagttctatagactggactcacctagatg 1774
gcagccaagggtggggcaggtgaagaatttatgcaaacagtgagtgaggtgaggccagctgggggattgg 1844
gcaggccaccatcagcccaggaaatggggtcgggggaacctactgagaggaggactggggagccaagagt 1914
gagtgctaagttcctgtaataatgttctcttgtcctcgtag TTCGGCAAGACCCATTGCTTGGGAGCAGC 1984
                                            R  Q  D  P  L  L  G  S  S   95
GAGACTTTCAACAGTTTCCTGCGTCGGGCACAACAGgtagggctttgggtgggaccaaggatttaaggaa 2054
E  T  F  N  S  F  L  R  R  A  Q  Q                                     107
aggaccttgctggaaggtcatgtcttttatttttttgagacagagtcttgctcttttgcacaggctggag 2124
tgcagtggtgcgatctcagctcactgcaatctccgcctcccggattcaagcgattctcctgcctcagcct 2194
cccaagtagctagaattacaggcgcctgcaccatgtccggctaatttttgtatttttttagtggagacggg 2264
gttttgccgtgttggcctggctggtctcgaactcctgacctcaggtgatcaccgctggctccaaagtgct 2334
gggattacggtgtgagccaccacacccagccaggccatgtcatcttaataacagaggagtcatctttttt 2404
taatggaatggaacttctattcttatgaaatcaaagaaatcacagcctatatttaagcagctgagaataa 2474
agacacacaagggacctagaaatagagttgggcataggtcaaggcaagggacaaggcagtgcttcccatc 2544
cccatgtgtgctcacaacag GAGACACAGCAGGTCCCCACAGAGGAAGTGTCCTTGGAAGTGCTGCTCAG 2614
                      E  T  Q  Q  V  P  T  E  E  V  S  L  E  V  L  L  S  124
CAACGGGCAGAAAGTTCTGGTCAACGTGCTAACTTCAGATCAGACTGAGGATGTCCTGGAGgtgaggcgc 2684
 N  G  Q  K  V  L  V  N  V  L  T  S  D  Q  T  E  D  V  L  E            144
ttgttcagcactgccccttcttcccctacatcctggggtcctgggcttggagaatgattggaaccagcca 2754
gtatgatgagctgtacttctatccctatccccag GCTGTAGCTGCAAAGCTGGATCTTCCAGATGACTTG 2824
                                    A  V  A  A  K  L  D  L  P  D  D  L   156
ATTGGATACTTTAGTCTATTCTTAGTTCGAGAAAAAGAGGATGGAGCCTTTTTCTTgtgagtttctctgga 2894
I  G  Y  F  S  L  F  L  V  R  E  K  E  D  G  A  F  S  F                175
cttgactgcagtacaagggtacttcagtgagtttgcagccccctaactccccaccccagaatgaacat 2964
tgcctcagggcacttctttcttgatttcccttcctttaaatctcactatattttcatttttttttttta 3034
acctgtag TTGTACGGAAGTTGCAAGAGTTTGAGCTGCCTTATGTGTCTGTCACCAGCCTTCGGAGTCAA 3104
         V  R  K  L  Q  E  F  E  L  P  Y  V  S  V  T  S  L  R  S  Q   195
GAGTATAAGATTGTGCTAAGGAAGAGgtcagggctgggcctggaaggggaggggtgggaggtgctgtgct 3174
E  Y  K  I  V  L  R  K  S                                              204
```

Continued on next page.

Figure 4.21 The complete *KIAA0064* genomic sequence and its putative translation product.

```
ggattggattattgggccacatattgagacagaataatctggacaaggggtgtagtgctgggtcgtttc 3244
tgccaggcctcccaagccttcctcccactataccattagccctgatagttccagattgctcctgttttgc 3314
ccatttccattctacctcttgccttaacccagcttagccccattacccctttccacccctggcctcac 3384
agTTATTGGGACTCTGCCTATGATGACGATGTCATGGAGAACCGGGTTGGCCTGAACCTGCTTTATGCTC 3454
   Y  W  D  S  A  Y  D  D  D  V  M  E  N  R  V  G  L  N  L  L  Y  A  Q 227
AGgtgagcttggagctgcctcagaacccttcccctgaagaaaataacgggtgactcactctcccaagaaa 3524
actgctgtcgtgtctttgttcccaatttgtggggaagttcctagaatactatcagtgctgctggcacaat 3594
atctactctccctacttttattaagctgacaagatcaaggacactgcagagaggtcgctcaagtgtgggg 3664
ctaggggtcaggctggacagaggtaatggtagacgctttccttggattgctgactgggacctcctactgc 3734
ctgccccttgtctctactatagACGGTATCAGATATTGAGCGTGGGTGGATCTTGGTCACCAAGGAACAG 3804
                        T  V  S  D  I  E  R  G  W  I  L  V  T  K  E  Q  243
CACCGGCAACTCAAATCTCTGCAAGAGAAAGTCTCCAAGAAGGAGgtgagccctgcctcctctctgtctt 3874
H  R  Q  L  K  S  L  Q  E  K  V  S  K  K  E                          258
cctctaagggcttgcagtggcgcgatcttggctcactgcaagctctgcctcccagattcatgccattctc 3944
ccgcctcagcctcccgagtagctgggactacaggcacccgccacgacgcccggctaatttttgtatttt 4014
tgggagagacagggtttcaccatgttagccaggatggtcttgagctcctgaccttctgatctgtctgcct 4084
ctgcctcctaaagtgctgggattatagccatgagccaccgccgccgaccgggttgcttttctgagctgc 4154
cccattctccctcctaatctacccccatgtgatgaccatttttcagTTCCTGAGACTGGCCCAGACGCTG 4224
                                                F  L  R  L  A  Q  T  L  266
CGGCACTATGGCTACTTGCGCTTTGATGCCTGTGTGGCTGACTTCCCAGAAAAGGACTGTCCTGTGGTGG 4294
R  H  Y  G  Y  L  R  F  D  A  C  V  A  D  F  P  E  K  D  C  P  V  V  V 290
TGAGCGCGGGCAACAGTGAGCTCAGCCTGCAGCTCCGCCTGCCTGGCCAGCAACTCCGAGAAGGCTCCTT 4364
  S  A  G  N  S  E  L  S  L  Q  L  R  L  P  G  Q  Q  L  R  E  G  S  F  313
CCGGGTCACCCGCATGCGATGCTGGCGGGTCACCTCCTCTgtgagtcgggttaggagggggaagggcctg 4434
 R  V  T  R  M  R  C  W  R  V  T  S  S                              326
ggttgggggccgcaagccttgagcttaggtatgggctgcatgctgggtcagggagctcagacatggttgg 4504
tcagagtgaactcaaccgaattccccttcctctcccagGTACCATTGCCCAGTGGAAGCACGAGCAGCCC 4574
                                        V  P  L  P  S  G  S  T  S  S  P  337
AGGCCGGGGCCGGGGTGAGGTGCGCCTGGAACTGGCTTTTGAATACCTCATGAGCAAGGACCGGCTACAG 4644
 G  R  G  R  G  E  V  R  L  E  L  A  F  E  Y  L  M  S  K  D  R  L  Q  360
TGGGTCACCATCACTAGCCCCCAGgtgtgaacctaccctcagccctcctctgggcaccttaaagtgtaga 4714
W  V  T  I  T  S  P  Q                                              368
gtttccagtactcaatatgggaggcattagtggtgctgggctcagagtgaccactctcatcaagctggac 4784
actcttcttgccctcagGCTATCATGATGAGCATCTGCTTGCAGTCCATGGTTGATGAACTGATGGTGAA 4854
                  A  I  M  M  S  I  C  L  Q  S  M  V  D  E  L  M  V  K  386
GAAATCTGGCGGCAGTATCAGGAAGgtaggcagcaagtgtggactgagcagtgagcaggtgtgctcccct 4924
 K  S  G  G  S  I  R  K                                              394
tgcctttgtcctagatgtgaggcttgtcttgagagagaggggaagtgatcctgccctcaccggcacctt 4994
gtgtctgtccccagATGCTGCGCCGGCGGGTGGGGGGTACTCTGAGACGCTCAGACAGCCAGCAAGCAGT 5064
               M  L  R  R  R  V  G  G  T  L  R  R  S  D  S  Q  Q  A  V  413
GAAGTCCCCACCACTGCTTgtaagtattacctcctggtcagaaccctggctctcagccctgcctcacctc 5134
 K  S  P  P  L  L                                                    419
tcctagtgagtttctgacacctctgcctcttcttccccagGAGTCACCTGATGCCACCCGGGAGTCTATG 5204
                                        E  S  P  D  A  T  R  E  S  M  429
GTCAAACTCTCAgtgagttccagcgttggtgaggttgctgtttgttgggggatcttgcacgcccaagatc 5274
V  K  L  S                                                           433
tctgaccccaccctgcctttgttacagAGTAAGCTGAGTGCCGTGAGCTTGCGGGGAATTGGCAGTCCCA 5344
                           S  K  L  S  A  V  S  L  R  G  I  G  S  P  S  448
GCACAGATGCCAGTGCCAGTGATGTCCACGGCAATTTCGCCTTCGAGGGCATTGGAGATGAGGATCTGTA 5414
  T  D  A  S  A  S  D  V  H  G  N  F  A  F  E  G  I  G  D  E  D  L  *  470
ATCTCCACTGCTTGGATGTCTGCCCTCTACCCCAGAGGAATTTACAGAAACTTGCCCTGTGCCTGTGTCC 5484
CCCATGCTAGGGGCGGAGGGGTCTTTTCCTTCTTCTTTCCTACCTACCCCTTTTCTCTTGGCCAGGGGCC 5554
TCGTATCCTACCTTTCCTTGTCCCCTGGGCTGGCTGCACAGAGGATTGCCCCTTCTCTTTTCAGAGCTGG 5624
CCCTCGATGCCAAATTAGCATTTAGTATTTTGCACAAAGTCTAAGGGACCATGGCTGCCTGCCTTGGGGA 5694
GGAACCATAGCTCCCTCTGGGCCGCTTCTGGCCTCTTGGAGCCATGGGCCAAAGGCCAAGGGGATGGGCA 5764
GAGGTCTGTGTTTGGTCTGGCCCAGTTCCCCATCATTAAACTCAGCCTGACTGCTGCCTacctctggttc 5834
cctctcactgcccctgcttcccccatcacagcatggactactatgctaagggtaaggccaaattgcctgc 5904
cattgccaattcagcatagacatggctttgttagtgtttcctttattataaagcactgaaataagttaaa 5974
taaacaggtgggaggctgggcagtcccccagccggtttgtccacagcccctggggcgagtgaggtgaata 6044
cagggcccttctcactgagctcgtgaagtgcctcagtcaaggcaaggtcccctggtccatatgggccccc 6114
ccgcccatgggttcccctgctcctt                                            6139
```

Figure 4.21 The complete *KIAA0064* genomic sequence and its putative translation product. The translation termination signal is marked with an asterisk.

Exon1-GCGGCGGCTCCGCCTACGTGgtgaggagcggccggagccg----------455 bp----------gactacttttgtgtcctcagGCCTATAACATTCACGTGAA-Exon2
Exon2-TCCTGGGGCTGCACGAGCAGgtgggactagcacccctgcc---------1.10 kb----------gtatctctttctctaaatagCTTCGGAAGGAGTATGGGGC-Exon3
Exon3-AGAGAAGTACATGCAAGCTGgtgagtggttgcaggaaact----------506 bp----------atgttctcttgtcctcgtagTTCGGCAAGACCCATTGCTT-Exon4
Exon4-TCCTGCGTCGGGCACAACAGgtagggctttgggtgggacc----------544 bp----------cccatgtgtgctcacaacagGAGACACAGCAGGTCCCCAC-Exon5
Exon5-AGACTGAGGATGTCCTGGAGgtgaggcgcttgttcagcac----------113 bp----------acttctatccctatccccagGCTGTAGCTGCAAAGCTGGA-Exon6
Exon6-AGAGGATGGAGCCTTTTCTTgtgagtttctctggacttga----------163 bp----------ttttttttttaacctgtagTTGTACGGAAGTTGCAAGAG-Exon7
Exon7-AAGATTGTGCTAAGGAAGAGgtcagggctgggcctggaag----------256 bp----------tccacccctggcctcacagTTATTGGGACTCTGCCTATG-Exon8
Exon8-TGAACCTGCTTTATGCTCAGgtgagcttggagctgcctca----------300 bp----------gccccttgtctctactatagACGGTATCAGATATTGAGCG-Exon9
Exon9-AGAAAGTCTCCAAGAAGGAGgtgagccctgcctcctctct----------351 bp----------atgtgatgaccatttttcagTTCCTGAGACTGGCCCAGAC-Exon10
Exon10-GCTGGCGGGTCACCTCCTCTgtgagtcgggttaggagggg----------138 bp----------aattcccctttcctctcccagGTACCATTGCCCAGTGGAAG-Exon11
Exon11-TCACCATCACTAGCCCCCAGgtgtgaacctaccctcagcc----------133 bp----------gacactcttcttgccctcagGCTATCATGATGAGCATCTG-Exon12
Exon12-CTGGCGGCAGTATCAGGAAGgtaggcagcaagtgtggact----------129 bp----------caccttgtgtctgtccccagATGCTGCGCCGGCGGGTGGG-Exon13
Exon13-TGAAGTCCCCACCACTGCTTgtaagtattacctcctggtc-----------91 bp----------ctctgcctcttcttccccagGAGTCACCTGATGCCACCCG-Exon14
Exon14-AGTCTATGGTCAAACTCTCAgtgagttccagcgttggtga-----------85 bp----------ccaccctgcctttgttacagAGTAAGCTGAGTGCCGTGAG-Exon15

Table 4.5 *KIAA0064* exon/intron splice junction sequences and intron sizes. Intron sequences at the 5'- and 3'-end of the exons are in lower case letters. Exon sequences are in capital letters.

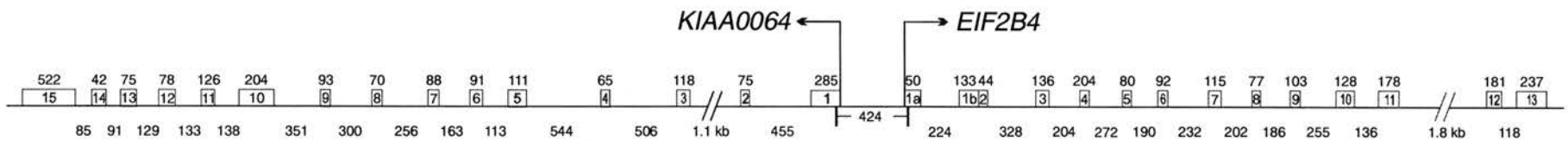


Figure 4.22 Genomic structures of *KIAA0064* and *EIF2B4*. These two genes are arranged in a "head to head" orientation with 424 bp between the start codon of *KIAA0064* and the start codon of *EIF2B4* exon 1a. Exons are indicated by boxes (containing the exon number). Exon sizes (bp) are indicated above the boxes and intron sizes (bp unless indicated as kb) are also shown.

4.3.3.3 cDNA analysis

A BLASTN search of the Genbank database using the *KIAA0064* cDNA sequence revealed two significant hits, the *Homo sapiens* Eph-like receptor kinase *EPHB1B* gene (accession number AF037332) and a *Homo sapiens* CpG island DNA genomic *Mse*I fragment, clone 84g11 (accession number Z63465). The *KIAA0064* and *EPHB1B* sequences were 97.5% identical (1840 out of 1887 nucleotides were identical) over a region of *KIAA0064* from nucleotides 142-2028 (the *KIAA0064* cDNA sequence is 2043 nucleotides in length; accession number D31764) and of the *EPHB1B* 3'UTR from nucleotides 5330-7221 – see Figure 4.23. Examination of the two sequences shows small deletions and additions of nucleotides in the *KIAA0064*-like sequence of the *EPHB1B* 3'UTR compared to *KIAA0064*. There are more deletions than additions, resulting in the *KIAA0064*-like sequence of the *EPHB1B* 3'UTR being 5 nucleotides shorter than the region of similarity with *KIAA0064*. Although the putative start codon of *KIAA0064* is retained in the *KIAA0064*-like sequence of the *EPHB1B* 3'UTR, there is no corresponding long open reading frame, due to a single C nucleotide deletion in the *EPHB1B* 3'UTR (corresponding to nucleotides 1038, Figure 4.21), which results in a stop codon (corresponding to nucleotides 1393-1395, Figure 4.21). Therefore, although the *KIAA0064*-like sequence of the *EPHB1B* 3'UTR is transcribed as part of the *EPHB1B* transcript, it appears not to be a protein coding sequence. This suggests the *KIAA0064*-like sequence in the *EPHB1B* 3'UTR to be a processed pseudogene of *KIAA0064*.

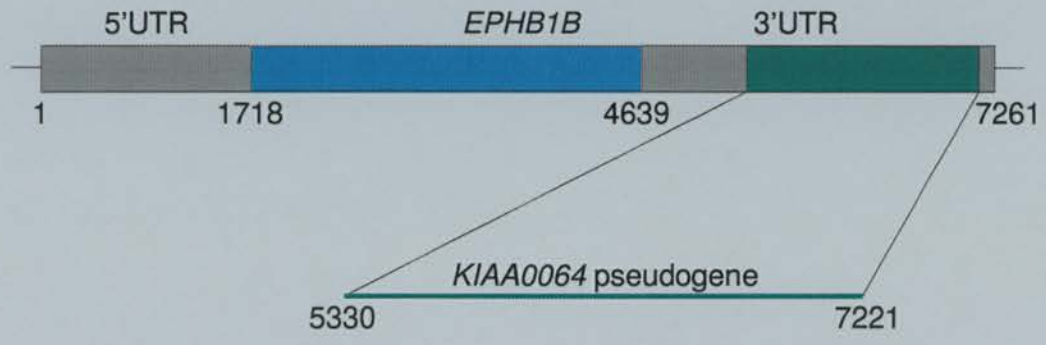


Figure 4.23 Location of a *KIAA0064* pseudogene in the 3'UTR of the *EPHB1B* gene. The diagram represents the *EPHB1B* cDNA (nucleotides 1-7261). The *EPHB1B* coding region (nucleotides 1718-4639) is represented by blue, and the *EPHB1B* 5'UTR and 3'UTR are shown as grey boxes. The *KIAA0064* pseudogene is represented by a green box (nucleotides 5330-7221).

Closer examination of the similarity between *KIAA0064* and the *Homo sapiens* CpG island DNA genomic *Mse*I fragment, clone 84g11, reveals that the region of similarity encompasses the whole of *KIAA0064* exon 2. Comparison of the *KIAA0064* exon 2 and intronic sequences (IVS1 and IVS2) with the clone 84g11 database sequence shows that the clone 84g11 database sequence corresponds to nucleotides 940-1155, Figure 4.21. This suggests that it is highly likely that the 84g11 clone originated from chromosome 2p23.3. Analysis of the *KIAA0064* genomic sequence obtained in the present study, using the GRAIL/CpG island program (Roberts, 1991) predicts nucleotides 1-777 to be a CpG island, Figure 4.21. This CpG island extends into the *KIAA0064-EIF2B4* intergenic region (described in Section 4.2).

4.3.3.4 Predicted KIAA0064 protein analysis

A BLASTP search of EMBL databases using the KIAA0064 amino acid sequence reveals that this protein shows greatest similarity to three hypothetical proteins: the two *D. melanogaster* hypothetical proteins, accession numbers AAF52870 and AAF46070 (40% and 26% identity, respectively), and the *C. elegans* hypothetical 54.2 kDa protein F17H10.3 in chromosome X, accession number CAA93650 (29% identity). There was also similarity to a lesser extent to another *C. elegans* hypothetical protein (accession number CAB02091), and to two hypothetical open reading frames in *S. cerevisiae* and *S. pombe* (accession numbers CAB02091 and CAA88260, respectively). No functional information is known about these hypothetical proteins, so that no predictions can be made concerning KIAA0064 function. However, two amino acid sequence motifs are recognisable within the KIAA0064 sequence.

The first motif is found at the active site of serine proteases and spans from amino acids 10 to 21 in KIAA0064 (Figure 4.24). The active site of an enzyme is the region that binds substrates and contains the residues that directly participate in the making and breaking of bonds. In the case of serine proteases, a highly reactive serine residue plays a critical role in catalysis that results in the hydrolysis of a peptide substrate. It is the nucleophilicity of the serine –OH at the serine protease active site that gives the serine residue the name of "active serine". The structure and mechanism of a serine protease enzyme was first defined in bovine chymotrypsin (Birktoft & Blow, 1972; Kraut, 1977) and has subsequently been defined in many other active serine enzymes.

The KIAA0064 serine protease active site motif is encoded entirely within *KIAA0064* exon 1. Although the sequence comparison shown in Figure 4.24 suggests there to be a serine protease active site in KIAA0064, experimental evidence is required to be certain that this a functional motif.

```
                                    *
KIAA0064            10 S R S G D S G G S A Y V^21
Plasminogen, human     S C Q^G D S G G V F A V
UPA, human             S C Q^G D S G G P L V C
Trypsin, rat           S C Q^G D S G G P V V C
TLP (hepsin), Human    A C Q^G D S G G P F V C
```

Figure 4.24 Comparison of the putative KIAA0064 serine protease active site to similar motifs found in various other serine proteases. The putative active serine residue is starred (that is, the highly reactive serine residue that acts as a nucleophile during catalysis). The putative KIAA0064 active serine was assigned by similarity of the KIAA0064 active site motif to known serine protease active site sequences. The symbol ^ indicates the location of a splice junction in the amino acid sequence. Abbreviations: UPA, urokinase plasminogen activator, TLP, trypsin-like protease.

BLAST analysis of the KIAA0064 peptide sequence also reveals similarity of amino acids 1-105 of KIAA0064 to a protein motif called the PX domain. The PX domain is found in a large group of eukaryotic proteins including NADPH oxidase subunits, sorting nexins (proteins that are thought to be involved in targeting cell surface receptors to the lysosome (Haft *et al.*, 1998), and phosphatidylinositol 3-kinases (Ponting, 1996). Figure 4.25 shows a consensus PX domain compared to the putative KIAA0064 PX domain. This shows that the KIAA0064 PX domain spans from amino acid residues 1-105, and contains several highly conserved residues including a two prolines (amino acids 57-60) that might represent a binding domain (proline-X-X-proline, where X is any amino acid) for the SH3 domain of other proteins (Kay *et al.*, 2000).

```
PX domain        GRIAQ.VVVLERETSGKDLGDSKHYYYVIEIETKTGSRWTVYRRYSDFYE
                 ++++++ ++++  +S       +  +Y++++I++++     +++ RYS++++
KIAA0064       1 MHFSIPETESRSGDS------GGSAYVAYNIHVNG--VLHCRVRYSQLLG 42

PX domain        LHEKLLERFPQEKFEKLRKRIIPPLPEKKLLGRKREEEEEGILSSLTVKY
                 LHE+L+++++          +++P +P+KKL++
KIAA0064      43 LHEQLRKEYG--------ANVLPAFPPKKLFS------------------ 66

PX domain        ASEPKLDEEFIEKRRRQLEKYLQRLLNHPELINEYRVSELVLEFLESS
                      L ++ +E+RR+QLEKY+Q++ ++P+L +    SE + +FL
KIAA0064      67 -----LTPAEVEQRREQLEKYMQAVRQDPLLGS----SETFNSFLRRA   105
```

Figure 4.25 Comparison of a consensus PX domain (Ponting, 1996) to the putative PX domain contained in KIAA0064 (amino acid residues 1-105). Matching residues are shown as bold type and in shaded boxes. This figure is adapted from that shown at the Kazusa web page (http://zearth.kazusa.or.jp/en/).

4.3.4 Discussion

4.3.4.1 Characterisation of *KIAA0064*

Section 4.3 describes the determination and analysis of the structure of the putative gene *KIAA0064*. This gene was identified to lie in the *GCKR-KHK* intergenic region by sequencing upstream of the *EIF2B4* gene (a gene characterised in Section 4.2) and also by mapping of the EST stSG149 (see Chapter 3, Figure 3.8). *EIF2B4* had been investigated due to its possible regulation by glycaemic status. The intimate "head to head" arrangement of *KIAA0064* with *EIF2B4* (see Section 4.3, Figure 4.22) inevitably raises questions about possible regulatory interactions between the two genes. There is less than 200 bp of DNA between the putative transcriptional start sites of *KIAA0064* and *EIF2B4*, suggesting that *KIAA0064* and *EIF2B4* may share promoter and/or regulatory elements. This makes it logical to have included an investigation of *KIAA0064* along with *EIF2B4*, which was selected for study because of evidence that it is glucose-regulated.

The *KIAA0064* cDNA had previously been cloned and shown to be 2043 nucleotides in size, containing an open reading frame encoding 470 amino acids (Nomura *et al.*, 1994). The characterisation of the *KIAA0064* genomic structure described in this chapter reveals *KIAA0064* to span approximately 6.2 kb of genomic DNA, consisting of 15 exons and 14 introns with the exons ranging in size from 42 bp to 522 bp and introns from 85 bp to 1.1 kb (Figure 4.21). The *KIAA0064* genomic sequence was found to be 100% identical to the CpG island clone 84g11 (accession number Z63465) in the region of *KIAA0064* exon 2, indicating the presence of a CpG island. Computer analysis of the *KIAA0064* genomic sequence confirms this, and shows that this CpG island spans the *EIF2B4-KIAA0064* intergenic region, indicating that the 5′ regions of both *KIAA0064* and *EIF2B4* are CG-rich. As CpG islands are often associated with promoters of genes (Bird, 1993; Delgado *et al.*, 1998), the location of this CpG island at the *KIAA0064–EIF2B4* intergenic region suggests that it is associated with the promoter sequences of both genes. As described in Chapter 3 (see Figure 3.8), a further 4 ESTs were mapped close to the *KIAA0064* and *EIF2B4* genes (stSG9174, WI-19785, EST1, and EST2, - refer to Table 3.4 in Chapter 3 for further EST information). Due to the high density of genes residing at the *GCKR-KHK* interval, it is likely that other CpG islands will reside at this genomic region.

4.3.4.2 Homology between *KIAA0064* and *EPHB1B*

At the cDNA level, *KIAA0064* is almost identical to a region of the *EPHB1B* 3'UTR, a human *EPH*-related receptor protein-tyrosine kinase gene (accession number AF037332). *EPHB1B* is the "long" isoform of *EPHB1* (accession number AF037331) (Tang *et al.*, 1995). Both *EPHB1*gene isoforms share identical coding regions but differ at their 5'UTRs and 3'UTRs. The striking difference between *EPHB1* and *EPHB1B* is the presence of a much longer 3'UTR in *EPHB1B* (2622 nucleotides compared to 695 nucleotides). Examination of the *EPHB1B* sequence reveals that the region of similarity with *KIAA0064* occurs only in the *EPHB1B* 3'UTR. Detailed examination of the region of similarity between *KIAA0064* and the *KIAA0064*-like region in the *EPHB1B* 3'UTR reveals, however, that they are not 100% identical. The *EPHB1B* sequence lacks a region of identity with the first 141 nucleotides of *KIAA0064* (part of the 5'UTR), but it does contain an equivalent to the *KIAA0064* start codon (*EPHB1B* nucleotides 5410 to 5412, Figure 4.23). However, comparison of the sequence downstream of this equivalent methionine residue to the *KIAA0064* sequence, reveals that there are a number of differences including single and multiple bp deletions/additions and substitutions that produce a shortened open reading frame that if translated would result in a 68 amino acid truncated protein. The absence of any corresponding *KIAA0064* full length open reading frame within the *EPHB1B* gene sequence, added to the fact that the similarity between *KIAA0064* and *EPHB1B* occurs only in the *EPHB1B* 3'UTR, strongly suggests that the *EPHB1B* 3'UTR contains a processed *KIAA0064* pseudogene.

4.3.4.3 Chromosomal mapping of *EPHB1b* and *KIAA0064*

The radiation hybrid mapping of *KIAA0064* to chromosome 2 (Nomura *et al.*, 1994) was confirmed by two methods. The screening of a human-rodent mono-chromosomal cell hybrid panel using stSG149 (designed from an EST within the *KIAA0064* sequence) confirmed that this STS mapped only to chromosome 2. The sequencing of the *KIAA0064* gene from a cosmid subclone of YAC 26BA11, a YAC mapped by FISH to chromosome 2p23.3 by FISH (Warner *et al.*, 1995) provided evidence for the co-localisation of *KIAA0064* with *EIF2B4*, *GCKR*, and *KHK*. Indeed, mapping of KIAA0064 to the *GCKR-KHK* physical contig described in Chapter 3, shows *KIAA0064* to reside ~150 kb from *GCKR* and ~350 kb from *KHK* (Figure 4.20).

The *EPHB1* gene, which contains the *KIAA0064* pseudogene, was known to map to human chromosome 3 (*EPHB1* and *EPHB1B* are encoded by the same gene and only differ at their

5′ termini; *EPHB1B* has a much longer 3′UTR which contains the *KIAA0064* pseudogene). This had been shown by both PCR screening of a human-rodent somatic cell hybrid panel and by FISH using a *EPHB1*-containing P1 clone (Tang *et al.*, 1995). The primers used to map the *EPHB1*gene were designed from the *EPHB1* 3′UTR sequence shared by both *EPHB1* isoforms, out-with the *KIAA0064* pseudogene, therefore mapping both isoforms to chromosome 3. It was intriguing to find that the sequence for primer stSG149a (stSG149 maps *KIAA0064* to chromosome 2) can be found within the *EPHB1B* 3′UTR and the stSG149b primer sequence also exists except that there is one cytosine nucleotide missing within the primer sequence. The screening of a human-rodent mono-chromosomal cell hybrid panel using stSG149 confirmed that it mapped only to chromosome 2. This proved that the one base deletion within the *EPHB1B* 3′UTR on chromosome 3 within the stSG149b primer sequence was enough to prevent any amplification of a PCR product from the *EPHB1B* gene.

Although the *KIAA0064* mouse homologue has not been mapped, examination of the Jackson map of mouse chromosome 5 in the region of conserved synteny corresponding to human chromosome 2p23 (Wightman *et al.*, 1997), reveals that an EST called *D5Ertd260e* (Ko *et al.*, 2000) co-localises with *Gckr* and *Khk*. This EST, an anonymous cDNA J0071F09 from a 3.5-dpc mouse blastocyst cDNA library, shows significant similarity to the human *KIAA0064* gene and is likely to be the *KIAA0064* mouse homologue. This suggests the co-localisation of *GCKR*, *KHK*, and *KIAA0064* to be evolutionarily conserved in human and mouse genomes and allows for the possibility of common *cis*-acting regulatory elements for these genes.

4.3.4.4 Protein motifs in KIAA0064

The analysis of the KIAA0064 peptide sequence reveals that the N-terminus shows homology to the active site of serine proteases (amino acids 10-21, see Figure 4.24) and a more extended region of homology to a PX domain (amino acids 1-105, see Figure 4.25). There are several residues well conserved between species within both these domains, possibly indicating that they have been conserved to carry out an important function. However, without investigation using biochemical techniques, it is impossible to say whether these motifs possess any functional significance.

The protein similarity search reveals that the closest known human protein to KIAA0064 was a PX domain-containing protein called sorting nexin 1 (SNX1). The SNX1 protein was

first identified using a yeast two-hybrid experiment to identify proteins that bound to the cytoplasmic domain of the epidermal growth factor (EGF) receptor (Kurten *et al.*, 1996). Homology between SNX1 and the yeast protein Mvp1p which is known to be involved in targeting hydrolases to the vacuole, led to the hypothesis that SNX1 may be involved in EGF receptor degradation in lysosomes (Ekena & Stevens, 1995). The identification and cloning of a further three novel proteins that contain a PX domain and show homology to SNX1 led to the formation of a family of sorting nexin proteins (Haft *et al.*, 1998). To date, 14 sorting nexin proteins have been cloned and a search of sequence databases suggests the existence of orthologues in yeast and *C. elegans*.

Studies of PX domain-containing proteins in yeast have revealed that they are often associated with processes involving the actin cytoskeleton, membranes and/or GTP binding proteins. For example, the *S. cerevisiae* protein Bem1 appears to co-ordinate rearrangement of the cortical cytoskeleton during cell polarisation in response to mating factors (Chenevert *et al.*, 1992; Peterson *et al.*, 1994). It is impossible to predict what other proteins the protein encoded by *KIAA0064* would interact without biochemical experimentation, for example by utilising the yeast two hybrid system.

It is also unclear without further experimentation whether genetic interactions between *KIAA0064* and the adjacent *EIF2B4* gene might occur. However, the evolutionary conservation of the intimate genomic location between *KIAA0064* with *EIF2B4* in human and mouse does suggest the possibility of interactions at the transcriptional level, as a result of the sharing of regulatory sequences.

4.4 Ribokinase (*RBSK*)

4.4.1 Introduction

4.4.1.1 Identification of a ribokinase-like EST

During the search for transcripts that map to *GCKR-KHK* genomic region that might have a related function to the GKRP (glucokinase regulatory protein) or KHK (ketohexokinase), an EST (Genbank accession number T69020) was identified that showed significant similarity (45% at the nucleotide level) to the *E. coli* ribokinase gene. A comparison of this ribokinase-like EST to the corresponding region of the *E. coli* ribokinase gene is shown in Figure 4.26. As ribokinase (RBSK) and KHK are both members of the same sugar kinase family (Bork *et al.*, 1993), this human ribokinase-like EST was investigated to examine a possible evolutionary link to *KHK*.

```
T69020   GTAAAAGAACTAATATTTGCAAAGAAAGGGG--ACGAGGACATTTTCTAATAAG-----C 53
         | |  |  |||   |  ||  | |  || ||  |  |   |   |  ||||||     |
E. coli  GAATCACCACTCGAAAGTGTGATGGCAGCGGCGAAAATCGCCCATCAAAATAAGACTATC 4613

T69020   ATTA-GCAGGAGCAGCCACCCCCAAGTACATTTTATTCCCAGGNATATTTATTTTGGGAC 112
          ||  ||   | | |   ||  |  |   | ||  |  | |       |     | | |
E. coli  GTTGCGCTTAACCCGGCTCCGGCTCGCGAACTTCCTGACGAA-CTGCTGGCGCTGGTGGA 4672

T69020   TAATAGCAATCAAAACAGA--GTAAGCGGAAGGTCTTTTTTGTAAGGGT----AAGATGA 166
         | ||  |    ||||  |  | |||| |||   ||     |||  ||    || ||||
E. coli  CATTATTACGCCAAACGAAACGGAAGCAGAAAAGCTCACCGGTATTCGTGTTGAAAATGA 4732

T69020   CTGTGTTCCTGCAGCCTGGACACTGACTGCTGCAATGAAATTGGATCTGTTGAGCA-TGT 225
             | | | ||      | | | |   |||||  |||  | ||| ||   | | ||
E. coli  TGAAGATGCAGCGAAGGCGGCGCAGG-TACTGCATGAAAAAGGTATCCGTACTGTACTGA 4791

T69020   CTTCCAAGGACAGATTTGGATAGTAAGCCAGGTAGAAGGCCAGAGCTCCCACAAAGCTNT 285
          | |    |  ||   |||    |  || ||   ||| |    || ||    |  ||
E. coli  TTACTTTAGGAAGTCGTGGTGTATGGGCTAGCGTGAATGGTGAAGGTC----A--GCGCG 4845

T69020   CACCAGCACCCGTGGTATCCACAGCCTTGACTTTCTCTGTGGGAATGTGCTTTGGCTCAG 345
           || | |  |  |||     | | |  ||  |  |||  |||     ||||  |   |
E. coli  TTCCTGGATTCCGGGTGCAGGCTGTCGATACCATTGCTGCCGGAGATACCTTTAAC---G 4902

T69020   GTTCTGTCTGTGACAGCACCAC----ACATCCTTCAGCCCCTAAGGNTAATGATTACCAC 401
         || | ||   | || |||   |    | |      ||   ||  |||  | |  ||    |
E. coli  GTGC-GTTAATCACGGCATTGCTGGAAGAAAAACCATTGCCAGAGGCGATTCGTTTTGCC 4961

T69020   CTGGGCAGCCCCTTTTCAAGGAGCACTAN-TGCAGCCTNCCCAGCATC--TTNCAG     454
         |  | | ||| ||      |  | |  |  |  || |   | | | ||  |  |
E. coli  CATG-CTGCCGCTGCGATTGCCGTAACACGTAAAGGCGCACAACCTTCCGTACCGT     5016
```

Figure 4.26 Sequence comparison of human EST T69020 sequence with the *E. coli* ribokinase gene (part of a large genomic sequence assigned Genbank accession number M13169). Numbering refers to database entry numbering.

The ribokinase-like EST was first identified by examination of the Genemap99 at NCBI (http:/www.ncbi.nlm.nih.gov/genemap/map.cgi?CHR=2), revealing that SHGC15128, an EST showing significant similarity to *E. coli* ribokinase, Genbank accession number T69020, was located between the markers D2S171 and D2S165 on chromosome 2p23. This places SHGC15128 within the same genomic region as *GCKR* and *KHK* (based upon radiation hybrid mapping data at NCBI). PCR screening of genomic clones located at the *GCKR-KHK* genomic region using SHGC15128 revealed that this EST maps to the YAC 29IH8, already known to contain *GCKR* (see Chapter 3, Figure 3.5). As ribokinase belongs to the same sugar kinase family as *KHK*, ~500 kb from *GCKR* (shown by FISH studies (Hayward *et al.*, 1996) and the *GCKR-KHK* physical contig described in Chapter 3, Figure 3.8), the ribokinase-like EST was chosen for further investigation.

4.4.1.2 Sugar kinases

The main source of carbon and energy used by cells is the metabolism of sugars. After transport into a cell, the first step of metabolism is to phosphorylate the sugar, preparing it for further chemical reactions, either catabolic or anabolic. The main catabolic pathway is glycolysis, a sequence of reactions that converts glucose into pyruvate with the concomitant production of ATP. For glucose to enter the glycolytic pathway, it is first phosphorylated by a hexokinase (see Chapter 1). Other abundant sugars, for example fructose and galactose, can be "funnelled" into the glycolytic pathway, but again the first step of their metabolism is the phosphorylation by their own specific kinase (fructokinase (KHK) and galactokinase respectively). Some sugars are so rare in the natural environment, for example ribulose, xylulose, or fucose, that they can only be processed by specialised micro-organisms.

Glucose, fructose, and galactose, all contain six carbon atoms in their structure and are processed as energy sources by entering the glycolytic pathway. Ribose is another abundant sugar. It however contains five carbon atoms in its structure and once phosphorylated, enters the pentose phosphate pathway, or is used in the biosynthesis of nucleotides, histidine, and tryptophan. The first step of ribose metabolism is the ATP-dependent phosphorylation of ribose to ribose-5-phosphate by ribokinase (see Figure 4.27).

Many cells have a greater requirement for NADPH for reductive biosynthesis than for ribose-5-phosphate for incorporation into nucleotides and nucleic acids. In this situation, ribose-5-phosphate is converted into the glycolytic intermediates glyceraldehyde 3-phosphate and fructose-6-phosphate by two enzymes called transketolase and transaldolase (Katz & Rognstad, 1967). These enzymes constitute a reversible link between the pentose

phosphate pathway and glycolysis. Therefore, excess ribose-5-phosphate formed by the pentose phosphate pathway can be completely converted into glycolytic intermediates.

4.4.1.3 Sugar kinase families

The examination of sugar kinase primary structures in combination with biochemical data reveals at least three distinct non-homologous sugar kinase families: hexokinases, ribokinases and galactokinases (Bork *et al.*, 1993). Each family appears to have strikingly different conserved sequences, suggesting each family to possess its own distinct three-dimensional structural motifs. As kinases catalyse chemically equivalent reactions on similar or identical substrates, this has been interpreted to suggest that the enzymatic function of sugar phosphorylation has evolved independently on three distinct structural frameworks by convergent evolution.

4.4.1.4 The ribokinase family

The ribokinase family of sugar kinases is a structurally distinct group of both prokaryotic and eukaryotic sequences including several fructokinases, the minor 6-phosphofructokinase from *E. coli*, 1-phosphofructokinases, and 6-phosphotagatokinases. The primary structure of every family member contains at least one of three conserved regions, identifying it as a member of the ribokinase family (Bork *et al.*, 1993). The creation of an evolutionary tree by multiple sequence alignment reveals that substrate specificity is the major organising principle. This suggests that in this family, divergence of substrate specificity occurred before species divergence, *ie.* that fructokinases branched off from ribokinases before the divergence of yeast and *E. coli* (Bork *et al.*, 1993).

Ribokinase (RBSK) and ketokexokinase (KHK) are part of the same ribokinase family of kinase proteins because they both contain at least one of the three conserved amino acid sequence motifs. Although RBSK and KHK are part of the same family, they phosphorylate different substrate sugars (ribose and fructose, respectively). Figure 4.27 shows the phosphorylation of ribose to ribose-5-phosphate and fructose to fructose-1-phosphate by ribokinase and ketohexokinase respectively. Comparison of the sugar and sugar-phosphate chemical structures shown in Figure 4.27 reveal that RBSK and KHK carry out similar functions by adding one phosphate group to a carbon atom outwith the carbon ring of each sugar (carbon atom "5" and "1" respectively). Although ribose and fructose are five and six carbon sugars respectively, both sugars contain a five carbon furanose ring (Figure 4.27). Therefore, RBSK and KHK phosphorylate structurally similar substrates. As ribokinases

and fructokinases contain conserved amino acid motifs, these motifs might be involved in kinase function and substrate specificity for furanose sugars. Non-conserved regions of amino acid sequence in ribokinase and fructokinase proteins might be involved in substrate specificity, that is distinguishing between ribose and fructose sugars.

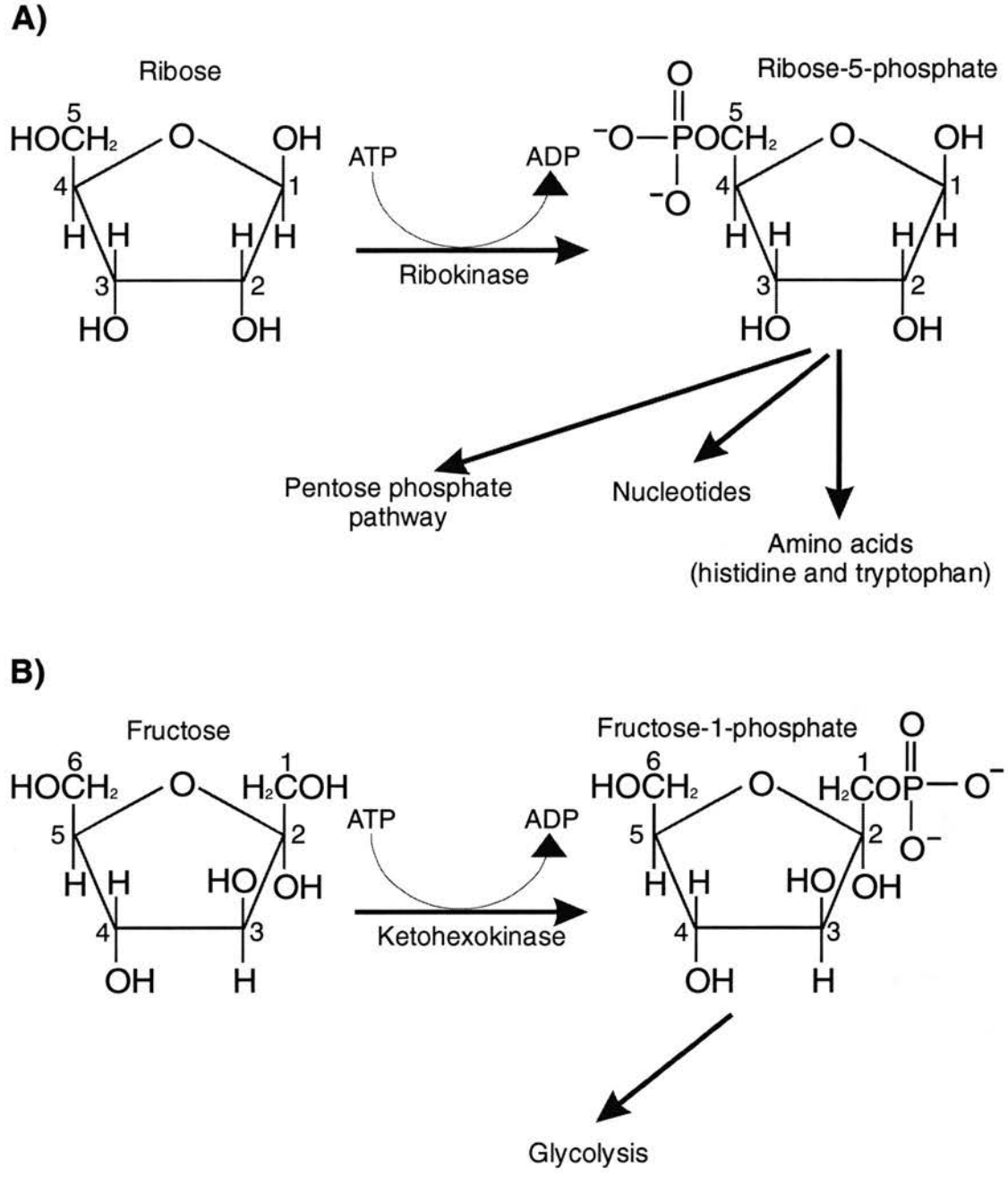


Figure 4.27 Role of ribokinase and ketohexokinase. A) Phosphorylation of ribose to ribose-5-phosphate by ribokinase. B) Phosphorylation of fructose to fructose-1-phosphate by ketohexokinase. The ribose and fructose carbon atoms are numbered 1-5 and 1-6, respectively.

4.4.1.5 Experimental aims

Ketohexokinase (KHK) is a member of the ribokinase family of proteins (Bork *et al.*, 1993). The identification of a ribokinase-like EST near to the *KHK* gene on chromosome 2p23.3, raises the possibility that the ribokinase-like EST and *KHK* gene originate from a common ancestral gene. However, based on the known evolutionary relationships between the ribokinase family members, this would be surprising, since as mentioned above, it is believed that divergence of substrate specificities within the family predates the prokaryotic-eukaryotic divergence. This section describes the cloning and characterisation of the novel ribokinase-like gene. To investigate the possibility that *KHK* and the ribokinase-like gene have evolved by divergent evolution from a common ancestral gene, their amino acid sequences and genomic structures were compared. Similar sequence motifs and similar genomic structural features, might suggest that these two genes have indeed originated from a common ancestral gene, providing a further interesting insight into the genomic evolution of chromosome 2p23.3.

4.4.2 Methods

4.4.2.1 cDNA cloning and sequencing

The EST SHGC15128 (designed from the cDNA clone Genbank accession number T69020) was used in a BLAST search of EMBL and Genbank EST databases. Overlapping cDNA clones were identified and a cDNA contig was thereby constructed that spanned 1196 nucleotides (Figure 4.28). These cDNA clones were obtained from the IMAGE consortium and sequencing performed on both strands using the Thermosequenase cycle sequencing kit (Amersham).

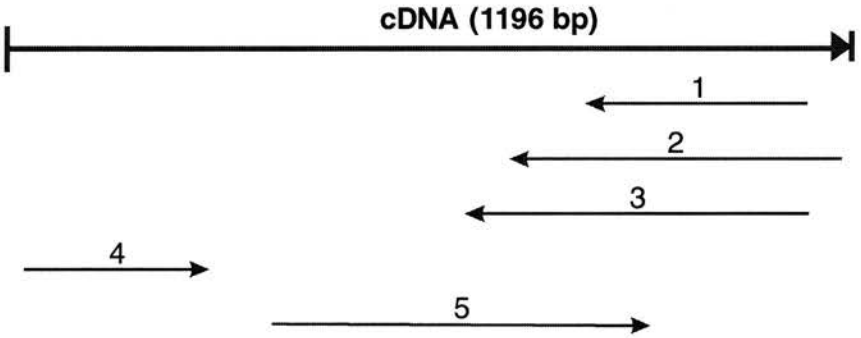


Figure 4.28 Ribokinase EST clone contig. See Table 4.6 for cDNA clone identification. Arrows indicate the extent of the original EST reads.

cDNA clone no.	IMAGE clone number	Size of insert (kb)	cDNA library
1	220049	1.1	Retina N2b4HR, Soares
2	484872	0.68	Pregnant uterus NbHPU, Soares
3	246380	0.8	Fetal liver and spleen 1NFLS, Soares
4	82322	1.1	Liver, Stratagene.
5	484876	0.6	Pregnant uterus NbHPU, Soares

Table 4.6 cDNA clone identification for Figure 4.28.

4.4.2.2 Isolation of genomic clones

The SHGC15128 primer pair (5′-dAAGAAAGGGGACGAGGACAT-3′ and 5′-dCTTACCCTTACAAAAAAGACC-3′; corresponding to nucleotides 940-960 and 1062-1081 respectively, Figure 4.29) was used to screen the PAC human genomic library RPCI1 by PCR (obtained from the HGMP Resource Centre, Hinxton, U.K.). Standard PCR conditions were used (1.5 mM $MgCl_2$) with the "hot start" PCR program 94°C, 5 min; 40 x

(94°C, 45s; 58°C, 45s; 72°C, 1min). Two positive PAC clones containing the ribokinase-like gene were obtained: F1622 and K13137.

During the investigation of the ribokinase-like gene genomic structure, sequencing problems were experienced at the 5′ end of the gene when using PAC DNA as template. To obtain alternative genomic clones for sequencing, cosmid clones were isolated for the ribokinase-like gene 5′ end by screening a chromosome 2 cosmid library (LL02NC02). This was carried out by radioactively labelling the oligonucleotide R15 (5′-dACCTTTGAGCGATGGCGG-3′; corresponding to *RBSK* nucleotides 1-18, Figure 4.29) and screening the cosmid filters by hybridisation (Filters AE1, AE2, and AE3, supplied from the HGMP Resource Centre, Hinxton, U.K.). Four positive R15 cosmid clones were obtained: 43K21, 46P3, 83O7, and 99K11.

4.4.2.3 Structural analysis

Oligonucleotides were designed from the cDNA sequence and used to identify exon-intron boundaries using as template DNA from the PAC clones F1622 and K13137, both of which contain SHSG15128 (designed from the *RBSK* 3′UTR). Sequencing using the PAC DNA as template proved troublesome, however, especially at the 5′ end of the *RBSK* gene. Therefore, the cosmid clone 43K21 was also used as template. Sizes of introns were determined by PCR amplification across each intron using PAC DNA as template and primers within adjacent exons. The fragments were analysed on a 1·5% agarose gel and sized against a Gibco BRL 1 kb ladder.

Although all exon-intron boundaries were identified in the ribokinase-like gene by sequencing using genomic PAC and cosmid clones, not all the intron sizes could be determined, due to unspecific PCR amplification. Alteration of the PCR conditions by raising the annealing temperature and addition of betaine or DMSO (to reduce template DNA secondary structure) did not improve the specificity of the PCR. A long-range PCR kit (Stratagene) was also used, in case the introns were large, but again no specific PCR product was obtained. Use of different primers and different genomic clones for template DNA did not improve the results.

4.4.2.4 Sequence analysis

Sequence analysis was performed using the MAGI sequence alignment program (ClustalW), based at the HGMP resource centre.

4.4.3 Results

4.4.3.1 The ribokinase-like cDNA and putative translation product

The sequencing of overlapping cDNA clones showing significant homology to a ribokinase-like gene revealed a 1196 bp cDNA sequence with a single 969 bp long open reading frame (corresponding to nucleotides 12-980; Figure 4.29), encoding a putative translation product of 322 amino acids with a molecular weight of 34.15 kDa. This ribokinase cDNA sequence was deposited in the EMBL sequence database and assigned the accession number AJ404857. The putative translation product possesses two conserved regions, characteristic of the ribokinase family, corresponding to amino acids 51-78 and 258-279 (underlined in Figure 4.29).

The putative translation product of the human ribokinase gene shown in Figure 4.29 was compared to ribokinase proteins from other species (Figure 4.30). This shows that homologous ribokinase proteins are highly conserved between *D. melanogaster*, *H. influenzae*, *E. coli*, *B. subtilis*, and *S. cerevisiae*, and display identity to the human RBSK of 44%, 37%, 35%, 35%, and 28% respectively. The significant similarity of the human ribokinase-like gene described in this section strongly supports the notion that this gene is the human ribokinase gene (*RBSK*). Figure 4.30 also reveals there to be several regions of highly conserved amino acid residues throughout all six ribokinase proteins examined. The two amino acid sequences highlighted in Figure 4.30 are indicative of ribokinase family members (Bork *et al.*, 1993).

4.4.3.2 Structural organisation of the human *RBSK* gene

To determine the genomic structure of the *RBSK* gene, primers that were designed from the *RBSK* cDNA sequence were used to sequence across the exon-intron boundaries, using *RBSK*-containing PAC clones as template DNA (isolated from PAC libraries using SHSG15128). The sequencing using the PAC DNA as template proved troublesome, especially at the 5′ end of the *RBSK* gene (probable secondary structure due to high G+C nucleotide content). Therefore the cosmid clone 43K21 was also used as sequencing template. The results of sequence analysis of PAC and cosmid clones are summarised in Figure 4.31. The *RBSK* gene consists of 8 exons and 7 introns. The exons vary in size from 63 bp to 390 bp. As shown in Figure 4.29, the first methionine codon of the open reading frame is located in exon 1 (nucleotides 12-14), whereas the stop codon is located in the last exon, exon 8 (nucleotides 978-980). Exon 8 contains 216 bp of 3′UTR sequence.

```
1     ACCTTTGAGCGATGGCGGCGTCTGGGGAACCCCAGAGGCAGTGGCAAGAGGAGGTGGCGG 60
                M  A  A  S  G  E  P  Q  R  Q  W  Q  E  E  V  A  A 17

61    CGGTGGTAGTGGTGGGCTCCTGCATGACCGACCTGGTCAGTCTTACTTCTCGTTTGCCAA 120
        V  V  V  V  G  S  C  M  T  D  L  V  S  L  T  S  R  L  P  K 37

121   AAACTGGAGAAACCATCCATGGACATAAGTTTTTTATTGGCTTTGGAGGGAAAGGTGCCA 180
        T  G  E  T  I  H  G  H  K  F  F  I  G  F  G  G  K  G  A  N 57

181   ACCAGTGTGTCCAAGCTGCTCGGCTTGGAGCAATGACGTCCATGGTGTGTAAGGTTGGCA 240
      Q  C  V  Q  A  A  R  L  G  A  M  T  S  M  V  C  K  V  G  K 77

241   AAGATTCTTTTGGCAATGATTATATAGAAAACTTAAAACAGAATGATATTTCTACAGAAT 300
      D  S  F  G  N  D  Y  I  E  N  L  K  Q  N  D  I  S  T  E  F 97

301   TTACATATCAGACTAAAGATGCTGCTACAGGAACTGCTTCTATAATTGTCAATAATGAAG 360
        T  Y  Q  T  K  D  A  A  T  G  T  A  S  I  I  V  N  N  E  G 117

361   GCCAGAATATCATTGTCATAGTGGCTGGAGCAAATTTACTTTTGAATACGGAGGATCTGA 420
        Q  N  I  I  V  I  V  A  G  A  N  L  L  L  N  T  E  D  L  R 137

421   GGGCAGCAGCCAATGTCATTAGCAGAGCCAAAGTCATGGTCTGCCAGCTCGAAATAACTC 480
        A  A  A  N  V  I  S  R  A  K  V  M  V  C  Q  L  E  I  T  P 157

481   CAGCAACTTCTTTGGAAGCCCTAACAATGGCCCGCAGGAGTGGAGTGAAAACCTTGTTCA 540
        A  T  S  L  E  A  L  T  M  A  R  R  S  G  V  K  T  L  F  N 177

541   ATCCAGCCCCTGCCATTGCTGACCTGGATCCCCAGTTCTACACCCTCTCAGATGTGTTCT 600
        P  A  P  A  I  A  D  L  D  P  Q  F  Y  T  L  S  D  V  F  C 197

601   GCTGCAATGAAAGTGAGGCTGAGATTTTAACTGGCCTCACGGTGGGCAGCGCTGCAGATG 660
        C  N  E  S  E  A  E  I  L  T  G  L  T  V  G  S  A  A  D  A 217

661   CTGGGGAGGCTGCATTAGTGCTCTTGAAAAGGGGCTGCCAGGTGGTAATCATTACCTTAG 720
        G  E  A  A  L  V  L  L  K  R  G  C  Q  V  V  I  I  T  L  G 237

721   GGGCTGAAGGATGTGTGGTGCTGTCACAGACAGAACCTGAGCCAAAGCACATTCCCACAG 780
        A  E  G  C  V  V  L  S  Q  T  E  P  E  P  K  H  I  P  T  E 257

781   AGAAAGTCAAGGCTGTGGATACCACGGGTGCTGGTGACAGCTTTGTGGGAGCTCTGGCCT 840
      K  V  K  A  V  D  T  T  G  A  G  D  S  F  V  G  A  L  A  F 277

841   TCTACCTGGCTTACTATCCAAATCTGTCCTTGGAAGACATGCTCAACAGATCCAATTTCA 900
      Y  L  A  Y  Y  P  N  L  S  L  E  D  M  L  N  R  S  N  F  I 297

901   TTGCAGCAGTCAGTGTCCAGGCTGCAGGAACACAGTCATCTTACCCTTACAAAAAAGACC 960
        A  A  V  S  V  Q  A  A  G  T  Q  S  S  Y  P  Y  K  K  D  L 317

961   TTCCGCTTACTCTGTTTTGATTGCTATTAGTCCCAAAATAAATATACCTGGGAATAAAAT 1020
        P  L  T  L  F  *                                          322

1021  GTACTTGGGGGTGGCTGCTCCTGGCTAATGCTTATTAGAAAATGTCCTCGTCCCCTTTCT 1080

1081  TTGCAAATATTAGTTCTTTTACGAAGTCATCCTCAAGCTTCAATTTATTTATAACGATGA 1140

1141  TTCTTTTGCTTTCCATGCATTTGCACAAAACAACCAGAATTAAAGATTCCACAACC      1196
```

Figure 4.29 cDNA sequence of human ribokinase (*RBSK*) gene; the start codon is located at nucleotides 12-14 and the stop codon at nucleotides 978-980. The underlined amino acid sequences are conserved regions belonging to ribokinase family members.

Figure 4.30 Comparison of homologous ribokinase proteins from various species. Shaded boxes indicate conservation of an amino acid residue between four out of six species. A consensus amino acid sequence is shown: bold letters represent complete conservation of a residue between all six species, non-bold type represents conservation of a residue between four out of six species. Blue shaded boxes represent ribokinase family amino acid motifs. Sequences were obtained from the Genbank and Swissprot sequence database. Accession numbers for ribokinase proteins are: *D. melanogaster*, CAA20884; *H. influenzae*, P44331; *E. coli*, P05054; *B. subtilis*, P36945; *S. cerevisiae*, P25332.

The intron sizes were investigated by using primers designed from within the exons and using PCR to try and amplify across the intronic sequence. For introns 1, 2, 3, 4, and 7, PCR amplification proved successful. However, PCR was unable to amplify a clean PCR product for introns 5 and 6. This could have been due to the large size of the introns or due to miss-priming by the oligonucleotides. The use of long range PCR, addition of DMSO or betaine to the PCR reaction (to help destroy secondary structure of the template DNA), alternative primers and template DNA, did not resolve this problem.

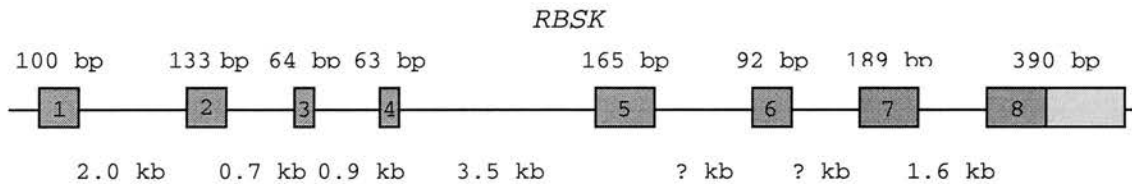


Figure 4.31 Genomic structure of ribokinase gene. Exons (1-8) are shown as shaded boxes, coding regions darker than untranslated regions. The diagram is not to scale; exact exon sizes (bp) and approximate intron sizes (kb) are indicated where known.

4.4.3.3 Amino acid sequence analysis

The human ribokinase gene (*RBSK*) is the third member of the ribokinase family to be cloned (the other two are ketohexokinase (*KHK*) (Bonthron *et al.*, 1994)and adenosine kinase (*ADK*) (Singh *et al.*, 1996a). Comparison of the encoded proteins of *RBSK* to *KHK* and *ADK* reveals 17% and 20% identity respectively. All three kinase amino acid sequences contain two conserved domains that are characteristic of ribokinase proteins (Figure 4.32), providing evidence that all three proteins belong to the same ribokinase family.

```
RBSK   51 FGGKGANQCVQAARLGA----MTSMVCKVGKD 78
KHK    39 RGGNASNSCTVLSLLGA----PCAFMGSMAPG 66
ADK    78 AGGSTQNSOKVAQWMIQQPHKAATFFGCIGID 109
             **.   *.       :          ::.   :. .

RBSK  258 KVKAVDTTGAGDSFVGALAFYL 279
KHK   247 PPRVVDTLGAGDTFNASVIFSL 268
ADK   246 QKEIIDTNGAGDAFVGGFLSQL 267
            . :** ****:* ...    *
```

Figure 4.32 Comparison of conserved ribokinase family domains in the human ribokinase (RBSK), ketohexokinase (KHK), and adenosine kinase (ADK) proteins. Numbering corresponds to amino acid number. Clustal alignment program nomenclature: "*" indicates fully conserved residue, ":" indicates a "strongly similar" residue is fully conserved, and "." indicates a "weakly similar" residue is fully conserved. Genbank accession numbers for KHK and ADH are CAA55347 and NP_006712, respectively.

The co-localisation of *RBSK* and *KHK* to chromosome 2p23.3 suggested the possibility these two genes might have originated from a common ancestral gene. Although comparison of amino acid sequences show these proteins to be related by sequence, gene structure can be more revealing than sequence similarity for the investigation of gene evolutionary linkage. To investigate the hypothesis of an evolutionary link between *RBSK* and *KHK*, the genomic structures of *RBSK* and *KHK* were compared. Figure 4.33 shows a comparison of RBSK to KHK amino acid sequences, with the position of the exon-intron boundaries indicated.

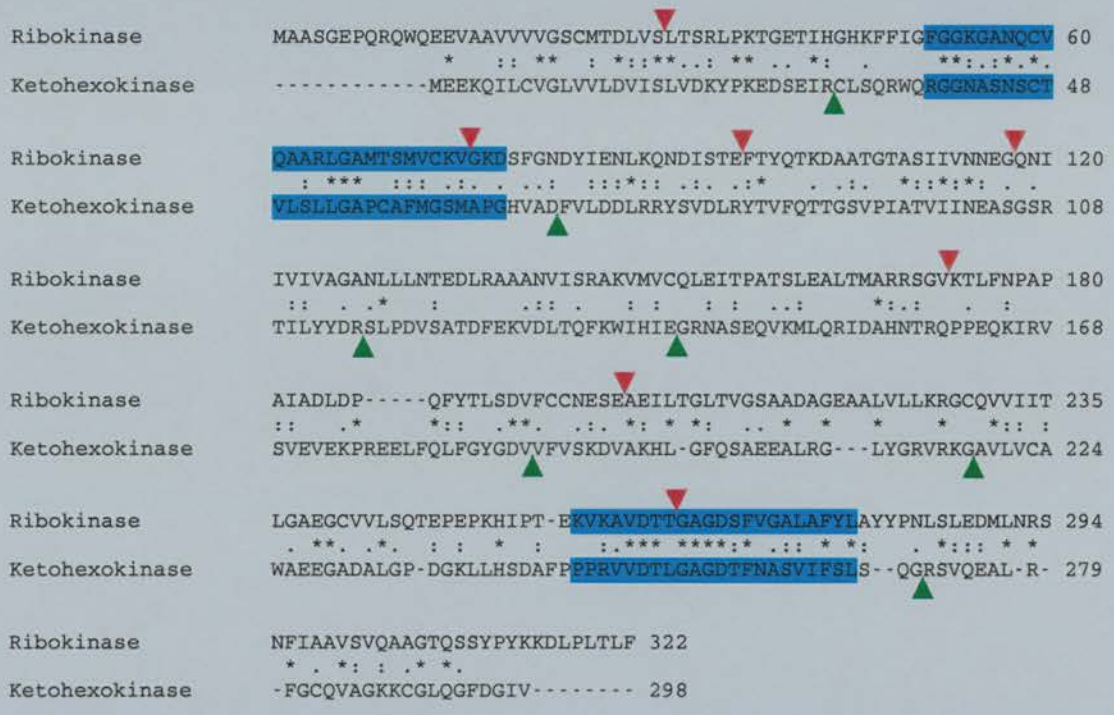


Figure 4.33 Comparison of ribokinase and ketohexokinase proteins. Numbering corresponds to amino acid number. Clustal alignment program nomenclature: "*" indicates fully conserved residue, ":" indicates a "strongly similar" residue is fully conserved, and "." indicates a "weakly similar" residue is fully conserved. Red and green triangles indicate positions of exon-intron boundaries in the amino acid sequence for ribokinase and ketohexokinase,respectively. Blue highlighted sequences represent conserved amino acid motifs of the ribokinase family. Genbank accession numbers for RBSK and KHK (3A form) are AJ404857 and CAA55347, respectively.

Although comparison of the RBSK and KHK amino acid sequences reveals 17% identity, comparison of the positions of their exon-intron boundaries show *RBSK* and *KHK* to possess completely distinct genomic structures (Figure 4.33). This is further discussed below.

4.4.4 Discussion

4.4.4.1 Cloning of the human ribokinase gene

This section describes the cloning, structural analysis, and chromosomal localisation of the human ribokinase (*RBSK*) gene. The cDNA sequence contains a 969 bp open reading frame encoding a predicted protein of 322 amino acids (34.15 kDa) – see Figure 4.29. For sequence comparison, there are presently only two other related human sugar kinase genes that have been cloned, ketohexokinase (*KHK*) and adenosine kinase (*ADK*). Comparison of RBSK to KHK and ADK at the amino acid level reveals 17% and 20% identity, respectively. Inspection of the RBSK, KHK, and ADH sequences reveal that all three possess two conserved amino acid motifs that are characteristic of ribokinase family members (Bork *et al.*, 1993).

The putative protein encoded by *RBSK* shows significant homology to other prokaryotic and eukaryotic ribokinases but similarity is also seen to other fructokinases and adenosine kinases. The overall homology of human RBSK with ribokinase proteins from *D. Melanogaster*, *H. influenzae*, *E.coli*, *B. subtilis*, and *S. cerevisiae* is 44%, 37%, 35%, 35%, and 28% respectively (Figure 4.30). The conservation of sequence between ribokinase proteins suggest that many of the conserved residues are essential for the three dimensional structure and ribose phosphorylation function. It is interesting that the *S. cerevisiae* ribokinase protein displays relatively low identity to the human ribokinase protein. Examination of Figure 4.30 reveals that this is mainly due to the insertion of several stretches of amino acids that are not present in any of the other homologous ribokinase proteins. This results in the *S. cerevisiae* ribokinase protein being 11 residues longer than the human ribokinase protein (333 amino acids compared to 322 amino acids, respectively). The function of these extra amino acids is unknown.

Mapping of *RBSK* to the same YAC (29IH8) as the *GCKR* gene (*GCKR* is located ~500 kb from *KHK*) reveals *RBSK* and *KHK* to be ~800 kb apart (estimated from the physical YAC contig of the *GCKR-KHK* genomic region described in Chapter 3,). Although the mouse homologue of *RBSK* has not been completely characterised, mouse ESTs exist in the databases that show significant homology to the human *RBSK* gene (for example,Genbank accession numbers AA560831 and AA260331). The physical mapping of human *RBSK* to between the *GCKR* and *KCNK3* genes on the 2p23.3 physical contig (gene order *KHK/GCKR/RBSK/KCNK3*), suggests that like *KHK*, *GCKR* and *KCNK3*, the *RBSK* gene will also prove to map to the region of conserved synteny on mouse chromosome 5.

The mapping of both *RBSK* and *KHK* to human chromosome 2p23.3 suggests that these two genes might have originated from a common ancestral gene. The comparison of *RBSK* to *KHK* at the amino acid level reveals 17% identity (no significant homology is seen at the nucleotide level). This amino acid sequence similarity places both RBSK and KHK within the same ribokinase family. However, such a low level of sequence similarity alone does not provide enough information to conclude that *RBSK* and *KHK* have originated from a common ancestral gene.

Comparison of the *RBSK* and *KHK* genomic structures reveals that they consist of 8 and 9 exons, respectively. However, *KHK* has two alternative exons (3a and 3c) that show enough similarity to each other to suggest that they arose through an intragenic duplication event within the *KHK* gene (Bonthron *et al.*, 1994; Hayward & Bonthron, 1998). Therefore, it is possible that before this intragenic exon duplication event within *KHK*, both *RBSK* and *KHK* would have consisted of 8 exons. However, comparison of the intron-exon boundary positions relative to the amino acid sequence alignment reveals that in fact both genes have very different genomic structures (Figure 4.33). This suggests that although both *RBSK* and *KHK* belong to the ribokinase family and both carry out the enzymatic function of furanose sugar phosphorylation (see Figure 4.27), their exon-intron structures have evolved independently. Indeed, none of the splice junctions can be perfectly aligned between the two genes. This would suggest that although the *RBSK* and *KHK* genes show amino acid sequence similarity, their divergence is an ancient one (predating the introduction of introns into these genes). Their chromosomal co-localization is therefore presumably coincidental, rather than reflecting genomic evolutionary history.

The *RBSK* gene was investigated because like *KHK*, it resides on chromosome 2p23.3 and both encoded proteins belong to the ribokinase family. The work described in previous sections in this chapter, though, was motivated by the metabolic relatedness of KHK and GKRP, and the idea that other candidate type 2 diabetes genes with metabolically linked functions might, also reside on chromosome 2p23.3. Although *RBSK* itself is an unlikely candidate type 2 diabetes gene, it is interesting to note that in times of excess ribose-5-phosphate (the product of ribose phosphorylation by RBSK), ribose-5-phosphate can be converted into the glycolytic pathway intermediates glyceraldehyde 3-phosphate and fructose-6-phosphate by the action of transketolase and transaldolase. As discussed in Chapter 1 (Section 1.5.4, Figure 1.2), fructose-6-phosphate has been shown to increase glucokinase inhibition by glucokinase regulatory protein (GKRP). In other words, the generation of excess ribose-5-phosphate could affect glucokinase activity through the

fructose phosphate effect on GKRP. This could act as a subtle mechanism for reducing glucose phosphorylation in times of excess ribose, therefore conserving glucose. This illustrates that a very detailed understanding of the relevant biochemical pathways and the complex interactions between them will be required to fully understand subtle disease processes such as the pathogenesis of type 2 diabetes.

Chapter 5

5 Candidate neurosensory non-syndromic recessive deafness 9 (*DFNB9*) genes

5.1 Introduction

5.1.1 Mapping of the *DFNB9* gene to chromosome 2p22-23

The *DFNB9* gene was first localised to 2p22-23 by genetic analysis of prelingual and fully penetrant deafness in a Sunnite consanguineous family living in an isolated village of North Lebanon (Chaib *et al.*, 1996) – this locus was first described as DFNB6 but this designation had already been assigned to the DFNB locus 3p14-p21 (Fukushima *et al.*, 1995). At this stage the *DFNB9* gene was defined to a 2 cM interval delimited by the markers D2S2303 and D2S174 at 2p23.1 (see Chapter 3, Figure 3.5). The identification of a second Middle Eastern kindred with autosomal recessive non-syndromic hearing loss segregating to the DFNB9 locus (Leal *et al.*, 1998) presented further evidence of a *DFNB* gene at the 2p22-23 locus. This second linkage study suggests that the locus containing the *DFNB9* gene is less than 1.08 cM, 95 % confidence interval (0-2.59 cM).

The study of genetic maps covering the DFNB9 region reveals that the genomic interval in which the causative gene is determined to reside, maps close to the *GCKR–KHK* genomic region (see Chapter 3, Figure 3.5). As much of this genomic region has already been cloned, the aim of this research is to identify and characterise candidate *DFNB9* genes and examine their role in the pathogenesis of sensorineural deafness.

5.1.2 Syndromic and non-syndromic deafness

Significant hearing impairment affects 1 in 1000 children, and of these cases, approximately 60% are genetically determined (Marazita *et al.*, 1993). Inherited hearing impairment can be classified as syndromic or non-syndromic, reflecting the presence or absence of inherited physical abnormalities, respectively. Approximately 30% of patients with prelingual deafness show syndromic hearing impairment, that is they have additional anomalies. Within the prelingual non-syndromic hearing impairment category, inheritance is most commonly autosomal recessive (75%-80%), followed by dominant (20%-25%) and X-linked (1%-1.5%) (Van Camp *et al.*, 1997). Numbered sequentially on the basis of time of

discovery, dominant loci carry the prefix DFNA, recessive loci the prefix DFNB, and X-linked loci the prefix DFN. Mitochondrial mutations are designated by the site of the mutation (for example, A1555G, T7445C). To date, 31 dominant, 28 recessive, 8 X-linked, and 4 mitochondrial loci for non-syndromic hearing loss have been identified. The genes that have been found to underlie syndromic and non-syndromic deafness have been shown to encode a large diversity of molecules, including extracellular matrix components, enzymes, transcriptional complex factors, cytoskeletal components and membrane components, as well as four different mitochondrial encoded proteins, three tRNA molecules and one rRNA molecule. The genetic locations of deafness loci and descriptions of genes cloned so far can be found at the Hereditary Hearing Loss Home Page (see http://dnalab-www.uia.ac.be/dnalab/hhh/).

The many different cell types in the inner ear and diverse range of genes shown to be expressed in the inner ear demonstrate the heterogeneity underlying the genetic basis of sensorineural deafness with estimates of over 100 genes involved in non-syndromic deafness with an autosomal recessive mode of inheritance (Morton, 1991). Localisation of the genes underlying NSHI initially proved difficult because in addition to the genetic heterogeneity of deafness, many cases of prelingual NSHI show an autosomal recessive mode of inheritance and assortative mating in the deaf community (where several deafness genes are introduced into the one pedigree) and this genetic heterogeneity precludes pooling of families for linkage studies.

The cloning of NSHI genes was made possible mainly due to gene localisation studies that were performed in large multigenerational inbred families from ethnically isolated regions. The high degree of consanguinity in these families meant that the families were large enough to generate significant LOD scores for linkage within a single pedigree. Ethnically isolated families also lacked assortitive mating. A recent advance was the homozygosity mapping of autosomal recessive deafness genes using small consanguineous multiplex families with ≥3 affected patients (Fukushima *et al.*, 1995).

5.1.3 Cloning genes for non-syndromic deafness

Once a gene for non-syndromic hearing loss has been localised to a genomic region, several strategies can be utilised to clone the relevant gene. It is advantageous to define the region of interest by fine mapping so that the search for candidate genes can be focused and any ESTs/genes mapping out with the defined interval rejected for further investigation. The

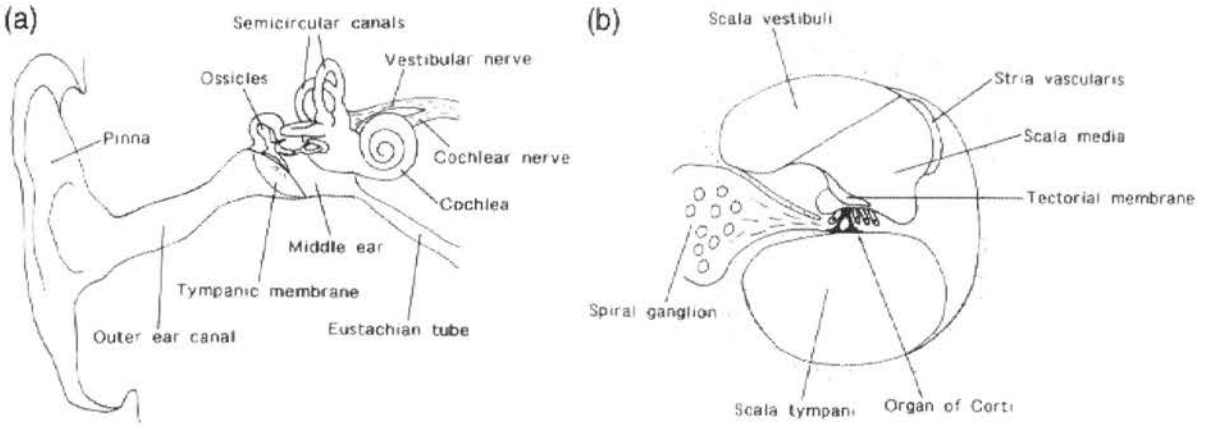
construction of a physical map covering the genomic region of interest using genomic clones such as YACs, BACs and PACs, allows sequence-tagged sites (STSs) including microsatellite repeat polymorphisms and expressed sequenced tags (ESTs) to be located conveniently by PCR screening. If the gene or EST content of the candidate region is unknown, it is necessary to identify novel open reading frames and this can be carried out by techniques such as cDNA selection, exon trapping or direct sequencing (see Chapter 3, Section 3.1.3).

Genes and ESTs known to be in the candidate region must be prioritised for further investigation. In the search for candidate non-syndromic deafness genes, it is useful to check that the gene or EST is expressed in a cochlear cDNA library. It should be noted however that the *PAX3* gene, which encodes a transcription factor that underlies a syndromic form of hearing loss called Waardenburg type 1/3 is not actually expressed in the ear at all, but in adjacent structures such as the developing neural tube (Baldwin *et al.*, 1995). Therefore one should be cautious about discarding genes as candidates because they are not expressed in the ear. Genes with functions similar to other genes that are known to play a role in the pathogenesis of deafness are also good candidates for further investigation.

Identifying good candidate sensorineural deafness genes requires the knowledge of the molecular interactions in the development and function of the inner ear. It is a highly intricate structure (Figure 5.1) and many of the molecular mechanisms involved in hearing are unknown or are just beginning to be understood (Steel & Brown, 1994). Briefly, endolymph- and perilymph-filled channels course around the cochlea, the auditory sense organ, and around the sacculus, utriculus and the three semicircular canals, which together form the vestibular part of the inner ear that detects head position and movement and hence aids balance. Several types of sensory epithelia form the auditory transduction apparatus of the cochlear duct: the organ of Corti; the maculae of the utricle and saccule; and the cristae ampullae of the semicircular canals. These structures consist of a highly organised array of supporting and sensory cells, the latter carrying a distinct bundle of actin-filled stiff microvilli, called stereocilia, on their apical surface. The neuroepithelia are covered in an acellular gelatinous membrane that can be displaced relative to the neuroepithelia by sound or head movement. This displacement provokes a deflection of the sensory hair cell stereociliary bundles, which in turn opens up the mechanotransduction channels located at the tip of the stereocilia. It has been proposed that the tip link, a filamentous connection attaching the tip of a stereocilium to the nearest taller stereocilium, is the gating spring for opening transduction channels. The resultant influx of potassium, from the potassium rich

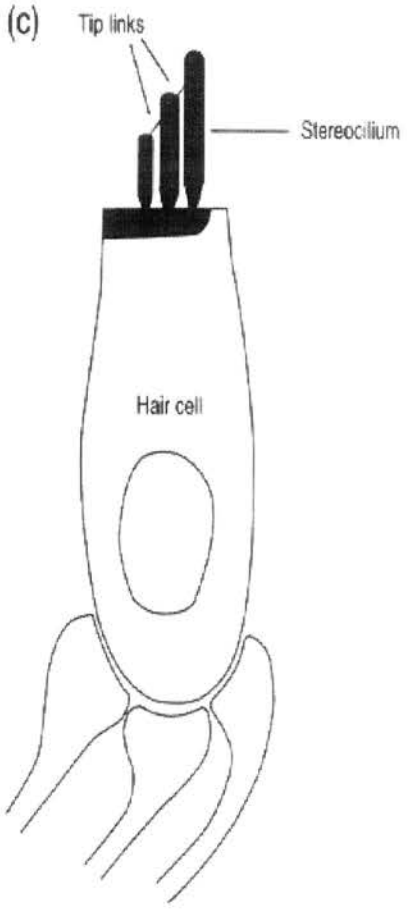
endolymph through the mechanotransduction channels, alters the membrane potential, which results in the release of a synaptic transmitter from the hair cell. On neurotransmitter release, an afferent nerve fibre at the base of the hair cell transmits to the brain a patter of action potentials encoding certain characteristics of the stimulus, such as intensity, frequency and time course. This knowledge of the intricate structure of the inner ear and complex molecular mechanisms that underlie hearing allow the ESTs/genes within the genomic interval for which a non-syndromic hearing loss gene has been localised, to be prioritised for further investigation. The next step is to clone and characterise the gene from which the EST originated.

The full length cDNA sequence can be derived by searching DNA sequence databases such as Genbank and constructing a cDNA contig covering the full length of the gene. If the cDNA contig is incomplete, techniques such as 3′ and 5′ RACE (rapid amplification of cDNA ends) can be used to produce the full length cDNA sequence. Once the full length cDNA sequence for the candidate deafness gene has been ascertained, mutation screening of the deafness patients with linkage to the correct genomic region can proceed in two ways. Directly, by cDNA screening if the gene is illegitimately transcribed in lymphoblastoids cell lines, or indirectly, by determining the genomic structure of the gene (see Section 5.3.2.3) and making PCR primers from the DNA sequence flanking the exons in order to amplify the individual exons. The search for mutations can be carried out using techniques such as heteroduplex analysis (Cotton, 1993; Grompe, 1993) and/or single strand conformational polymorphism analysis (Orita *et al.*, 1989), complemented by direct sequencing. This allows comparison of the gene sequence from deafness patients to that of the control population. Once a nucleotide change has been identified, it must be studied to establish whether it represents a benign polymorphism or a disease-causing mutation. Critical factors to consider include the frequency of this change in the general population, its impact at the amino acid level, the importance of the involved amino acid in the protein, especially cross-species, and, ultimately, the impact of the change on transcription, translation and protein function.



(a) The human outer, middle and inner ear. The outer ear canal leads to the tympanic membrane (ear drum), and sound vibrations are transmitted from there to the inner ear via the three middle ear ossicles, the malleus, incus and stapes. The stapes inserts into the oval window at the base of the cochlea.

(b) One turn of the cochlear duct. During sound stimulation, the organ of Corti vibrates and the sensory hair cells are stimulated by a shearing movement between the top surface of the organ of Corti and the tectorial membrane. Most afferent neurons innervate inner hair cells (shown here on the left of the black pillar cells) and their cell bodies form the main part of the spiral ganglion, which in turn leads to the cochlear nerve. The three or more rows of outer hair cells (to the right of the pillar cells) receive primarily efferent innervation, and are believed to act as motor cells, enhancing the response of the cochlea. The stria vascularis on the lateral wall of the cochlea duct has a key role in generating the high resting endocochlear potential in the endlymph that fills the scala media, and also controls the high K^+ and low Na^+ concentation in the endolymph. Scala vestibuli and scala tympani both contain perilymph, a fluid that is much more like a normal extracellular fluid, with low K^+ and high Na^+ levels and no resting high resting potential. The stippled area represents bone.



(c) A typical sensory hair cell. The actin-filled stereocilia are arranged in rows of graded heights at the top of the hair cell. Tip links connect the shorter stereocilia to the adjacent taller stereocilia, and it is the tip links that is thought to activate transduction channels of the hair cell, leading to depolarization and initiation of synaptic activity at the base of the cell.

Figure 5.1 The structure of the ear. Figure adapted from Steel and Brown, 1994.

5.1.4 Potential mouse homologues for non-syndromic hearing impairment

An alternative method to identify potential human deafness loci is to identify genes responsible for hearing loss in the mouse. The structure and development of the inner ear, and the pathology leading to hearing impairment is very similar between mice and humans (Steel & Brown, 1994) therefore the mouse is a good model animal for the study of deafness. In fact it was the genetic analysis of both Snell's waltzer and shaker-1 mice that identified the role of unconventional myosin genes found to underlie deafness in mice and humans (Hasson, 1997); see Section 5.3.1.2. With over 60 mouse mutations having been mapped in the mouse genome that cause abnormalities in the inner ear, the identification and analysis of these genes should provide an extremely useful insight into the pathology leading to deafness. Further more, the identification of human homologues to these mouse genes will provide new candidate genes for the study of human hearing impairment.

5.1.5 Aim

This chapter describes the search for candidate *DFNB9* genes. During this search, three genes were identified as candidate *DFNB9* genes: *MPV17*, *KIF3C*, and *KCNK3*. These genes all map close to, or within the DFNB9 interval on human chromosome 2p23 and encode proteins with potentially important functions within the inner ear. One other gene called *CRMP1* was identified as a potential deafness gene although mapping data ruled out this gene as a candidate *DFNB9* gene. Briefly, investigation of these genes involved mutational analysis of *MPV17*, cloning and characterisation of *KIF3C*, and mapping of *KCNK3* and *CRMP1*.

5.2 The *MPV17* gene

5.2.1 Introduction

5.2.1.1 *Mpv17* – a candidate kidney disease and deafness gene

The *Mpv17* gene was first identified during an experiment to produce transgenic mouse strains by defective retrovirus insertion into the germline (Weiher *et al.*, 1990). Mice homozygous for one of the integrations were found to develop a phenotype of glomerulosclerosis at age of three months or older and die later as a consequence of renal failure. The *Mpv17* gene was identified at the integration site and its expression was found to be completely abolished in the homozygous mice (Weiher *et al.*, 1990) thus implying *Mpv17* to be a recessive kidney disease gene.

Further study of the *Mpv17* (-/-) mice was performed by examining the inner ear. This work was carried out because the kidney and the inner ear are known to both have specialised epithelia involved in active ion transport (Mizuta *et al.*, 1995). Other links between the kidney and inner ear are suggested by numerous drugs that have both nephrotoxic and ototoxic side effects (Begg & Barclay, 1995), and congenital anomalies exist that can cause lesions in both organs (Bergstrom *et al.*, 1979). Electron microscopic investigations performed on the inner ears of *Mpv17* (-/-) mice revealed degeneration of the stria vascularis and spiral ligament, loss of cochlear neurons and degeneration of the organ of Corti (Meyer zum Gottesberge *et al.*, 1996). These alterations observed were similar to those described for Alport's syndrome, which is a hereditary kidney disorder associated with deafness (Alport, 1927). In Alport's syndrome, it is mutations in the basement membrane collagen genes *COL4A3*, *COL4A4* and *COL4A5* which are mainly found in the basilar membrane, parts of the spiral ligament, and stria vascularis that underlie the sensorineural hearing loss (Barker *et al.*, 1990; Mochizuki *et al.*, 1994). Although the mechanism of hearing loss is unknown, in the glomerulus there is focal thinning of the basement membrane. The observation of similar abnormalities in the inner ear of the *Mpv17* (-/-) mouse means that this mutant mouse was a valuable genetic model for both renal disorders and deafness.

In contrast to transgenic mouse or rat models where the disorder phenotype is induced by expression of a foreign transgene (Dressler *et al.*, 1993; Kopp *et al.*, 1992), the *Mpv17* (-/-) mouse defines an endogenous gene whose loss of function correlates with the phenotype. At first it could not be excluded that a second gene distinct from the *Mpv17* gene was also

affected by the retrovirus insertion and that the non-function of the second gene was causing the *Mpv17* homozygous phenotype. However in a later experiment, the functional rescue of the glomerulosclerosis phenotype in the *Mpv17* (-/-) mice by transgenesis with the human *MPV17* homologue (Schenkel *et al.*, 1995) proved definitively that mutation of the *Mpv17* gene led to kidney disease. However, there was inconclusive evidence for the rescue of the deafness phenotype (Schenkel, personal communication).

Mapping of *Mpv17* and *MPV17* added further circumstantial support to the evidence for *MPV17* as a candidate deafness gene. The *Mpv17* gene mapped in mice to chromosome 5 and in humans to the region of conserved synteny on chromosome 2p21-23 (Karasawa *et al.*, 1993). The studies described in Chapter 2 further showed that *MPV17* in fact maps within the *KHK-GCKR* contig. The mapping of a sensorineural non-syndromic recessive deafness gene (*DFNB9*) to the same region on chromosome 2p22-23 (Chaib *et al.*, 1996) made *MPV17* an excellent candidate DFNB9 gene and thus warranted further investigation.

5.2.1.2 The *MPV17* gene

Both human *MPV17* and mouse *Mpv17* encode a 176 amino acid protein. Comparison of the two amino acid sequences reveals 92% identity, corresponding to 14 amino acid differences (Karasawa *et al.*, 1993; Weiher *et al.*, 1990). One feature of the mouse *Mpv17* gene is the presence of a B1 and B2 repetitive element located in the 3'UTR (no repeat elements are found in the human *MPV17* gene 3'UTR). The determination of the *MPV17* genomic structure reveals it to contain 8 exons and 7 introns with the exon sizes varying in size from 24 bp to 484 bp and the introns varying in size from 0.09 kb to 5.3 kb. The start codon was found to be located within exon 2 and the termination codon within exon 8.

With the establishment of the intron-exon junction sequences, primers were designed within the intron sequences (Karasawa *et al.*, 1993), allowing PCR amplification of the protein coding regions from human genomic DNA from blood samples or biopsy material. This would enable the screening of the *MPV17* gene in patients with renal disorders and/or deafness by PCR and subsequent analysis by sequencing or SSCP analysis.

Several cases of familial glomerulosclerosis have been screened for mutations in *MPV17*, including patients suffering from congenital nephrotic syndrome of the Finnish type (CNF), a recessive disorder present in the Finnish population. Mutation screening of *MPV17* in individuals with CNF revealed no alterations within the *MPV17* gene (Schenkel *et al.*, 1995) At around the same time, analysis of CNF families (CNF is now called Nephrosis-1,

NPHS1), mapped the gene underlying NPHS1 to chromosome 19q13.1 (Kestila *et al.*, 1994). This gene was later cloned and named "nephrin" (Kestila *et al.*, 1998). The *MPV17* gene remains a good candidate gene for renal disorders. Furthermore, screening of deafness patients for alterations within the *MPV17* gene has not been reported, therefore *MPV17* cannot be ruled out as a candidate gene for deafness.

5.2.1.3 Function of MPV17

When the *Mpv17* gene was first cloned in the mouse, it was noticed that two hydrophobic regions were present that suggested that the protein may be membrane associated although both regions did not constitute a typical transmembrane domain (Weiher *et al.*, 1990). At this point there was no information concerning the Mpv17 function but an EMBL database search revealed significant similarity to the peroxisomal membrane protein Pmp22 (Kaldi *et al.*, 1993). The rat Pmp22 protein was found to be identical to Mpv17 in >25% of all positions and although *Pmp22* and *Mpv17* are not orthologues, the amino acid sequence similarity hinted that Mpv17 might also be a peroxisomal protein.

The investigation of Mpv17 subcellular localisation was performed by raising antibodies to bacterially produced Mpv17 proteins, followed by the use of immunofluorescence techniques (Zwacka *et al.*, 1994). This study showed Mpv17 to co-localise with catalase, a marker for peroxisomal localisation. Thus, Mpv17 is indeed a peroxisomal protein.

The peroxisomes are small, membrane limited cytoplasmic organelles that contain enzymes that degrade fatty acids and amino acids, generating ATP and heat by catabolism. A by-product of these reactions is hydrogen peroxide, a corrosive substance that oxidises many amino acid side-chains. To counter the potentially harmful effects of the hydrogen peroxide, peroxisomes also contain copious amounts of the enzyme catalase, which degrades hydrogen peroxide to water and oxygen. The localisation of Mpv17 to the peroxisomes represents a novel link between the peroxisome and glomerular disease.

Examination of the peroxisomes in *Mpv17* (-/-) mice revealed that there was no structural abnormality, no deficiency in enzymic activity and no change in membrane permeability (Zwacka *et al.*, 1994). As the peroxisomes are a site of detoxification of O_2^- radicals and peroxides, the intermediates of these pathways were investigated in mutant and normal cells. The production of O_2^- radicals and other reactive oxygen species (ROS) was measured in primary skin fibroblasts from mutant and non-mutant mice by loading the cells with hydroethidine which reacts with the ROS to form the fluorescent dye ethidium. The

intracellular ROS was measured by FACS (fluorescent activated cell sorting) analysis and this revealed that loss of Mpv17 protein led to a reduced ability to produce ROS. The evidence suggests a beneficial effect of ROS production on glomerular function.

To test whether Mpv17 has a direct involvement in ROS production, overproduction of the *Mpv17* transcript in transfected 3T3 murine fibroblast cells was investigated (Zwacka *et al.*, 1994). This resulted in dramatically enhanced levels of intracellular ROS and indicated that Mpv17 is either directly involved in cellular ROS production or at least accomplishes a rate-limiting step in the process. Widespread expression of the *Mpv17* gene (Weiher et al., 1990) contrasts with the apparent specific kidney and inner ear phenotype, and suggests that lack of *Mpv17* expression might alter the regulation of other genes leading to the *Mpv17* (-/-) phenotype. As both the kidney and inner ear show morphological defects in the basement membrane, the Mpv17 protein might play an important role in a regulatory mechanism of enzymes involved in basement membrane metabolism.

The matrix metalloproteinase 2 (MMP-2) protein plays a critical role in the basement membrane turnover within the glomerulus and this led to the study of this protein in *Mpv17* (-/-) mice. A direct relationship between Mpv17 and MMP-2 was found in the *Mpv17* (-/-) mice in which there was over-expression of MMP-2 at both mRNA and protein level (Reuter *et al.*, 1998). Moreover, when the human *MPV17* gene was introduced into *Mpv17*-negative cells, *MMP-2* expression was repressed. This suggests that MMP-2 is likely to be a common mediator of both glomerulosclerosis and deafness. However, the molecular mechanism by which *Mpv17* expression controls *MMP-2* is unknown. This mechanism maybe complex as it was also found that the transcription factor *c-jun* expression paralleled the *MMP-2* expression and that there was stronger expression of the tissue-specific inhibitor of metalloproteinase 2 (TIMP-2) in the *Mpv17* negative tissue culture cells (Reuter et al., 1998).

5.2.1.4 Aim

The *MPV17* gene is an excellent candidate gene for kidney disease and deafness. This section describes the screening for mutations/polymorphisms in the *MPV17* gene from two patients with both glomerulosclerosis and deafness, and a further glomerulosclerosis patient.

5.2.2 Methods

5.2.2.1 Mapping of *MPV17*

The gene encoding MPV17 was localised to the *GCKR-KHK* genomic region by PCR screening the YAC, PAC and cosmid contig (see Chapter 3, Figure 3.8) using the primers designed to amplify exons 4-5 (Table 5.1).

5.2.2.2 PCR amplification of *MPV17* coding region

Six primer pairs (see Table 5.1) designed from the *MPV17* intronic sequence (Karasawa et al., 1993) were used for exon amplification from patient DNA using standard PCR conditions and a hot start PCR program: 5 min 94°C; 40x(94°C, 1 min; X°C, 1 min; 72°C, 2 min). The PCR products ranged in size from 160 to 448 nucleotides and were sequenced on both strands using the Thermosequenase cycle sequencing kit (Amersham).

Exon number	Primer sequence (5′-3′)	Annealing temperature (°C)	Expected size (bp)
1	dGGGTCTCTCACAGAGTGGGTG	55	233
	dGTGGGCACTCATGGCTTCGAC		
2	dCTTATCGTGGAGAGGGACGGT	55	206
	dAGGAAGTGAGGGCGGCGCAG		
3	dTGTCCCTCCCTCCTTGAATGG	55	240
	dGAACTAAGACCACTGTTGAGC		
4-5	dGATACTTGGGGCAGGGAGCTT	58	426
	dAGCCCGCCAGCCAGAGACATT		
6-7	dCGCAAGTGTTAATTTGTTCCT	55	402
	dTTTGTCTCCAACTGTTGGTAA		
8	dATCTCCAGCCCTTGCTCACTG	58	160
	dAAACGATGGAGTGAGGCAGG		

Table 5.1 Primers and annealing temperatures for *MPV17* exon amplification. Expected size for each PCR product is indicated. The second exon 8 primer was designed within the 3′UTR so that only the protein coding region of exon 8 was amplified.

5.2.3 Results

5.2.3.1 Isolation of genomic clones

PCR screening and sequence analysis of genomic clones map *MPV17* to YACs 26BA11 and 29IH8, BAC NH0538J11, the PAC 13L13, and cosmid B2 (see Figure 5.2, and Chapter 3, Figure 3.8). More detailed PCR analysis of the YAC 26BA11 and the YAC 26BA11 cosmid subclone B2 using the primer pairs shown in Table 5.1, reveal that *MPV17* exons 1 to 5 are present and exons 6-8 are missing. Although a YAC contig spanning this genomic region using YACs 26BA11, 29IH8, 3AG3, and 18AG7, had previously been constructed by Hayward *et al*, 1996, STS content analysis revealed that it was not possible to assemble an internally consistent contig (Chapter 3, Figure 3.1, discussed in Section 3.1.2.1).

The identification of PAC 13L13 and BAC NH0538J11, that both contain *MPV17*, and the construction of a complete physical contig at the *GCKR-KHK* interval (Chapter 3, Figure 3.8), shows that the inconsistency in the original YAC contig was due to a deletion in YAC 26BA11. This deletion starts from near *MPV17* exon five and includes the *KHK* gene. A simplified contig of the *GCKR-KHK* interval is shown in Figure 5.2. Mapping of *MPV17* places this gene within the *GCKR-KHK* interval, ~300 kb from *GCKR*. The construction of a detailed cosmid contig (described in Chapter 3), added to the discovery that *MPV17* exons 6 to 8 are absent from YAC 26BA11 and cosmid B2, reveals that the *MPV17* gene is in the opposite orientation to *GCKR* and *MPV17*.

5.2.3.2 Mutation screening

The 8 exons from *MPV17* were PCR amplified from genomic DNA from two patients with both glomerulosclerosis and deafness, and a further glomerulosclerosis patient.. Each PCR product was sequenced in both directions and the exon sequences generated compared to the published *MPV17* sequence. No differences were found within the coding region of the patient *MPV17* gene to that of the published sequence. This suggests that mutations in the *MPV17* does not underlie the kidney disorder and deafness phenotype seen in these patients.

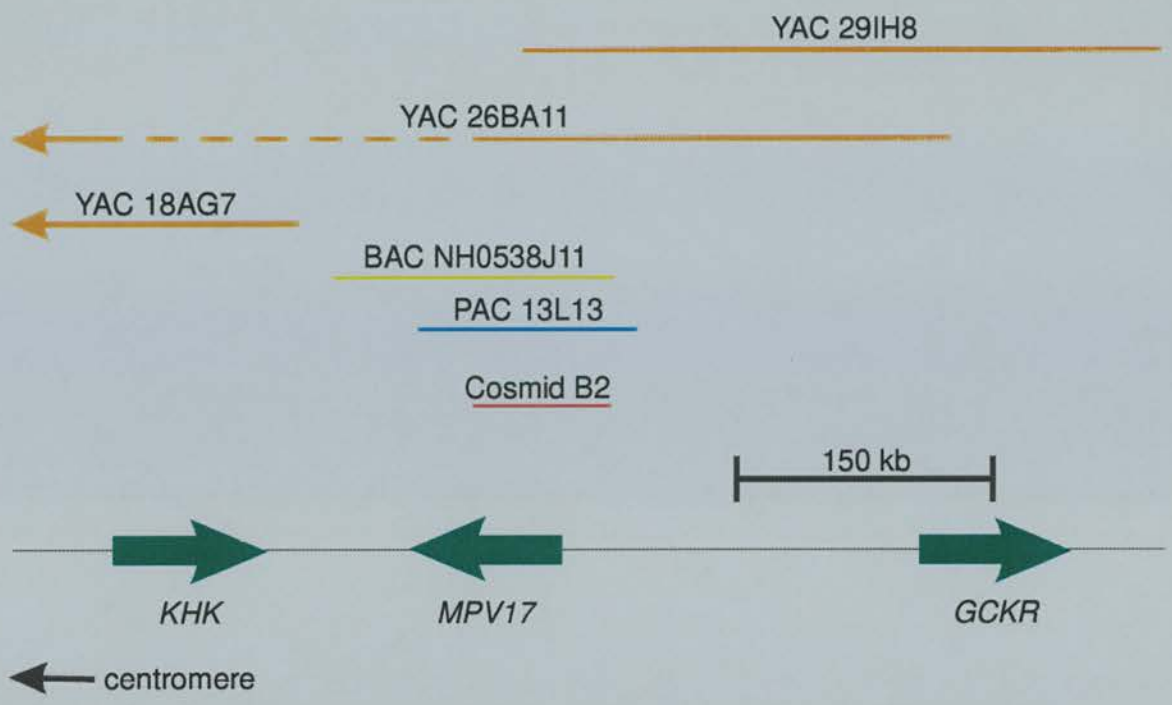


Figure 5.2 Co-localisation of *MPV17* with *GCKR* and *KHK*. The arrows represent genes, gene orientation (5'to 3') is indicated by the arrow direction. Genomic clones containing *MPV17* are indicated above the *EIF2B4* arrow. This diagram is a simplified version of Figure 3.8 shown in Chapter 3.

5.2.4 Discussion

5.2.4.1 The role of *MPV17* in glomerulosclerosis and deafness

Screening the coding region of *MPV17* and the 5'UTR (exon 1) from patients with glomerulosclerosis and/or deafness revealed no mutations or polymorphisms. This suggests that in these patients, *MPV17* is unlikely to play a causal role in glomerulosclerosis or deafness. Further screening of patients with kidney disease and/or deafness who show linkage to chromosome 2p21-23 region should be carried out before this gene can be fully ruled out from involvement in the pathogenesis of glomerulosclerosis and deafness in humans. Even if further screening reveals no mutations, the *MPV17* gene will remain an interesting candidate gene for glomerulosclerosis and deafness. Further study of MPV17 function will provide an important insight into the molecular interactions in the kidney and inner ear, and may help in understanding the mechanisms underlying the development of both glumerulosclerosis and deafness.

5.2.4.2 Another gene located upstream of *MPV17*

The cloning of the mouse and human urocortin genes (*Ucn* and *UCN*) and characterisation of their genomic locations led to the discovery that the *Mpv17* gene is located adjacent to both mouse and human urocortin genes in the 5' upstream region (Donaldson *et al.*, 1996; Zhao et al., 1998). The identification of the *Mpv17* gene approximately 2.1 and 1.3 kb upstream of exon 1 of the *Ucn* and *UCN* genes respectively, and transcription in the same orientation raises the possibility that expression of *Ucn* is also disrupted in the *Mpv17* (-/-) mice. If the retroviral insertion affected both *Mpv17* and *Ucn*, it is possible that disruption of *Ucn* contributes to the *Mpv17* (-/-) mouse phenotype.

Although the glomerulosclerosis and deafness phenotype in *Mpv17* (-/-) mice was rescued by transgenesis with the human *Mpv17* homologue (Schenkel et al., 1995), it is unknown whether the other accompanying aspects of the phenotype were rescued. Other aspects of the *Mpv17* (-/-) phenotype include hypertension, and visible symptoms such as inactivity, weight loss and pallor.

The neuroregulator function of urocortin lends itself to an intriguing role in the other aspects of the *Mpv17* (-/-) mouse phenotype. The urocortin is a member of the corticotropin-releasing hormone family that plays a role in the mammalian stress response and is related to fish urotensin and corticotropin-releasing factor (CRF) (Vaughan *et al.*,

1995). The actions of CRF and related peptides are mediated by seven transmembrane domain G-protein-coupled receptors (CRF-Rs). Synthetic urocortin has been shown to have potent biological actions on biological events mediated by both CRF-R1 (stimulation of pituitary ACTH release, increase in arousal and anxiety) and CRF-R2 (vasodilation (Vaughan et al., 1995), cardiac ionotropism (Parkes *et al.*, 1997), reduction of vascular permeability (Turnbull *et al.*, 1996) and suppression of appetite (Spina *et al.*, 1996)). The specific involvement of urocortin in these biological processes suggests possible links to the hypertension, weight loss and inactivity of *Mpv* (-/-) mice, perhaps as a result of diminished *Ucn* expression. For the elucidation of any urocortin role in the *Mpv17* (-/-) mouse phenotype, extensive study of the *Ucn* expression in the *Mpv17* (-/-) mouse must be carried out.

5.3 The *KIF3C* gene

5.3.1 Introduction

5.3.1.1 Identification of an EST showing similarity to *KIF3B*

The examination of ESTs that map to the DFNB9 interval on chromosome 2p23 revealed the presence of an EST (stSG4510, designed from the cDNA clone Genbank accession number R43988) that shows similarity to the *KIF3B* gene, a member of the kinesin family. Sequence comparison of this EST with *KIF3B* reveals 52% identity at the nucleotide level (see Figure 5.3). Kinesins are microtubule-based motor proteins and share a similar function to the myosins, actin-based motor proteins. Unconventional myosin genes underlie some forms of syndromic and non-syndromic hearing impairment (discussed in Section 5.3.1.2). Furthermore, it has been shown that a kinesin protein can suppress a myosin defect in yeast (Lillie & Brown, 1992), (discussed in Section 5.3.1.3). This information suggested that the novel kinesin-like EST that maps to the DFNB9 interval was a good candidate deafness gene for further investigation.

```
R43988 GTAGAAATTTACTTATC--ACTTGAGATACCTAGAGACATTTTGGGCCATCACAAAGGAA 71
KIF3B  GCTGAAACTCAAGCATCTTATTATAGAAAACTTTATCCCTCTGGA---AGAAAAAAGTAA 1960

R43988 GGTAAGGAGTATCCCCCT----------AGGAACCAATTTGCGTAACTAGTGAATAAAG 120
KIF3B  AATTATGAATAGAGCCTTCTTTGATGAAGAGGAAGATCATTG-GAAACTACATCCTATAA 2019

R43988 GATCTATTGTCAACAAAATAATACCTTAAAGATGCAATTCAGAAACAGGGG---TAACAG 177
KIF3B  CCAGACTGGAGAACCAGC--AGATGATGAAGCGGCCAGTCTCAGCCGTGGGATATAAGAG 2077

R43988 GCAA---AGCTGGNAA--AAGATTGTGCCTGGGCTTCTGTTTCCTGTGACAGATGAAGGG 232
KIF3B  ACCATTGAGCCAGCACGCAAGAATGTCCATGATGATTCGTC-CAGAGGCCCGATACAGGG 2136

R43988 AA-AAAGCAATA-GATGTTAATATCTTCGTTTAGCGAGGGGTAG-ANT---GACATTGAC 286
KIF3B  CAGAAAACATTGTGCTGTTAG-AGCTGGACATGCCCAGCCGGACCACCAGAGACTATGAG 2195

R43988 --TCCTCCCACTGTGNAACGGGGTCTAGG--GCAGCTGCAGGGAGAANTGATCGAAGCAG 342
KIF3B  GGTCCAGCCATTGCCCC-CAAGGTCCAGGCTGCAT-TGGATGCGGCTCTGCAGGATGAAG 2253

R43988 GCAGGGAGCAGCCT 356
KIF3B  ATGAGATACAGGTG 2267
```

Figure 5.3 Sequence comparison of EST R43988 sequence with *KIF3B* (Genbank accession number NM_004798). Numbering refers to database entry numbering.

5.3.1.2 Role of unconventional myosin genes in non-syndromic hearing impairment

Myosins are molecular motors that use the energy from ATP hydrolysis to generate force and move along actin filaments (Mooseker & Cheney, 1995). Whereas conventional myosin, or myosin 2, has the specialised ability to form bipolar filaments that form an essential component for the process of muscle contraction, unconventional myosins are non-filament forming (Hasson, 1997). Unconventional myosins are primarily found within non-muscle cells, share structurally conserved heads (the motor domain), and have divergent tails presumably to move different macromolecular structures relative to actin filaments. Phylogenetic analysis of the conserved head domain allow the myosin protein family to be divided up into 15 classes, 9 of which are present in mammals (reviewed in Sellers, 2000). These nine classes (1-7, 9, 10, and 15) can contain multiple members and to date, 26 different myosin genes (10 conventional and 16 unconventional) have been identified.

In addition to the motor domain, myosins contain regulatory or light-chain-binding domains and distinctive C-terminal tail domains. In the unconventional myosins studied so far, the regulatory domain has served as a binding site for calcium binding proteins such as calmodulin, which have been shown to regulate the motor activity in a calcium-dependent fashion (Houdusse *et al.*, 1996). The tail domain is unique to each myosin, functioning in protein-protein, protein-membrane interactions, dimerisation and subcellular targeting (Bement *et al.*, 1994). One of the functions of unconventional myosins is to transport cargo within the cell and it is the tail domain that is thought to bind vesicles containing the cargo (Prekeris & Terrian, 1997).

The involvement of unconventional myosins in the pathogenesis of deafness was first demonstrated by the analysis of mice with hearing impairment. Both Snell's waltzer and shaker-1 mice (Deol, 1956; Mikaelin, 1964) exhibit head tossing, hyperactivity and circling behaviour, which are due to vestibular dysfunction, and rapid progressive hearing loss accompanied by neuroepithelial degeneration (Deol & Green, 1966; Green, 1960). Positional cloning techniques revealed that both these mouse phenotypes were caused by mutations in two unconventional myosin genes. The deafness genes *Myo6* (Avraham *et al.*, 1995)and *Myo7a* (Gibson *et al.*, 1995) were identified in the Snell's mouse and in the shaker-1 mouse, respectively. Although both these genes are expressed in the inner ear, their exact function is unknown.

Recent research has shown that unconventional myosins are expressed in the stereocilia of hair cells, the sensory cells of the inner ear (Hasson *et al.*, 1995). Stereocilia are actin-based projections on the apical surface of hair cells, which are required to convert mechanical forces such as sound waves and gravity into electrical signals (Figure 5.1 (c)). A specialised linkage, the tip link, joins the tip of the shorter stereocilia to the adjacent taller stereocilia to form a stereocilium bundle. Stretching of this linkage on bundle deflection leads to the opening or closing of stretch-gated transduction channels. Opening of transduction channels causes an influx of endolymphatic K^+ and Ca^{2+} ions, which leads to depolarisation of the hair cell and neurotransmitter release. After opening, the channels are reset, in a process termed "adaptation", whereby the tension on the tip link is modulated by movement of the transduction apparatus up or down the actin filaments of the stereocilium. It has been suggested that both myosin 6 and myosin 7a are components of the adaptation motor (Avraham et al., 1995; Hasson et al., 1995).

Recent studies of the inner ear epithelia by indirect immunofluorescence and immuno-electron microscopy techniques have suggested that it is another unconventional myosin, myosin 1β, that is a better candidate for the adaptation motor and that myosin 6 and myosin 7a perform other functions in the inner ear (Hasson *et al.*, 1997). Myosin 1β protein was found to be enriched at the tips of the stereocilia, the site of the adaptation process and myosin 7a was found along the length of the stereocilia associating with the linkages that join adjacent stereocilia to their neighbours. However, myosin 6 was found outwith the stereocilium and located in the cuticular plate, an actin rich region below the stereocilium bundle in the apical domain of the hair cell (Hasson et al., 1997). The location of the unconventional myosins 1b, 7a, and 6 in the hair cell and stereocilium of the inner ear is summarised in Figure 5.4.

It is thought that myosin 6 plays a role in anchoring the stereocilia bundle into the cuticular plate. In addition, all three unconventional myosins were localised to a new subcellular domain of the hair cell, termed the "pericuticular necklace" (Hasson et al., 1997). Found between the zona adherens and the cuticular plate, the pericuticular necklace is rich in membrane vesicles and is the site of microtubule ends. It has been suggested that the pericuticular necklace may represent a release point for vesicles carrying cargo such as myosins in transit between the microtubule arrays in the cell body and the actin arrays in either the cuticular plate or stereocilium. Both myosin 6 and myosin 7a are implicated in membrane trafficking, so, perhaps it is these actin based movements that are truly essential for inner ear function.

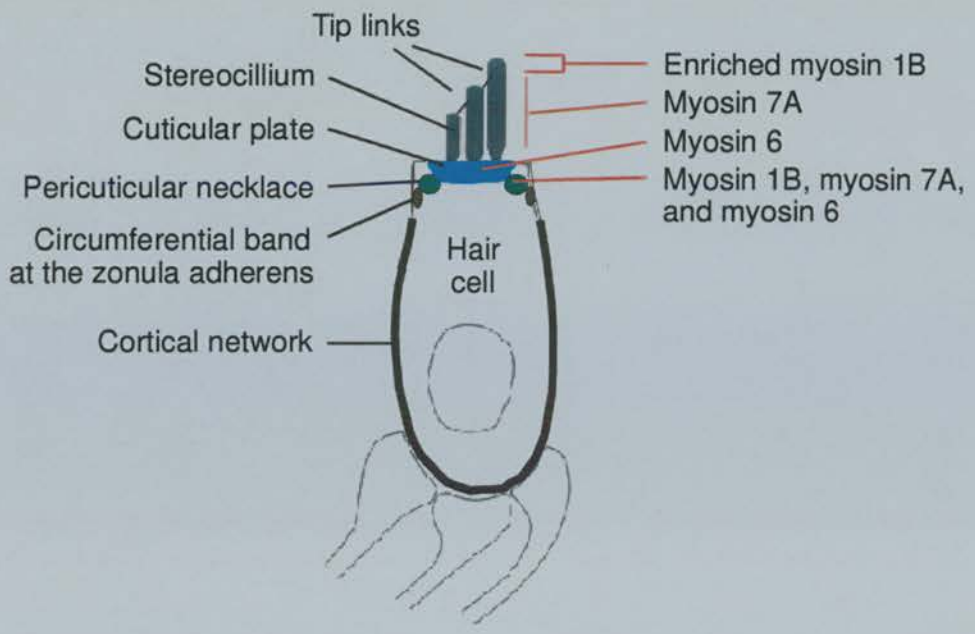


Figure 5.4 Location of unconventional myosins (myosins 1b, 7a, and 6) in the hair cell and stereocilium of the inner ear.

So far, the screening of unconventional myosin genes in humans with non-syndromic hearing impairment has implicated two myosin genes: *MYO7A* as the gene corresponding to both the recessive DFNB2 and dominant DFNA11 loci (Liu *et al.*, 1997), and *MYO15* at the DFNB3 locus (Wang *et al.*, 1998). The *MYO7A* gene has also been determined to be responsible for Usher syndrome Type 1b, a recessively inherited disease characterised by congenital deafness, vestibular dysfunction and retinitis pigmentosa (Weil *et al.*, 1995). There are three other unconventional myosins that are candidates for human deafness genes, the recently cloned *MYO6* and myosin 1β (*MYO1C*) genes, and myosin 1D (*MYO1F*). While *MYO6* does not map to any previously identified human deafness loci, *MYO1C* does map to the DFNB3 interval and *MYO1F* potentially maps to the DFNB15 locus. The cloning of new unconventional myosin genes will undoubtedly provide more candidates for deafness genes.

5.3.1.3 Suppression of a myosin defect by a kinesin-related gene

Yeast studies revealed that the *SMY1* gene, a member of the kinesin superfamily (see Section 5.3.1.4), when present at high-copy number would suppress the defects in a temperature-sensitive myosin mutant (*myo2-66*) in *S. cerevisiae* (Lillie & Brown, 1992). At the restrictive temperature the *myo2-66* mutation does not impair DNA, RNA, or protein

biosynthetic activity, but produces unbudded, enlarged cells (Johnston *et al.*, 1991). In an experiment to look for "multicopy suppressors" for the *myo2-66* mutant phenotype, that is, heterologous genes that, when over expressed, could correct the temperature sensitivity of the *myo2-66* mutant, one suppressor was identified (*SMY1*). This gene encoded a predicted polypeptide sharing sequence similarity with the motor portion of proteins in the kinesin superfamily. In *SMY1* deletion mutants, no obvious defects were observed, therefore the function of Smy1p must either be redundant or non-essential under the conditions used. However, a cross of the myo2-66 mutant with the *SMY1* deletion mutant produced no live double mutants. This "synthetic lethality" was evidence that Smy1p, a putative microtubule based motor, can interact or substitute for Myo2p, a putative actin based motor. The exact function of Myo2p is unknown, but Yeast studies involving the temperature-sensitive mutation (*myo2-66*) have implicated the Myo2 protein (Myo2p) in the process of polarised secretion in *S. cerevisiae* (Johnston et al., 1991). Possible functions that have been suggested include that Myo2p could organise or stabilise the actin cytoskeleton, or that Myo2p attaches to secretory vesicles and carries them along actin filaments to the bud (Johnston et al., 1991; Lillie & Brown, 1992). If the Myo2p mutant causes a defect in vesicle transport, the Smy1p might correct this by carrying the vesicles along microtubules. However this is unlikely as it has been demonstrated in yeast that microtubules are not required for delivery of secretory vesicles to the bud, nor do they substitute in actin mutants. One possibility might be that the microtubules could normally provide a minor pathway that could compensate for a partial defect in myosin at the nominally permissive temperature. Such transport of vesicles along microtubules by Smy1p would represent the first clear-cut example of functional redundancy between actin and microtubules. Another possibility is that Smy1p suppresses the myosin defect by carrying vesicles along actin filaments. Again this seems unlikely, given the sequence similarity of Smy1p to kinesin.

Analysis of the microtubule interaction site of a kinesin motor using alanine-scanning mutagenesis, reveals that the microtubule-interacting residues are located in three loops that cluster in a patch on the motor surface (Woehlke *et al.*, 1997). Crystallographic studies have also shown that the core of the microtubule-binding interface corresponds topologically to the major actin-binding domain of myosin. The nucleotide binding pockets of kinesin and myosin also show similarity to the corresponding region of G proteins, a group of "molecular switches" that exhibit nucleotide-dependent binding interactions with a variety of target proteins (Sack *et al.*, 1999). This evidence suggests that kinesin and myosin and possibly G proteins evolved from a common ancestral protein. The similarity between the kinesin and myosin binding sites for microtubules and actin respectively may support the idea that there

could be some functional redundancy between kinesin like proteins and myosins, for example Myo2p and Smy1p.

To summarise, the kinesin-like gene located on chromosome 2p23.3 can be considered a good candidate deafness gene because 1)defects in other molecular motor proteins have been shown to play a role in the pathogenesis of hearing impairment, 2) both kinesins and atypical myosins have an intraneuronal transport function, and 3) genetic defects in some atypical myosins and kinesins can cross-complement in yeast (Lillie & Brown, 1992).

5.3.1.4 The kinesin superfamily

The kinesin superfamily was defined by the motor protein kinesin, which was first found in the axoplasm of squid where it is thought to play a role in axonal transport (Brady, 1985; Vale *et al.*, 1985). Subsequent research has shown that kinesin is but one member of an extended superfamily of microtubule motor proteins that all share in common a motor domain of approximately 350 amino acids (for either anterograde or retrograde transport), a putative ATP-binding site and a microtubule-binding site.

Characterisation of the kinesin superfamily first began with a systematic search for novel, putative motors for organelle transport by cloning and sequencing cDNAs encoding both ATP-binding and microtubule-binding consensus sequences homologous to the kinesin heavy chain domain from a murine cDNA library (Aizawa *et al.*, 1992). Three types of kinesin superfamily proteins (KIFs) have been defined: N-terminal-motor-domain type, such as *KIF1*, *KIF3*, *KIF4*, *KIF5*, central-motor-domain type, such as *KIF2*, and C-terminal-motor-domain type such as *KIFC1*, *KIFC2* and *KIFC3* (reviewed in Goldstein, 1993). The kinesin superfamily manage to carry out a wide range of functions by differences in its members' direction of transport, velocity, tissue expression and type of organelle transported. It has also been found that while some KIFs can function as monomeric motors (KIF1A and KIF1B), others work as homodimers (kinesin, KIF2, KIF4, and KIF5), and heterodimeric motors (KIF3A and KIF3B) with associated proteins, for example KAP3. Further research is required to understand how these changes in configuration might alter function but it has been suggested that changes in KIF configuration may be important in determining the nature of the cargo and rate of organelle transport.

While tissue expression can give an indication of KIF function, the identification of the cargo transported by the KIFs is required for a better understanding of their cellular function.

Binding domains unique for each individual KIF protein are thought to specify the type of cargo they attach to, for example organelles like mitochondria and vesicles. Working out KIF function will depend on identification of their cargoes and in the case of vesicles, what they carry within. This can be carried out by subcellular fractionation and immunoprecipitation experiments. Additional functional information can be obtained by using antisense oligonucleotides to suppress expression of individual KIF in cell cultures (Ferreira *et al.*, 1992), microinjection of blocking antibodies in other cell types, and "knock-in" and "knock-out" mouse mutants, including the introduction of point mutations within binding domains of the KIF protein. The effects on cellular morphology can be examined and the KIF function deduced.

In understanding the function of KIFs, the issue of functional redundancy must be considered. One clear example of functional redundancy is shown between the kinesin superfamily *CIN8* and *KIP1* genes in *S. cerevisiae*. Deletion of either gene causes little or no phenotype on its own, but the double mutant is lethal (Hoyt *et al.*, 1992; Roof *et al.*, 1992). This situation becomes even more complex when *KAR3* mutations are introduced, which suppress the *CIN8 KIP1* abnormalities (Hoyt *et al.*, 1993).

Direct redundancy might involve two different motors attaching to, and moving the same cargo to the same destination. "Bypass redundancy" might involve a logic similar to that of metabolic pathways or bypass suppressors. Bypass redundancy could occur if more than one pathway or movement could achieve the same outcome as far as the cell is concerned. If the cargo and destination for each motor can be ascertained, exploitation of functional redundancy could become very useful as a treatment for diseases caused by a malfunctioning KIF. By switching on an alternative pathway, it may be possible to bypass a faulty gene product and therefore restore the healthy phenotype. In the analysis of disease phenotypes, it should be considered that the phenotype exhibited in each tissue could be dependent on gene expression for that tissue and whether there is any functional redundancy between the proteins that are present in each different tissue. Why eukaryotes have functional redundancy is unknown but in the case of yeast, it may give an evolutionary advantage for survival in certain extreme environmental conditions (Thomas, 1993). In complex organisms, the movement of organelles for specialised cellular functions for example in the human inner ear, might require many different KIFs and although there is some functional redundancy, each KIF may retain its own specialised function.

Although much progress has been made recently concerning how KIFs generate directional force from the chemical energy in ATP, there is still much to learn about what dictates how any specific individual cargo is transported to a specific destination and the mechanisms controlling the bidirectional transport of organelles by KIFs within the cell.

5.3.2 Methods

5.3.2.1 cDNA cloning and sequencing

Two strategies were employed to obtain the full length *KIF3C* cDNA sequence:

1/ The EMBL and Genbank DNA sequence databases were searched using the EST sequence from IMAGE cDNA clone 28784 (Lennon *et al.*, 1996) that contained stSG4150 (Genbank accession number R43988), to look for overlapping cDNA clones. This process was continued until no new overlapping clones were found. The longer cDNA clones that spanned the whole cDNA contig were chosen for sequencing on both DNA strands (IMAGE clone numbers 261794, 261794, and 28784).

2/ Oligonucleotides were designed from the sequence of IMAGE clone 28784 (5′-dGGAGATCCAGGACCAGCATG-3′ and 5′-dGCTTGTCCAATCGCATGAGCC-3′) and a 442 bp cDNA probe (corresponding to nucleotides 1953 to 2394 of human *KIF3C*; Figure 5.6) was amplified by PCR, radiolabelled using a random primer DNA labelling kit (Boehringer) and used to screen a human fetal brain cDNA library (Clontech). Sequence analysis of a clone containing a 4 kb insert revealed the cDNA to be a hybrid of an unrelated human sequence fused to a kinesin-like open reading frame (Figure 5.6, nucleotides 874-2320). To determine the 5′ end of the *KIF3C* cDNA sequence, the method of RACE amplification was used –see Chapter 6, Section 6.2.3.3.

The products of 5′ RACE reactions were analysed by Southern blot hybridisation with ^{32}P-labelled primer (5′-dGGAGAGAGGCCTAAGGAAG-3′; Figure 5.6, nucleotides 1003-1021) and an 820 bp product was cloned into the TA vector (Invitrogen) and sequenced with vector- and kinesin-specific oligonucleotide primers. Multiple PCR-derived cDNA clones were used to check for errors in the nucleotide sequence produced by amplification procedures. The remaining 357 bp from the 5′ end and 1181 bp from the 3′-end of the gene were sequenced from PAC genomic DNA. DNA sequencing was performed on both strands of cloned cDNA using either the Thermosequenase cycle sequencing kit (Amersham) or AmpliTaq cycle sequencing kit (Perkin Elmer) and the ABI Prism 377 DNA sequencer.

The BLAST algorithm (Altschul et al., 1990) was used to search for homologies or identities between sequences identified and sequences entered in the Genbank database. The cDNA sequence has been submitted to Genbank and assigned the accession number AF035621.

5.3.2.2 Isolation of genomic clones

The stSG4150 primer pair (5′-dCCTAGAGACATTTGGGCCA-3′ and 5′-dTTGCCTGTTACCCCTGTTTC-3′; Figure 5.6, nucleotides 3678-3697 and 3557-3576 respectively) was used to PCR screen the ICI YAC genomic library (Anand et al., 1990) and the PAC human genomic library RPCI1 (obtained from the HGMP Resource Centre, Hinxton, U.K.). Standard PCR conditions were used (1.5 mM $MgCl_2$) with the hot start PCR program 94°C, 5min; 40 x (94°C,45 s; 58°C,45 s; 72°C, 1 min). Four positive genomic clones containing *KIF3C* were obtained: two YAC clones 14IC3 and 35HB12 and the two PAC clones 14H17 and 97K3.

5.3.2.3 Structural analysis

To determine the *KIF3C* genomic structure, oligonucleotides were designed from the cDNA contig and used to sequence directly from the PAC clones 14H17 and 97K3. The exon-intron boundaries were sequenced across from both directions and the exact junctions determined by comparison of the cDNA sequence to the genomic sequence and looking for the exon-intron consensus sequence. Sizes of introns were determined by amplification across each intron using PAC genomic DNA as template and primers located within adjacent exons. The fragments amplified were analysed on a 1.5 % agarose gel and sized against a GibcoBRL 1 kb DNA ladder. Genomic sequences have been submitted to the EMBL sequence database and assigned accession numbers AJ002223-AJ002229.

5.3.2.4 Interspecific backcross mapping

To determine the location of the mouse *Kif3c* gene, linkage analysis was performed using the Jackson laboratory interspecific backcross panel – see Chapter 2, Section 2.1.6 for more information about this backcross panel. Searching of the Genbank EST database with the human *KIF3C* sequence revealed several ESTs showing significant similarity to the human *KIF3C* sequence including EST W82835, highly similar to the 3′-untranslated region of *KIF3C* (nucleotides 4349-4913, Figure 5.6). This suggests a high degree of sequence conservation between human and mouse *KIF3C*. A 458 bp genomic fragment of the 3′UTR corresponding to part of EST W82835 was amplified from *Mus spretus* and C57BL/6J using the primers MkinF (5′-dCTACCCTCACAGTCTTATAGC-3′) and MkinR (5′-dAAAGGCCTCCATCCTAAACCA-3′), but no variants were found on sequencing.

The primers Kin3 (5′-dGGAGATGCAGCAGGAGATG-3′) and Kin2 (5′-dGGGTCTGCTCGTTCTGCG-3′), corresponding to nucleotides 1815-1833 and 2026-2009 respectively in Figure 5.6, were found to amplify fragments of ~1.0 kb from C57BL/6J and *M. spretus* DNA. Sequence analysis of this PCR product revealed that the sequence from the Kin2 primer crosses the splice acceptor site of intron 3, and within this intron a polymorphic $(CT)_n$ repeat region is found (see Figure 5.5). A third primer Mkin2R (5′-dGCATTCCATCAGTTCTCTTTCAG-3′), on the other side of the $(CT)_n$ repeat was used in conjunction with the Kin2 primer (annealing temperature 63°C) to type this polymorphism in 94 animals from the Jackson laboratory interspecific backcross panel: Mkin2R and Kin2 amplify a ~350 bp fragment from C57BL/6J and ~400 bp fragment from *M. spretus*.

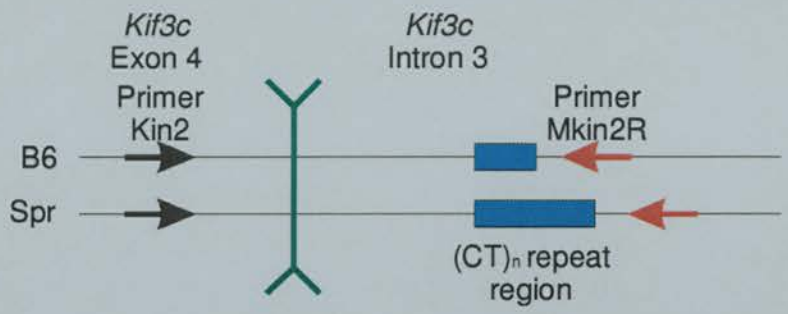


Figure 5.5 Diagram showing polymorphic $(CT)_n$ repeat region within C57BL/6J (B6) and SPRET/Ei *Kif3c* intron 3. The primers Kin2 (black arrow) and Mkin2R (red arrow) were used to PCR across the polymorphic $(CT)_n$ repeat region (blue box), amplifying products of ~350 and ~400 bp from C57BL/6J and SPRET/Ei, respectively. This PCR assay was used to type the Jackson Laboratory interspecific backcross panel (C57BL/6JEi x SPRET/Ei) x SPRET/Ei).

5.3.2.5 Northern blot analysis

To examine expression of the human *KIF3C* gene, a multiple tissue Northern blot was probed using a ^{32}P-labeled 818 bp PCR product encoding the motor domain (nucleotides 360-1177, Figure 5.6), performed by Dr L. Telford. Approximately 1 µg of total RNA from various adult human tissues (brain, salivary gland, oesophagus, trachea, heart, lung, and stomach) was tested.

5.3.2.6 Fluorescence in situ hybridisation

To map *KIF3C* by FISH, *KIF3C*-containing YAC clones were identified by screening the ICI YAC library by PCR with the stSG4510-specific primers. Two YACs were identified, YAC 35HB12 and YAC 14IC3. The YACs 14IC3 and 35HB12 were shown by PCR to contain the marker *D2S2144*. Total yeast DNA was extracted for the YAC 35HB12, and the DNA labelled by nick-translation with digoxigenin-11-dUTP and used for FISH analysis (Pinkel *et al.*, 1986). FISH analysis was performed by Dr J. Leek.

5.3.3 Results

5.3.3.1 Sequencing *KIF3C* cDNA

The human IMAGE clone 28784 (Genbank accession number R14361) displayed significant sequence homology to the 68 kDa and the 95 kDa kinesin-like proteins of *Drosophila melanogaster* and the sea-urchin *Strongylocentrus purpuratus* respectively. Nucleotide sequence analysis of the entire 1917 bp insert reveals that it contains a 720 bp open reading frame encoding a kinesin-like carboxy-terminal tail domain and 1197 bp of 3′UTR sequence. A composite full-length cDNA (4913 bp) was assembled from a human fetal brain cDNA clone, from 5′ RACE products which were generated from total fibroblast RNA, and from cDNA clones identified by database searching (see Methods).

The full-length *KIF3C* cDNA is characterised by a single open reading frame of 2382 bp encoding a predicted protein of 793 amino acids (Figure 5.6). The overall organisation of the predicted protein is similar to members of the KIF3 family of KRPs and contains an amino-terminal motor domain (residues 1-389), a central rod domain (residues 390-599) and a carboxy-terminal tail domain (residues 600-793). A putative polyadenylation signal is located nucleotides 2354-2359 nucleotides downstream from the translation termination codon. The cDNA clones have a poly(A) tail added 20 nucleotides downstream of this (2378 nucleotides downstream of the termination codon; Figure 5.6).

At the time of cloning *KIF3C*, searching the Genbank sequence database revealed that among the kinesin superfamily proteins described, this novel sequence was most similar to a 195 bp partial cDNA encoding the murine *KIF3C* motor domain (Genbank accession number AB001433; corresponding to nucleotides 430-624 in Figure 5.6), the human and partial mouse sequence sharing 91% and 98% identities in nucleotide and amino acid sequence, respectively. The mouse and rat *Kif3c* genes have subsequently been cloned (Genbank accession numbers AF013116 and AF083330, respectively). Comparison of human *KIF3C* to the mouse and rat *Kif3c* genes reveals 89% identity at the nucleotide level over the protein coding region. When compared to human KIF3C, the mouse and rat Kif3c proteins exhibit 94% and 95% identity respectively.

KIF3C cDNA sequence and putative translation product

Figure 5.6 The complete *KIF3C* cDNA sequence and its putative translation product. The termination signal is marked with an asterisk and the poly(A) addition signal is underlined green typeface. The conserved ATP/GTP-binding motif and kinesin motor domain signature are blue typeface. The first base of exons 2-8 is underlined red typeface.

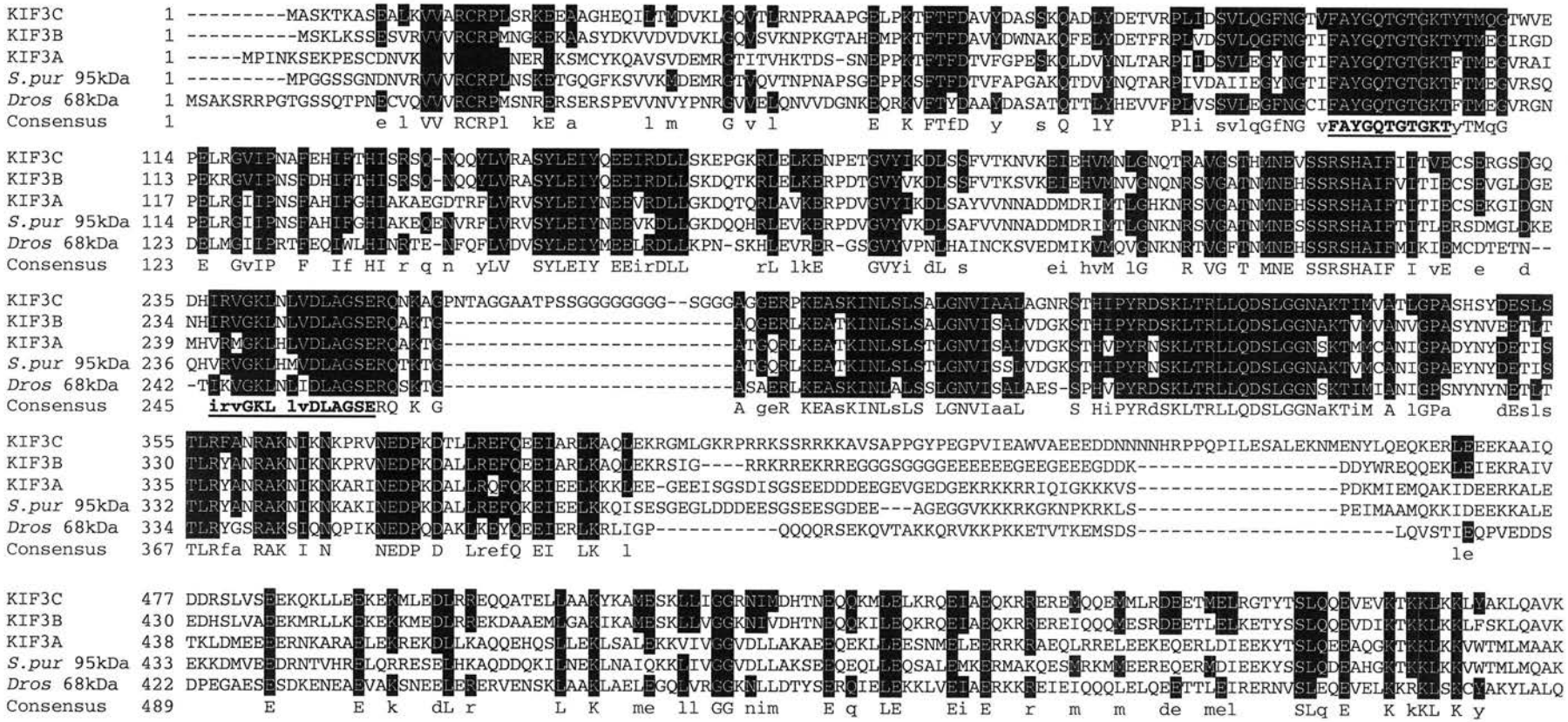


Figure 5.7 Comparison of human KIF3C amino-terminal motor domain (residues 1-389) and central rod domain (residues 390-599) with homologous KIF3 proteins; KIF3B (NP_004789), KIF3A (NP_008985), sea urchin 95 kDa (P46871), and Drosophila 68 kDa (P46867). Sequences were aligned using the ClustalW sequence alignment program. Black box shading indicates positions which have a single, fully conserved residue. The ATP/GTP-binding motif and kinesin motor domain signature are in underlined bold typeface.

Figure 5.7 shows the human KIF3C amino-terminal motor domain and central rod domain amino acid sequence compared to various KIF3-like proteins (KIF3A, KIF3B, *Drosophila melanogaster* 68 kDa and *Strongylocentrus purpuratus* 95 kda kinesin-like proteins). This reveals that the amino-terminal motor domain (residues 1-389) is highly conserved with many positions showing a single, fully conserved residue. This includes the ATP/GTP-binding motif and kinesin motor domain signature which are highly conserved even between species. Comparison of the amino acid sequences of KIF3 motor domains also reveals that the human KIF3C sequence encodes a 24 amino acid glycine-rich domain which is not conserved in human KIF3A, KIF3B, *Drosophila melanogaster* 68 kDa and *Strongylocentrus purpuratus* 95 kda KIF3-like proteins (Figure 5.7, residues 255-292). However, comparison of human, mouse and rat KIF3C sequences reveals that the glycine-rich domain is conserved in these three proteins (Figure 5.8). Although it is not known why KIF3C contains this glycine-rich domain, the fact that it is conserved in human, mouse, and rat KIF3C proteins does suggest it to possess some functional significance.

```
Human KIF3C (255-292) KAGPNTAGGAATPSSGGGGGGGGSGGGAG--GERPKEASK
Mouse KIF3C (255-294) KAGPNAAGGPATQPTAGGGSGSGSASGSASSGERPKEASK
Rat KIF3C   (255-294) KAGPNTPGGPATQSTAGGGGGGGGTSGSGSSGERPKEASK
Consensus             KAGPNTPGG       GGG G G   G    GERPKEASK
```

Figure 5.8 Comparison of KIF3C glycine-rich region in human, mouse, and rat (Genbank accession numbers JC5831, O14782, and O55165 respectively. Glycine residues are shaded and a consensus sequence is shown.

Although the central rod domain (residues 390-599) is not so highly conserved, a high proportion of residues are fully conserved between these five proteins. The carboxy-terminal tail domain (residues 600-793, not shown in Figure 5.7) is the least conserved domain. An explanation for the differences in sequence conservation for the three KIF protein domains in various KIF3 proteins is the role each domain plays in KIF function. The motor and rod domains are conserved because of their conserved functional role in KIF proteins (the motor domain uses the energy from ATP hydrolysis to generate force and move along microtubules, the rod domain is thought to be the region of dimerisation). The carboxy-terminal domain shows little similarity between different KIF proteins because it is where accessory polypeptides and/or vesicles containing KIF-specific cargo are attached.

5.3.3.2 Structural organisation of the human *KIF3C* gene

To determine the exon-intron structure of the novel *KIF3C* gene, PAC clones containing KIF3C were isolated and analysed by direct sequencing and exon-to-exon PCR. The results of sequence analysis of PAC clones are summarised in Figure 5.10 and Table 5.2. The human *KIF3C* gene spans approximately 12 kb of genomic DNA and consists of 8 exons and 7 introns, the exons ranging in size from 2472 bp to 109 bp. The first methionine codon of the open reading frame is located in exon 1, whereas the stop codon and poly(A) addition signal (Gil *et al.*, 1987) are located in the last exon, exon 8. Exon 8 contains 2378 bp of untranslated sequence. All introns have the consensus sequence (C/T/A)AG-exon-GT(G/A) at their boundaries (Table 5.2).

The present gene structure is the first described for any kinesin family member. The ATP/GTP-binding site motif and the kinesin motor domain signature (nucleotides (nucleotides 442 to 465 and 871 to 906 respectively, Figure 5.6) are both located in exon 1. This exon includes the whole of the N-terminal motor and over half of the rod domain. Exons 2 and 3 encode the remaining part of the rod domain and exons 4-8 the C-terminal tail.

5.3.3.3 Expression of human *KIF3C*

The tissue distribution of human *KIF3C* mRNA was determined by analysis of a multi-tissue northern blot (see Figure 5.9). A transcript of approximately 5.0 kb in size was found to be expressed in brain but could not be detected in any of the other tissues tested (salivary gland, oesophagus, trachea, heart, lung, and stomach). Smaller transcripts of approximately 1.5 kb, 1.3 kb, and 1.0 kb were identified in all tissues with similar levels of expression.

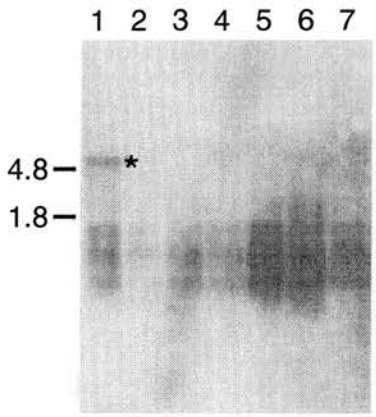


Figure 5.9 Northern blot analysis of *KIF3C* expression in different human tissues. Approximately 1µg of total RNA from various adult human tissues were loaded: brain (lane 1), salivary gland (lane 2), oesophagus (lane 3), trachea (lane 4), heart (lane 5), lung (lane 6), stomach (lane 7). The size of the 28S (4.8 kb) and 18S (1.8 kb) ribosomal RNA subunits are shown. The largest transcript, expressed only in brain is marked with an asterisk. This Northern blot was performed by Dr L. Telford.

5.3.3.4 Mapping of human *KIF3C*

Following hybridisation of the *KIF3C*-containing YAC (35HB12) to normal human chromosomes, doublet signals on the short arm of chromosome 2p23 were observed in 25 cells (Figure 5.11). The distribution on 2p was as follows: 1(4), 2(21), 3(0), 4(0) chromatids per cell. This FISH result independently confirms the *KIF3C* localisation obtained by radiation hybrid mapping of stSG4510 to chromosome 2p23.3 by the Sanger Centre. STS mapping of the YACs 35HB12 and 14IC3 also revealed the presence of the STS *D2S2144* on both YACs.

Exon1-TGCTTGCGGCCAAGTACAAGgtaagggccccagaggagct----1.0 kb----acccctctgtgtgtccccagGCCATGGAGAGCAAGCTCCT-Exon2
Exon2-GGCAGGAGATTGCCGAGCAGgtagggcctccaggtgccag----0.6 kb----actgcccgtccttggcctagAAACGTCGTGAGCGGGAGAT-Exon3
Exon3-AAACCAAGAAACTCAAGAAGgtgagacgctgcagcaggac----1.3 kb----tggcacctgtccccacccagCTCTACGCCAAGCTGCAGGC-Exon4
Exon4-ACCCGCGAACTCAAGCTCAAgtagggcccgcagctctttt----2.6 kb----tccttcattcctgctcccagGTACCTAATCATCGAGAACT-Exon5
Exon5-CCACTGGTGCCAGCCGGCGTgtagtctctaacccagctgt----1.5 kb----atatagcctctttcctacagCAGTAGCAGCCAGATGAAGA-Exon6
Exon6-GGTCCCACCCCAGGTACAGGgtaagaagcggagaggggag----0.4 kb----tggcatgatattcccaccagGCTGAAAACATAATGTTTCT-Exon7
Exon7-GTCCGAAAGTCCAGATCCTGgtcagtacctccatggtccc--0.231 kb----ctgcgtcatctcccctgcagGTGCCAGAGTCCTCAGCGGC-Exon8

Table 5.2 *KIF3C* exon-intron splice junction sequences. Intron sequences are in lowercase letters and exon sequences are in capital letters.

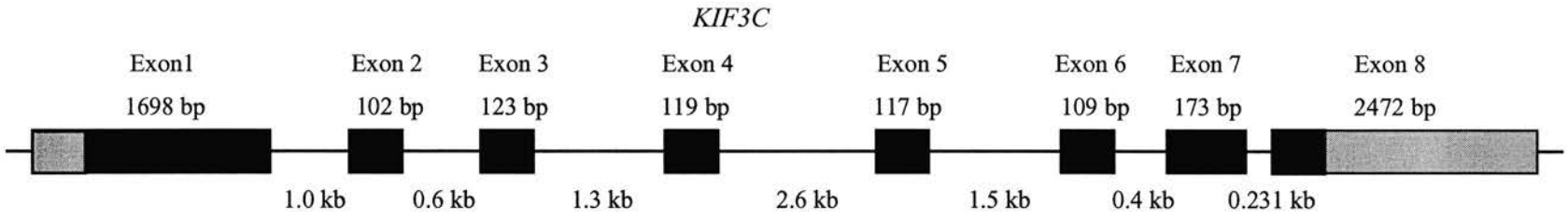


Figure 5.10 Genomic organisation of *KIF3C*. Exons are shown as shaded boxes, coding regions darker than untranslated regions. The diagram is not drawn to scale; exact exon sizes (bp) and approximate intron sizes (kb) are indicated.

Figure 5.11 Chromosome mapping of human *KIF3C* by fluorescence *in situ* hybridisation. A signal is present on both chromatids of chromosome 2 at a location corresponding to p23. This experiment was performed by Dr J. Leek.

5.3.3.5 Mapping of mouse *Kif3c*

The *Kif3c* gene was genotyped as described in the Methods section. The products from the PCR assay used to type the alleles from the Jackson interspecific backcross mapping panel were visualised by electrophoresis through agarose gel (Figure 5.12). Mapping of murine *Kif3c* using the Jackson laboratory interspecific backcross panel placed the murine gene on chromosome 12 between *D12Mit44* and *D12Mit182* (Figure 5.13). The mouse *ApoB* gene also maps to this interval. (Human *APOB* is localised to chromosome 2p23-24).

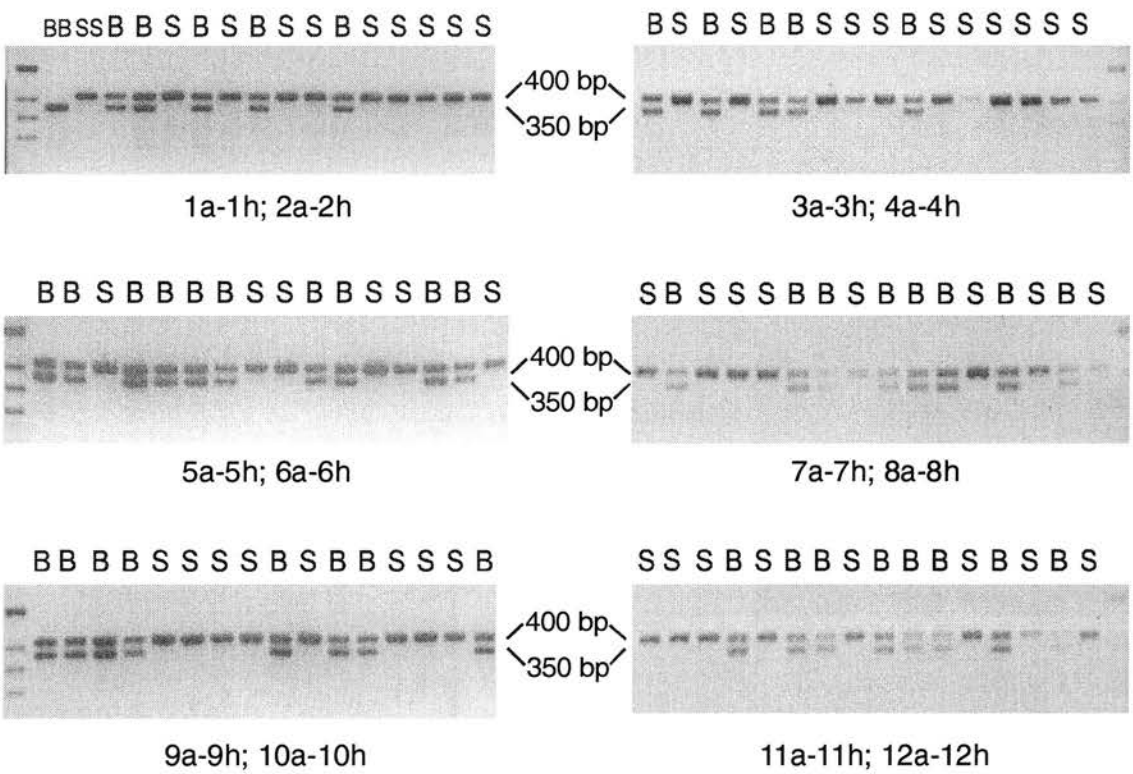


Figure 5.12 *Kif3c* allele typing. The typing of each interspecific backcross animal is indicated above each gel image (B:heterozygote C57BL/6JEi / SPRET/Ei type; S: homozygous SPRET/Ei type. Lane 1A contains parental C57BL/6JEi homozygous alleles (BB) and lane 1B contains parental SPRET/Ei homozygous alleles (SS). The size of each allele is shown (C57BL/6JEi: 350 bp and SPRET/Ei: 400 bp). The end lanes of each gel contain 1 kb ladder.

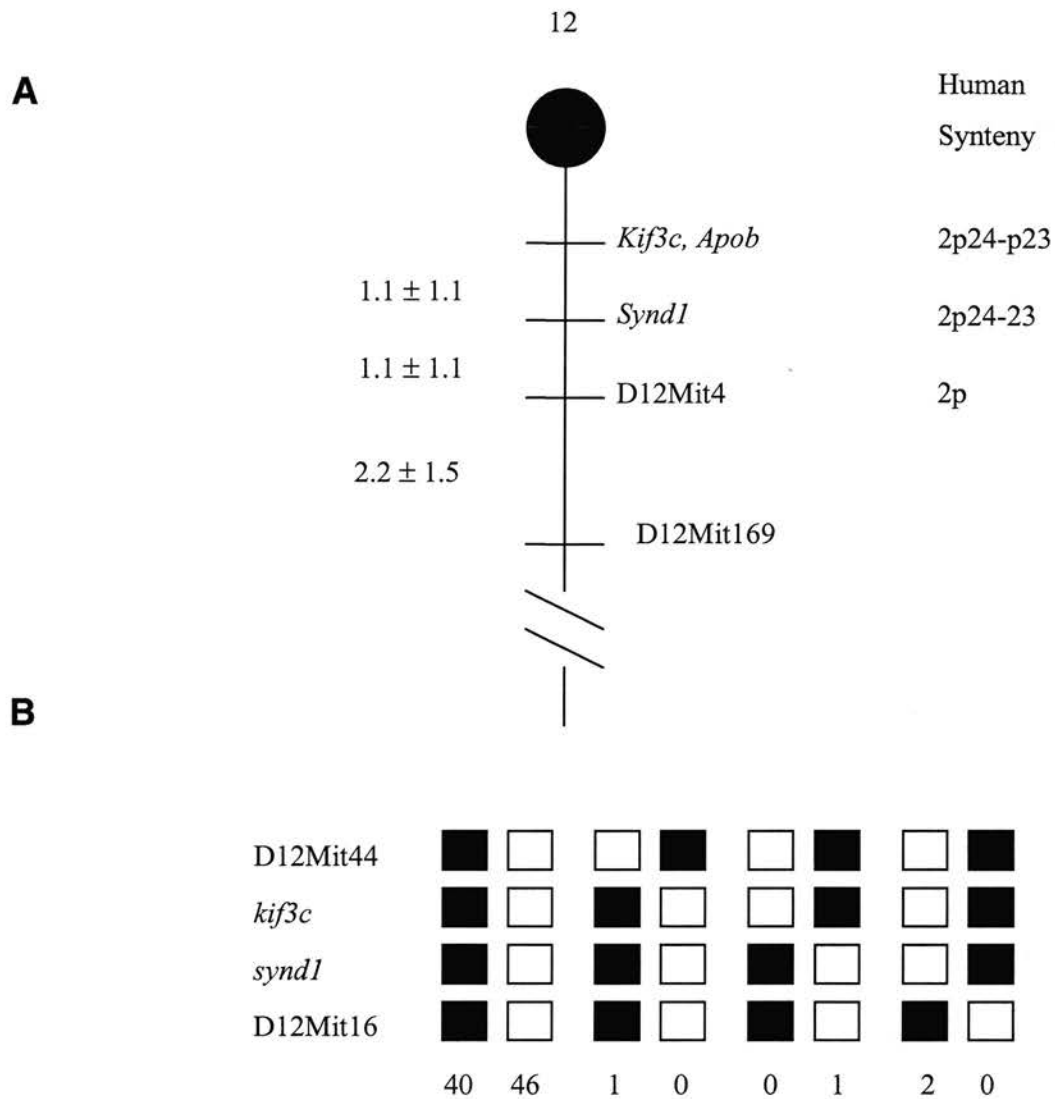


Figure 5.13 Localisation of *Kif3c* on mouse chromosome 12.

A) Map figure showing the proximal part of chromosome 12, with markers linked to *Kif3c*. The map is depicted with the centromere (black circle) toward the top. The gene order and relative positions of markers used in this study are shown. Map distances in centimorgans are shown on the left. The proposed positions of loci in human chromosomes are shown to the right of the chromosome map.

B) Haplotype figure showing loci linked to *Kif3c* on chromosome 12. Loci are listed in order with the most proximal on top. The black boxes represent the C57BL/6J allele, and the white boxes the SPRET/Ei allele. The number of animals with each haplotype is given at the bottom of each column of boxes.

5.3.4 Discussion

5.3.4.1 Cloning of *KIF3C*

This section describes the cloning, structural analysis, expression and localisation of a novel human kinesin-like gene, *KIF3C*. The *KIF3C* gene is highly conserved between human, mouse and rat, showing 89% identity at the nucleotide level within the protein coding region. Comparison of human KIF3C with mouse and rat kif3c protein sequences reveals 94% and 95% identity. The KIF3C protein is most similar to members of the KIF3 subfamily of kinesins which includes the sea urchin 95 kDa (Cole *et al.*, 1993), *Drosophila* 68 kDa (Stewart *et al.*, 1991), mouse *Kif3a* (Kondo *et al.*, 1994), mouse *Kif3b* (Yamazaki *et al.*, 1995), human *KIF3A* (Kondo et al., 1994), and human *KIF3B* (Yamazaki et al., 1995). The overall identity of these proteins with the human sequence is 59%, 43%, 45% 67%, 44%, and 66% respectively, and that relative to the motor domain (amino acid residues 1-389, Figure 5.6) increases to 69%, 51%, 57%, 72%, 57% and 72%.

A striking feature of *KIF3C* with respect to the other KIF3 family members is the insertion of a 24-residue glycine-rich stretch in the otherwise conserved amino-terminal domain (Figure 5.7). Comparison of KIF3C glycine-rich region between human, mouse, and rat reveals that this insertion is conserved (Figure 5.8). This insertion immediately precedes the L11 loop (Sablin *et al.*, 1996), a region that is thought to play a crucial role in microtubule binding, forming an arm that protrudes into the groove between microtubule protofilaments (Sosa *et al.*, 1997). As discussed below, experiments studying KIF3 proteins in the rat, have revealed Kif3c, Kif3b and Kif3a to form heterodimers (Muresan *et al.*, 1998). Therefore, whether the structural difference between KIF3C proteins and other KIF3 family members is of functional significance might be ascertained by comparing the microtubule and nucleotide binding kinetics of KIF3C to KIF3A and KIF3B.

5.3.4.2 The KIF3 superfamily

Although genes of the kinesin family are ubiquitously expressed, some of its members display a restricted tissue distribution, including Kif1, Kif3 and Kif5, which are expressed almost exclusively in murine brain (Aizawa et al., 1992), suggesting that certain kinesins may perform a role in tissue-specific functions. Northern blot analysis, using a probe encoding the *KIF3C* motor domain, identified a transcript of approximately 5.0 kb which was expressed in brain but not in any of the other tissues tested. This is in agreement with studies showing mouse *Kif3c* to be expressed mainly in neural tissues such as brain, spinal

cord and retina (Yang & Goldstein, 1998). The fact that RACE products from near the 5' end of the gene could be generated from total fibroblast RNA suggests a basal level of expression of the full-length transcript in tissue other than brain which is detectable by RT-PCR but not Northern hybridisation. Smaller transcripts of 1.0 kb, 1.3 kb and 1.5 kb appear clearly in this Northern blot after stringent wash conditions and were expressed at comparable levels in all tissues. It is possible that these smaller transcripts represent cross-reacting kinesin related proteins (KRPs).

KIF3 subfamily proteins have been reported to be plus-end-directed microtubule motors with roles in anterograde axonal transport for membraneous organelles in neurons (Kondo et al., 1994). Immunoprecipitation assays have shown that some members, including mouse Kif3a and Kif3b, assemble heterotrimeric complexes comprising two homologous but distinct KRPs associated with a non-kinesin polypeptide subunit which has been proposed to function as an adapter for cargo attachment (Cole et al., 1993; Yamazaki et al., 1995). Immunoprecipitation experiments have now shown that rat Kif3c associates with Kif3a but not Kif3b, but also that a significant fraction of Kif3c is not in association with Kif3a, suggesting that rat Kif3c is part of two different motor complexes (Muresan et al., 1998). Proteins with which human KIF3C interact and the nature of the cargo that it transports have yet to be determined. A recent study found that *KIF3C* expression was upregulated in several cell lines undergoing growth arrest, suggesting a possible functional correlation of KIF3C with growth control (Cabibbo *et al.*, 1998).

5.3.4.3 *KIF3C* as a candidate *DFNB9* gene

The localisation of the human *KIF3C* gene to the genomic interval for sensorineural non-syndromic recessive deafness, DFNB9 (Chaib et al., 1996), suggested *KIF3C* as a good candidate gene for DFNB9. This was supported by the fact that non-syndromic deafness at the DFNB2 locus (11q13) and the syndromic form of deafness Usher syndrome Ib result from mutations of the myosin 7A (*MYO7A*) gene (Weil *et al.*, 1997). *MYO7A* encodes an atypical myosin which like kinesins, has an intraneuronal transport function. Genetic defects of some atypical myosins and kinesins can even cross-complement in yeast (Lillie & Brown, 1992). As *KIF3C* is also mainly expressed in neural tissues including the cochlea, this gene could play an important intraneuronal transport role within the inner ear. Further investigation of *KIF3C* expression within the inner ear, the proteins it may interact with, and the cargo that this kinesin-like protein may transport, might give clues as to KIF3C function. However, the identification of *OTOF* as the *DFNB9* gene has eliminated *KIF3C* from further consideration in the context of *DFNB9*.

5.4 The *CRMP1* gene

5.4.1 Introduction

Genes that co-localise with *Gckr* and *Khk* on mouse chromosome 5 may also map to the region of conserved synteny on human chromosome 2p23 and this may be a useful approach to identifying possible candidate DFNB9 genes. A search of the mouse chromosome 5 region surrounding the *Gckr* and *Khk* locus for genes that could encode proteins with functions within the inner ear identified the collapsin response mediator protein-1 gene (*Crmp1*).

The *Crmp1* gene was cloned during the analysis of cDNA clones obtained by the subtraction of a 2-day-old mouse cochlear cDNA library by mouse liver cDNA (Cohen-Salmon *et al.*, 1997b) and was found to be specifically expressed in the brain, retina, and in the cochlea, mainly during the development of the nervous system. The *Crmp1* gene was also later found to be expressed in the testes and was mapped to mouse chromosome 5 near the *Gckr* and *Khk* locus (Taketo *et al.*, 1997).

Sequence analysis reveals *Crmp1* to show 87% homology to its human homologue, *CRMP1*. *CRMP1* had been cloned during a study to identify genes that encode proteins related to dihydropyrimidase (Hamajima *et al.*, 1996). *CRMP1* was shown to be highly expressed in the brain. Crmp1 is also related to the chick *CRMP-62* gene and *C. elegans Unc-33* gene, a nematode gene involved in the co-ordination of axonal outgrowth (Hedgecock *et al.*, 1987; Li *et al.*, 1992; McIntire *et al.*, 1992). This family of proteins is thought to play a role in mediating "collapsin" activity during neuronal development. Collapsin, a member of the semaphorin family, is involved in the development of the nervous system by acting as a repulsive cue toward specific neuronal populations (Luo *et al.*, 1995; Puschel *et al.*, 1995).

The investigation of collapsin signalling identified chick *CRMP-62*, a gene expressed exclusively in the developing chick nervous system (Goshima *et al.*, 1995). Introduction of anti-CRMP-62 antibodies into dorsal root ganglion neurons blocked collapsin-induced growth cone collapse, suggesting CRMP-62 to be one of the components of the signaling cascade initiated by the fixation of collapsin to an yet unidentified receptor (Goshima et al., 1995).

The mapping of *Crmp1* to the same region of mouse chromosome 5 as *Gckr* and *Khk* suggests that its human homologue *CRMP1* might map, like *GCKR* and *KHK*, to the region of conserved synteny on human chromosome 2p23.3. If *CRMP1* did map to human chromosome 2p23 like *GCKR* and *KHK*, it could be considered an excellent candidate *DFNB9* gene because of its role in regulating neuronal growth and its high expression levels during innervation of the developing cochlea.

5.4.2 Methods

To map the human homologue *of Crmp1*, an STS was designed from the partial human *CRMP1* gene sequence (Genbank accession number U17278) and used to screen a monochromosomal somatic cell hybrid DNA panel (supplied by HGMP). The primers CRMP1F (5′-dTTAGTTTGGTGCTGATGGAG-3′) and CRMP1R (5′-dTCTGAGTGTGAACCTGGCT-3′) were used to amplify a PCR product of size 657 bp using the hot start PCR program 94°C, 5min; 30x(94°C,45s;58°C,45s;72°C,1min), 50ng of template and standard PCR buffer (1.5 mM $MgCl_2$). PCR products were run on a 1.5% agarose gel.

5.4.3 Results

Screening of the monochromosomal somatic cell hybrid DNA panel (Kelsell *et al.*, 1995) for *CRMP1* gave positive results for chromosome 4 and chromosome 20 (Figure 5.14). Examination of the data sheet for the DNA panel revealed that the chromosome 20 cell hybrid also contained part of chromosome 4 therefore it is most likely that *CRMP1* resides on human chromosome 4. Since *CRMP1* does not map to chromosome 2p23, its structure was not further investigated.

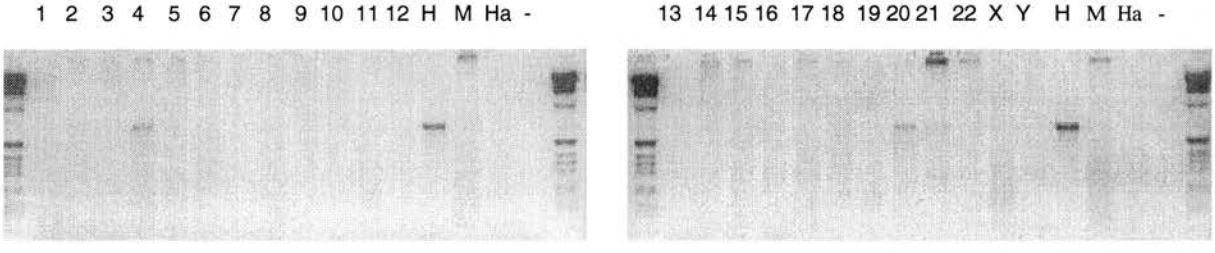


Figure 5.14 PCR screening of a monochromosomal somatic cell hybrid DNA panel. First half of gel sample order: 1 kb ladder; cell hybrid chromosomes 1-12; human; mouse; hamster; negative control; 1 kb ladder. Second half of gel sample order: 1kb ladder; cell hybrid chromosomes 13-22, X, Y; human; mouse; hamster; negative control; 1 kb ladder.

5.4.4 Discussion

The genes surrounding *Gckr* and *Khk* locus on the proximal part of mouse chromosome 5 map to three regions of conserved synteny within the human genome: *Gbx1* and *En2* map to human chromosome 7q36; *Gckr* and *Khk* map to human chromosome 2p23; *Htr5a*, *Hmx*1 and *Msx*1 map to human chromosome 4 (*Msx1* maps to chromosome 4p26), gene order starting closest to the centromere *Gbx1/En2 - Gckr/Khk-Htr5a/Hmx1/Msx1*. Mapping of *CRMP1* to human chromosome 4 (Figure 5.14) suggests that it too will most likely reside close to the *Hmx1* and *Msx1* genes, possibly on chromosome 4p16. The result discounts its candidacy as a *DFNB9* gene. Nonetheless, other research carried out to investigate *Crmp1* expression in 2-day old mouse cochlea by *in situ* hybridisation revealed a highly specific signal in the spiral ganglia, which contain the neurons innervating the sensory hair cells of the organ of Corti (Cohen-Salmon *et al.*, 1997a). The high expression of *Crmp1* during innervation of the developing cochlea suggests that although the *CRMP1* gene is not the *DFNB9* gene, it remains an excellent candidate deafness gene.

5.5 The *KCNK3* gene

5.5.1 Introduction

5.5.1.1 Chromosomal mapping of *KCNK3*

During the search for candidate *DFNB9* genes, ESTs that map to chromosome 2p23 by radiation hybrid mapping (see http://www.ncbi.nlm.nih.gov/genemap) were used to PCR screen the YAC physical contig constructed at the DFNB9 interval (See Chapter 3, Figure 3.5). One EST was found to be positive for YAC 7DD11, a YAC that maps within the DFNB9 interval. This EST had been designed from the 3'UTR of the human potassium ion channel gene *KCNK3* (*TASK*), Genbank accession number AF006823. The mapping of *KCNK3* to within the DFNB9 interval, added to the important role that potassium ion channels play in the ear (discussed later in this section), makes *KCNK3* a good candidate deafness gene.

5.5.1.2 Potassium channels

Potassium channels represent the largest and most diverse group of ion channels (Rudy, 1988). This diversity originates partly from the large number of genes coding for K^+ channel principal subunits, but also from other processes such as alternative splicing, heteromeric assembly of different principal subunits, and as possible RNA editing and post-translational modifications. By determining and modulating the membrane potential, potassium channels play a major role in neuronal integration, muscular excitability and hormone secretion (Edwards & Weston, 1995).

It has recently been shown that mutations in potassium channel genes are responsible for a number of inherited human diseases. Heterozygous mutations in *KCNQ1* cause autosomal dominant long QT syndrome (LQTS) (Wang *et al.*, 1996). LQTS is electrocardiographically characterised by a prolonged QT interval and polymorphic ventricular arrhythmias (Romano, 1963; Ward, 1964). These cardiac arrhythmias may result in recurrent syncope, seizures, or sudden death. Mutations in *KCNQ1*, when present on both alleles, were also shown to underlie Jervell and Lange-Nielsen (JLN) syndrome (Neyroud *et al.*, 1997), whose symptoms include deafness in addition to cardiac arrhythmias. Mutations in a second gene, *KCNE1*, have also been shown to result in JLN syndrome (Schulze-Bahr *et al.*, 1997). *KCNE1* encodes a transmembrane protein known to associate with KCNQ1, together

forming a delayed rectifier potassium channel (Barhanin *et al.*, 1996; Sanguinetti *et al.*, 1996).

Mutations in either *KCNQ2* or *KCNQ3*, cause benign familial neonatal convulsions (Biervert *et al.*, 1998; Charlier *et al.*, 1998; Singh *et al.*, 1998). *KCNQ2* and *KCNQ3* encode potassium ion channel subunits that can associate with each to form functional potassium ion channels. Interestingly, KCNE1 which is mutated in JLN syndrome, can also associate with KCNQ2 and KCNQ3 (Yang *et al.*, 1998). Mutations in another potassium ion channel gene, *KCNQ4*, can cause autosomal dominant deafness in DFNA2-affected families (Kubisch *et al.*, 1999).

The role of potassium channels in deafness is not surprising as in the cochlea, the transduction current through the sensory cells is carried by potassium and depends on the high concentration of that ion in the endolymph (Delpire *et al.*, 1999). The presence of potassium channels in most cell types suggests that mutations in these genes will lead to deafness by various mechanisms. Indeed, it has been shown that mutations in *KCNQ1* cause deafness by affecting endolymph secretion (Neyroud et al., 1997) but that the mechanism leading to *KCNQ4*-related hearing loss is intrinsic to the outer hair cells (Kubisch et al., 1999). Further evidence for the importance of K^+ in the inner ear is the finding that mutations in the gene for connexin 26, which encodes a gap-juctional protein, are responsible for a surprisingly large fraction of non-syndromic human deafness (Estivill *et al.*, 1998; Kelsell *et al.*, 1997). Gap junctions could have a vital role in the maintenance of K^+ homeostasis, as it has been suggested that K^+ ions are recycled from endolymph through the sensory epithelium and back to the stria vascularis by moving from cell to cell through gap junctions.

5.5.1.3 The *KCNK3* gene

The *KCNK3* (*TASK*) gene encodes a potassium ion channel that is part of the same family of potassium channel proteins as TWIK-1 and TREK-1 (Duprat *et al.*, 1997). Although these three potassium channels have different functional properties, they all possess four transmembrane segments and two P domains (a motif that forms an essential element of the potassium ion-selective filter of the aqueous pore). TWIK-1 gives rise to weakly inward rectifier potassium ion currents (Lesage *et al.*, 1996) while TREK-1 produces outward rectifier potassium ion currents (Meadows *et al.*, 2000). The outward rectification seen in the KCNK3 channel is a result of asymmetric concentrations of potassium ions on both sides of the membrane suggesting that it lacks intrinsic voltage sensitivity (Duprat et al., 1997). In

other words, KCNK3 behaves like a potassium ion selective "hole". Therefore, KCNK3 can be classified as a "background" potassium ion channel.

Background potassium ion channels are open at all membrane potentials and probably play a pivotal role in the control of the resting membrane potential and in the modulation of electrical activity of both neurons and cardiac cells. *KCNK*3, the first mammalian potassium ion channel to be identified as a background potassium ion channel (Duprat et al., 1997), is extremely sensitive to extracellular pH, fully opening or closing within a range of only 0.5 pH unit around the physiological pH (7.4). The sensitivity of KCNK3 to external protons probably has important implications for the physiological function of the KCNK3 channel, for example in modulating neuronal activity.

The mapping of *KCNK3* to chromosome 2p23 , the role of other potassium channel genes in the pathogenesis of deafness, for example *KCNQ1*, *KCNQ4*, and *KCNE1*, and the high sensitivity of KCNK3 to pH in cells like neurons, make this an excellent candidate *DFNB9* gene. To confirm *KCNK3* as a candidate *DFNB9* gene, a more precise localisation of *KCNK3* was performed to see if this gene indeed mapped to the DFNB9 interval.

5.5.2 Methods

5.5.2.1 Isolation of genomic clones

The primer pair (5′-dGTCCTCAGAGACCCTGCTG-3′ and 5′-dCTCCAGTGCGACCATTCTGC-3′), designed from the *KCNK3* 3′UTR sequence (Genbank accession number AF006823), amplifies a PCR product of 376 bp. This primer pair was used to PCR screen the ICI human genomic YAC library (Anand et al., 1990) and other YACs that had been used to construct a physical contig spanning the DFNB9 interval (see Chapter 3, Figure 3.5). To identify other genomic clones containing *KCNK3*, the same primer pair was used to screen a human genomic PAC library (RPCI1) – see Chapter 3 Methods section for details of PAC library screening strategy. Standard PCR conditions were used, with an annealing temperature of 58°C.

5.5.2.2 Radiation hybrid mapping

The primer pair (as described above) was used to PCR screen the Genebridge 4 (GB4) radiation hybrid DNA panel (Gyapay *et al.*, 1996), supplied by the HGMP Resource Centre, Hinxton. Radiation hybrid mapping makes use of a panel of somatic cell hybrids, with each cell line containing a random set of fragments of irradiated human genomic DNA in a hamster background. For the GB4 panel, the X-ray dosage was 3000 rad. The GB4 panel consists of 93 somatic cell hybrids, with an average fragment size of 25 Mb and an effective resolution of 1Mb.

The results were analysed using the Radiation Hybrid Mapping Environment (RhyME) bioinformatics application based at the HGMP resource centre. RhyME takes the results of the typed marker and analyses the typing data in relation to the 1998 International Gene Map (NCBI) using the RADMAP program (HGMP). The marker in question is assigned to the best position on the Genebridge 4 (GB4) radiation hybrid map framework (http://www.ncbi.nlm.nih.gov/genemap98/).

5.5.3 Results

5.5.3.1 Mapping to genomic clones within physical contig

PCR screening of the ICI YAC library identified two KCNK3-containing YAC clones, 7DD11 and 1CA11. Further PCR screening of the YAC physical contig spanning the DFNB9 interval also revealed *KCNK3* to map to the CEPH mega-YAC 964D8 (see Chapter 3, Figure 3.5). PCR screening of the RPCI1 PAC library identified one PAC containing *KCNK3*, 249B24. All three *KCNK3*-containing YACs were also positive for the microsatellite marker D2S174. These genomic clones were placed on the physical contig (see Chapter 3, Figure 3.5).

5.5.3.2 Radiation hybrid mapping

The Genebridge 4 radiation hybrid mapping panel was PCR screened using the *KCNK3*-specific primer pair and the PCR products were analysed on 1.5% agarose gels (Figure 5.15). The presence of a PCR product was scored as a "+" for positives, "-" for negatives and "2" for unknowns/uncertains.

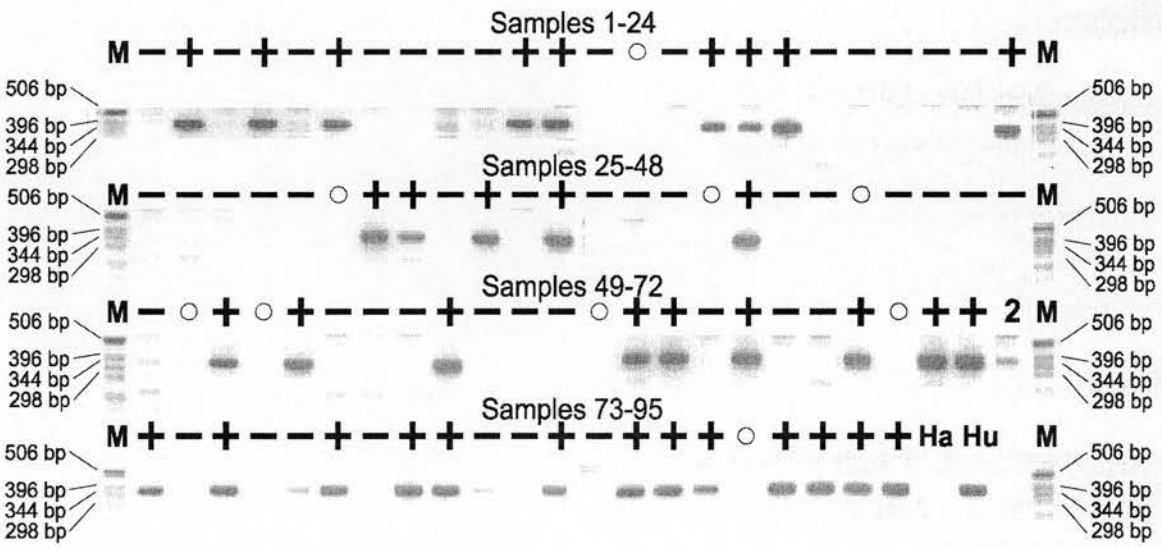


Figure 5.15 Radiation hybrid mapping of *KCNK3*. PCR samples 1 to 95 correspond to the Genebridge 4 radiation hybrid panel, 96 well plate format. Samples 94 (Ha) and 95 (Hu) are hamster and human genomic DNA respectively. The presence of a PCR product was scored as a "+" for positives, "-" for negatives and "2" for unknowns/uncertains. Circles indicate a missing sample (not supplied) and these were also scored as unknowns. The outer lanes contain 1 kb ladder; the sizes of the DNA fragments are indicated.

The "scored" data (or "vector") was analysed using the RADMAP program. Two markers are considered linked if they have vectors of statistically significant similarity (defined as a LOD score), and a measure of their separation is obtained from the analysis of the degree of difference between two vectors. A LOD score >3 is statistically significant. Table 5.3 shows a list of the chromosome 2 markers that vector analysis has determined to be statistically linked to the *KCNK3* marker. This reveals *KCNK3* to be most highly linked to D2S392 on the GB4 radiation hybrid map. The location of *KCNK* relative to other markers in the Genebridge 4 radiation hybrid panel is shown in Figure 5.16.

Marker	Theta	LOD
AFM347ya5 (D2S392)	0.326	7.964
AFM234ya9 (D2S165)	0.390	6.717
AFM296vg9 (D2S352)	0.422	5.956
AFM242yd8 (D2S171)	0.437	5.741
AFM267zc9 (D2S177)	0.469	5.060

Table 5.3 Chromosome 2 markers linked to *KCNK3*.

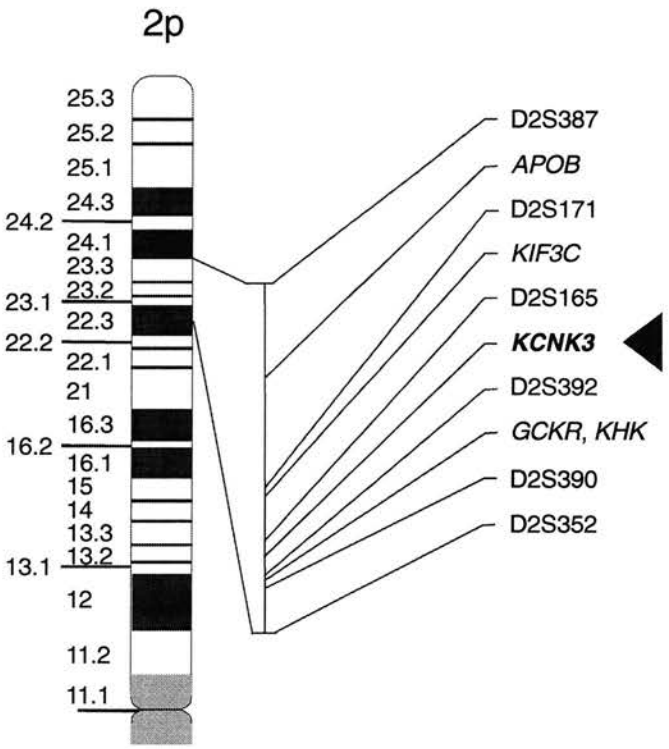


Figure 5.16 Ideogram of human G-banded chromosome 2p. The location of *KCNK3* (indicated by an arrowhead) is shown relative to markers in the Genebridge 4 radiation hybrid panel. Mapping information was obtained from the NCBI database (http://www.ncbi.nlm.nih.gov/)

5.5.4 Discussion

Potassium ion channel genes have a proven role in underlying human syndromes including, long-QT syndrome (Wang et al., 1996), Jervell and Lange-Nielsen (JLN) syndrome(Neyroud et al., 1997), and autosomal dominant deafness (DFNA2) (Coucke *et al.*, 1999). Furthermore, high levels of potassium ions within the inner ear suggest that potassium ion channels are essential for proper functioning of the inner ear. The location of *KCNK3*, a potassium channel gene mapping to chromosome 2p23, suggested that it would be a good candidate DFNB9 gene.

The aim of this section was to determine whether the *KCNK3* gene resided within the DFNB9 interval and therefore could be considered a candidate *DFNB9* gene. This was carried out by physical mapping of *KCNK3*-containing genomic clones and by radiation hybrid mapping. Three YAC clones (7DD11, 1CA11, and 964D8) and one PAC clone (249B24) were shown to contain *KCNK3* by PCR screening using *KCNK3*-specific primers. The three YACs were used to form part of the physical YAC contig located on chromosome 2p23.3. Further PCR screening revealed that the microsatellite marker D2S174 also mapped to these YACs. The DFNB9 interval is delimited by the markers D2S2303 and D2S174 at 2p23.1, so that this physical mapping data places *KCNK3* close to the DFNB9 region.

Radiation hybrid mapping places *KCNK3* between the chromosomal markers D2S392 and D2S165 on chromosome 2p23 in the Genebridge 4 (GB4) radiation hybrid mapping panel (Figure 5.16). This mapping data confirms the assignment of *KCNK3* to human chromosome 2p23, and also confirms FISH studies that locate *KCNK3* to chromosome 2p23.3. (Leek, 1998 unpublished). The data agrees with other radiation mapping data carried out while these studies were in progress (Lesage & Lazdunski, 1998). Recently, FISH analysis has again confirmed the location of *KCNK3* on chromosome 2p23 (Manjunath *et al.*, 1999). Physical mapping of *KCNK3* also shows this gene to map close to D2S165 (Chapter 3, Figure 3.5).

The assignment to the proximal region of mouse chromosome 5 of a mouse cardiac two pore background potassium channel gene (*Kcnk4*) (Fujita et al., 1998) that was highly similar to *KCNK3* (79% and 93% at the nucleotide and amino acid level respectively), suggests that *Kcnk4* is the mouse orthologue of *KCNK3*. The *Kcnk4* gene has now been renamed *Kcnk3* and analysis by FISH confirms this mouse gene to map to chromosome 5 (Manjunath et al., 1999). This shows *KCNK3* to be further example of a gene residing in the region of conserved synteny that lies on human chromosome 2p23 and mouse chromosome 5.

The mapping data described in this section place *KCNK3* very close to the DFNB9 interval. However, the *KCNK3* gene was not further investigated as a DFNB9 candidate, since the identity of DFNB9 with a gene encoding another novel protein was established by others (Yasunaga, 1999). With the proven role of potassium channel genes in human diseases such as LQTS, JLN syndrome and both syndromic and non-syndromic deafness, however, further research into the role of KCNK3 within the heart and inner ear is nonetheless warranted.

5.6 The DFNB9 gene

5.6.1 Cloning of *OTOF*

A combination of the candidate gene and positional cloning strategies was employed to identify the gene underlying DFNB9 (Yasunaga *et al.*, 1999). The segregation of the DFNB9 locus in a consanguineous family living in northern Lebanon led to the initial assignment of the DFNB9 gene to a 2 cM interval delimited by D2S2303 and D2S174 (Chaib et al., 1996). Further linkage studies in families with profound senorineural prelingual hearing loss led to the identification of three additional unrelated consanguineous DFNB9-affected families from Lebanon and this allowed the refinement of the DFNB9 interval to between D2S158 and D2S174 (in a region of ~1 cM).

With the construction of a physical contig from YACs, BACs, and PACs covering the DFNB9 interval, Yasunaga et al. estimated the size of the DFNB9 interval to be less than 700 kb. This contig allowed the assignment of genes and ESTs to the candidate region. Two genes mapping to the DFNB9 interval including *HADHB* (encoding trifunctional protein β subunit) and *CENPA* (encoding centromeric protein A) were not considered as candidate genes for deafness due to the putative functions of their encoded proteins. Several novel ESTs that were mapped to the DFNB9 region were submitted to rounds of 5′ RACE on total fetus mRNA and the predicted amino acid sequences they encoded were compared to other amino acid sequences derived from clones that had been isolated from two subtracted mouse cochlear cDNA libraries (Cohen-Salmon et al., 1997b; Verpy *et al.*, 1999). One of the clones (RH12053) predicted 89.7% amino acid identity and 97.1% similarity to a mouse clone, suggesting that they were orthologous genes. The full length cDNA was constructed by 5′ RACE and an open reading frame encoding 1230 amino acids was identified. Due to sequence homology between the predicted amino acid sequence and the *C. elegans* spermatogenesis factor FER-1 (Achanzar & Ward, 1997), the human protein was named otoferlin (OTOF).

5.6.2 Mutation detection

The screening of all 28 *OTOF* exons from the four DFNB9-affected families identified one stop mutation in exon 18 that would lead to a truncated protein of 729 amino acids in the affected patients. This mutation was homozygous in all affected individuals (21

individuals), heterozygous in their parents (11 individuals) and not detected in any of the unaffected controls screened (106 individuals from Lebanon). These results identified *OTOF* as the causative gene for DFNB9.

5.6.3 Function of OTOF

A search for proteins showing homology to OTOF revealed that apart from the *C. elegans* FER-1 protein, the human protein dysferlin (DYSF) also showed significant similarity. The *DYSF* gene has been reported to underlie Miyoshi myopathy (MM) and limb-girdle muscular dystrophy type 2B (Bashir *et al.*, 1998; Liu *et al.*, 1998). Inspection of the amino acids sequence reveals that both these proteins have a highly hydrophobic C-terminus and further sequence analysis suggests that both are C-terminal membrane-anchored cytosolic proteins containing C2 elements (two four-stranded β sheets thought to bind calcium ions).

The exact function of the FER-1 like family is unknown but clues have been drawn from *C. elegans Fer-1* mutants. During the maturation of spermatids to motile spermatozoa, large vesicles called membranous organelles fuse with the spermatid plasma membrane. In *Fer-1* mutants, the fusion between the large vesicles and the spermatid plasma membrane is found to be defective (Achanzar & Ward, 1997).

It is also known that C2 domain-containing proteins interact with phospholipids and proteins (Rizo & Sudhof, 1998), so it has been hypothesised that otoferlin is involved in Ca^{2+}-triggered vesicle membrane fusions. With *in situ* hybridisation of mouse *Otof* in the inner ear revealing expression in the inner hair cells and vestibular type I sensory hair cells, a possible function of otoferlin maybe in synaptic vesicular trafficking within the synapses of these cells.

5.7 Summary

During this study, three candidate genes: *MPV17*, *KIF3C* and *KCNK3*, were identified that all map close to, or within the DFNB9 interval on human chromosome 2p23 and encode proteins with potentially important functions within the inner ear. However, the gene underlying DFNB9 was eventually identified as *OTOF* (Yasunaga et al., 1999), cloned from an EST mapping to within the DFNB9 interval. Mutation screening of the *OTOF* gene reveals it to contain the same nonsense mutation in four unrelated families of Lebanese origin affected with non-syndromic prelingual deafness.

In the autosomal DFNA2 families, the *DFNA2* gene has been mapped to chromosome 1p34, and fine mapping using different DFNA2 families indicates non-overlapping candidate regions (Van Hauwe *et al.*, 1999). It is now thought that there are two (possibly three) deafness genes that map to the same locus. Therefore, in the case of DFNB9, if the mutation screening of further DFNB9 families reveals no mutations in *OTOF*, the other candidate DFNB9 genes identified in this Chapter would have to be reconsidered for mutation screening. It is now important to screen all families with non-syndromic prelingual deafness and linkage to chromosome 2p23 for mutations in *OTOF*. This work is now in progress.

Chapter 6

6 Summary and future research

6.1 Introduction

This chapter summarises the work carried out in previous chapters and discusses how the genes identified as part of this research could now be subjected to further biochemical and genetic analysis. Also included in this chapter is a brief discussion concerning the impact of recent advances in the Human Genome Project and new technologies on the type of research that is described in this thesis.

6.2 Summary of research and future work

6.2.1 Aim

The candidate type 2 diabetes genes *GCKR* and *KHK* and the gene underlying non-syndromic recessive sensorineural deafness *DFNB9* map to the same genomic region of chromosome 2p23.3. As *GCKR* and *KHK* have previously been shown to encode metabolically connected proteins (see Chapter 1), this raised the possibility that other genes encoding functionally related proteins to GKRP and KHK might also map to the *GCKR-KHK* intergenic region. Further circumstantial evidence for the possibility of *GCKR* and *KHK* co-ordinate regulation was obtained by the mapping of both *Gckr* and *Khk* to mouse chromosome 5, even though the surrounding genes on either side of the *GCKR-KHK* genomic region map to mouse chromosome 12 (see Chapter 2). This thesis describes the characterisation of human chromosome 2p23.3 to search for novel candidate genes for type 2 diabetes mapping to the *GCKR-KHK* intergenic region and also the identification of candidate *DFNB9* genes within the DFNB9 interval.

6.2.2 Assembly of a physical contig

To confirm the intimate co-localisation of *GCKR* and *KHK* to a 500 kb genomic region on chromosome 2p23.3, that had previously been suggested by fluorescent *in situ* hybridisation (Hayward *et al*, 1996), and to aid the search for novel transcripts within the *GCKR-KHK* genomic region, a physical contig consisting of YAC, BAC, PAC, and cosmid clones was constructed using a combination of STS/EST mapping and cosmid (see Chapter 3). The *GCKR-KHK* physical contig showed that *GCKR* and *KHK* were indeed ~500 kb apart. As

the *DFNB9* interval was known to map close to the *GCKR-KHK* genomic region, the *GCKR-KHK* physical contig was extended to include the *DFNB9* interval.

The complete physical contig on chromosome 2p23.3 was estimated to span ~2 Mb (see Figure 3.5) and the search for transcripts by sequence analysis and PCR screening for cDNA markers identified 14 known genes and 15 ESTs that mapped to this contig (described in Chapter 3). The high gene density on chromosome 2p23.3 posed a dilemma as to which genes to further investigate. To prioritise genes for further investigation, candidate type 2 diabetes genes were chosen based on the possibility of involvement in biochemical pathways relating to carbohydrate metabolism (described in Chapter 4), and candidate deafness genes were chosen based on their location within the *DFNB9* interval and potential function of the encoded protein in the inner ear (described in Chapter 5).

6.2.3 Transcripts located within the *GCKR-KHK* genomic region

In Chapter 4, three genes were identified for further investigation: *EIF2B4*, encoding the delta subunit of the initiation factor eIF2B – a protein complex known to play a key role in the regulation of protein synthesis, the activity of which has been shown to be stimulated by both glucose and sugar phosphates; a novel gene called *KIAA0064* that was shown to be intimately located in a "head to head" arrangement with *EIF2B4*; and *RBSK*, the gene encoding the human ribokinase protein – a member of the same kinase family as KHK. After mapping and cDNA cloning of these transcripts, the genomic structures for these genes were elucidated.

In the case of *EIF2B4*, two isoforms were identified that differed at their 5′ ends. Investigation of the *EIF2B4* isoform expression by RT-PCR showed that both isoforms were expressed in brain, heart, liver, and muscle. However, alternative splice forms of *EIF2B4* were identified only in the brain. Although the function of the two *EIF2B4* isoforms and brain specific alternative splice forms is unknown, their existence might correlate with a complex regulatory system that acts on the eIF2B complex and control of protein synthesis. The function of the two *EIF2B4* isoforms and the brain isoforms requires further investigation.

Possible ways to investigate the EIF2B4 isoform function in the eIF2B complex would be to remove expression of one isoform and examine the effect on eIF2B activity to external cellular stimuli such as glucose. This experiment could be conducted using anti-sense RNA oligonucleotides that can block expression of one of the *EIF2B4* isoforms in cell tissue

culture. Activity of the endogenous eIF2B complex could be compared to activity of the eIF2B complex lacking one of the EIF2B4 isoforms. Whereas in the past, phosphorothioate based anti-sense RNA oligonucleotide experiments have suffered from degradation by nucleases, unpredictable targeting, poor sequence specificity, and multiple non-antisense activities, the use of "morpholino" based anti-sense RNA oligonucleotides has been found to be resistant to nucleases, has predictable targeting, excellent sequence specificity and minimal non-antisense activity. (Reviewed in Summerton and Weller, 1997). The advances in antisense RNA oligonucleotide design make it an attractive tool for blocking mRNA translation especially in tissue culture but has also been shown to work effectively in *Xenopus laevis* and *Dario rerio*.

An investigation into the functionality of the brain *EIF2B4* alternative splice forms could be carried out by cloning the individual splice forms into a protein expression vector and co-expressing each splice form with the other four eIF2B subunits. Comparison of the guanine nucleotide exchange activity each eIF2B protein complex containing the different brain *EIF2B4* splice forms could give an indication of the difference in functionality of each brain *EIF2B4* splice form.

To assess whether eIF2B4 is a legitimate candidate type 2 diabetes gene, a genetic screen for polymorphisms and mutations in the *EIF2B4* gene in individuals with type 2 diabetes should be carried out. Any polymorphisms or mutations identified could be introduced into recombinant EIF2B4 protein and the effect examined on eIF2B activity in response to presence and absence of glucose. This experiment would utilise the same protein expression system described for the investigation of the function of *EIF2B4* brain splice forms.

With the identification of the *KIAA0064* transcript adjacent to the *EIF2B4* gene in a intimate "head to head arrangement", these two genes could share promoter elements. Although previous work suggests KIAA0064 to be ubiquitously expressed (Nomura *et al*, 1994), it would be of interest to investigate whether the expression of *KIAA0064* and *EIF2B4* are in some way linked, for example, by up-regulation through glucose presence. This experiment could be conducted in tissue culture cells, for example the rat insulin secreting (INS-1) cell line which was established from cells isolated from an x-ray-induced rat transplantable insulinoma and which behave similarly to pancreatic β-cells (Asfari *et al*, 1992). It would also be of interest to perform *in situ* hybridisation experiments for both *EIF2B4* and *KIAA0064* to investigate whether these genes share any subcellular localisation. As KIAA0064 has no known function, a yeast two hybrid experiment could be conducted to

isolate any proteins that interact with KIAA0064, which may provide clues as to the function of KIAA0064 itself.

The ribokinase gene (*RBSK*) was identified as a transcript on chromosome 2p23.3 for further investigation because RBSK belongs to the same kinase family as KHK (see Chapter 4, Section 4.4). If evidence could be gathered that suggested *RBSK* and *KHK* to have evolved from a common ancestral gene, it would provide an interesting insight to the evolutionary history of the genomic region on chromosome 2p23.3. Although the peptide sequences of both RBSK and KHK do show similarity, investigation and comparison of the genomic structures for the *RBSK* and *KHK* genes revealed distinct genomic structures. This suggests that the *RBSK* and *KHK* genes had not evolved from a common ancestral gene. Another indication as to whether *RBSK* and *KHK* may have evolved from a common ancestral gene could be gained by performing X-ray crystallography on purified RBSK and KHK protein. This would allow the comparison of the tertiary protein structures for RBSK and KHK and give an alternative insight into the evolutionary history of the genes encoding these two proteins.

6.2.4 Candidate *DFNB9* genes

The search for candidate deafness genes located in the *DFNB9* interval identified three genes: *MPV17*, *KIF3C*, and *KCNK3* (see Chapter 5), before the actual *DFNB9* gene was identified to be Otoferlin (*OTOF*) by Yasuga *et al.*, 1999. *MPV17* was chosen for further investigation because the *Mpv17* (-/-) mouse has abnormalities in the inner ear similar to Alport's syndrome and is a mouse model for both deafness and renal disease (Meyer zum Gottesberge, 1996). Although mutation screening of *MPV17* revealed no mutations in two individuals with deafness and glomerulosclerosis, a much larger number of patient samples is required for mutation screening to give a better indication as to whether any mutations exist in this gene that cause deafness and/or renal disorder.

The *KIF3C* was chosen for further investigation because it is an intracellular motor protein, having a similar function to myosin proteins which have a proven role in the pathogenesis of deafness (Hasson, 1997). Expression analysis of *KIF3C* revealed a brain specific 5 kb transcript suggesting a intraneural function for KIF3C. To further investigate the function of KIF3C, an interesting experiment would be to identify the cargo that is attached to KIF3C. This could be carried out by immunoprecipitation experiments in which antibodies raised against KIF3C are used to pull down KIF3C and its cargo. Proteins found to be associated

with KIF3C could be separated by two dimensional electrophoresis and micro-sequenced to reveal the protein's identity.

The *KCNK3* gene, encoding a potassium channel protein, was mapped to the DFNB9 interval by physical and radiation hybrid mapping. This *KCNK3* gene was a good candidate deafness gene because of the high concentration of potassium ions in the endolymph of the inner ear and the proven role of other genes encoding potassium ion channels in the pathogenesis of deafness (Neyroud, 1997). Although this gene was not further investigated since the *DFNB9* gene had been identified as *OTOF*, the proven role of potassium channel genes in diseases such as long QT syndrome, Jervell and Lange-Nielson syndrome, and autosomal dominant deafness (DFNA2), makes *KCNK3* an interesting gene for further investigation. Any polymorphisms in *KCNK3* identified by genetic screening of the general population could be introduced into recombinant KCNK3 proteins and the effect on KCNK3 function investigated by use of the patch clamp technique (Neher *et al*, 1978). The patch clamp technique can measure the current flowing through a single ion channel and give an indication to when the ion channel is open or closed, and if mutations affect the normal function of the ion channel.

The *DFNB9* gene was identified as otoferlin (Yasunaga *et al*, 1999), a gene encoding a protein of unknown function. However OTOF does show similarity to the *C. elegans* FER-1 protein (see Chapter 5, Section 5.6). Clues to the function of OTOF have been drawn from the *C. elegans Fer-1* mutants, which show defects in the process of vesicle fusion with the spermatid plasma membrane (Achanzer & Ward, 1997). As OTOF is a C2 domain-containing protein (the C2 domain often interacts with phospholipids and other proteins, and has also been shown to bind Ca^{2+} ions), it has been hypothesised that OTOF is involved in Ca^{2} ion triggered vesicle membrane fusions.

Mutations in *OTOF* underlie neurosensory non-syndromic recessive deafness 9 (*DFNB9*) in humans. Therefore it would be of much interest to understand the function of this protein, the role it plays in the inner ear, and how mutations in this gene are involved in the pathogenesis of deafness. An insight into the pathogenic mechanism of deafness caused by mutations in *OTOF* might be gained by use of a model organism such as the mouse. Antibodies to OTOF could be used in *in situ* hybridisation and immunocytochemistry experiments, carried out using inner ear tissue from normal mice to determine when and where *Otof* and its encoded protein are expressed. The production of a mouse *Otof* knock-out would allow further insight into Otof function. Firstly, the auditory function of *Otof* (-/-)

mice can be examined which should confirm that mutations in *Otof* can cause deafness. An *Otof* knock-out mouse would facilitate the investigation of what actually happens *in-vivo* when no Otof protein is present. Electron microscopy of mouse inner ear tissue from the mutant mouse should indicate whether Otof is actually involved in the process of vesicle fusion and if it is, an investigation into the vesicle components could be carried out.

6.3 Impact of the Human Genome Project and technological advances on functional genomics

6.3.1 Use of the Human Genome Sequence

The first draft human genome sequence has now been deposited in the sequence databases and the expected date for completion of the entire human genome is 2003. With over 30 fully sequenced genomes already in the public domain, the emphasis of genome research is now shifting from description of genome sequence to assignment of gene function and how the proteins encoded by genes interact, for example in the organisation and control of genetic pathways that come together to make up the physiology of an organism. This section discusses how the Human Genome Project, in addition to advances in new technologies for investigation of gene function that are available to the research scientist, has changed the methodology for the type of project that is described in this thesis.

This thesis describes the characterisation of the genomic region surrounding the genes *GCKR* and *KHK* on chromosome 2p23.3. This was carried out by creating a physical contig of genomic clones spanning the genomic region of interest and identifying transcripts by sequence analysis and PCR screening for EST sequences. This is a time consuming and labour intensive process. Therefore one obvious advantage of the Human Genome Project is the availability of sequence data for the whole genome. This means that once a genomic region of interest has been identified, the complete annotated DNA sequence for that region will be available in the sequence database. With ESTs and transcripts already mapped to the DNA sequence, the investigator can immediately make a decision on what genes require further investigation.

The sequencing of many genomes from different organisms has facilitated the development of computational methods for establishing functional linkages between proteins. One such method is the production of a phylogenetic profile for a particular gene. The presence of a gene sequence can be searched for in genome sequences from different organisms and the

presence or absence of that gene sequence is used to produce a phylogenetic profile. If two genes are found to have identical phylogenetic profiles, their encoded proteins might have a functional link and could be engaged in a common pathway or complex. As the number of genomes sequenced from different organisms increases, the phylogenetic profile method will become more accurate for establishing functional linkages.

The mapping of transcripts to chromosome 2p23.3 identified 14 known genes and 15 EST sequences. The examination of putative functions of the encoded proteins identified no obvious functional links for these transcripts. Therefore, the method of phylogenetic profiling for the genes mapping to the *GCKR-KHK* genomic region might identify functional links between the encoded proteins that were not apparent by homology searching.

Another computational method that can reveal functional linkages from genome sequences is the gene neighbour method (Overbeek *et al*, 1999). As more complete genome sequences are deposited in databases, genes that lie in regions of conserved synteny in many different organisms might indicate functional linkages between the encoded proteins. This method has proven very successful in prokaryote genomes, where operons are common, and may also be applicable to eukaryote genomes. As described in Chapter 2, both *GCKR* and *KHK* were both mapped to the same region of mouse chromosome 5. If *GCKR* and *KHK* are shown to be adjacent to each other in other genomes, this might suggest a functional link between their encoded proteins.

6.3.2 DNA arrays

Computer analysis of genome sequences can only infer a functional linkage between two proteins and it should be pointed out that computer predictions should always be backed up with genetic and biochemical experimental evidence. One way to gain an insight into the function of a protein is to examine when, where and to what extent its encoding gene is expressed. Whereas in the past, the expression of a gene would be examined by RT-PCR or Northern blotting, the recent advances in DNA array technology now allow expression analysis to be carried out on a genome-wide scale. Significant numbers of arrays can now be made which have 250,000 different oligonucleotide probes or 10,000 different cDNAs per square centimetre (Lipshutz *et al*, 1999). The small size of an array means that only a small amount of probe, for example mRNA from a specific cell type, is required to perform the screen. A great advantage of DNA arrays is the ability to monitor the expression of thousands of known and novel genes in response to altering environmental stimuli,

development, and in disease. In this way, links between genes can be found that would never have been found by analysing the expression of genes one at a time. In addition, analysis of the multi-gene expression patterns can provide clues about regulatory mechanisms, broader cellular functions, and biochemical pathways.

6.3.3 Proteomics

The investigation of gene expression can be highly informative about cell state and the activity of genes, but mRNA is only an intermediate on the way to a functional protein product. Although in many cases, gene expression correlates with the translation of mRNA into a functional protein, in the functional genomic era, it is important to investigate where a protein is expressed, its abundance and whether it is functional.

Although the analysis of proteins has proven more difficult, less sensitive, and lower throughput than RNA-based methods, recent technological advances have allowed the large scale analysis of proteins. This field which has been termed "proteomics", can be divided into three main areas: 1) the characterisation of proteins and their post-translational modifications, 2) "differential display" proteomics for the comparison of protein levels with the potential application in a wide range of diseases, and 3) studies of protein-protein interactions using techniques such as mass spectrometry (Shevchenko *et al*, 2000) or the yeast two hybrid system (Fields and Song, 1989). The advances in techniques like mass spectrometry of gel purified proteins has revolutionised proteomics. The purification and analysis of proteins and protein complexes, and the identification of interacting proteins will be essential in the ultimate goal of understanding cellular physiology and the pathogenic mechanisms of disease processes.

6.3.4 Identification of genetic determinants

Whereas the identification of single gene genetic disorders has proven fruitful, it has become apparent that methods used to identify genes that underlie polygenic diseases are inadequate. An example of this is the search for genes that underlie type 2 diabetes. Using the positional gene method which utilises linkage analysis for DNA markers with the disease allele in families with the genetic disorder, five genes have so far been identified that underlie the monogenic MODY form of type 2 diabetes (discussed in Chapter 1). However, linkage analysis of polygenic disorders have so far been largely unsuccessful. Genome-wide screens to identify markers associated with the disease allele in sib pairs have only identified three

loci for late-onset type 2 diabetes. This lack of success has been attributed to the polygenic nature of the disease and also to different factors having more important affects in different ethnic populations.

One method that has offered hope for identifying the genetic determinants in polygenic disorders has been the identification and utilisation of single nucleotide polymorphisms (SNPs). SNPs are present throughout the human genome with an average frequency of approximately 1 per 1000 bp (Brookes, 1999). Much effort is being put into the identification of SNPs, with the SNP consortium aiming for a target of over 300 000 SNPs. The resultant SNP map will enable disease genes to be mapped by linkage disequilibrium (Kruglyak, 1998). This works on the basis that linkage disequilibrium occurs when haplotype combinations of alleles at different loci occur more frequently than would be expected from random association; the linkage disequilibrium will decay over generations in proportion to the recombination fraction between the loci. Therefore, consecutive SNP variations that are in linkage disequilibrium and associated with a disease phenotype can "mark" the position on the chromosome where a susceptibility gene is located.

Once enough SNPs have been identified and a cost effective high throughput method for SNP screening has been obtained, the accurate assignment of chromosomal regions containing susceptibility genes for polygenic disorders can be carried out (Zhao *et al*, 1998). However, until a high density SNP map (one SNP every 10 kb) is available, the time consuming and expensive individual testing of multiple candidate genes found to be located within a linkage interval for a disease will continue. The benefits of SNP mapping can be demonstrated using as an example the search for the gene underlying non-syndromic sensorineural deafness *DFNB9* (see chapter 5). The *DFNB9* gene had been mapped by linkage to a 2cM region of chromosome 2p23.3 (Chaib *et al*, 1996) that was shown to contain a large number of transcripts. To identify the DFNB9 gene, candidates had to be chosen for mutation analysis based on expression analysis and putative function of the encoded protein. This is a very time consuming and haphazard method. The DFNB9 gene was eventually identified as encoding a novel protein called *OTOF* (Yasunaga *et al*, 1999) but many other candidate gene had to be examined before its identification. This time consuming process could have been much shorter if SNP mapping had been available. SNP mapping in the 2 cM *DFNB9* interval in the *DFNB9* families would have identified a much smaller genomic region showing SNP linkage disequilibrium associated with the *DFNB9* phenotype. The genomic region could be relatively quickly sequenced and a much smaller

number of candidate *DFNB9* genes identified by sequence analysis. Therefore SNP mapping would have been much less time consuming and more cost effective.

Looking to the future, SNP mapping might have an important role in pharmacogenetics. This is the study of how genetic differences influence the variability in patients responses to drugs. An SNP profile might correlate to a good or bad response to a drug for a certain disease (McCarthy and Hilfiker, 2000). By identifying patients who will respond to drug treatments will provide better treatments for the patient in addition to much cost saving on wasted treatments.

6.4 Summary

As discussed in this chapter, the progress of the Human Genome Project has resulted in greater research emphasis on functional genomics. With the recent technological advances in the field of molecular biology and proteomics for example DNA arrays and mass spectrometry, in projects such as described in this thesis, a much more in depth analysis of gene function, protein function, and *in vivo* protein interactions can now be performed. The use of SNP maps and linkage disequilibrium should allow genes underlying both monogenic and polygenic disorders to be elucidated, for example in the identification of type 2 diabetes DFNB9 loci. Functional genomics will play an essential role in working out the pathogenic mechanisms of these genetic diseases and in the long term, identify targets for disease treatment by both drug and gene therapy.

Chapter 7

7 Materials and Methods

7.1 Materials

7.1.1 Chemicals and reagents

All chemicals for general use were supplied by Sigma (Sigma-Aldrich Company Ltd) or BDH (Merck Ltd) unless otherwise stated, and were of molecular biology grade. Bacterial media were supplied by Difco U.K. Ltd.

7.1.2 Radiochemicals

The radioactive isotopes [α^{32}P] dCTP at a specific activity of 3000 Ci/mmol and [γ^{32}P] ATP at a specific activity of 1415 Ci/mmol were supplied by Amersham Life Science. The four Redivue™ ^{33}P-labelled dideoxynucleotide terminators were also supplied by Amersham Life Science.

7.1.3 Enzymes

Restriction enzymes, polynucleotide kinase and T4 DNA ligase were supplied either by Gibco BRL Life Technologies or New England Biolabs. Klenow fragment of *E. coli* DNA polymerase I was provided with the Rediprime™ DNA labelling kit (Amersham Life Science). Desiccated proteinase K and RNase A were supplied by Sigma. *Taq* DNA polymerase was supplied at a concentration of 5 U/µl by Gibco BRL Life Technologies. Thermo Sequenase™ DNA polymerase (a site-directed mutant of *Taq* polymerase with improved affinity for dideoxynucleotides) was included with the Thermo Sequenase™ radiolabelled terminator cycle sequencing kit (Amersham Life Science) and AmpliTaq™ DNA polymerase was supplied with the ABI PRISM® BigDyeTM terminator cycle sequencing kit (PE Applied Biosystems).

7.1.4 Nucleic acids, vectors and markers

Salmon sperm DNA was supplied by Sigma Chemical Company Ltd. pBluescript™ was purchased from Stratagene. Subcloning-Efficiency™ DH5α™ competent *E. coli* cells were

purchased from Gibco BRL Life Technologies. The 1 kb ladder DNA marker was obtained from Gibco BRL Life Technologies.

7.1.5 Electrophoretic and DNA transfer materials

Standard grade agarose was supplied by Sigma Chemical Company Ltd. Acrylamide / bis-acrylamide in proportions 19:1 were supplied as a 40% solution by BDH chemicals. Hybond-N and HybondN$^+$ hybridisation membranes were supplied by Amersham Life Science. Hyperfilm MP™ X-ray film was supplied by Amersham Life Science.

7.1.6 Solutions and buffers

Unless otherwise stated, solutions and buffers were prepared using distilled and deionised water and were stored at room temperature (i.e. between 15 and 25°C). Sterilisation was by autoclaving at 15 psi 121°C (30 min). The components of general solutions and buffers are presented in appendix 6.A.1.

7.1.7 Fluorescent *in situ* hybridisation (FISH) materials and solutions

All solutions for fluorescent *in situ* hybridisation (FISH) were prepared using distilled and deionised water and stored at room temperature (unless otherwise stated). For YAC probes, the clone was propagated on synthetic dextrose (SD) medium to obtain a pure, single colony which was grown to saturation in SD broth. A total yeast DNA extract was obtained using standard techniques. The probe was labelled by nick-translation with digoxigen-11-dUTP and 200 ng was used for FISH analysis as described by (Pinkel *et al.*, 1986). Chromosomes were identified with 4,6-diamidino-2-phenylindole-dihydrochloride (DAPI). Microscopy was performed using a Zeiss Axioskop fluorescence microscope coupled to a CCD camera and image analysis system (Vysis, UK).

FISH analysis was carried out by Dr Judy Fantes and Dr Jack Leek.

7.2 Standard Methods

This section details the generalised methods for the molecular biology techniques used as part of the work presented in this thesis. More specialised methodologies are described in the chapters of this thesis where they have been used to generate the data described in those sections.

7.2.1 Preparation of closed circle DNA from bacteria

The general method of alkaline lysis (Birnboim & Doly, 1979) was used for the preparation of plasmid DNA (including cosmids, PACs, and P1 clones) from bacteria. For small scale preparation (mini-prep) of plasmid and cosmid DNA, the following method was used:

A single colony was used to inoculate 10 ml of LB-broth (see appendix 6.A.2), containing the appropriate antibiotics (appendix 6.A.3). This culture was incubated overnight at 37°C with vigorous shaking (225 rpm). The cells were pelleted by centrifugation at 13000 *g* (5 min) and resuspended in solution I (100 µl) to create a completely homogenous suspension. Freshly made solution II (200 µl) was added and the mixture gently mixed by swirling. Solution III (150 µl) was added and the mixture shaken vigorously. The lysate was spun at 13000 *g* (10 min), and the supernatant removed. An equal volume of saturated phenol/chloroform was added and the tube shaken vigorously (1 min). The phases were separated by centrifugation at 13000 *g* (5 min), and the bottom layer discarded. This process was repeated once with saturated phenol/chloroform and again with chloroform. The DNA was recovered by addition of absolute ethanol (3 vol.) and centrifugation at 13000 *g* (20 min, room temperature). The supernatant was discarded and the DNA pellet washed with 70% ethanol. The pellet was dissolved in TE (50 µl, pH8.0) and stored at -20°C.

The quality of DNA produced by this method was adequate for sequencing using the ^{33}P cycle sequencing method (Amersham, see section 6.2.17). However, it was found that for automated cycle sequencing, use of the S.N.A.P.™ (a *s*imple *n*ucleic *a*cid *p*rep) kit (Invitrogen) gave better quality DNA as well as being a quicker method when handling large numbers of samples. For large scale preparation of plasmid and cosmid DNA, the Qiagen plasmid maxi kit was used according to supplied instructions.

For the isolation of high quality PAC DNA, a modified version of the alkaline lysis method was used:

A single colony was used to inoculate LB (5 ml) containing kanamycin (25 μg/ml) and incubated at 37°C for 4-7 hr with vigorous shaking (225 rpm). The cells were pelleted by centrifugation at 6000 *g* (20 min) and the supernatant discarded. The pellet was resuspended in Solution I (5 ml) and left at room temperature (5 min). The cells were lysed by addition of Solution II (5 ml) with gentle mixing and incubated at room temperature (10 min). Solution III (5 ml) was added and mixed before placing on ice (10 min). The cell debris was removed by centrifugation at 13000 *g* (20 min, 4°C) in Corex™ tubes. The supernatant was filtered through cheese cloth into fresh Corex tubes and ice cold isopropanol (0.6 vol.) added. After mixing well, the mixture was centrifuged at 13000 *g* (20 min). The supernatant was poured off and the pellet washed with 70% ethanol. After air drying, the pellet was resuspended in TE (1.5 ml, pH8.0). Ice cold 5 M LiCl (1.5 ml) was added and after mixing, the mixture was centrifuged at 13000 *g* (15 min, 4°C). The supernatant was transferred into fresh Corex tubes and an equal volume of isopropanol added. After mixing, the mixture was centrifuged at 13000 *g* (20 min) and the supernatant discarded before washing the pellet with 70% ethanol. After air drying, the pellet was resuspended in TE (250 μl, pH8.0) containing RNAse A (20 μg/ml) and incubated at room temperature (30 min). To this solution, 1.6 M NaCl containing 13% (w/v) polyethylene glycol (250 μl) was added and after mixing, the PAC DNA recovered by centrifugation at 13000 *g* (15 min, 4°C). The supernatant was removed by aspiration and the pellet resuspended in TE (200 μl, pH8.0).

To determine the concentration of the DNA, the absorption at 260 nm (A_{260}) was measured. The spectrophotometer was first blanked using TE before the plasmid DNA sample (5-10 μl of the miniprep) absorbance was measured. The formula below was used to estimate the DNA concentration:

$$[DNA] = (A_{260})(0.05 \text{ mg/ml}) \times D$$

where D is the dilution factor.

7.2.2 Preparation of RNA from tissue culture cells - Guanidinium Thiocyanate Method

The medium was removed from the culture dish and 0.5 ml of solution D (see Appendix A) was added directly to the dish before scraping the cells into 2 ml tubes. (At this point the cells can be frozen at -70°C if necessary). The cells were homogenised by passing the lysate 10 times through a 1 ml pipette tip. The following solutions were sequentially added: 2 M

sodium acetate (0.05 ml), phenol (0.5 ml), chloroform isoamyl alcohol mixture 24:1 ratio (0.1 ml). The mixture was mixed thoroughly by inversion after addition of each reagent. After shaking vigorously for 10 s, the mixture was cooled on ice (15 min) before centrifuging at 10,000 *g* at 4°C (20 min). The aqueous phase (containing the RNA) was transferred to a fresh tube and mixed with isopropanol (0.5 ml) and cooled at -20°C for at least 1 hr. After centrifuging at 10,000 *g* (20 min), the supernatant was removed and the pellet redissolved in solution D (0.3 ml) and transferred to a 1.5 ml Eppendorf tube. The RNA was precipitated with 1 volume of isopropanol and cooled at -20°C (1 hr) before centrifuging at 4°C (10 min). The pellet was resuspended in 75% ethanol and centrifuged. The pellet was vacuum dried (15 min) and dissolved in 0.5% SDS (50 µl) before incubating at 65°C (10 min). The purified RNA was stored at -70°C.

7.2.3 The polymerase chain reaction (PCR)

7.2.3.1 Oligonucleotides

Oligonucleotide primers were generally designed by selecting a sequence of about 20 nucleotides, with approximately equal [G+C] to [A+T] content, taking care to ensure that the two primers' 3′ ends were not complementary. The primers were checked using the "PRIMER" computer programme (PRIMER is copyright 1991 by The Whitehead Institute for Biomedical Research) to ensure that they were not chosen from sequences which are likely to be repetitive, and that they have approximately equal predicted melting temperatures. Oligonucleotides were purchased from Sigma-Genosys and were supplied as lyophilised products at a nominal synthesis scale of 0.03 µmol. The oligonucleotide sequences used and their respective annealing temperatures are described in their respective chapters.

7.2.3.2 The polymerase chain reaction (PCR)

The PCR reactions were carried out in a Hybaid Omnigene™ Thermal Cycler. The Gilson pipettes, tips, and 0.5 ml PCR tubes were kept separate from those used in other experiments. Reactions were also set up in a separate room to the one used for electrophoresis of PCR products. These precautions minimised the possibility of extraneous DNA contamination. PCR was carried out in 1 x PCR buffer, with 200 µM deoxynucleoside triphosphates (dNTPs), 15 pmol primers, template DNA (0.1 µg genomic DNA, 0.1-1 ng plasmid DNA), and a concentration of Mg^{2+} particular to each set of primers (usually 1-3 mM) in a volume of 50 µl. After a denaturation step of 95°C (5 min) and the addition of 0.25 units *Taq*

polymerase, the PCR conditions were 94°C for 1 min, the appropriate annealing temperature for 1 min, and 72°C for 1 min (if product is <500 bp), 3 min (if product is >500 bp). The number of cycles depended on the type of DNA template: 25 cycles for plasmid, 30 cycles for cosmid, 35 cycles for PAC, and 40 cycles for YAC and human genomic DNA. Reaction mixes were overlaid with 50-75 µl mineral oil. The annealing temperatures used are described along with oligonucleotide sequences in the respective chapters of this thesis.

7.2.3.3 RT-PCR and 5′ RACE

For RT-PCR, the GibcoBRL RT-PCR kit was used.

In chapter 5, the method of 5′ RACE (Frohman *et al.*, 1988) was used to generate cDNA sequence for the 5′ end of *KIF3C*. Briefly, 5 µg of total RNA from human fibroblasts was mixed with the reverse *KIF3C*-specific primer (5′-dGGCCCCAGTGTGGCTACC-3′; Figure 5.6, nucleotides 1167-1184) and reverse transcribed with 200 U Superscript II™ reverse transcriptase (GibcoBRL) at 42°C for 50 minutes. A poly(dA) tail was added to the 3′-end of the cDNA and tailed cDNA was amplified with a $(dT)_n$-adapter primer and the reverse primer (5′-dGTGTGGCTACCATGATGGTC-3′; Figure 5.6, nucleotides 1158-1177). The products of 5′ RACE reactions were analysed by Southern blot hybridisation with ^{32}P-labelled primer (5′-dGGAGAGAGGCCTAAGGAAG-3′; Figure 5.6, nucleotides 1003-1021) and an 820 bp product was cloned into the TA vector (Invitrogen) and sequenced with vector- and kinesin-specific oligonucleotide primers. Multiple PCR-derived cDNA clones were used to check for errors in the nucleotide sequence produced by amplification procedures.

7.2.3.3.1 First strand cDNA synthesis

In a 30 µl reaction volume, the following components were added to a nuclease-free microcentrifuge tube: 5 µl RACE 2 primer (3 µM), 1-5 µg total RNA, sterile, distilled water to 12 µl. The mixture was heated at 80°C (3 min) and chilled quickly on ice. The contents of the tube were collected by brief centrifugation before adding: 6 µl 5 x first strand buffer (250 mM Tris-HCl, pH 8.3, 375 mM KCl, 15 mM $MgCl_2$), 3 µl DTT (0.1 M), 3 µl 20 mM dNTP mix (20 mM each dATP, dGTP, dCTP and dTTP at neutral pH).

The contents of the tube were mixed gently and incubated at 42°C (60 min) before adding 1 µl (200 units) of Superscript II and mixing by pipetting gently up and down. After

incubation at 25°C (10 min), the mixture was incubated at 42°C (50 min) before inactivating the reaction by heating at 70°C (15 min).

NOTE: The cDNA can now be used as a template for amplification in PCR. However, amplification of some PCR targets (those >1 kb) may require the removal of RNA complementary to the cDNA. To remove RNA complementary to the cDNA, 1 µl (2 units) of *E. coli* RNase H is added, followed by incubation at 37°C for 20 min.

7.2.3.3.2 PCR Reaction

10% of the first strand reaction was used for PCR (adding larger amounts of the first strand reaction may not increase amplification and may result in decreased amounts of PCR product). The PCR reaction was set up as described in section 6.2.3.2 and PCR performed by following the PCR program: 5 min, 94°C, 40 x (94°C, 1 min; X°C, 1 min; 72°C, 2 min) – where X is the annealing temperature for the specific primer pair used in the reaction.

7.2.4 Restriction enzyme digestion of DNA

Restriction endonuclease digestions were performed under the appropriate conditions as recommended by the manufacturers, and with the buffers supplied with the enzymes. In general, 5 µg of genomic DNA was mixed with 4µl 10x restriction enzyme buffer and 30 units of enzyme, in a total volume of 40 µl (made up with sterile water) and incubated (5-16 hr) at the optimal temperature for the enzyme. Restriction enzyme digestion of plasmid, cosmid and PAC DNA was generally carried out using 2-5 µg of DNA in a volume of 20 µl. If the DNA was to be used for another manipulation, the enzyme was heat inactivated (if heat labile) at 65°C (15 min) or phenol/chloroform extracted and ethanol precipitated.

7.2.5 Agarose gel electrophoresis

Digested DNA was size-fractionated by agarose gel electrophoresis with gels prepared to concentrations of between 0.8 and 1.5% w/v agarose (unless stated otherwise) in the appropriate volume of 1 x TAE or 0.5 x TBE buffer (Appendix 6.A.1). Prior to electrophoresis, DNA samples were combined with one-tenth volume DNA loading buffer. Electrophoresis was performed at 50-150 V until dye markers had migrated an appropriate distance, depending on the size of DNA to be visualised. Estimation of the size of fractionated DNA was achieved by running 0.5 µg of 1 kb ladder (Gibco BRL Life

Technologies) alongside the DNA samples. DNA was visualised by viewing under ultraviolet (UV) transillumination and photographed using an electronic imager (Appligene).

7.2.6 Phenol/chloroform extraction of DNA

An equal volume of buffer-saturated phenol/chloroform/isoamyl alcohol (25:24:1) was added to the DNA solution. After mixing well (most DNA solutions can be vortexed for 10 s, except for high molecular weight DNA, which should be gently rocked), the mixture was microcentrifuged (3 min). The aqueous layer was carefully removed (being careful to avoid the interface) and placed in a new tube. This phenol/chloroform extraction can be repeated until an interface is no longer visible. To remove traces of phenol, an equal volume of chloroform was added to the aqueous layer and the mixture microcentrifuged (3 min). The aqueous layer was removed placed in new tube before ethanol precipitating the DNA.

7.2.7 Ethanol precipitation of DNA

The volume of the DNA sample was measured and the salt concentration adjusted by adding 1/10 volume of sodium acetate (3 M; pH 5.2), or an equal volume of ammonium acetate (5 M). After mixing well, 2 to 2.5 volumes of cold 100% ethanol (calculated after salt addition) was added and again mixed well. If small amounts of DNA were to be recovered, 5-10 μg tRNA was added to increase recovery from dilute DNA solutions. The mixture was placed on ice or at -20°C for >20 minutes. After spinning at maximum speed in a microfuge (10-15 min), the supernatant was carefully decanted off. The pellet was washed in 70% ethanol (1 ml) and spun briefly. The supernatant was carefully decanted off before air drying or briefly vacuum drying the pellet and resuspending in the appropriate volume of TE (pH8.0) or water.

7.2.8 Method for purifying DNA from agarose gels

The DNA band of interest was excised from a TAE gel and its volume estimated by weighing the gel slice. After dissolving the gel slice in 3 volumes of NaI/Na_2SO_3 solution by heating at 65°C for 5 min or until the gel dissolved, 8 μl of glass milk (thoroughly vortex glass milk before use) was added and the tube placed on ice (15 min). The mixture was microcentrifuged and the pellet washed with ice cold ethanol (200 μl). The mixture was again microcentrifuged and the supernatant discarded. This washing process was repeated a further two times before resuspending the glass milk in 20 μl TE. After heating at 50°C (10

min), the mixture was microcentrifuged at full speed (30 sec) and TE removed (now containing eluted DNA). This elution process was repeated by addition of a further 20 µl of TE, heating at 50°C (10 min) and microcentrifuging at full speed (30 sec).

7.2.9 Subcloning

7.2.9.1 Restriction digestion and dephosphorylation of DNA

Restriction enzyme digestion of plasmid DNA was carried out essentially as described in section 6.2.4, using 2-5 µg of DNA in a reaction volume of 20 µl. The terminal phosphate of linearised vector DNA was removed using calf intestinal phosphatase (CIP) to prevent self-ligation of the ends in subsequent cloning experiments. The vector DNA (2-5 µg of pBluescript™; Pharmacia) was digested with the restriction enzyme of choice in the appropriate buffer for the enzyme. CIP (0.2 units) was added and incubated at 37°C (1 hr) after which the CIP was inactivated by heating at 85°C (15 min). The CIP was removed by phenol / chloroform extraction and the DNA was recovered by ethanol precipitation. The pellet was washed in 70% ethanol, dried and resuspended in a suitable volume of TE.

7.2.9.2 Ligation of DNA

The insert DNA was digested with the appropriate restriction enzyme and ligated to the appropriate cut and phosphatased vector in a volume of 20 µl. A three fold molar excess of insert DNA : vector DNA was used for most cloning experiments. Reactions were performed in 1 x T4 DNA ligase buffer and 1 unit of T4 DNA ligase at 14°C overnight.

7.2.9.3 Preparation of competent *E. coli*

Competent cells were prepared using the calcium chloride procedure as described by Sambrook *et al.* (1989). The host cells used in the cloning process were the *E. coli* strain DH5α. These cells are a restriction-deficient strain of *E.coli*, which are also endonuclease deficient and recombination deficient, ensuring stability of any inserts cloned within the system. This bacterial strain contains the F′ episome and also a defective copy of the *lacZ* gene. The latter can be activated to generate an active β-galactosidase enzyme by α-complementation from a peptide encoded on many common cloning vectors. These plasmid vectors' cloning sites interrupt the peptide coding region, so that α–complementation is lost by ligation of an insert. On a medium containing the chromogenic lactose analogue 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal),

and the *lac* operon inducer isopropyl-β-D-galactopyranoside (IPTG), colonies carrying uninterrupted plasmids are blue. Most insert-bearing plasmids fail to complement the defective *lacZ* and are white (or pale blue).

7.2.9.4 Transformation of competent *E. coli*

For transformation, generally up to 10 µl ligation reaction (50-100 ng) was added to a 200 µl aliquot of competent cells, cooled on ice (40 min), heat shocked at 42°C (90 sec) and cooled on ice (2 min). LB-medium (500 µl) without antibiotic was added to the transformation reaction and the mixture incubated at 37°C (1 hr), to allow for pre-expression of the antibiotic resistance. The transformation mix was then plated onto LB-agar plates with the appropriate antibiotic and incubated at 37°C overnight. When using the blue/white selection screening, 50 µl of X-Gal (2% solution in dimethylformamide) and 20 µl of 100 mM IPTG were added to the LB-agar before pouring the plates. The transformation efficiency was tested by transforming with 10 ng of undigested vector DNA. The efficiency of the CIP reaction was examined by transforming with a control ligation performed using dephosphorylated vector without insert DNA.

7.2.9.5 Electrotransformation of competent *E. coli*

For transformation, generally 1-2 µl ligation reaction (50-100 ng) was added to a 40 µl aliquot of competent cells in an electroporation cuvette. The Gene Pulser apparatus (Bio-Rad) was set to 25µF, 2.5kV, 200Ω. The cells were pulsed to a time constant of 4.5-5 msec and immediately resuspended in 1ml of ice-cold LB-medium. The cell suspension was transferred to a 17 x 100 mm polypropylene tube and incubated at 37°C, 225rpm for 1 hour. The transformation mix was subsequently treated according to the method described above.

7.2.10 Cosmid Fingerprinting

Fluorescent fingerprinting was performed using the method described by Carrano *et al.*, 1989. A single reaction mixture was used to restriction digest the cosmid DNA and ligate fluorescent primers to the restriction fragment ends (see Chapter 3, Section 3.2.2). The mixture consisted of: 1 µg cosmid DNA, 1 µl fluorescent dye labelled universal primer (ratio of primer to insert ends, 0.5-2), 1 µl complementary synthetic oligonucleotide (at the same ratio as the dye primer), 10 units of restriction enzyme, 1µl 10x restriction enzyme buffer, 2 µl T4 DNA ligase (0.5 unit/µl), 1 µl 8.8 mM ATP, 1.4 µl 100mM DTT, and water to bring

the total volume to 10 µl. The mixture was incubated at 37°C overnight (12-14 hr) before addition of 1 µl *Hin*fI restriction enzyme (40 units/µl) and a further 4 hr incubation at 37°C. Each tube was cooled on ice, and 0.5 µl of 500 mM EDTA added. The sample was precipitated with ammonium acetate (0.75 M), glycogen (1µl of 20 mg/ml stock, and 2.5 vol. 95% ethanol). The precipitate was washed in ethanol and resuspended in 4 µl of TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 7.4). After at least 60 min of resuspension, 4 µl of de-ionised formamide and 0.5 µl size standard (GENSCAN TAMRA 500) were added. The DNA was denatured at 95°C for 5 min before loading 2.5 µl into a single lane of a denaturing (8 M urea) polyacrylamide gel and run using the Genscan software package on the ABI 377 Fluorescent DNA sequencer. The cosmid fragments were sized by comparison to the internal lane size standard and the cosmids assembled into a contig according to the number of shared fragments of equal size.

7.2.11 Southern blotting

Transfer of DNA from agarose to nylon membrane was performed according to the method originally described by Southern, 1975. Prior to DNA transfer, the gels were treated in depurination solution (10 min, room temperature) with gentle agitation (this step is necessary if target sequences are greater than 10 kb in size), denaturation solution (25 min, room temperature) with gentle agitation, and neutralisation solution (30 min, room temperature) with gentle agitation. The DNA was transferred from the gel to a Hybond-N^+ nylon membrane using capillary blotting apparatus that consisted of a reservoir of 10 x SSC (the transfer buffer), Whatman 3MM paper wicks and a blotting platform. The DNA was allowed to transfer for at least 16 hr. After transfer, the membrane was rinsed four times with 2 x SSC (5 min). The membrane was either baked at 80°C (2 hr) or UV cross-linked to fix the DNA to the membrane.

7.2.12 Northern blotting

RNA agarose gels must be run in gel tanks specifically for RNA work. The gel tank was washed with 0.1% SDS and rinsed well with Nanopure™ water before using. The RNA gel was prepared by first autoclaving agarose (1.5 g), 20 x MOPS (5 ml), and water (80 ml). After boiling and allowing to cool to ~60°C, formaldehyde (15 ml of a 37% solution) was added (this procedure was carried out in a fume cupboard). The RNA samples were prepared using 10-20 µg RNA = 1 volume + 3 volume loading buffer. After heating the

samples at 65°C (10 min), the samples were loaded onto the gel and electrophoresed at 100 V for 4 hrs in 1 x MOPS buffer. Then the RNA gel was washed in water before washing in 20 x SSC. The gel was blotted onto Hybond-N nylon membrane using 20 x SSC as transfer buffer (the blotting apparatus consisted of the same set-up as described for Southern blotting – section 6.2.11). The blot was washed in 50 mM sodium phosphate (pH 7.2) and baked at 80°C (1 hr). The filter was pre-hybridised in Church buffer for at least 1 hr before hybridising in 10 ml of Church buffer with a labelled probe overnight. After hybridisation, the filter was washed three times (30 min per wash) at 65°C in 50 mM sodium phosphate, pH 7.2 / 1% SDS. The filter was air dried, sealed in a polythene bag and exposed overnight to x-ray film.

To examine expression of the human *KIF3C* gene, a multiple tissue Northern blot (Clontech) was hybridised at 65°C overnight in a solution consisting of 0.5M sodium phosphate, 1% BSA, 7% SDS, 1mM EDTA and 50µg/ml denatured herring sperm DNA. The probe was a ^{32}P-labelled 818 bp PCR product encoding the motor domain (nucleotides 360-1177, Figure 5.6). Following hybridisation, the blot was washed twice in 2 x SSC, 0.1% SDS at 65°C for 40 minutes.

7.2.13 Use of Radiolabelled DNA probes

7.2.13.1 Radiolabelling DNA probes

Probes were labelled by the random hexanucleotide priming method (Feinberg & Vogelstein, 1983) using the Rediprime™ DNA labelling system (Amersham) and with [α^{32}P] dCTP (3000 Ci/mmol) as the radioisotope. The DNA probe (25 ng) was made up to 45 µl volume using distilled deionised water and denatured by boiling (5 min). Klenow enzyme, random 9mer primers and a buffered solution of dATP, dGTP, and dTTP were all supplied in a single tube in a dried stabilised form (1 per reaction) in the labelling kit. Following denaturation of the probe and snap cooling on ice, the probe was centrifuged at 13000 *g* (30 sec) and the DNA added to the tube of labelling mix. Following the addition of 20 µCi [α^{32}P] dCTP (2 µl) the reaction was incubated at 37°C for 30 min and then centrifuged through a Sephadex G-50 column at 1500 *g* (1 min) to remove unincorporated nucleotides.

7.2.13.2 Hybridisation of DNA probes to Hybond N^+ nylon membranes

Hybridisations were performed at 65°C in hybridisation bottles in a volume of 15 ml hybridisation buffer per 20 x 20 cm filter. Hybond N+ nylon membranes were pre-hybridised in hybridisation buffer at 65°C (1 hr). Radiolabelled probes were first denatured along with sonicated salmon sperm DNA (250 µl 10mg/ml solution per 10 ml of hybridisation buffer) by boiling (5 min). Probe and salmon sperm were added directly to the hybridisation bottle containing the membrane and pre-hybridisation buffer, and hybridised at 65°C (16-24 hr).

7.2.13.3 Post-hybridisation washing and radioactive signal detection

Following hybridisation, residual non-specifically bound probe was removed by washing twice in 2 x SSC (10 min, room temperature), followed by two washes in 2 x SSC / 0.1% SDS at 65°C (15 min). When greater stringencies were required, membranes were washed with 0.1 x SSC / 0.1% SDS at 65°C for varying lengths of time. For radioactive signal detection, membranes were exposed to autoradiographic X-ray film (Hyperfilm; Amersham) in light proof cassettes at –70°C with intensifying screens for between 16 hours and 14 days.

7.2.13.4 Removal of radiolabelled probe

To remove radiolabelled probes for subsequent hybridisation experiments, membranes were submerged in boiling 0.1 x SSC / 0.1% SDS in a plastic sandwich box, and left to cool to room temperature. Membranes were exposed to autoradiographic film as previously described to ensure complete removal of probe.

7.2.14 Sequencing

7.2.14.1 Sequencing kits

Sequencing of DNA was performed using either the Thermo Sequenase™ radiolabelled terminator cycle sequencing protocal (Amersham Life Science) or the BigDye™ terminator cycle sequencing protocol (PE Applied Biosystems). It was found that the radioactive Thermo Sequenase protocol worked well for DNA templates such as PCR products, plasmid, cosmid and PAC, but that the automated BigDye™ terminator cycle sequencing protocol worked well only for only PCR products and plasmid template. However, the automated fluorescent method was more efficient and less time consuming for sequencing large numbers of samples.

7.2.14.2 Thermo Sequenase radiolabelled terminator cycle sequencing

7.2.14.2.1 Method for sequencing clones or plasmid

One reaction mixture was prepared for each sequencing primer consisting of: reaction buffer (2µl), template DNA (25-250 fmol), primer (2pmol), H_2O (to adjust total volume to 20 µl), and Thermo Sequenase polymerase (2µl) which was added last. Note: in most cases, half reaction volumes produced adequate results.

7.2.14.2.2 Sequencing of PCR products

Before sequencing, PCR products were treated with a combination of exonuclease I and shrimp alkaline phosphatase (SAP). The exonuclease I removes residual single-stranded primers and any extraneous single stranded DNA produced by the PCR. The shrimp alkaline phosphatase removes the remaining dNTPs from the PCR mixture to prevent any interference with the sequencing reactions. The reaction mix consisted of: 5 µl PCR product, 1 µl exonuclease I, and 1 µl SAP. The reaction was incubated at 37°C (15 min) before inactivating the enzymes at 80°C (15 min). The treated PCR product (1 µl; equivalent to 10-100 fmol) can now be used directly for sequencing using the Thermo Sequenase radiolabelled terminator cycle sequencing kit.

7.2.14.3 BigDye™ terminator cycle sequencing

The BigDye™ terminator cycle sequencing reaction mixes were set up in PCR tubes (0.5 ml) and consisted of: terminator ready reaction mix (4 µl), primer (3.2 pmol), template DNA

(single-stranded DNA: 25-50 ng; double-stranded DNA: 100-250 ng; PCR product DNA: 15-45 ng), and deionised water (total volume: 10 µl). (Note: before sequencing PCR products, the PCR products were cleaned by using the Qiagen PCR product clean up system according to manufacturers instructions.) The reaction mix was mixed well and microcentrifuged briefly. The PCR tubes were transferred to a thermal cycler (MJ225) and 25 cycles of the following program run: rapid thermal ramp (1°C/sec) to 96°C, 96°C for 30 sec, rapid thermal ramp to 50°C, 50°C for 15 sec, rapid thermal ramp to 60°C, 60°C for 4 min, rapid thermal ramp to 4°C and held until ready to purify. The contents of the tubes were spun down in a microcentrifuge and the sequencing products purified by ethanol precipitation. After drying, the pellets can be stored at -20°C or resuspended in loading dye (3 µl) in preparation for sample loading. Samples were denatured at 95°C (5 min) prior to loading 2 µl of sample onto an ABI sequencing gel.

7.A Materials and Methods Appendices

Appendix 7.A.1 General solutions and buffers

1 x TE	10 mM Tris-HCl, 1mM EDTA, pH 8.0.
10 x TAE	0.4 M Tris, 10 mM EDTA, 9 ml/L glacial acetic acid.
10 x TBE	1 M Boric acid, 1 M Tris, 20 mM EDTA.
10% SDS	10% w/v sodium dodecyl sulphate.
20 x MOPS	0.2 M MOPS, 0.05 M sodium acetate, 0.01 M EDTA.
20 x SSC	3 M NaCl, 0.3 M trisodium citrate.
20 x SSPE	3.6 M NaCl, 0.2 M sodium phosphate, 0.02 M EDTA pH7.0.
5 x T4 DNA ligase buffer	250 mM Tris-HCl pH7.6, 50 mM $MgCl_2$, 5 mM ATP, 5 mM DTT, 25% w/v polyethylene glycol 8000.
50 x Denhardt's solution	1% w/v BSA, 1% w/v Ficoll, 1% w/v polyvinylpyrrolidone.
Acrylamide 30% stock (19:1)	28.5% acrylamide, 1.5% N,N' – methylenebisacrylamide.
Church buffer	0.5 M sodium phosphate pH7.2, 7% SDS, 1 mM EDTA.
Denaturation solution	1.5 M NaCl, 0.5 M NaOH.
Depurination solution	250mM HCl.
DNA loading buffer	0.25% bromophenol blue, 15% Ficoll.
Ethanol wash solution	50% ethanol, 0.1 M NaCl, 10 mM Tris-HCl pH7.5, 1 mM EDTA. Store at -20°C.
Glass milk	Mix Silica, 325 Mesh (a powdered flint glass obtainable from ceramic stores) with TE buffer and store as a 50% slurry.
Hybridisation buffer	10% w/v dextran sulphate, 6 x SSC, 4 x Denhardt's solution, 0.5% SDS, 1% PPi.
NaI/ Na_2SO_3 solution	Dissolve 90.8 g NaI and 1.5 g Na_2SO_3 in 100 ml H_2O. Filter through Whatman paper before adding 0.5 g Na_2SO_3. Solution should be saturated. Store in the dark.
Neutralisation solution	1.5 M NaCl, 0.5 M Tris-HCl, pH adjusted to 7.5.
10 x PCR buffer	500 mM KCl, 250 mM Tricine, 125 mM KOH, 15 mM $MgCl_2$, pH 8.3.
5 x Long range PCR buffer #1	425 mM Potassium acetate, 125 mM Tricine (pH8.7), 40% glycerol, 5% DMSO, 6 mM $MgCl_2$.

10 x Polynucleotide kinase buffer	0.7 M Tris-HCl pH8.0, 0.1 M $MgCl_2$, 50 mM DTT.
Quick oligonucleotide hybridisation mix	5 x SSC, 2.5 x Denhardt's solution, 0.1% SDS, 0.1% PPi.
RNA loading buffer	20 x MOPS (50µl), formamide (500µl), 37% formaldehyde (150µl), 10 mg/ml ethidium bromide (3µl), DNA loading buffer (100µl).
RNase	10 mg/ml RNase Ain 10mM Tris-HCl pH7.5, 15 mM NaCl. Boil the solution for 15 minutes to destroy DNase activity and allow to cool slowly to room temperature. Store at -20°C.
Solution D	4M guanidinium thiocyanate, 25 mM sodium citrate, pH7; 0.5% sarcosyl, 0.1 M 2-mercaptoethanol.

Maxi- and miniprep solutions

Solution I (resuspension solution)	50 mM glucose, 25mM Tris-HCl pH8.0, 10mM Na_2EDTA pH8.0, 100 µg/ml RNase A. Filter sterilise. Store at 4°C.
Solution II (cell lysis solution)	0.2 M sodium hydroxide, 1% SDS. Filter sterilise. Store at room temperature.
Solution III (neutralisation solution)	5 M potassium acetate (60ml), glacial acetic acid (11.5ml). Make up to 100ml. Autoclave. Store at 4°C. (The resulting solution III is 5 M with respect to acetate and 3 M with respect to potassium).

Appendix 7.A.2 Growth media for bacterial cultures

Luria-Bertani (LB) Medium	10 g/l Bacto-tryptone, 5 g/l Bacto-yeast extract, 5 g/l NaCl.
LB agar	10 g/l Bacto-tryptone, 5 g/l Bacto-yeast extract, 5 g/l NaCl, 15 g/l Bacto-agar.

Appendix 7.A.3 Antibiotics

Ampicillin	50 mg/ml in H_2O stock solution, 20 μg/ml working concentration.
Kanamycin	10 mg/ml in H_2O stock solution, 10 μg/ml working concentration.

Antibiotics dissolved in H_2O were sterilised by filtration through a 0.22μm filter.

References

Achanzar, W. E. & Ward, S. (1997). A nematode gene required for sperm vesicle fusion. *J Cell Sci* 110(Pt 9), 1073-81.

Aizawa, H., Sekine, Y., Takemura, R., Zhang, Z., Nangaku, M. & Hirokawa, N. (1992). Kinesin family in murine central nervous system. *J Cell Biol* 119(5), 1287-96.

Aizawa, T., Sato, Y., Ishihara, F., Taguchi, N., Komatsu, M., Suzuki, N., Hashizume, K. & Yamada, T. (1994). ATP-sensitive K+ channel-independent glucose action in rat pancreatic beta-cell. *Am J Physiol* 266(3 Pt 1), C622-7.

Alport, A. C. (1927). Hereditary familial congenital hemorrhagic nephritis. *Brit. Med. J.* 1, 504-506.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* 215(3), 403-10.

Anand, R., Riley, J. H., Butler, R., Smith, J. C. & Markham, A. F. (1990). A 3.5 genome equivalent multiaccess yac library - construction, characterization, screening and storage. *Nuc Acid Res* 18(8), 1951-1956.

Andreone, T. L., Printz, R. L., Pilkis, S. J., Magnuson, M. A. & Granner, D. K. (1989). The amino acid sequence of rat liver glucokinase deduced from cloned cDNA. *J Biol Chem* 264(1), 363-9.

Antequera, F. & Bird, A. (1993). CpG islands. *Exs* 64, 169-85.

Antequera, F., Macleod, D. & Bird, A. P. (1989). Specific protection of methylated CpGs in mammalian nuclei. *Cell* 58(3), 509-17.

Asfari, M., Janjic, D., Meda, P., Li, G., Halban, P. A. & Wollheim, C. B. (1992). Establishment of 2-mercaptoethanol-dependent differentiated insulin- secreting cell lines. *Endocrin* 130(1), 167-78.

Ashcroft, F. M. & Gribble, F. M. (1999). ATP-sensitive K+ channels and insulin secretion: their role in health and disease. *Diabetol* 42(8), 903-919.

Ashcroft, S. J., Bassett, J. M. & Randle, P. J. (1972). Insulin secretion mechanisms and glucose metabolism in isolated islets. *Diab* 21(2):Suppl), 538-45.

Avner, P., Amar, L., Dandolo, L. & Guenet, J. L. (1988). Genetic analysis of the mouse using interspecific crosses. *Trends Genet* 4(1), 18-23.

Avraham, K. B., Hasson, T., Steel, K. P., Kingsley, D. M., Russell, L. B., Mooseker, M. S., Copeland, N. G. & Jenkins, N. A. (1995). The mouse Snell's waltzer deafness gene encodes an unconventional myosin required for structural integrity of inner ear hair cells. *Nat Genet* 11(4), 369-75.

Bailey, D. W. (1971). Recombinant-inbred strains. An aid to finding identity, linkage, and function of histocompatibility and other genes. *Transpl* 11(3), 325-7.

Baldwin, C. T., Hoth, C. F., Macina, R. A. & Milunsky, A. (1995). Mutations in PAX3 that cause Waardenburg syndrome type I: ten new mutations and review of the literature. *Am J Med Genet* 58(2), 115-22.

Barhanin, J., Lesage, F., Guillemare, E., Fink, M., Lazdunski, M. & Romey, G. (1996). K(V)LQT1 and lsK (minK) proteins associate to form the I(Ks) cardiac potassium current [see comments]. *Nat* 384(6604), 78-80.

Barker, D. F., Hostikka, S. L., Zhou, J., Chow, L. T., Oliphant, A. R., Gerken, S. C., Gregory, M. C., Skolnick, M. H., Atkin, C. L. & Tryggvason, K. (1990). Identification of mutations in the COL4A5 collagen gene in Alport syndrome. *Sci* 248(4960), 1224-7.

Barnett, A. H., Eff, C., Leslie, R. D. & Pyke, D. A. (1981a). Diabetes in identical twins. A study of 200 pairs. *Diabetol* 20(2), 87-93.

Barnett, A. H., Spiliopoulos, A. J., Pyke, D. A., Stubbs, W. A., Burrin, J. & Alberti, K. G. (1981b). Metabolic studies in unaffected co-twins of non-insulin-dependent diabetics. *Br Med J (Clin Res Ed)* 282(6277), 1656-8.

Bashir, R., Britton, S., Strachan, T., Keers, S., Vafiadaki, E., Lako, M., Richard, I., Marchand, S., Bourg, N., Argov, Z., Sadeh, M., Mahjneh, I., Marconi, G., Passos-Bueno, M. R., Moreira, E. d. S., Zatz, M., Beckmann, J. S. & Bushby, K. (1998). A gene related to Caenorhabditis elegans spermatogenesis factor fer-1 is mutated in limb-girdle muscular dystrophy type 2B. *Nat Genet* 20(1), 37-42.

Bassett, D. E., Jr., Boguski, M. S., Spencer, F., Reeves, R., Kim, S., Weaver, T. & Hieter, P. (1997). Genome cross-referencing and XREFdb: implications for the identification and analysis of genes mutated in human disease. *Nat Genet* 15(4), 339-44.

Begg, E. J. & Barclay, M. L. (1995). Aminoglycosides--50 years on. *Br J Clin Pharmacol* 39(6), 597-603.

Bement, W. M., Wirth, J. A. & Mooseker, M. S. (1994). Cloning and mRNA expression of human unconventional myosin-IC. A homologue of amoeboid myosins-I with a single IQ motif and an SH3 domain. *J Mol Biol* 243(2), 356-63.

Bennett, P. H. (1999). Type 2 diabetes among the Pima Indians of Arizona: an epidemic attributable to environmental change? *Nutr Rev* 57(5 Pt 2), S51-4.

Bergstrom, L., Thompson, P. & Wood, R. P. d. (1979). New patterns in genetic and congenital otonephropathies. *Laryn* 89(2 Pt 1), 177-94.

Biervert, C., Schroeder, B. C., Kubisch, C., Berkovic, S. F., Propping, P., Jentsch, T. J. & Steinlein, O. K. (1998). A potassium channel mutation in neonatal human epilepsy. *Sci* 279(5349), 403-6.

Bird, A. P. (1993). Functions for DNA methylation in vertebrates. *Cold Spring Harb Symp Quant Biol* 58, 281-5.

Bird, A. P. (1996). The relationship of DNA methylation to cancer. *Cancer Surv* 28, 87-101.

Bird, A. P., Taggart, M. H., Nicholls, R. D. & Higgs, D. R. (1987). Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. *EMBO J* 6(4), 999-1004.

Birkenmeier, E. H., Schneider, U. & Thurston, S. J. (1992). Fingerprinting genomes by use of PCR with primers that encode protein motifs or contain sequences that regulate gene expression [published erratum appears in Mamm Genome 1993;4(2):133]. *Mamm Gen* 3(10), 537-45.

Birktoft, J. J. & Blow, D. M. (1972). Structure of crystalline -chymotrypsin. V. The atomic structure of tosyl- -chymotrypsin at 2 A resolution. *J Mol Biol* 68(2), 187-240.

Birnboim, H. C. & Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nuc Acid Res* 7(6), 1513-23.

Bonthron, D. T., Brady, N., Donaldson, I. A. & Steinmann, B. (1994). Molecular basis of essential fructosuria: molecular cloning and mutational analysis of human ketohexokinase (fructokinase). *Hum Mol Genet* 3(9), 1627-31.

Bork, P., Sander, C. & Valencia, A. (1993). Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Prot Sci* 2(1), 31-40.

Brady, S. T. (1985). A novel brain ATPase with properties expected for the fast axonal transport motor. *Nat* 317(6032), 73-5.

Breen. (1994). Towards high resolution maps of the mouse and human genomes--a facility for ordering markers to 0.1 cM resolution. European Backcross Collaborative Group. *Hum Mol Genet* 3(4), 621-7.

Brookes, A. J. (1999). The essence of SNPs. *Gene* 234(2), 177-86.

Brown, K. S., Kalinowski, S. S., Megill, J. R., Durham, S. K. & Mookhtiar, K. A. (1997). Glucokinase regulatory protein may interact with glucokinase in the hepatocyte nucleus. *Diab* 46(2), 179-86.

Burant, C. F., Sivitz, W. I., Fukumoto, H., Kayano, T., Nagamatsu, S., Seino, S., Pessin, J. E. & Bell, G. I. (1991). Mammalian glucose transporters: structure and molecular regulation. *Rec Prog Horm Res* 47, 349-87.

Bushman, J. L., Asuru, A. I., Matts, R. L. & Hinnebusch, A. G. (1993). Evidence that GCD6 and GCD7, translational regulators of GCN4, are subunits of the guanine nucleotide exchange factor for eIF-2 in Saccharomyces cerevisiae. *Mol Cell Biol* 13(3), 1920-32.

Cabibbo, A., Consalez, G. G., Sardella, M., Sitia, R. & Rubartelli, A. (1998). Changes in gene expression during the growth arrest of HepG2 hepatoma cells induced by reducing agents or TGFbeta1. *Onco* 16(22), 2935-43.

Cahill, G. F. J., Ashmore, J., Earle, A. S., Zottu, S. (1958a). Glucose penetration into liver. *Am. J. Physiol.* 192, 491-496.

Cahill, G. F. J., Hastings, A. B., Ashmore, J., Earle, A. S., Zottu, S. (1958b). Studies on carbohydrate metabolism in liver slices. X. Factors in the regulation of pathways of glucose metabolism. *J. Biol. Chem.* 230, 125-135.

Carrano, A. V., Lamerdin, J., Ashworth, L. K., Watkins, B., Branscomb, E., Slezak, T., Raff, M., de Jong, P. J., Keith, D., McBride, L., Meister, S., and Kronick, M. (1989). A high-resolution, fluorescence-based, semiautomated method for DNA fingerprinting. *Genomics* 4(2), 129-36.

Celli, J., van Beusekom, E., Hennekam, R. C., Gallardo, M. E., Smeets, D. F., de Cordoba, S. R., Innis, J. W., Frydman, M., Konig, R., Kingston, H., Tolmie, J., Govaerts, L. C., van Bokhoven, H. & Brunner, H. G. (2000). Familial syndromic esophageal atresia maps to 2p23-p24. *Am J Hum Genet* 66(2), 436-44.

Chaib, H., Place, C., Salem, N., Chardenoux, S., Vincent, C., Weissenbach, J., ElZir, E., Loiselet, J. & Petit, C. (1996). A gene responsible for a sensorineural nonsyndromic recessive deafness maps to chromosome 2p22-23. *Hum Mol Genet* 5(1), 155-158.

Charlier, C., Singh, N. A., Ryan, S. G., Lewis, T. B., Reus, B. E., Leach, R. J. & Leppert, M. (1998). A pore mutation in a novel KQT-like potassium channel gene in an idiopathic epilepsy family [see comments]. *Nat Genet* 18(1), 53-5.

Chen, C., Hosokawa, H., Bumbalo, L. M. & Leahy, J. L. (1994). Regulatory effects of glucose on the catalytic activity and cellular content of glucokinase in the pancreatic beta-cell - study using cultured rat islets. *J Clin Invest* 94(4), 1616-1620.

Chenevert, J., Corrado, K., Bender, A., Pringle, J. & Herskowitz, I. (1992). A yeast gene (BEM1) necessary for cell polarization whose product contains two SH3 domains. *Nat* 356(6364), 77-9.

Chevre, J. C., Hani, E. H., Boutin, P., Vaxillaire, M., Blanche, H., Vionnet, N., Pardini, V. C., Timsit, J., Larger, E., Charpentier, G., Beckers, D., Maes, M., Bellanne-Chantelot, C., Velho, G. & Froguel, P. (1998). Mutation screening in 18 Caucasian families suggest the existence of other MODY genes. *Diabetol* 41(9), 1017-23.

Cigan, A. M., Bushman, J. L., Boal, T. R. & Hinnebusch, A. G. (1993). A protein complex of translational regulators of GCN4 mRNA is the guanine nucleotide-exchange factor for translation initiation factor 2 in yeast. *Proc Natl Acad Sci U S A* 90(11), 5350-4.

Cigan, A. M., Foiani, M., Hannig, E. M. & Hinnebusch, A. G. (1991). Complex formation by positive and negative translational regulators of GCN4. *Mol Cell Biol* 11(6), 3217-28.

Cohen-Salmon, M., Crozet, F., Rebillard, G. & Petit, C. (1997a). Cloning and characterization of the mouse collapsin response mediator protein-1, Crmp1. *Mamm Genomics* 8(5), 349-51.

Cohen-Salmon, M., El-Amraoui, A., Leibovici, M. & Petit, C. (1997b). Otogelin: a glycoprotein specific to the acellular membranes of the inner ear. *Proc Natl Acad Sci U S A* 94(26), 14450-5.

Cole, D. G., Chinn, S. W., Wedaman, K. P., Hall, K., Vuong, T. & Scholey, J. M. (1993). Novel heterotrimeric kinesin-related protein purified from sea urchin eggs. *Nat* 366(6452), 268-70.

Collins, J. E., Cole, C. G., Smink, L. J., Garrett, C. L., Leversha, M. A., Soderlund, C. A., Maslen, G. L., Everett, L. A., Rice, K. M., Coffey, A. J., Gregory, S. G., Gwillian, R., Dunham, A., Davies, A. F., Hassock, S., Todd, C. M., Lehrach, H., Hulsebos, T. J. M., Weissenbach, J., Morrow, B., Kucherlapati, R. S., Wadey, R., Scambler, P. J., Kim, U., Simon, M. I., Peyrard, M., Xie, Y., Carter, N. P., Durbin, R., Dumanski, J. P., Bentley, D. R., and Dunham, I. (1995). A high-density YAC contig map of human chromosome 22. *Nat* 377(6547 Suppl), 367-79.

Copeland, N. G. & Jenkins, N. A. (1991). Development and applications of a molecular genetic linkage map of the mouse genome. *Trends Genet* 7(4), 113-8.

Cotton, R. G. (1993). Current methods of mutation detection. *Mutat Res* 285(1), 125-44.

Coucke, P. J., Van Hauwe, P., Kelley, P. M., Kunst, H., Schatteman, I., Van Velzen, D., Meyers, J., Ensink, R. J., Verstreken, M., Declau, F., Marres, H., Kastury, K., Bhasin, S., McGuirt, W. T., Smith, R. J., Cremers, C. W., Van de Heyning, P., Willems, P. J., Smith, S. D. & Van Camp, G. (1999). Mutations in the KCNQ4 gene are responsible for autosomal dominant deafness in four DFNA2 families. *Hum Mol Genet* 8(7), 1321-8.

Courtois, G., Morgan, J. G., Campbell, L. A., Fourel, G. & Crabtree, G. R. (1987). Interaction of a liver-specific nuclear factor with the fibrinogen and alpha-1-antitrypsin promoters. *Sci* 238(4827), 688-692.

Cross, S. H. & Bird, A. P. (1995). CpG islands and genes. *Curr Opin Genet Dev* 5(3), 309-14.

de la Iglesia, N., Veiga-da-Cunha, M., Van Schaftingen, E., Guinovart, J. J. & Ferrer, J. C. (1999). Glucokinase regulatory protein is essential for the proper subcellular localisation of liver glucokinase. *FEBS Lett* 456(2), 332-8.

DeBry, R. W. & Seldin, M. F. (1996). Human/mouse homology relationships. *Genomics* 33(3), 337-51.

DeFronzo, R. A., Bonadonna, R. C. & Ferrannini, E. (1992). Pathogenesis of NIDDM. A balanced overview. *Diabs C* 15(3), 318-68.

Delgado, S., Gomez, M., Bird, A. & Antequera, F. (1998). Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J* 17(8), 2426-35.

Delpire, E., Lu, J., England, R., Dull, C. & Thorne, T. (1999). Deafness and imbalance associated with inactivation of the secretory Na-K-2Cl co-transporter. *Nat Genet* 22(2), 192-5.

Deol, M. S. (1956). The anatomy and development of the mutants pirouette, shaker-1 and waltzer in the mouse. *Proc. R. Soc. Lond. B. Biol. Sci.* 145, 206-213.

Deol, M. S. & Green, M. C. (1966). Snell's waltzer, a new mutation affecting behaviour and the inner ear in the mouse. *Genet Res* 8(3), 339-45.

Detheux, M., Vandekerckhove, J. & Van Schaftingen, E. (1993). Cloning and sequencing of rat liver cDNAs encoding the regulatory protein of glucokinase [published erratum appears in FEBS Lett 1994 Feb 21;339(3):312]. *FEBS Lett* 321(2-3), 111-5.

Detheux, M. & Vanschaftingen, E. (1994). Heterologous expression of an active-rat regulatory protein of glucokinase. *FEBS Lett* 355(1), 27-29.

Dholakia, J. N., Francis, B. R., Haley, B. E. & Wahba, A. J. (1989). Photoaffinity labeling of the rabbit reticulocyte guanine nucleotide exchange factor and eukaryotic initiation factor 2 with 8-azidopurine nucleotides. Identification of GTP- and ATP-binding domains. *J Biol Chem* 264(34), 20638-42.

Dietrich, W., Katz, H., Lincoln, S. E., Shin, H. S., Friedman, J., Dracopoli, N. C. & Lander, E. S. (1992). A genetic map of the mouse suitable for typing intraspecific crosses. *Genet* 131(2), 423-47.

DiPietro, D. L., Sharma, C., Weinhouse, S. (1962). Studies on glucose phosphorylation in rat liver. *Biochem* 1, 455-462.

Doliana, R., Canton, A., Bucciotti, F., Mongiat, M., Bonaldo, P. & Colombatti, A. (2000). Structure, chromosomal localization, and promoter analysis of the human elastin microfibril interfase located proteIN (EMILIN) gene [In Process Citation]. *J Biol Chem* 275(2), 785-92.

Donaldson, C. J., Sutton, S. W., Perrin, M. H., Corrigan, A. Z., Lewis, K. A., Rivier, J. E., Vaughan, J. M. & Vale, W. W. (1996). Cloning and characterization of human urocortin [published erratum appears in Endocrinoogy 1996 Sep;137(9):3896]. *Endocrin* 137(5), 2167-70.

Donaldson, I. A., Doyle, T. C. & Matas, N. (1993). Expression of rat liver ketohexokinase in yeast results in fructose intolerance. *Biochem J* 291(Pt 1), 179-86.

Dressler, G. R., Wilkinson, J. E., Rothenpieler, U. W., Patterson, L. T., Williams-Simons, L. & Westphal, H. (1993). Deregulation of Pax-2 expression in transgenic mice generates severe kidney abnormalities. *Nat* 362(6415), 65-7.

Dunne, M. J. (2000). Ions, genes and insulin release: from basic science to clinical disease. Based on the 1998 R. D. Lawrence Lecture. *Diabet Med* 17(2), 91-104.

Duprat, F., Lesage, F., Fink, M., Reyes, R., Heurteaux, C. & Lazdunski, M. (1997). TASK, a human background K+ channel to sense external pH variations near physiological pH. *EMBO J* 16(17), 5464-71.

Duyk, G. M., Kim, S. W., Myers, R. M. & Cox, D. R. (1990). Exon trapping: a genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. *Proc Natl Acad Sci U S A* 87(22), 8995-9.

Edwards, G. & Weston, A. H. (1995). The role of potassium channels in excitable cells. *Diabetes Res Clin Pract* 28 Suppl, S57-66.

Ekena, K. & Stevens, T. H. (1995). The Saccharomyces cerevisiae MVP1 gene interacts with VPS1 and is required for vacuolar protein sorting. *Mol Cell Biol* 15(3), 1671-8.

Estivill, X., Fortina, P., Surrey, S., Rabionet, R., Melchionda, S., D'Agruma, L., Mansfield, E., Rappaport, E., Govea, N., Mila, M., Zelante, L. & Gasparini, P. (1998). Connexin-26 mutations in sporadic and inherited sensorineural deafness [see comments]. *Lancet* 351(9100), 394-8.

Farrelly, D., Brown, K. S., Tieman, A., Ren, J., Lira, S. A., Hagan, D., Gregg, R., Mookhtiar, K. A. & Hariharan, N. (1999). Mice mutant for glucokinase regulatory protein exhibit decreased liver glucokinase: a sequestration mechanism in metabolic regulation. *Proc Natl Acad Sci U S A* 96(25), 14511-6.

Feinberg, A. P. & Vogelstein, B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal Biochem* 132(1), 6-13.

Fernandez-Novell, J. M., Castel, S., Bellido, D., Ferrer, J. C., Vilaro, S. & Guinovart, J. J. (1999). Intracellular distribution of hepatic glucokinase and glucokinase regulatory protein during the fasted to refed transition in rats. *FEBS Lett* 459(2), 211-4.

Ferreira, A., Niclas, J., Vale, R. D., Banker, G. & Kosik, K. S. (1992). Suppression of kinesin expression in cultured hippocampal neurons using antisense oligonucleotides. *J Cell Biol* 117(3), 595-606.

Fields, S. & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nat* 340(6230), 245-6.

Francis, F., Zehetner, G., Hoglund, M. & Lehrach, H. (1994). Construction and preliminary analysis of the ICRF human P1 library. *Genet Anal Tech Appl* 11(5-6), 148-57.

Froguel, P., Vaxillaire, M., Sun, F., Velho, G., Zouali, H., Butel, M. O., Lesage, S., Vionnet, N., Clement, K., Fougerousse, F., Tanizawa, Y., Weissenbach, J., Beckmann, J. S., Lathrop, G. M., Passa, P., Permutt, M. A. & Cohen, D. (1992). Close linkage of glucokinase locus on chromosome-7p to early-onset non-insulin-dependent diabetes-mellitus. *Nat* 356(6365), 162-164.

Froguel, P. & Velho, G. (1999). Molecular genetics of maturity-onset diabetes of the young. *Trends Endocrin Metab* 10(4), 142-146.

Frohman, M. A., Dush, M. K. & Martin, G. R. (1988). Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci U S A* 85(23), 8998-9002.

Fujita, A., Horio, Y., Copeland, N. G., Gilbert, D. J., Jenkins, N. A. & Kurachi, Y. (1998). Assignment of mouse cardiac two-pore background K+ channel gene (Kcnk4) to the proximal region of mouse chromosome 5. *Genomics* 54(1), 183-4.

Fukushima, K., Ramesh, A., Srisailapathy, C. R., Ni, L., Wayne, S., O'Neill, M. E., Van Camp, G., Coucke, P., Jain, P., Wilcox, E. R., Smith, S. D., Kenyon, J. B., Zbar, R. I. S., and Smith, R. J. H. (1995). An autosomal recessive nonsyndromic form of sensorineural hearing loss maps to 3p-DFNB6. *Genome Res* 5(3), 305-8.

Gaspar, N. J., Kinzy, T. G., Scherer, B. J., Humbelin, M., Hershey, J. W. & Merrick, W. C. (1994). Translation initiation factor eIF-2. Cloning and expression of the human cDNA encoding the gamma-subunit. *J Biol Chem* 269(5), 3415-22.

German, M. S. (1993). Glucose sensing in pancreatic-islet beta-cells - the key role of glucokinase and the glycolytic-intermediates. *Proc Natl Acad Sci U S A* 90(5), 1781-1785.

Ghosh, S., Watanabe, R. M., Hauser, E. R., Valle, T., Magnuson, V. L., Erdos, M. R., Langefeld, C. D., Balow, J., Jr., Ally, D. S., Kohtamaki, K., Chines, P., Birznieks, G., Kaleta, H. S., Musick, A., Te, C., Tannenbaum, J., Eldridge, W., Shapiro, S., Martin, C., Witt, A., So, A., Chang, J., Shurtleff, B., Porter, R., Kudelko, k., Unni, A., Segal, L., Sharaf, R., Blaschak-Harvan, J., Eriksson, J., Tenkula, T., Vidgren, G., Ehnholm, C., Tuomilehto-Wolf, E., Hagopian, W., Buchanan, T. A., Tuomilehto J., Bergman, R. N., Collins, F. S., and Boehnke, M.. (1999). Type 2 diabetes: evidence for linkage on chromosome 20 in 716 Finnish affected sib pairs. *Proc Natl Acad Sci U S A* 96(5), 2198-203.

Gibson, F., Walsh, J., Mburu, P., Varela, A., Brown, K. A., Antonio, M., Beisel, K. W., Steel, K. P. & Brown, S. D. (1995). A type VII myosin encoded by the mouse deafness gene shaker-1. *Nat* 374(6517), 62-4.

Gidhjain, M., Takeda, J., Xu, L. Z., Lange, A. J., Vionnet, N., Stoffel, M., Froguel, P., Velho, G., Sun, F., Cohen, D., Patel, P., Lo, Y. M. D., Hattersley, A. T., Luthman, H., Wedell, A., Stcharles, R., Harrison, R. W., Weber, I. T., Bell, G. I. & Pilkis, S. J. (1993). Glucokinase mutations associated with non-insulin-dependent (type-2) diabetes-mellitus have decreased enzymatic-activity - implications for structure-function-relationships. *Proc Natl Acad Sci U S A* 90(5), 1932-1936.

Gil, A. & Proudfoot, N. J. (1987). Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation. *Cell* 49(3), 399-406.

Gilligan, M., Welsh, G. I., Flynn, A., Bujalska, I., Diggle, T. A., Denton, R. M., Proud, C. G. & Docherty, K. (1996). Glucose stimulates the activity of the guanine nucleotide-exchange factor eIF-2B in isolated rat islets of Langerhans. *J Biol Chem* 271(4), 2121-5.

Goldstein, L. S. (1993). With apologies to scheherazade: tails of 1001 kinesin motors. *Annu Rev Genet* 27, 319-51.

Goshima, Y., Nakamura, F., Strittmatter, P. & Strittmatter, S. M. (1995). Collapsin-induced growth cone collapse mediated by an intracellular protein related to UNC-33. *Nat* 376(6540), 509-14.

Green, M. C. (1960). New mutant - Snell's Waltzer - sv. *Mouse News Lett.* 23, 34.

Greenberg, M. L., Myers, P. L., Skvirsky, R. C. & Greer, H. (1986). New positive and negative regulators for general control of amino acid biosynthesis in Saccharomyces cerevisiae. *Mol Cell Biol* 6(5), 1820-9.

Grimsby, J., Coffey, J. W., Dvorozniak, M. T., Magram, J., Li, G., Matschinsky, F. M., Shiota, C., Kaur, S., Magnuson, M. A. & Grippo, J. F. (2000). Characterization of glucokinase regulatory protein-deficient mice. *J Biol Chem* 275(11), 7826-31.

Grompe, M. (1993). The rapid detection of unknown mutations in nucleic acids. *Nat Genet* 5(2), 111-7.

Gyapay, G., Schmitt, K., Fizames, C., Jones, H., Vega-Czarny, N., Spillett, D., Muselet, D., Prud'Homme, J. F., Dib, C., Auffray, C., Morissette, J., Weissenbach, J. & Goodfellow, P. N. (1996). A radiation hybrid map of the human genome. *Hum Mol Genet* 5(3), 339-46.

Habener, J. F. & Stoffers, D. A. (1998). A newly discovered role of transcription factors involved in pancreas development and the pathogenesis of diabetes mellitus. *Proc Assoc Am Phys* 110(1), 12-21.

Haft, C. R., de la Luz Sierra, M., Barr, V. A., Haft, D. H. & Taylor, S. I. (1998). Identification of a family of sorting nexin molecules and characterization of their association with receptors. *Mol Cell Biol* 18(12), 7278-87.

Haldane, J. B. S., Sprunt, A. D., Haldane, N. M.,. (1915). Reduplication in Mice. *J. Genet.* 5, 133-135.

Hamajima, N., Matsuda, K., Sakata, S., Tamaki, N., Sasaki, M. & Nonaka, M. (1996). A novel gene family defined by human dihydropyrimidinase and three related proteins with differential tissue distribution. *Gene* 180(1-2), 157-63.

Hani, E. H., Stoffers, D. A., Chevre, J. C., Durand, E., Stanojevic, V., Dina, C., Habener, J. F. & Froguel, P. (1999). Defective mutations in the insulin promoter factor-1 (IPF-1) gene in late-onset type 2 diabetes mellitus. *J Clin Invest* 104(9), R41-8.

Hanis, C. L., Boerwinkle, E., Chakraborty, R., Ellsworth, D. L., Concannon, P., Stirling, B., Morrison, V. A., Wapelhorst, B., Spielman, R. S., GogolinEwens, K. J., Shephard, J. M., Williams, S. R., Risch, N., Hinds, D., Iwasaki, N., Ogata, M., Omori, Y., Petzold, C., Rietzsch, H., Schroder, H. E., Schulze, J., Cox, N. J., Menzel, S., Boriraj, V. V., Chen, X., Lim, L. R., Lindner, T., Mereu, L. E., Wang, Y. Q., Xiang, K., Yamagata, K., Yang, Y. & Bell, G. I. (1996). A genome-wide search for human non-insulin dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nat Genet* 13(2), 161-166.

Harashima, S. & Hinnebusch, A. G. (1986). Multiple GCD genes required for repression of GCN4, a transcriptional activator of amino acid biosynthetic genes in Saccharomyces cerevisiae. *Mol Cell Biol* 6(11), 3990-8.

Hasson, T. (1997). Unconventional myosins, the basis for deafness in mouse and man [editorial]. *Am J Hum Genet* 61(4), 801-5.

Hasson, T., Gillespie, P. G., Garcia, J. A., MacDonald, R. B., Zhao, Y., Yee, A. G., Mooseker, M. S. & Corey, D. P. (1997). Unconventional myosins in inner-ear sensory epithelia. *J Cell Biol* 137(6), 1287-307.

Hasson, T., Heintzelman, M. B., Santos-Sacchi, J., Corey, D. P. & Mooseker, M. S. (1995). Expression in cochlea and retina of myosin VIIa, the gene product defective in Usher syndrome type 1B. *Proc Natl Acad Sci U S A* 92(21), 9815-9.

Hayward, B. E. & Bonthron, D. T. (1998). Structure and alternative splicing of the ketohexokinase gene. *Eur J Biochem* 257(1), 85-91.

Hayward, B. E., Dunlop, N., Intody, S., Leek, J. P., Markham, A. F., Warner, J. P. & Bonthron, D. T. (1998). Organization of the human glucokinase regulator gene GCKR. *Genomics* 49(1), 137-42.

Hayward, B. E., Fantes, J. A., Warner, J. P., Intody, S., Leek, J. P., Markham, A. F. & Bonthron, D. T. (1996). Co-localization of the ketohexokinase and glucokinase regulator genes to a 500-kb region of chromosome 2p23. *Mamm Gen* 7(6), 454-8.

Hayward, B. E., Warner, J. P., Dunlop, N., Fantes, J., Intody, S., Leek, J., Markham, A. F. & Bonthron, D. T. (1997). Molecular genetics of the human glucokinase regulator-fructokinase (GCKR-KHK) region of chromosome 2p23. *Biochem Soc Trans* 25(1), 140-5.

Hedgecock, E. M., Culotti, J. G., Hall, D. H. & Stern, B. D. (1987). Genetics of cell and axon migrations in Caenorhabditis elegans. *Devel* 100(3), 365-82.

Heiss, N. S., Rogner, U. C., Kioschis, P., Korn, B. & Poustka, A. (1996). Transcription mapping in a 700-kb region around the DXS52 locus in Xq28: isolation of six novel transcripts and a novel ATPase isoform (hPMCA5). *Genome Res* 6(6), 478-91.

Henderson, R. A., Krissansen, G. W., Yong, R. Y., Leung, E., Watson, J. D. & Dholakia, J. N. (1994). The delta-subunit of murine guanine nucleotide exchange factor eIF-2B. Characterization of cDNAs predicts isoforms differing at the amino- terminal end. *J Biol Chem* 269(48), 30517-23.

Hendrich, B. & Bird, A. (1998). Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol* 18(11), 6538-47.

Hershey, J. W. (1991). Translational control in mammalian cells. *Annu Rev Biochem* 60, 717-55.

Hershey, J. W. (1994). Expression of initiation factor genes in mammalian cells. *Biochim* 76(9), 847-52.

Hinnebusch, A. G. (1986). The general control of amino acid biosynthetic genes in the yeast Saccharomyces cerevisiae. *CRC Crit Rev Biochem* 21(3), 277-317.

Horikawa, Y., Iwasaki, N., Hara, M., Furuta, H., Hinokio, Y., Cockburn, B. N., Lindner, T., Yamagata, K., Ogata, M., Tomonaga, O., Kuroki, H., Kasahara, T., Iwamoto, Y. & Bell, G. I. (1997). Mutation in hepatocyte nuclear factor-1 beta gene (TCF2) associated with MODY [letter]. *Nat Genet* 17(4), 384-5.

Houdusse, A., Silver, M. & Cohen, C. (1996). A model of Ca(2+)-free calmodulin binding to unconventional myosins reveals how calmodulin acts as a regulatory switch. *Struct* 4(12), 1475-90.

Hoyt, M. A., He, L., Loo, K. K. & Saunders, W. S. (1992). Two Saccharomyces cerevisiae kinesin-related gene products required for mitotic spindle assembly. *J Cell Biol* 118(1), 109-20.

Hoyt, M. A., He, L., Totis, L. & Saunders, W. S. (1993). Loss of function of Saccharomyces cerevisiae kinesin-related CIN8 and KIP1 is suppressed by KAR3 motor domain mutations. *Genet* 135(1), 35-44.

Iynedjian, P. B. (1993). Mammalian glucokinase and its gene. *Biochem J* 293(Pt 1), 1-13.

Johnson, J. H., Newgard, C. B., Milburn, J. L., Lodish, H. F. & Thorens, B. (1990). The high Km glucose transporter of islets of Langerhans is functionally similar to the low affinity transporter of liver and has an identical primary sequence. *J Biol Chem* 265(12), 6548-51.

Johnston, G. C., Prendergast, J. A. & Singer, R. A. (1991). The Saccharomyces cerevisiae MYO2 gene encodes an essential myosin for vectorial transport of vesicles. *J Cell Biol* 113(3), 539-51.

Kaldi, K., Diestelkotter, P., Stenbeck, G., Auerbach, S., Jakle, U., Magert, H. J., Wieland, F. T. & Just, W. W. (1993). Membrane topology of the 22 kDa integral peroxisomal membrane protein. *FEBS Lett* 315(3), 217-22.

Karasawa, M., Zwacka, R. M., Reuter, A., Fink, T., Hsieh, C. L., Lichter, P., Francke, U. & Weiher, H. (1993). The human homolog of the glomerulosclerosis gene Mpv17: structure and genomic organization. *Hum Mol Genet* 2(11), 1829-34.

Karinch, A. M., Kimball, S. R., Vary, T. C. & Jefferson, L. S. (1993). Regulation of eukaryotic initiation factor-2B activity in muscle of diabetic rats. *Am J Physiol* 264(1 Pt 1), E101-8.

Katz, J. & Rognstad, R. (1967). The labeling of pentose phosphate from glucose-14C and estimation of the rates of transaldolase, transketolase, the contribution of the pentose cycle, and ribose phosphate synthesis. *Biochem* 6(7), 2227-47.

Kay, B. K., Williamson, M. P. & Sudol, M. (2000). The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. *FASEB J* 14(2), 231-41.

Kelsell, D. P., Rooke, L., Warne, D., Bouzyk, M., Cullin, L., Cox, S., West, L., Povey, S. & Spurr, N. K. (1995). Development of a panel of monochromosomal somatic cell hybrids for rapid gene mapping. *Ann Hum Genet* 59(Pt 2), 233-41.

Kelsell, D. P., Dunlop, J., Stevens, H. P., Lench, N. J., Liang, J. N., Parry, G., Mueller, R. F. & Leigh, I. M. (1997). Connexin 26 mutations in hereditary non-syndromic sensorineural deafness [see comments]. *Nat* 387(6628), 80-3.

Kestila, M., Lenkkeri, U., Mannikko, M., Lamerdin, J., McCready, P., Putaala, H., Ruotsalainen, V., Morita, T., Nissinen, M., Herva, R., Kashtan, C. E., Peltonen, L., Holmberg, C., Olsen, A. & Tryggvason, K. (1998). Positionally cloned gene for a novel glomerular protein--nephrin--is mutated in congenital nephrotic syndrome. *Mol Cell* 1(4), 575-82.

Kestila, M., Mannikko, M., Holmberg, C., Gyapay, G., Weissenbach, J., Savolainen, E. R., Peltonen, L. & Tryggvason, K. (1994). Congenital nephrotic syndrome of the Finnish type maps to the long arm of chromosome 19. *Am J Hum Genet* 54(5), 757-64.

Kida, Y., Espositodelpuente, A., Bogardus, C. & Mott, D. M. (1990). Insulin resistance is associated with reduced fasting and insulin-stimulated glycogen-synthase phosphatase-activity in human skeletal-muscle. *J Clin Invest* 85(2), 476-481.

Kimball, S. R., Fabian, J. R., Pavitt, G. D., Hinnebusch, A. G. & Jefferson, L. S. (1998). Regulation of guanine nucleotide exchange through phosphorylation of eukaryotic initiation factor eIF2alpha. Role of the alpha- and delta- subunits of eiF2b. *J Biol Chem* 273(21), 12841-5.

Kimball, S. R., Mellor, H., Flowers, K. M. & Jefferson, L. S. (1996). Role of translation initiation factor eIF-2B in the regulation of protein synthesis in mammalian cells. *Prog Nucleic Acid Res Mol Biol* 54, 165-96.

King, H. & Rewers, M. (1993). Global estimates for prevalence of diabetes mellitus and impaired glucose tolerance in adults. WHO Ad Hoc Diabetes Reporting Group [see comments]. *Diab Care* 16(1), 157-77.

Ko, M. S., Kitchen, J. R., Wang, X., Threat, T. A., Hasegawa, A., Sun, T., Grahovac, M. J., Kargul, G. J., Lim, M. K., Cui, Y., Sano, Y., Tanaka, T., Liang, Y., Mason, S., Paonessa, P. D., Sauls, A. D., DePalma, G. E., Sharara, R., Rowe, L. B., Eppig, J., Morrell, C. & Doi, H. (2000). Large-scale cDNA analysis reveals phased gene expression patterns during preimplantation mouse development. *Dev* 127(8), 1737-49.

Ko, M. S., Wang, X., Horton, J. H., Hagen, M. D., Takahashi, N., Maezaki, Y. & Nadeau, J. H. (1994). Genetic mapping of 40 cDNA clones on the mouse genome by PCR. *Mamm Gen* 5(6), 349-55.

Kondo, S., Sato-Yoshitake, R., Noda, Y., Aizawa, H., Nakata, T., Matsuura, Y. & Hirokawa, N. (1994). KIF3A is a new microtubule-based anterograde motor in the nerve axon. *J Cell Biol* 125(5), 1095-107.

Kopp, J. B., Klotman, M. E., Adler, S. H., Bruggeman, L. A., Dickie, P., Marinos, N. J., Eckhaus, M., Bryant, J. L., Notkins, A. L. & Klotman, P. E. (1992). Progressive glomerulosclerosis and enhanced renal accumulation of basement membrane components in mice transgenic for human immunodeficiency virus type 1 genes. *Proc Natl Acad Sci U S A* 89(5), 1577-81.

Kraut, J. (1977). Serine proteases: structure and mechanism of catalysis. *Annu Rev Biochem* 46, 331-58.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 22(2), 139-44.

Kubisch, C., Schroeder, B. C., Friedrich, T., Lutjohann, B., El-Amraoui, A., Marlin, S., Petit, C. & Jentsch, T. J. (1999). KCNQ4, a novel potassium channel expressed in sensory outer hair cells, is mutated in dominant deafness. *Cell* 96(3), 437-46.

Kurten, R. C., Cadena, D. L. & Gill, G. N. (1996). Enhanced degradation of EGF receptors by a sorting nexin, SNX1. *Sci* 272(5264), 1008-10.

Leal, S. M., Apaydin, F., Barnwell, C., Iber, M., Kandogan, T., Pfister, M., Braendle, U., Cura, O., Schwalb, M., Zenner, H. P. & Vitale, E. (1998). A second middle eastern kindred with autosomal recessive non-syndromic hearing loss segregates DFNB9. *Eur J Hum Genet* 6(4), 341-4.

Leek, J., Wightman, P. J., Bonthron, D. T., Lench, N. (1998). Assignment of the potassium channel gene KCNK3 to human chromosome band 2p23.3 by fluorescent in situ hybridisation. unpublished.

Lehto, M., Tuomi, T., Mahtani, M. M., Widen, E., Forsblom, C., Sarelin, L., Gullstrom, M., Isomaa, B., Lehtovirta, M., Hyrkko, A., Kanninen, T., Orho, M., Manley, S., Turner, R. C., Brettin, T., Kirby, A., Thomas, J., Duyk, G., Lander, E., Taskinen, M. R. & Groop, L. (1997). Characterization of the MODY3 phenotype. Early-onset diabetes caused by an insulin secretion defect. *J Clin Invest* 99(4), 582-91.

Lennon, G., Auffray, C., Polymeropoulos, M. & Soares, M. B. (1996). The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* 33(1), 151-2.

Lesage, F., Guillemare, E., Fink, M., Duprat, F., Lazdunski, M., Romey, G. & Barhanin, J. (1996). TWIK-1, a ubiquitous human weakly inward rectifying K+ channel with a novel structure. *EMBO J* 15(5), 1004-11.

Lesage, F. & Lazdunski, M. (1998). Mapping of human potassium channel genes TREK-1 (KCNK2) and TASK (KCNK3) to chromosomes 1q41 and 2p23. *Genomics* 51(3), 478-9.

Li, W., Herman, R. K. & Shaw, J. E. (1992). Analysis of the Caenorhabditis elegans axonal guidance and outgrowth gene unc-33. *Genet* 132(3), 675-89.

Liang, Y., Najafi, H., Smith, R. M., Zimmerman, E. C., Magnuson, M. A., Tal, M. & Matschinsky, F. M. (1992). Concordant glucose induction of glucokinase, glucose usage, and glucose- stimulated insulin release in pancreatic islets maintained in organ culture. *Diab* 41(7), 792-806.

Lillie, S. H. & Brown, S. S. (1992). Suppression of a myosin defect by a kinesin-related gene. *Nat* 356(6367), 358-61.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. & Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nat Genet* 21(1 Suppl), 20-4.

Liu, J., Aoki, M., Illa, I., Wu, C., Fardeau, M., Angelini, C., Serrano, C., Urtizberea, J. A., Hentati, F., Hamida, M. B., Bohlega, S., Culper, E. J., Amato, A. A., Bossie, K., Oeltjen, J., Bejaoui, K., McKenna-Yasek, D., Hosler, B. A., Schurr, E., Arahata, K., de Jong, P. J. & Brown, R. H., Jr. (1998). Dysferlin, a novel skeletal muscle gene, is mutated in Miyoshi myopathy and limb girdle muscular dystrophy. *Nat Genet* 20(1), 31-6.

Liu, X. Z., Walsh, J., Mburu, P., Kendrick-Jones, J., Cope, M. J., Steel, K. P. & Brown, S. D. (1997). Mutations in the myosin VIIA gene cause non-syndromic recessive deafness. *Nat Genet* 16(2), 188-90.

Love, J. M., Knight, A. M., McAleer, M. A. & Todd, J. A. (1990). Towards construction of a high resolution map of the mouse genome using PCR-analysed microsatellites. *Nuc Acid Res* 18(14), 4123-30.

Lovett, M., Kere, J. & Hinton, L. M. (1991). Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc Natl Acad Sci U S A* 88(21), 9628-32.

Luo, Y., Shepherd, I., Li, J., Renzi, M. J., Chang, S. & Raper, J. A. (1995). A family of molecules related to collapsin in the embryonic chick nervous system [published erratum appears in Neuron 1995 Nov;15(5):following 1218]. *Neur* 14(6), 1131-40.

Macfarlane, W. M., Shepherd, R. M., Cosgrove, K. E., James, R. F., Dunne, M. J. & Docherty, K. (2000). Glucose modulation of insulin mRNA levels is dependent on transcription factor PDX-1 and occurs independently of changes in intracellular Ca2+. *Diab* 49(3), 418-23.

Magnuson, M. A. & Shelton, K. D. (1989). An alternate promoter in the glucokinase gene is active in the pancreatic beta cell. *J Biol Chem* 264(27), 15936-42.

Mahtani, M. M., Widen, E., Lehto, M., Thomas, J., McCarthy, M., Brayer, J., Bryant, B., Chan, G., Daly, M., Forsblom, C., Kanninen, T., Kirby, A., Kruglyak, L., Munnelly, K., Parkkonen, M., Reeve-Daly, M. P., Weaver, A., Brettin, T., Duyk, G., Lander, E. S. & Groop, L. C. (1996). Mapping of a gene for type 2 diabetes associated with an insulin secretion defect by a genome scan in Finnish families [see comments]. *Nat Genet* 14(1), 90-4.

Malaisse, W. J., Malaisselagae, F., Davies, D. R., Vandercammen, A. & Vanschaftingen, E. (1990). Regulation of glucokinase by a fructose-1-phosphate-sensitive protein in pancreatic-islets. *Eur J Biochem* 190(3), 539-545.

Manjunath, N. A., Bray-Ward, P., Goldstein, S. A. & Gallagher, P. G. (1999). Assignment of the 2P domain, acid-sensitive potassium channel OAT1 gene KCNK3 to human chromosome bands 2p24.1-->p23.3 and murine 5B by in situ hybridization. *Cytogenet Cell Genet* 86(3-4), 242-3.

Marazita, M. L., Ploughman, L. M., Rawlings, B., Remington, E., Arnos, K. S. & Nance, W. E. (1993). Genetic epidemiological studies of early-onset deafness in the U.S. school-age population. *Am J Med Genet* 46(5), 486-91.

Matschinsky, F. M. (1990). Glucokinase as glucose sensor and metabolic signal generator in pancreatic beta-cells and hepatocytes. *Diab* 39(6), 647-652.

Mayer, B. J. & Eck, M. J. (1995). SH3 domains. Minding your p's and q's. *Curr Biol* 5(4), 364-7.

McCarthy, J. J. & Hilfiker, R. (2000). The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nat Biotechnol* 18(5), 505-8.

McCarthy, L. C., Terrett, J., Davis, M. E., Knights, C. J., Smith, A. L., Critcher, R., Schmitt, K., Hudson, J., Spurr, N. K. & Goodfellow, P. N. (1997). A first-generation whole genome-radiation hybrid map spanning the mouse genome. *Genome Res* 7(12), 1153-61.

McIntire, S. L., Garriga, G., White, J., Jacobson, D. & Horvitz, H. R. (1992). Genes necessary for directed axonal elongation or fasciculation in C. elegans. *Neur* 8(2), 307-22.

Meadows, H. J., Benham, C. D., Cairns, W., Gloger, I., Jennings, C., Medhurst, A. D., Murdock, P. & Chapman, C. G. (2000). Cloning, localisation and functional expression of the human orthologue of the TREK-1 potassium channel. *Pflugers Arch* 439(6), 714-22.

Medici, F., Hawa, M., Ianari, A., Pyke, D. A. & Leslie, R. D. G. (1999). Concordance rate for Type II diabetes mellitus in monozygotic twins: actuarial analysis. *Diabetol* 42(2), 146-150.

Menzel, R., Kaisaki, P. J., Rjasanowski, I., Heinke, P., Kerner, W. & Menzel, S. (1998). A low renal threshold for glucose in diabetic patients with a mutation in the hepatocyte nuclear factor-1alpha (HNF-1alpha) gene. *Diab Med* 15(10), 816-20.

Meyer zum Gottesberge, A. M., Reuter, A. & Weiher, H. (1996). Inner ear defect similar to Alport's syndrome in the glomerulosclerosis mouse model Mpv17. *Eur Arch Oto* 253(8), 470-4.

Mikaelin, D. O. a. R., R.J. (1964). Hearing degeneration in Shaker-1 mouse. Correlation of physiological observations with behavioural responses and with cochlear anatomy. *Arch. Otolaryngol* 80, 418-430.

Mizuta, K., Iwasa, K. H., Tachibana, M., Benos, D. J. & Lim, D. J. (1995). Amiloride-sensitive Na+ channel-like immunoreactivity in the luminal membrane of some non-sensory epithelia of the inner ear [published erratum appears in Hear Res 1996 Sep 1;98(1-2):180]. *Hear Res* 88(1-2), 199-205.

Mochizuki, T., Lemmink, H. H., Mariyama, M., Antignac, C., Gubler, M. C., Pirson, Y., Verellen-Dumoulin, C., Chan, B., Schroder, C. H., Smeets, H. J., and Reeders, S. T. (1994). Identification of mutations in the alpha 3(IV) and alpha 4(IV) collagen genes in autosomal recessive Alport syndrome. *Nat Genet* 8(1), 77-81.

Mooseker, M. S. & Cheney, R. E. (1995). Unconventional myosins. *Annu Rev Cell Dev Biol* 11, 633-75.

Morton, N. E. (1991). Genetic epidemiology of hearing impairment. *Ann N Y Acad Sci* 630, 16-31.

Mukhtar, M., Stubbs, M. & Agius, L. (1999). Evidence for glucose and sorbitol-induced nuclear export of glucokinase regulatory protein in hepatocytes. *FEBS Lett* 462(3), 453-8.

Muresan, V., Abramson, T., Lyass, A., Winter, D., Porro, E., Hong, F., Chamberlin, N. L. & Schnapp, B. J. (1998). KIF3C and KIF3A form a novel neuronal heteromeric kinesin that associates with membrane vesicles. *Mol Biol Cell* 9(3), 637-52.

Neher, E., Sakmann, B. & Steinbach, J. H. (1978). The extracellular patch clamp: a method for resolving currents through individual open channels in biological membranes. *Pflugers Arch* 375(2), 219-28.

Neyroud, N., Tesson, F., Denjoy, I., Leibovici, M., Donger, C., Barhanin, J., Faure, S., Gary, F., Coumel, P., Petit, C., Schwartz, K. & Guicheney, P. (1997). A novel mutation in the potassium channel gene KVLQT1 causes the Jervell and Lange-Nielsen cardioauditory syndrome [see comments]. *Nat Genet* 15(2), 186-9.

Nishigori, H., Yamada, S., Kohama, T., Tomura, H., Sho, K., Horikawa, Y., Bell, G. I., Takeuchi, T. & Takeda, J. (1998). Frameshift mutation, A263fsinsGG, in the hepatocyte nuclear factor-1beta gene associated with diabetes and renal dysfunction. *Diab* 47(8), 1354-5.

Nomura, N., Nagase, T., Miyajima, N., Sazuka, T., Tanaka, A., Sato, S., Seki, N., Kawarabayasi, Y., Ishikawa, K. & Tabata, S. (1994). Prediction of the coding sequences of unidentified human genes. II. The coding sequences of 40 new genes (KIAA0041-KIAA0080) deduced by analysis of cDNA clones from human cell line KG-1. *DNA Res* 1(5), 223-9.

Oettinger, M. A., Stanger, B., Schatz, D. G., Glaser, T., Call, K., Housman, D. & Baltimore, D. (1992). The recombination activating genes, RAG-1 and RAG-2, are on chromosome-11p in humans and chromosome-2p in mice. *Immunogen* 35(2), 97-101.

Ohno, S. (1973). Ancient linkage groups and frozen accidents. *Nat* 244(5414), 259-62.

Orahilly, S., Choi, W. H., Patel, P., Turner, R. C., Flier, J. S. & Moller, D. E. (1991). Detection of mutations in insulin-receptor gene in NIDDM patients by analysis of single-stranded conformation polymorphisms. *Diab* 40(6), 777-782.

Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K. & Sekiya, T. (1989). Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc Natl Acad Sci U S A* 86(8), 2766-70.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* 96(6), 2896-901.

Parkes, D. G., Vaughan, J., Rivier, J., Vale, W. & May, C. N. (1997). Cardiac inotropic actions of urocortin in conscious sheep. *Am J Physiol* 272(5 Pt 2), H2115-22.

Pavitt, G. D., Yang, W. & Hinnebusch, A. G. (1997). Homologous segments in three subunits of the guanine nucleotide exchange factor eIF2B mediate translational regulation by phosphorylation of eIF2. *Mol Cell Biol* 17(3), 1298-313.

Pawson, T. & Gish, G. D. (1992). SH2 and SH3 domains: from structure to function. *Cell* 71(3), 359-62.

Peterson, J., Zheng, Y., Bender, L., Myers, A., Cerione, R. & Bender, A. (1994). Interactions between the bud emergence proteins Bem1p and Bem2p and Rho- type GTPases in yeast. *J Cell Biol* 127(5), 1395-406.

Pfeifer, G. P., Tanguay, R. L., Steigerwald, S. D. & Riggs, A. D. (1990). In vivo footprint and methylation analysis by PCR-aided genomic sequencing: comparison of active and inactive X chromosomal DNA at the CpG island and promoter of human PGK-1. *Genes Dev* 4(8), 1277-87.

Pinkel, D., Straume, T. & Gray, J. W. (1986). Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proc Natl Acad Sci U S A* 83(9), 2934-8.

Ponting, C. P. (1996). Novel domains in NADPH oxidase subunits, sorting nexins, and PtdIns 3- kinases: binding partners of SH3 domains? *Prot Sci* 5(11), 2353-7.

Poulsen, P., Kyvik, K. O., Vaag, A. & BeckNielsen, H. (1999). Heritability of Type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance - a population-based twin study. *Diabetol* 42(2), 139-145.

Prekeris, R. & Terrian, D. M. (1997). Brain myosin V is a synaptic vesicle-associated motor protein: evidence for a Ca2+-dependent interaction with the synaptobrevin-synaptophysin complex. *J Cell Biol* 137(7), 1589-601.

Price, N. & Proud, C. (1994). The guanine nucleotide-exchange factor, eIF-2B. *Biochim* 76(8), 748-60.

Price, N. T., Francia, G., Hall, L. & Proud, C. G. (1994). Guanine nucleotide exchange factor for eukaryotic initiation factor-2. Cloning of cDNA for the delta-subunit of rabbit translation initiation factor-2B. *Biochim Biophys Acta* 1217(2), 207-10.

Price, N. T., Mellor, H., Craddock, B. L., Flowers, K. M., Kimball, S. R., Wilmer, T., Jefferson, L. S. & Proud, C. G. (1996). eIF2B, the guanine nucleotide-exchange factor for eukaryotic initiation factor 2. Sequence conservation between the alpha, beta and delta subunits of eIF2B from mammals and yeast. *Biochem J* 318(Pt 2), 637-43.

Prochazka, M., Mochizuki, H., Baier, L. J., Cohen, P. T. W. & Bogardus, C. (1995). Molecular and linkage analysis of type-1 protein phosphatase catalytic beta-subunit gene - lack of evidence for its major role in insulin-resistance in pima-indians. *Diabetol* 38(4), 461-466.

Proud, C. G. (1992). Protein phosphorylation in translational control. *Curr Top Cell Regul* 32, 243-369.

Proud, C. G., Colthurst, D. R., Ferrari, S. & Pinna, L. A. (1991). The substrate specificity of protein kinases which phosphorylate the alpha subunit of eukaryotic initiation factor 2. *Eur J Biochem* 195(3), 771-9.

Proud, C. G. & Denton, R. M. (1997). Molecular mechanisms for the control of translation by insulin. *Biochem J* 328(Pt 2), 329-41.

Puschel, A. W., Adams, R. H. & Betz, H. (1995). Murine semaphorin D/collapsin is a member of a diverse gene family and creates domains inhibitory for axonal extension. *Neur* 14(5), 941-8.

Randle, P. J. (1993). Glucokinase and candidate genes for type-2 (non-insulin-dependent) diabetes-mellitus. *Diabetol* 36(4), 269-275.

Razin, A. & Cedar, H. (1994). DNA methylation and genomic imprinting. *Cell* 77(4), 473-6.

Reuter, A., Nestl, A., Zwacka, R. M., Tuckermann, J., Waldherr, R., Wagner, E. M., Hoyhtya, M., Meyer zum Gottesberge, A. M., Angel, P. & Weiher, H. (1998). Expression of the recessive glomerulosclerosis gene Mpv17 regulates MMP- 2 expression in fibroblasts, the kidney, and the inner ear of mice. *Mol Biol Cell* 9(7), 1675-82.

Rickles, R. J., Botfield, M. C., Weng, Z., Taylor, J. A., Green, O. M., Brugge, J. S. & Zoller, M. J. (1994). Identification of Src, Fyn, Lyn, PI3K and Abl SH3 domain ligands using phage display libraries. *EMBO J* 13(23), 5598-604.

Rizo, J. & Sudhof, T. C. (1998). C2-domains, structure and function of a universal Ca2+-binding domain. *J Biol Chem* 273(26), 15879-82.

Roberts, L. (1991). GRAIL seeks out genes buried in DNA sequence [news]. *Sci* 254(5033), 805.

Romano, C. G., G. Pongiglione, R. (1963). Aritmie cardiache rare dell' eta pediatrica. II. Accessi sincopali per fibrillazione ventricolare parossistica. (Presentazione del primo caso della letteratura pediatrica Italiana.). *Clin. Pediat.* 45, 656-683.

Roof, D. M., Meluh, P. B. & Rose, M. D. (1992). Kinesin-related proteins required for assembly of the mitotic spindle. *J Cell Biol* 118(1), 95-108.

Rowe, L. B., Nadeau, J. H., Turner, R., Frankel, W. N., Letts, V. A., Eppig, J. T., Ko, M. S., Thurston, S. J. & Birkenmeier, E. H. (1994). Maps from two interspecific backcross DNA panels available as a community genetic mapping resource [published erratum appears in Mamm Genome 1994 Jul;5(7):463]. *Mamm Gen* 5(5), 253-74.

Rowlands, A. G., Panniers, R. & Henshaw, E. C. (1988). The catalytic mechanism of guanine nucleotide exchange factor action and competitive inhibition by phosphorylated eukaryotic initiation factor 2. *J Biol Chem* 263(12), 5526-33.

Rudy, B. (1988). Diversity and ubiquity of K channels. *Neurosci* 25(3), 729-49.

Saadat, M., Kakinoki, Y., Mizuno, Y., Kikuchi, K. & Yoshida, M. C. (1994). Chromosomal localization of human, rat, and mouse protein phosphatase type 1 beta catalytic subunit genes (PPP1CB) by fluorescence in situ hybridization. *Jpn J Genet* 69(6), 697-700.

Sablin, E. P., Kull, F. J., Cooke, R., Vale, R. D. & Fletterick, R. J. (1996). Crystal structure of the motor domain of the kinesin-related motor ncd [see comments]. *Nat* 380(6574), 555-9.

Sack, S., Kull, F. J. & Mandelkow, E. (1999). Motor proteins of the kinesin family. Structures, variations, and nucleotide binding sites. *Eur J Biochem* 262(1), 1-11.

Sacks, D. B. & McDonald, J. M. (1996). The pathogenesis of type II diabetes mellitus. A polygenic disease. *Am J Clin Pathol* 105(2), 149-56.

Sambrook, J., Fritsch, E. F., Maniatis, T. (1989). *Molecular Cloning. A Laboratory Manual Second Edition.*

Samuel, C. E. (1993). The eIF-2 alpha protein kinases, regulators of translation in eukaryotes from yeasts to humans. *J Biol Chem* 268(11), 7603-6.

Sanguinetti, M. C., Curran, M. E., Zou, A., Shen, J., Spector, P. S., Atkinson, D. L. & Keating, M. T. (1996). Coassembly of K(V)LQT1 and minK (IsK) proteins to form cardiac I(Ks) potassium channel [see comments]. *Nat* 384(6604), 80-3.

Schapira, F., Schapira, G. and Dreyfus, J.-C. (1961-1962). La lesion enzymatique de la fructisurie benigne. *Enzymol. Biol. Clin.* 1, 170-175.

Schenkel, J., Zwacka, R. M., Rutenberg, C., Reuter, A., Waldherr, R. & Weiher, H. (1995). Functional rescue of the glomerulosclerosis phenotype in Mpv17 mice by transgenesis with the human Mpv17 homologue. *Kid Int* 48(1), 80-4.

Scheper, G. C., Mulder, J., Kleijn, M., Voorma, H. O., Thomas, A. A. & van Wijk, R. (1997). Inactivation of eIF2B and phosphorylation of PHAS-I in heat-shocked rat hepatoma cells. *J Biol Chem* 272(43), 26850-6.

Scheper, G. C., Thomas, A. A. & van Wijk, R. (1998). Inactivation of eukaryotic initiation factor 2B in vitro by heat shock. *Biochem J* 334(Pt 2), 463-7.

Schulze-Bahr, E., Wang, Q., Wedekind, H., Haverkamp, W., Chen, Q., Sun, Y., Rubie, C., Hordt, M., Towbin, J. A., Borggrefe, M., Assmann, G., Qu, X., Somberg, J. C., Breithardt, G., Oberti, C. & Funke, H. (1997). KCNE1 mutations cause jervell and Lange-Nielsen syndrome [letter]. *Nat Genet* 17(3), 267-8.

Sellers, J. R. (2000). Myosins: a diverse superfamily. *Biochim Biophys Acta* 1496(1), 3-22.

Serikawa, T., Montagutelli, X., Simon-Chazottes, D. & Guenet, J. L. (1992). Polymorphisms revealed by PCR with single, short-sized, arbitrary primers are reliable markers for mouse and rat gene mapping. *Mamm Gen* 3(2), 65-72.

Shevchenko, A., Loboda, A., Ens, W. & Standing, K. G. (2000). MALDI quadrupole time-of-flight mass spectrometry: a powerful tool for proteomic research. *Anal Chem* 72(9), 2132-41.

Shiota, C., Coffey, J., Grimsby, J., Grippo, J. F. & Magnuson, M. A. (1999). Nuclear import of hepatic glucokinase depends upon glucokinase regulatory protein, whereas export is due to a nuclear export signal sequence in glucokinase. *J Biol Chem* 274(52), 37125-30.

Singh, B., Hao, W., Wu, Z., Eigl, B. & Gupta, R. S. (1996a). Cloning and characterization of cDNA for adenosine kinase from mammalian (Chinese hamster, mouse, human and rat) species. High frequency mutants of Chinese hamster ovary cells involve structural alterations in the gene. *Eur J Biochem* 241(2), 564-71.

Singh, L. P., Denslow, N. D. & Wahba, A. J. (1996b). Modulation of rabbit reticulocyte guanine nucleotide exchange factor activity by casein kinases 1 and 2 and glycogen synthase kinase 3. *Biochem* 35(10), 3206-12.

Singh, L. P. & Wahba, A. J. (1995). Allosteric activation of rabbit reticulocyte guanine nucleotide exchange factor activity by sugar phosphates and inositol phosphates. *Biochem Biophys Res Commun* 217(2), 616-23.

Singh, N. A., Charlier, C., Stauffer, D., DuPont, B. R., Leach, R. J., Melis, R., Ronen, G. M., Bjerre, I., Quattlebaum, T., Murphy, J. V., McHarg, M. L., Gagnon, D., Rosales, T. O., Peiffer, A., Anderson, V. E. & Leppert, M. (1998). A novel potassium channel gene, KCNQ2, is mutated in an inherited epilepsy of newborns [see comments]. *Nat Genet* 18(1), 25-9.

Sosa, H., Dias, D. P., Hoenger, A., Whittaker, M., Wilson-Kubalek, E., Sablin, E., Fletterick, R. J., Vale, R. D. & Milligan, R. A. (1997). A model for the microtubule-Ncd motor protein complex obtained by cryo- electron microscopy and image analysis. *Cell* 90(2), 217-24.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* 98(3), 503-17.

Spina, M., Merlo-Pich, E., Chan, R. K., Basso, A. M., Rivier, J., Vale, W. & Koob, G. F. (1996). Appetite-suppressing effects of urocortin, a CRF-related neuropeptide. *Sci* 273(5281), 1561-4.

Steel, K. P. & Brown, S. D. (1994). Genes and deafness. *Trends Genet* 10(12), 428-35.

Stewart, R. J., Pesavento, P. A., Woerpel, D. N. & Goldstein, L. S. (1991). Identification and partial characterization of six members of the kinesin superfamily in Drosophila. *Proc Natl Acad Sci U S A* 88(19), 8470-4.

Stoffers, D. A., Ferrer, J., Clarke, W. L. & Habener, J. F. (1997). Early-onset type-II diabetes mellitus (MODY4) linked to IPF1 [letter]. *Nat Genet* 17(2), 138-9.

Stoye, J. P. & Coffin, J. M. (1988). Polymorphism of murine endogenous proviruses revealed by using virus class-specific oligonucleotide probes [published erratum appears in J Virol 1988 Jul;62(7):2530]. *J Virol* 62(1), 168-75.

Summerton, J. & Weller, D. (1997). Morpholino antisense oligomers: design, preparation, and properties. *Antisense Nucleic Acid Drug Dev* 7(3), 187-95.

Taketo, M. M., Araki, Y., Matsunaga, A., Yokoi, A., Tsuchida, J., Nishina, Y., Nozaki, M., Tanaka, H., Koga, M., Uchida, K., Matsumiya, K., Okuyama, A., Rochelle, J. M., Nishimune, Y., Matsui, M. & Seldin, M. F. (1997). Mapping of eight testis-specific genes to mouse chromosomes. *Genomics* 46(1), 138-42.

Tang, X. X., Biegel, J. A., Nycum, L. M., Yoshioka, A., Brodeur, G. M., Pleasure, D. E. & Ikegaki, N. (1995). cDNA cloning, molecular characterization, and chromosomal localization of NET(EPHT2), a human EPH-related receptor protein-tyrosine kinase gene preferentially expressed in brain. *Genomics* 29(2), 426-37.

Taylor, B. A. (1978). Recombinant inbred strains: use in gene mapping. pp. 423-38. *In: Morse HC 3d, ed. Origins of inbred mice. New York, Academic Press.*

Thomas, J. H. (1993). Thinking about genetic redundancy. *Trends Genet* 9(11), 395-9.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nuc Acid Res* 22(22), 4673-80.

Toyoda, Y., Miwa, I., Kamiya, M., Ogiso, S., Nonogaki, T., Aoki, S. & Okuda, J. (1995). Tissue and subcellular distribution of glucokinase in rat liver and their changes during fasting-refeeding. *Histochem Cell Biol* 103(1), 31-8.

Turnbull, A. V., Vale, W. & Rivier, C. (1996). Urocortin, a corticotropin-releasing factor-related mammalian peptide, inhibits edema due to thermal injury in rats. *Eur J Pharmacol* 303(3), 213-6.

Vale, R. D., Reese, T. S. & Sheetz, M. P. (1985). Identification of a novel force-generating protein, kinesin, involved in microtubule-based motility. *Cell* 42(1), 39-50.

Van Camp, G., Willems, P. J. & Smith, R. J. (1997). Nonsyndromic hearing impairment: unparalleled heterogeneity. *Am J Hum Genet* 60(4), 758-64.

Van Hauwe, P., Coucke, P. J., Declau, F., Kunst, H., Ensink, R. J., Marres, H. A., Cremers, C. W., Djelantik, B., Smith, S. D., Kelley, P., Van de Heyning, P. H. & Van Camp, G. (1999). Deafness linked to DFNA2: one locus but how many genes? [letter; comment]. *Nat Genet* 21(3), 263.

Van Schaftingen, E. (1989). A protein from rat liver confers to glucokinase the property of being antagonistically regulated by fructose 6-phosphate and fructose 1- phosphate. *Eur J Biochem* 179(1), 179-84.

Vanschaftingen, E. & Vandercammen, A. (1989). Stimulation of glucose phosphorylation by fructose in isolated rat hepatocytes. *Eur J Biochem* 179(1), 173-177.

Vaughan, J., Donaldson, C., Bittencourt, J., Perrin, M. H., Lewis, K., Sutton, S., Chan, R., Turnbull, A. V., Lovejoy, D., Rivier, C., Rivier, J., Sawchenko, P. E., and Vale, W. (1995). Urocortin, a mammalian neuropeptide related to fish urotensin I and to corticotropin-releasing factor. *Nat* 378(6554), 287-92.

Vaxillaire, M., Rouard, M., Yamagata, K., Oda, N., Kaisaki, P. J., Boriraj, V. V., Chevre, J. C., Boccio, V., Cox, R. D., Lathrop, G. M., Dussoix, P., Philippe, J., Timsit, J.,

Charpentier, G., Velho, G., Bell, G. I. & Froguel, P. (1997). Identification of nine novel mutations in the hepatocyte nuclear factor 1 alpha gene associated with maturity-onset diabetes of the young (MODY3). *Hum Mol Genet* 6(4), 583-586.

Vaxillaire, M., Vionnet, N., Vigouroux, C., Sun, F., Espinosa, R., 3rd, Lebeau, M. M., Stoffel, M., Lehto, M., Beckmann, J. S., Detheux, M, Passa, P., Cohen, D., Van Schaftingen, E., Velho, G., Bell, G., and Froguel, P. (1994). Search for a third susceptibility gene for maturity-onset diabetes of the young. Studies with eleven candidate genes. *Diab* 43(3), 389-95.

Vazquez de Aldana, C. R. & Hinnebusch, A. G. (1994). Mutations in the GCD7 subunit of yeast guanine nucleotide exchange factor eIF-2B overcome the inhibitory effects of phosphorylated eIF-2 on translation initiation. *Mol Cell Biol* 14(5), 3208-22.

Velho, G., Petersen, K. F., Perseghin, G., Hwang, J. H., Rothman, D. L., Pueyo, M. E., Cline, G. W., Froguel, P. & Shulman, G. I. (1996). Impaired hepatic glycogen synthesis in glucokinase-deficient (MODY-2) subjects. *J Clin Invest* 98(8), 1755-61.

Verpy, E., Leibovici, M. & Petit, C. (1999). Characterization of otoconin-95, the major protein of murine otoconia, provides insights into the formation of these inner ear biominerals. *Proc Natl Acad Sci U S A* 96(2), 529-34.

Vinuela, E., Salas, M., Sols, A. (1963). Glucokinase and hexokinase in liver in relation to glycogen synthesis. *J. Biol. Chem.* 238, PC1175-PC1177.

Vionnet, N., Hani, E. H., Lesage, S., Philippi, A., Hager, J., Varret, M., Stoffel, M., Tanizawa, Y., Chiu, K. C., Glaser, B., Permutt, M. A., Passa, P., Demenais, F. & Froguel, P. (1997). Genetics of NIDDM in France: studies with 19 candidate genes in affected sib pairs. *Diab* 46(6), 1062-8.

Vionnet, N., Stoffel, M., Takeda, J., Yasuda, K., Bell, G. I., Zouali, H., Lesage, S., Velho, G., Iris, F., Passa, P., Froguel, P. & Cohen, D. (1992). Nonsense mutation in the glucokinase gene causes early-onset non-insulin-dependent diabetes-mellitus. *Nat* 356(6371), 721-722.

Wang, A., Liang, Y., Fridell, R. A., Probst, F. J., Wilcox, E. R., Touchman, J. W., Morton, C. C., Morell, R. J., Noben-Trauth, K., Camper, S. A. & Friedman, T. B. (1998). Association of unconventional myosin MYO15 mutations with human nonsyndromic deafness DFNB3 [see comments]. *Sci* 280(5368), 1447-51.

Wang, Q., Curran, M. E., Splawski, I., Burn, T. C., Millholland, J. M., VanRaay, T. J., Shen, J., Timothy, K. W., Vincent, G. M., de Jager, T., Schwartz, P. J., Toubin, J. A., Moss, A. J., Atkinson, D. L., Landes, G. M., Connors, T. D. & Keating, M. T. (1996). Positional cloning of a novel potassium channel gene: KVLQT1 mutations cause cardiac arrhythmias. *Nat Genet* 12(1), 17-23.

Ward, O. C. (1964). A new familial cardiac syndrome in children. *J. Irish Med. Assoc.* 54, 103-106.

Warner, J. P., Leek, J. P., Intody, S., Markham, A. F. & Bonthron, D. T. (1995). Human glucokinase regulatory protein (GCKR): cDNA and genomic cloning, complete primary structure, and chromosomal localization. *Mamm Gen* 6(8), 532-6.

Webb, B. L. & Proud, C. G. (1997). Eukaryotic initiation factor 2B (eIF2B). *Int J Biochem Cell Biol* 29(10), 1127-31.

Weber, J. L. & May, P. E. (1989). Abundant class of human dna polymorphisms which can be typed using the polymerase chain-reaction. *Am J Hum Genet* 44(3), 388-396.

Weiher, H., Noda, T., Gray, D. A., Sharpe, A. H. & Jaenisch, R. (1990). Transgenic mouse model of kidney disease: insertional inactivation of ubiquitously expressed gene leads to nephrotic syndrome. *Cell* 62(3), 425-34.

Weil, D., Blanchard, S., Kaplan, J., Guilford, P., Gibson, F., Walsh, J., Mburu, P., Varela, A., Levilliers, J., Weston, M. D. Kelley, P. M., Kimberling, W.J., Wagenaar, M., Leviacobas, F., Largetpiet, D., Munnich, A., Steel, K. P., Brown, S. D. M., and Petit, C. (1995). Defective myosin VIIA gene responsible for Usher syndrome type 1B. *Nat* 374(6517), 60-1.

Weil, D., Kussel, P., Blanchard, S., Levy, G., Levi-Acobas, F., Drira, M., Ayadi, H. & Petit, C. (1997). The autosomal recessive isolated deafness, DFNB2, and the Usher 1B syndrome are allelic defects of the myosin-VIIA gene. *Nat Genet* 16(2), 191-3.

Welsh, G. I., Loughlin, A. J., Foulstone, E. J., Price, N. T. & Proud, C. G. (1997a). Regulation of initiation factor eIF-2B by GSK-3 regulated phosphorylation. *Biochem Soc Trans* 25(2), 191S.

Welsh, G. I. & Proud, C. G. (1993). Glycogen synthase kinase-3 is rapidly inactivated in response to insulin and phosphorylates eukaryotic initiation factor eIF-2B. *Biochem J* 294(Pt 3), 625-9.

Welsh, G. I., Stokes, C. M., Wang, X., Sakaue, H., Ogawa, W., Kasuga, M. & Proud, C. G. (1997b). Activation of translation initiation factor eIF2B by insulin requires phosphatidyl inositol 3-kinase. *FEBS Lett* 410(2-3), 418-22.

Wightman, P. J., Hayward, B. E. & Bonthron, D. T. (1997). The genes encoding glucokinase regulatory protein and ketohexokinase co- localize to mouse chromosome 5. *Mamm Gen* 8(9), 700-1.

Woehlke, G., Ruby, A. K., Hart, C. L., Ly, B., Hom-Booher, N. & Vale, R. D. (1997). Microtubule interaction site of the kinesin motor. *Cell* 90(2), 207-16.

Yamagata, K., Furuta, H., Oda, N., Kaisaki, P. J., Menzel, S., Cox, N. J., Fajans, S. S., Signorini, S., Stoffel, M. & Bell, G. I. (1996a). Mutations In the hepatocyte nuclear factor-4 alpha gene in maturity-onset diabetes of the young (MODY1). *Nat* 384(6608), 458-460.

Yamagata, K., Oda, N., Kaisaki, P. J., Menzel, S., Furuta, H., Vaxillaire, M., Southam, L., Cox, R. D., Lathrop, G. M., Boriraj, V. V., Chen, X. N., Cox, N. J., Oda, Y., Yano, H., LeBeau, M. M., Yamada, S., Nishigori, H., Takeda, J., Fajans, S. S., Hattersley, A. T., Iwasaki, N., Hansen, T., Pedersen, O., Polonsky, K. S., Turner, R. C., Velho, G., Chevre, J. C., Froguel, P. & Bell, G. I. (1996b). Mutations in the hepatocyte nuclear factor-1 alpha gene in maturity-onset diabetes of the young (MODY3). *Nat* 384(6608), 455-458.

Yamazaki, H., Nakata, T., Okada, Y. & Hirokawa, N. (1995). KIF3A/B: a heterodimeric kinesin superfamily protein that works as a microtubule plus end-directed motor for membrane organelle transport. *J Cell Biol* 130(6), 1387-99.

Yang, W. & Hinnebusch, A. G. (1996). Identification of a regulatory subcomplex in the guanine nucleotide exchange factor eIF2B that mediates inhibition by phosphorylated eIF2. *Mol Cell Biol* 16(11), 6603-16.

Yang, W. P., Levesque, P. C., Little, W. A., Conder, M. L., Ramakrishnan, P., Neubauer, M. G. & Blanar, M. A. (1998). Functional expression of two KvLQT1-related potassium channels responsible for an inherited idiopathic epilepsy. *J Biol Chem* 273(31), 19419-23.

Yang, Z. & Goldstein, L. S. (1998). Characterization of the KIF3C neural kinesin-like motor from mouse. *Mol Biol Cell* 9(2), 249-61.

Yasunaga, S., Grati, M., Cohen-Salmon, M., El-Amraoui, A., Mustapha, M., Salem, N., El-Zir, E., Loiselet, J. & Petit, C. (1999). A mutation in OTOF, encoding otoferlin, a FER-1-like protein, causes DFNB9, a nonsyndromic form of deafness [see comments]. *Nat Genet* 21(4), 363-9.

Zhao, L., Donaldson, C. J., Smith, G. W. & Vale, W. W. (1998). The structures of the mouse and human urocortin genes (Ucn and UCN). *Genomics* 50(1), 23-33.

Zhao, L. P., Aragaki, C., Hsu, L. & Quiaoit, F. (1998). Mapping of complex traits by single-nucleotide polymorphisms. *Am J Hum Genet* 63(1), 225-40

Zwacka, R. M., Reuter, A., Pfaff, E., Moll, J., Gorgas, K., Karasawa, M. & Weiher, H. (1994). The glomerulosclerosis gene Mpv17 encodes a peroxisomal protein producing reactive oxygen species. *EMBO J* 13(21), 5129-34.

The genes encoding glucokinase regulatory protein and ketohexokinase co-localize to mouse Chromosome 5

Patrick J. Wightman, Bruce E. Hayward, David T. Bonthron

Human Genetics Unit, University of Edinburgh, Molecular Medicine Centre, Western General Hospital, Edinburgh EH4 2XU, UK

Received: 24 March 1997 / Accepted: 8 May 1997

Species: Mouse

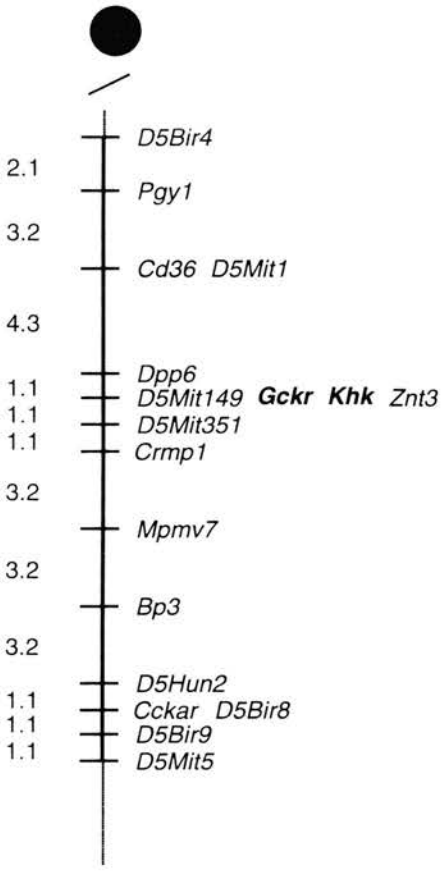


Fig. 1. Co-localization of *Gckr* and *Khk* on mouse Chr 5. The gene order and relative positions of markers used in this study are shown. Map distances in centimorgans are shown on the left.

Loci names: Glucokinase regulatory protein and ketohexokinase
Loci symbols: *Gckr* and *Khk*
Map position: *Gckr* and *Khk* map to the proximal part of Chromosome (Chr) 5, cosegregating with a number of other loci including *D5Mit149*. *D5Mit1*–3.19 ± 1.81–*D5Bir5*–1.06 ± 1.06–*Nos3/Dpp6/Fgl2*–1.06 ± 1.06–*Htr5a/Gbx1/En2/Nkx1–1/Plk–ps1/D5Mit149/D5Wsu178e/D5Xrf391/D5Bir6/Znt3/Gckr/Khk*–1.06 ± 1.06–*D5Mit351/D5Xrf47*–1.06 ± 1.06–*Crmp1/D5Bir7/Msx1*–2.13 ± 1.49–*Bapx1* (Fig. 1).
Method of mapping: Using The Jackson Laboratory Backcross DNA Panel Mapping Resource, 94 backcross animals from the interspecific cross (C57BL/6JEi × SPRET/Ei) × SPRET/Ei were typed with a PCR/restriction digest assay to reveal the variants.
Database deposit information: Typing data is deposited under accession number MGD-JNUM-37599 at http://www.jax.org/resources/documents/cmdata.
Molecular reagents: For *Gckr*, a ~4-kb fragment corresponding to human exons 2–8 [1] was amplified, and the PCR product ends were sequenced. The C57BL/6 and *Mus spretus* sequences diverge 145 bp upstream of the IVS7-exon 8 splice junction, where a poly(A) sequence (on the sense strand) is found in C57BL/6J but the start of a B1 repetitive element-like sequence is present in *M. spretus*. (The C57BL/6J PCR product is also 200–300 bp larger than that of *M. spretus*.) A single nt variant 55 bp upstream of the splice junction alters a *Xcm*I cutting site (present in *M. spretus* but absent in C57BL/6). A new primer GreSB (5′-dCTTGGTTGAGGAATCTATTTCTAG-3′) within intron 7 was used with the exon 8 primer Gre7R (5′-dGCTCACTGGATTGAAGCCAACC-3′) to type the polymorphism by PCR with standard buffers (1.5 mM $MgCl_2$) and an annealing temperature of 55°C. For *Khk*, primers KhkM4 (5′-dTGAGGGGCTTGTACAGTCGAG-3′) and KhkR9 (5′-dCCACCTGGCACCCGAATCTC-3′), designed from the sequences of rat [2] and mouse [3] *Khk* cDNA, were used to amplify a ~400-bp genomic fragment corresponding to exons 6–8, with an annealing temperature of 64°C.
Allele detection: GreSB and Gre7R amplify a ~200-bp *Gckr* fragment that, if derived from an *M. spretus* allele, cuts with *Xcm*I to yield fragments of size ~150 bp and ~50 bp. *Mbo*I digestion of the *Khk* ~400-bp genomic fragment produces a fragment of 300 bp from the C57BL/6J *Khk* allele and 220 bp from *M. spretus*.
Previously identified homologs: Human GCKR and KHK co-localize to Chr. 2p23.2-23.3 [4].
Discussion: *Gckr* encodes the regulatory protein of glucokinase, which binds to and inhibits glucokinase in liver and probably pancreatic islet [5,6]. This inhibitory interaction is promoted by fructose-6-phosphate and relieved by fructose-1-phosphate, the product of ketohexokinase (KHK). The postulated metabolic link between KHK and GCKR therefore makes the co-localization of their genes in human, and as shown here also in mouse, noteworthy. YAC contig mapping indicates the phosphatase gene PPP1CB to lie <400 kb from human KHK (gene order PPP1CB–KHK–GCKR) [1,4]. However, the mouse homolog of human PPP1CB maps, like several other human genes located on 2p, to mouse Chr 12D [7]. The co-localization of *Gckr* and *Khk* despite their non-syntenic relationship with PPP1CB, which is adjacent in human, adds circumstantial support to the possibility of co-ordinate regulation of *Gckr* and *Khk*.

Acknowledgments: We are grateful to Dr. C.M. Abbott for advice on use of the backcross resource. Work in the authors' laboratory is supported by grants from the Medical Research Council (G9403693MB) and the Wellcome Trust (046130).

References

1. Hayward BE, Warner JP, Dunlop N, Fantes J, Intody S, Leek J, Markham AF, Bonthron DT (1997) Biochem Soc Trans 25, 140–145
2. Donaldson IA, Doyle TC, Matas N (1993) Biochem J 291, 179–186
3. Hayward BE, Bonthron DT. Submitted; accession number Y09335.
4. Hayward BE, Fantes JA, Warner JP, Intody S, Leek JP, Markham AF, Bonthron DT (1996) Mamm Genome 7, 454–458
5. Van Schaftingen E (1989) Eur J Biochem 179, 179–184
6. Malaisse WJ, Malaisse-Lagae F, Davies DR, Vandercammen A, Van Schaftingen E (1990) Eur J Biochem 190, 539–545
7. Saadat M, Kakinoki Y, Mizuno Y, Kikuchi K, Yoshida MC (1994) Jpn J Genet 69, 697–700

Correspondence to: P.J. Wightman

BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS **242,** 407–412 (1998)
ARTICLE NO. RC977977

cDNA Cloning, Genomic Organization, and Chromosomal Localization of a Novel Human Gene That Encodes a Kinesin-Related Protein Highly Similar to Mouse Kif3C

Elizabeth A. R. Telford,[1] Patrick Wightman,* Jack Leek, Alexander F. Markham, Nicholas J. Lench, and David T. Bonthron*

*Molecular Medicine Unit, Clinical Sciences Building, St. James's University Hospital, University of Leeds, Leeds, LS9 7TF, United Kingdom; and *Human Genetics Unit, University of Edinburgh, Molecular Medicine Centre, Western General Hospital, Edinburgh, EH4 2XU, United Kingdom*

Received December 1, 1997

We report the cloning and characterization of a novel human kinesin-like gene with strong homology to the mouse kinesin *Kif3c*. The full-length cDNA contains an open reading frame of 2382 nucleotides encoding a predicted 793 amino acid peptide that includes a 389 amino acid motor domain conserved among other kinesins. PCR and DNA sequence analysis of PAC clones containing the human *KIF3C* sequence revealed that the gene contains 8 exons. All introns have the conserved GT and AG dinucleotides present at their donor and acceptor sites, respectively. We have localized *KIF3C* to chromosome band 2p23 by fluorescence in situ hybridization.

Kinesin heavy chain and kinesin-related proteins (KRPs) constitute a superfamily of molecular motors that utilize energy derived from the hydrolysis of ATP to translocate vesicles and organelles along microtubules (for review see ref 1). They thus play important roles in various intracellular transport events. Kinesin is a tetramer composed of two identical heavy chains (110-120KD) and two identical light chains (60-70KD) (2-3). Members of the KRP family share a similar structure which can be subdivided into three domains (4). The globular amino-terminal region forms a motor domain, which contains consensus sites for putative ATP- and microtubule-binding. The central region folds into a long α-helical coiled coil and forms the rod domain which allows heterodimerization (5). Finally, the carboxy-terminal globular domain interacts with the light chains and possibly other cytoplasmic components. This is assumed to be the binding site for specific "cargo" proteins. Amino acid sequence analysis of different KRPs from a variety of species revealed that the motor domain is highly conserved, usually 35-45% identical between all kinesin superfamily members (6), whereas the rod and tail domains are more divergent. Phylogenetic analysis of the conserved motor domains groups the kinesin proteins into a number of subfamilies, the members of which have more closely related motor domain sequences, exhibit sequence similarity outside of the motor domain and share common molecular organization and cellular function (7).

Here we report the isolation, characterisation, pattern of tissue expression and chromosomal localization of a novel human gene that encodes a protein similar to members of the KIF3 family of kinesin-related proteins.

METHODS

cDNA Cloning and sequencing. An I.M.A.G.E clone 28784 (8) containing sequence from EST stSG4510 was obtained from Research Genetics and sequenced. Oligonucleotides were synthesized from this sequence (dGGAGATCCAGGACCAGCATG; dGCTGTCCAATCGCATGAGCC, respectively) and a 442 bp cDNA probe (corresponding to nucleotides 1953 to 2394 of human *KIF3C;* Figure 1) was amplified by PCR, radiolabeled using a random primer DNA labeling kit (Boehringer) and used to screen a human fetal brain cDNA library (Clontech). Sequence analysis of a clone containing a 4kb insert revealed the cDNA to be a hybrid of an unrelated human sequence fused to a kinesin-like open reading frame (Figure 1, nt 874-2320).

To determine the 5′ end of the cDNA sequence RACE amplifications were performed as previously described (9). 5μg of total RNA from human fibroblasts was mixed with the reverse *KIF3C*-specific primer (dGGCCCCAGTGTGGCTACC, nt 1167-1184) and reverse transcribed with 200U Superscript II reverse transcriptase (Gibco BRL) at 42°C for 50 minutes. A poly(A) tail was added to the 3′-end of the cDNA and tailed cDNA was amplified with a d(T)-adapter

[1] Corresponding author. Fax: 044 113 2444475. E-mail: mmeeart@stjames.leeds.ac.uk.

0006-291X/98 $25.00

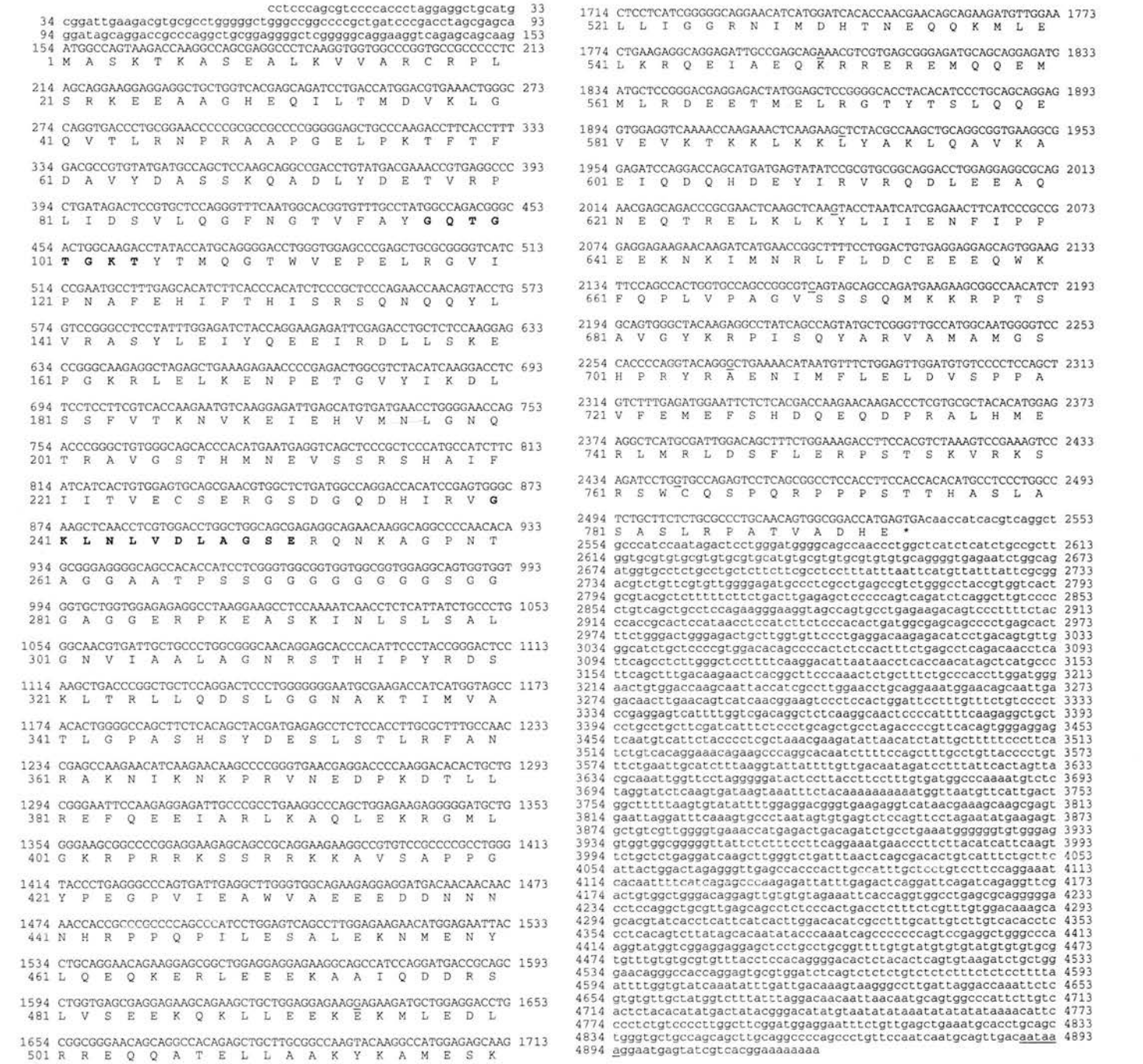


FIG. 1. The complete *KIF3C* cDNA sequence and its putative translation product. The termination signal is marked with an asterisk and the polyA addition signal is underlined. The conserved ATP/GTP-binding motif and kinesin motor domain signature are boldface. The first base of exons 2–8 is underlined.

primer and the reverse primer (dGTGTGGCTACCATGATGGTC, nt 1158-1177). The products of 5′ RACE reactions were analysed by Southern blot hybridization with ^{32}P-labelled primer (dGGAGAGAG-GCCTAAGGAAG, nt 1003-1021) and an 820 bp product was cloned into the TA vector (Invitrogen) and sequenced with vector- and kinesin-specific oligonucleotide primers. Multiple PCR-derived cDNA clones were used to check for errors in the nucleotide sequence produced by amplification procedures. The remaining 357 bp from the 5′ end and 1181bp from the 3′ end of the gene were sequenced from PAC genomic DNA.

DNA sequencing was performed on both strands of cloned cDNA using either the Thermosequenase cycle sequencing kit (Amersham) or AmpliTaq cycle squencing kit (Perkin Elmer) and the ABI Prism 377 DNA sequencer.

```
mokif3b     ~~~~~~~~~~ MSKLKSSESV RVVVRCRPMN GKEKAASYDK VVDVDVKLGQ VSVKNPKGTS HEMPKTFTFD AVYDWNAKQF ELYDETFRPL VDSVLQGFNG
S.pur95K    ~~~~~~~~~~ MSK.KSAETV KVVVRCRPMN SKEISQGHKR IVEMDNKRGL VEVTNPKGPP GEPNKSFTFD TVYDWNSKQI DLYDETFRSL VESVLQGFNG
hukif3c     ~~~~~~~~~M ASKTKASEAL KVVARCRPLS RKEEAAGHEQ ILTMDVKLGQ VTLRNPRAAP GELPKTFTFD AVYDASSKQA DLYDETVRPL IDSVLQGFNG
mokif3a     ~~~~~MPINK SEKPESCDNV KVVVRCRPLN EREKSMCYRQ AVSVDEMRGT ITV.HKTDSS NEPPKTFTFD TVFGPESKQL DVYNLTARPI IDSVLEGYNG
Dros68D     MSAKSRRPGT GSSQTPNECV QVVVRCRPMS NRERSERSPE VVNVYPNRGV VELQNVVDGN KEQRKVFTYD AAYDASATQT TLYHEVVFPL VSSVLEGFNG
consensus   ---------- ---------- -VVVRCRP-- --E------- --------G- ---------- -E----FT-D --------Q- --Y------- --SVL-G-NG

mokif3b     TIFAYGQTGT GKTYTMEGVR GDPEKRGVIP NSFDHIFTHI SRSQ.NQQYL VRASYLEIYQ EEIRDLLSKD QTKRLELKER PDTGVYVKDL SSFVTKSVKE
S.pur95K    TIFAYGQTGT GKTFTMEGVR SNPELRGVIP NSFEHIFTHI ARTQ.NQQFL VRASYLEIYQ EEIRDLLAKD QKKRLDLKER PDTGVYVKDL SSFVTKSVKE
hukif3c     TVFAYGQTGT GKTYTMQGTW VEPELRGVIP NAFEHIFTHI SRSQ.NQQYL VRASYLEIYQ EEIRDLLSKE PGKRLELKEN PETGVYIKDL SSFVTKNVKE
mokif3a     TIFAYGQTGT GKTFTMEGVR AVPGLRGVIP NSFAHIFGHI AKAEGDTRFL VRVSYLEIYN EEVRDLLGKD QTQRLEVKER PDVGVYIKDL SAYVVNNADD
Dros68D     CIFAYGQTGT GKTFTMEGVR GNDELMGIIP RTFEQIWLHI NRTE.NFQFL VDVSYLEIYM EELRDLL.KP NSKHLEVRER .GSGVYVPNL HAINCKSVED
consensus   --FAY**GQTGT** **GKT**-TM-G-- ------G-IP --F-----HI ---------L V--SYLEIY- EE-RDLL--- ----L---E- ---GVY---L ----------

mokif3b     IEHVMNVGNQ NRSVGATNMN EHSSRSHAIF VITIECSEVG LDGENHIRVG KLNLVDLAGS ERQAKT.... ..... ..... GAQGERLKEA
S.pur95K    IEHVMTVGNN NRSVGSTNMN EHSSRSHAIF IITIECSELG VDGENHIRVG KLNLVDLAGS ERQAKT.... ..... ..... GATGDRLKEA
hukif3c     IEHVMNLGNQ TRAVGSTHMN EVSSRSHAIF IITVECSERG SDGQDHIRVG KLNLVDLAGS ERQNKAGPNT AGGAATPSSG GGGGGGGSGG GAGGERPKEA
mokif3a     MDRIMTLGHK NRSVGATNMN EHSSRSHAIF TITIECSEKG VDGNMHVRMG KLHLVDLAGS ERQAKT.... ..... ..... GATGQRLKEA
Dros68D     MIKVMQVGNK NRTVGFTNMN EHSSRSHAIF MIKIEMCDTE T...NTIKVG KLNLIDLAGS ERQSKT.... ..... ..... GASAERLKEA
consensus   ----M--G-- -R-VG-T-MN E-SSRSHAIF -I--E----- ---------**G KLNL-DLAGS E**RQ-K----- ---------- ---------- GA---R-KEA

mokif3b     TKINLSLSAL GNVISALVDG KSTHIPYRDS KLTRLLQDSL GGNAKTVMVA NVGPASYNVE ETLTTLRYAN RAKNIKNKPR VNEDPKDALL REFQEEIAR
S.pur95K    TKINLSLSAL GNVISALVDG KSSHIPYRDS KLTRLLQDSL GGNAKTVMVA NMGPASYNFD ETITTLRYAN RAKNIKNKPK INEDPKDALL REFQEEISR
hukif3c     SKINLSLSAL GNVIAALAGN RSTHIPYRDS KLTRLLQDSL GGNAKTIMVA TLGPASHSYD ESLSTLRFAN RAKNIKNKPR VNEDPKDTLL REFQEEIAR
mokif3a     TKINLSLSTL GNVISALVDG KSTHVPYRNS KLTRLLQDSL GGNSKTMMCA NIGPADYNYD ETISTLRYAN RAKNIKNKAR INEDPKDALL RQFQKEIEE
Dros68D     SKINLALSSL GNVISALAES .SPHVPYRDS KLTRLLQDSL GGNSKTIMIA NIGPSNYNYN ETLTTLRYGS RAKSIQNQPI KNEDPQDAKL KEYQEEIER
consensus   -KINL-LS-L GNVISAL--- -S-H-PYR-S KLTRLLQDSL GGN-KT-M-A --GP------ E---TLRY-- RAK-I-N--- -NEDP-D--L ---QEEI--
```

FIG. 2. Comparison of the human KIF3C motor domain with homologous KIF3 proteins; mouse KIF3A (P28741), KIF3B (D26077), sea urchin 95K (P46871), and Drosophila 68D(P46867). The ATP/GTP-binding motif and kinesin motor domain signature are again boldface.

The BLAST algorithm (10) was used to search for homologies or identities between sequences identified and sequences entered in the GenBank database.

The cDNA sequence has been submitted to GenBank and assigned accession number AF035621.

Structural analysis. A PAC human genomic library (obtained from the HGMP Resource Centre, Hinxton, U.K) was screened by PCR using the stSG4510-specific primers (dCCTAGAGACATTTGGGCCA and dTTGCCTGTTACCCCTGTTTC; nt 3678-3697 and 3557-3576 respectively) and two positive clones were obtained. Oligonucleotides were designed from the cDNA and used to obtain sequence data across the exon/intron boundaries from PAC genomic DNA. Sizes of introns were determined by amplification across each intron using PAC genomic DNA as template and primers located within adjacent exons. The fragments amplified were analysed on a 1.5 % agarose gel and sized against a GibcoBRL 1 kb DNA ladder. Genomic sequences have been submitted to the EMBL sequence database and assigned accession numbers AJ002223-AJ002229.

Northern blot analysis. To examine expression of the human KIF3C gene, a multiple tissue Northern blot was hybridized at 65°C overnight in a solution consisting of 0.5M sodium phosphate buffer pH 7.2, 1% BSA, 7% SDS, 1mM EDTA and 50μg/ml denatured herring sperm DNA with a ^{32}P-labeled 818 bp PCR product encoding the motor domain (nt 360-1177, Figure 1). Following hybridization, the blot was washed twice in 2 × SSC, 0.1% SDS at 65°C for 40 minutes.

Fluorescence in situ hybridization. YAC clones containing the human *KIF3C* sequence were identified by screening the ICI YAC library (11) by PCR with the stSG4510-specific primers. These YACs, 14I C3 and 35H B12, were shown also to contain the marker *D2S2144.* Clone 35H B12 was propagated on synthetic dextrose (SD) medium to obtain a pure, single colony which was grown to saturation in SD broth. A total yeast DNA extract was obtained using standard techniques. The probe was labeled by nick-translation with Digoxigenin-11-dUTP and 300ng used for FISH analysis as described (12). Chromosomes were identified with 4,6-diamidino-2-phenylindole-dihydrochloride (DAP1). Microscopy was performed using a Zeiss Axioskop fluorescence microscope coupled to a CCD camera and image analysis system (Vysis, UK).

Interspecific backcross mapping. Primers Kin3: dGGAGATGCAGCAGGAGATG; and Kin2 : dGGGTCTGCTCGTTCTGCG were used to amplify fragments of ~1.0 kb from C57BL/6J and *M. spretus* DNA. A polymorphic $(CT)_n$ repeat was found adjacent to the Kin2 (exon 4) end of this fragment. A third primer MKin2R (dGCATTCCATCAGTTCTCTTTCAG), on the other side of the $(CT)_n$ repeat was used in conjunction with Kin2 to characterize this polymorphism: MKin2R and Kin2 amplify a ~350 bp fragment from C57BL/6J and a ~ 400 bp fragment from *M. spretus.*

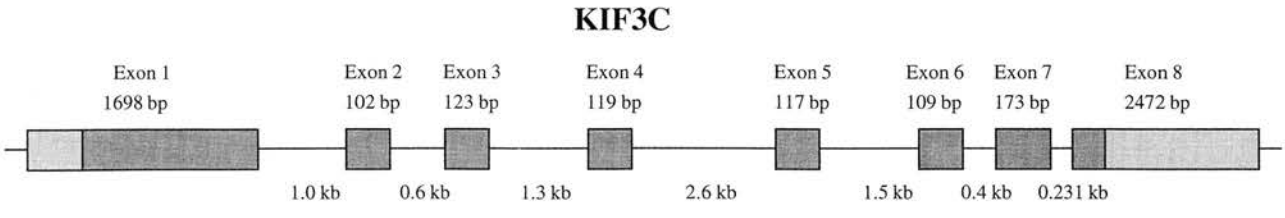


FIG. 3. Genomic organization of *KIF3C.* Exons are shown as shaded boxes, coding regions darker than untranslated regions. The diagram is not to scale; exact exon sizes (bp) and approximate intron sizes (kb) are indicated.

TABLE 1

KIF3C Exon/Intron Splice Junction Sequences

Exon1-TGCTTGCGGCCAAGTACAAGgtaagggccccagaggagct-	1.0 kb	-accctctgtgtgtccccagGCCATGGAGAGCAAGCTCCT-Exon2
Exon2-GGCAGGAGATTGCCGAGCAGgtagggcctccaggtgccag-	0.6 kb	-actgcccgtccttggcctagAAACGTCGTGAGCGGGAGAT-Exon3
Exon3-AAACCAAGAAACTCAAGAAGgtgagacgctgcagcaggac-	1.3 kb	-tggcacctgtccccacccagCTCTACGCCAAGCTGCAGGC-Exon4
Exon4-ACCCGCGAACTCAAGCTCAAgtagggcccgcagctctttt-	2.6 kb	-tccttcattcctgctcccagGTACCTAATCATCGAGAACT-Exon5
Exon5-CCACTGGTGCCAGCCGGCGTgtagtctctaacccagctgt-	1.5 kb	-atatagcctctttcctacagCAGTAGCAGCCAGATGAAGA-Exon6
Exon6-GGTCCCACCCCAGGTACAGGgtaagaagcggagaggggag-	0.4 kb	-tggcatgatattcccaccagGCTGAAAACATAATGTTTCT-Exon7
Exon7-GTCCGAAAGTCCAGATCCTGgtcagtacctccatggtccc-	0.231 kb	-ctgcgtcatctcccctgcagGTGCCAGAGTCCTCAGCGGC-Exon8

Note. Intron sequences are in lowercase and exon sequences in capital letters.

RESULTS

Sequencing KIF3C cDNA

The human EST clone 28784 (GenBank Accession No. R14361) displayed significant sequence homology to the 68D and the 95K kinesin-like proteins of *Drosophila melanogaster* and *Strongylocentrus purparatus* respectively. Nucleotide sequence analysis of the entire 1917 bp insert revealed a 720 bp ORF encoding a kinesin-like carboxy-terminal tail domain and 1197 bp of 3′UTR sequence. A composite full-length cDNA (4913 bp) was assembled from a human fetal brain cDNA clone, from 5′ RACE products which were generated from total fibroblast RNA and from genomic sequence of the large 3′UTR-containing exon. (This latter genomic sequence was shown to consist of a single exon by comparison to a large cDNA contig assembled from dbEST sequences spanning the whole region.)

The full-length cDNA is characterized by a single open reading frame of 2382 bp encoding a predicted protein of 793 amino acids (Figure 1). The overall organization of the predicted protein is similar to members of the KIF3 family of KRPs and contains an amino-terminal motor domain (residues 1-389), a central rod domain (residues 390-599) and a carboxy-terminal tail domain (residues 600-793). A putative polyadenylation signal is located 2354-2359 nucleotides downstream from the translation termination codon. The cDNA clones have a poly(A) tail added 20 nucleotides downstream of this (2378 bp downstream of the termination codon; see Figure 1).

Among the kinesin superfamily proteins described thus far, this novel sequence is most similar to a 195 bp partial cDNA encoding the murine KIF3C motor domain (GenBank Accession No AB001433; corresponding to residues 93-257 in Figure 1), the human and mouse sharing 91% and 98% identities in nucleotide and amino acid sequence respectively.

Comparison of the amino acid sequences of KIF3 motor domains revealed that the human KIF3C sequence encodes a 24 amino acid glycine-rich domain which is not conserved in other KIF3 proteins (Figure 2).

Structural Organization of the Human KIF3c Gene

The results of DNA sequence analysis of PAC clones are summarized in Figure 3 and Table 1. The human KIF3C gene spans approximately 12 kb of genomic DNA and consists of 8 exons and 7 introns, the exons ranging in size from 2472 bp to 109 bp. The first methionine codon of the open reading frame is located in exon 1, whereas the stop codon and poly(A) addition signal are located in the last exon-exon 8. Exon 8 contains 2378 bp of untranslated sequence. All introns have the consensus sequence C/T/A/AG-exon-GT(G/A) at their boundaries (Table 1).

The present gene structure is the first described for any kinesin family member. The ATP/GTP-binding site motif A (442 to 465 nt) and the kinesin motor domain signature (871 to 906 nt) are both located in exon 1. This exon includes the whole of the *N*-terminal motor and over half of the rod domain. Exons 2 and 3 encode the remaining part of the rod domain and exons 4-8 the *C*-terminal tail.

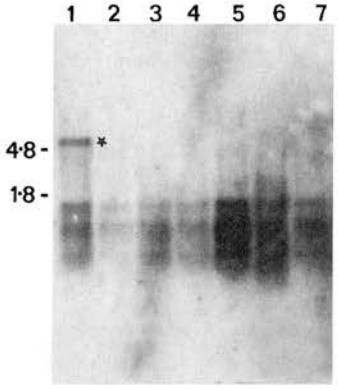


FIG. 4. Northern blot analysis of *KIF3C* expression in different human tissues. Approximately 1μg of total RNA from various adult human tissues were loaded: brain (lane 1), salivary gland (lane 2), oesophagus (lane 3), trachea (lane 4), heart (lane 5), lung (lane 6), stomach (lane 7). The size of the 28S (4.8 kb) and 18S (1.8 kb) ribosomal RNA subunits are shown. The largest transcript, expressed only in brain is marked with an asterisk.

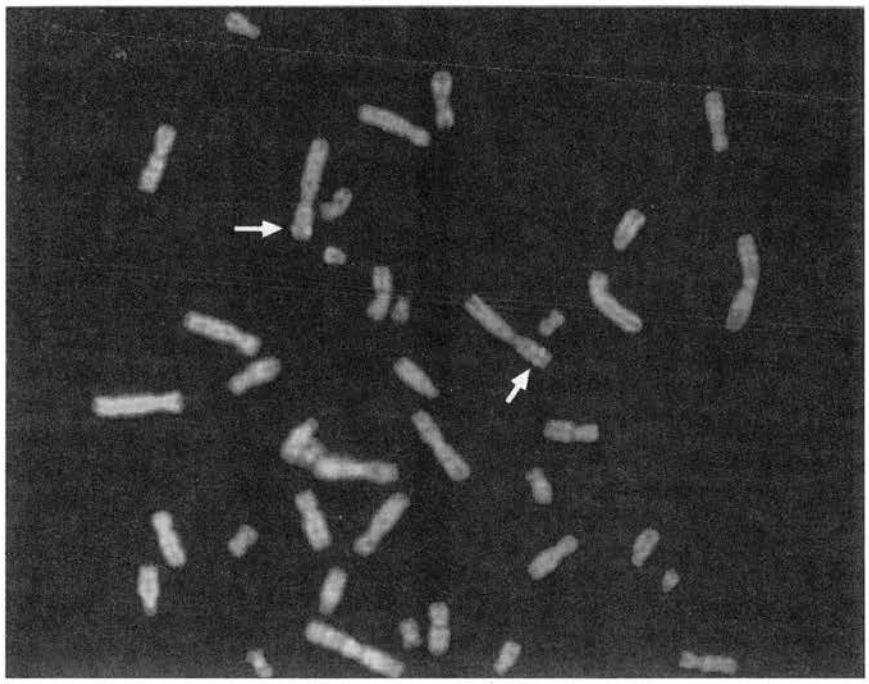


FIG. 5. Chromosome mapping of human *KIF3C* by fluorescence in situ hybridization. A signal is present on both both chromatids of chromosome 2 at a location corresponding to p23.

Expression of Human KIF3C

The tissue distribution of human KIF3C mRNA was determined by analysis of a multi-tissue Northern blot using an 818 bp motor domain PCR fragment as a probe. A transcript of approximately 5.0 kb in size was expressed in brain but could not be detected in any of the other tissues tested (Figure 4). Smaller transcripts of approximately 1.5, 1.3 and 1.0 kb were identified in all tissues with similar levels of expression.

Chromosomal Mapping of Human and Mouse KIF3C

Following hybridization of a digoxigenin-labeled *KIF3C* probe to normal human chromosomes, doublet signals at human chromosome 2p23 were observed in 25 cells (Figure 5). The distribution on 2p was as follows: 1(4), 2(21), 3(0), 4(0) chromatids per cell. This FISH result independently confirms the *KIF3C* localization obtained by radiation hybrid mapping of stsG4510 and by the co-localization with *D2S2144* in YACs 14I C3 and 35H B12.

To determine the location of the corresponding mouse gene, linkage analysis was performed. A database search was performed to identify a murine cDNA clone (W82835) highly similar to the 3′ untranslated region (nt 4349-4913) of *KIF3C*. A 458 bp genomic fragment corresponding to part of this clone was amplified from *Mus spretus* and C57BL/6J, but no polymorphic differences were seen on sequencing. Therefore, primers designed against the human *KIF3C* exon 3 and exon 4 sequences, were used to amplify fragments of ~1.0 kb from C57BL/6J and *M. spretus* DNA and the PCR products' ends sequenced. The sequence from the Kin2 primer end crossed the splice acceptor site of intron 3 and within this intron a polymorphic $(CT)_n$ repeat region was found. Using the Jackson Laboratory Backcross DNA Panel Mapping Resource, 94 animals from the interspecific cross (C57BL/6JEi × SPRET/Ei) × SPRET/Ei were genotyped. The resulting linkage data placed the murine gene on chromosome 12 between *D12Mit44* and *D12Mit182*. The mouse *ApoB* gene also maps to this interval. (Human *APOB* is localized to 2p23-p24). Therefore, *KIF3C/Kif3c* map to a region of conserved human-mouse synteny.

DISCUSSION

In this report we describe the cloning, structural analysis, expression and localization of a novel human kinesin-like gene. The protein encoded is most similar to members of the KIF3 subfamily of kinesins which includes the sea urchin 95K, *Drosophila* 68D, and mouse *Kif3a* and *Kif3b* KRPs (13-16). The overall homology of these proteins with the human sequence is 64%, 47%, 48% and 72% respectively and that relative to the motor domain (amino acid residues 93-251, Figure 1) increases to 80%, 56%, 63% and 82%. The strongest homology was to a sequence encoding 65 amino acids of the mouse Kif3c motor domain (17-18). The mouse and human peptides share 98% identity and differ only at a single residue. DNA and protein sequence

homology and localization of the human kinesin-like gene to chromosome 2p23, a region syntenic with mouse chromosome 12, suggests that the novel gene described in this paper is the human homologue of mouse *Kif3c*.

Although genes of the kinesin family are ubiquitously expressed, some of its members display a restricted tissue distribution, including Kif1, Kif3 and Kif5, which are expressed almost exclusively in murine brain (19), suggesting that certain kinesins may perform tissue-specific functions. Northern blot analysis, using a probe encoding the *KIF3C* motor domain, identified a transcript of approximately 5.0 kb which was expressed in human brain but not in any of the other tissues tested. This is in agreement with studies showing mouse *Kif3c* to be expressed mainly in neural tissues such as brain, spinal cord and retina (18). The fact that RACE products from near the 5′ end of the gene could be generated from total fibroblast RNA suggests a basal level of expression of the larger transcript in tissue other than brain, which is detectable by RT-PCR but not Northern hybridization. Smaller transcripts of approximately 1.0, 1.3 and 1.5 kb appear clearly in this Northern blot after stringent wash conditions and were expressed at comparable levels in all tissues. It is possible that these smaller transcripts represent cross-reacting KRPs.

KIF3 subfamily proteins have been reported to be plus-end-directed microtubule motors with roles in anterograde axonal transport for membraneous organelles in neurons. Immunoprecipitation assays have shown that some members, including mouse Kif3a and 3b, assemble heterotrimeric complexes comprising two homologous but distinct KRPs associated with a non-kinesin polypeptide subunit which has been proposed to function as an adapter for cargo attachment (13, 16). Proteins with which mouse Kif3c interacts and the nature of the cargo that it transports have yet to be determined. However, localization of a gene for sensorineural non-syndromic recessive deafness, *DFNB9,* to the same chromosomal region, 2p22-23 (20), as the human KIF3C gene suggests *KIF3C* as a good candidate gene for *DFNB9*. This is supported by the fact that both non-syndromic deafness at the *DFNB2* locus (11q13) and the syndromic form of deafness Usher syndrome IB result from mutations in the myosin VIIA (*MYO7A*) gene (21-25). *MYO7A* encodes an atypical myosin which like kinesins, has an intraneuronal transport function. Genetic defects of some atypical myosins and kinesins can even cross-complement in yeast (26). Investigation of the possible role of *KIF3C* in human sensorineural deafness is currently in progress.

ACKNOWLEDGMENTS

Research in the authors' laboratories is supported by the MRC, Wellcome Trust, YCRC and NYRHA.

REFERENCES

1. Moore, J. D., and Endow, S. A. (1996) *Bioessays* **18,** 207–219.
2. Bloom, G. S., Gagner, M. C., Pfister, K. K., and Brady, S. T. (1988) *Biochemistry* **27,** 409–416.
3. Kuznetzov, S. A., Vaisberg, Y. A., Shanina, N. A., Magretova, N. N., Chernyak, V. Y., and Gelfand, V. I. (1988) *EMBO J.* **7,** 353–356.
4. Yang, J. T., Laymon, R. A., and Goldstein, L. S. B. (1989) *Cell* **56,** 879–889.
5. de Cuevas, M., Tao, T., and Goldstein, L. S. B. (1992) *J. Cell Biol.* **116,** 957–965.
6. Goldstein, L. S. B. (1993) *Annu. Rev. Genet.* **27,** 319–351.
7. Goodson, H. V., Kang, S. J., and Endow, S. A. (1994) *J. Cell Sci.* **107,** 1875–1884.
8. Lennon, G., Auffray, C., Polymeropoulos, M., and Soares, M. B. (1996) *Genomics* **33,** 151–152.
9. Frohman, M. A., Dush, M. K., and Martin, G. R. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 8998–9002.
10. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
11. Anand, R., Riley, J. H., Butler, R., Smith, J. C., and Markham, A. F. (1990) *Nucleic Acids Res.* **18,** 1951–1956.
12. Pinkel, D., Straume, T., and Gray, J. W. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 2934–2938.
13. Cole, D. G., Chinn, S. W., Wedaman, K. P., Hall, K., Vuong, T., and Scholey, J. M. (1993) *Nature* **366,** 268–270.
14. Stewart, R. J., Pesavento, P. A., Woerpal, D. N., and Goldstein, L. S. B. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 8470–8474.
15. Kondo, S., Sato-Yoshitake, R., Noda, Y., Aizawa, H., Nakata, T., Matsuura, Y., and Hirokawa, N. (1994) *J. Cell Biol.* **125,** 1095–1107.
16. Yamazaki, H., Nakata, T., Okada, Y., and Hirokawa, N. (1995) *J. Cell Biol.* **130,** 1387–1399.
17. Nakagawa, T., Tanaka, Y., Matsuoka, E., Kondo, S., Okada, Y., Noda, Y., Kanai, Y., and Hirokawa, N. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 9654–9659.
18. Yang, Z., Hanlon, D. W., Marszalek, J. R., and Goldstein, L. S. B. (1997) *Genomics* **45,** 123–131.
19. Aizawa, H., Sekine, Y., Takemura, R., Zhang, Z., Nangaku, M., and Hirokawa, N. (1992) *J. Cell Biol.* **119,** 1287–1296.
20. Chaib H., Place, C., Salem, N., Chardenoux, S., Vincent, C., Weissenbach, J., El- Zir, E., Loiselet, J., and Petit, C. (1996) *Hum. Mol. Genet.* **5,** 155–158.
21. Weil, D., Blanchard, S., Kaplan, J., Guildford, P., Gibson, F., Walsh, J., Mburu, P., Varela, A., Levilliers, J., Weston, M. D., Kelley, P. M., Kimberling, W. J., Wagenaar, M., Levi-Acobas, F., Larget-Plet, D., Munnich, A., Steel, K. P., Brown, S. D. M., and Petit, C. (1995) *Nature* **374,** 60–61.
22. Liu, X-Z., Walsh, J., Mburu, P., Kendrick-Jones, J., Cope, M. J. T. V., Steel, K. P., and Brown, S. D. M. (1997) *Nature Genetics* **16,** 188–190.
23. Weil, D., Kussel, P., Blanchard, S., Levy, G., Levi-Acobas, F., Drira, M., Ayadi, H., and Petit, C. (1997) *Nature Genetics* **16,** 191–193.
24. Avraham, K. B., Hasson, T., Steel, K. P., Kingsley, D. M., Russell, L. B., Mooseker, M. S., Copeland, N. G., and Jenkins, N. A. (1995) *Nature Genetics* **11,** 369–375.
25. Gibson, F., Walsh, J., Mburu, P., Varela, A., Brown, K. A., Antonio, M., Beisel, K. W., Steel, K. P., and Brown, S. D. M. (1995) *Nature* **374,** 62–64.
26. Lillie, S. H., and Brown, S. S. (1992) *Nature* **356,** 358–361.