Stellingen behorende bij het proefschrift

Locus Heterogeneity and the Molecular Basis of Tuberous Sclerosis

Bart Janssen

Rotterdam, 3 mei 1995

Locus Heterogeneity and the Molecular Basis of Tuberous Sclerosis

Gedrukt bij Offsetdrukkerij Ridderprint B.V., Ridderkerk

ISBN 90-9008185-2

Front cover: H. Douben
N. Janssen
R. Koppenol

© Copyright: L.A.J. Janssen, BMJ Publishing Group, New York Academy of Sciences, S. Karger AG, Springer-Verlag GmbH & CO. KG, Cell Press. No part of this book may be reproduced or transmitted in any form by any means, electronic or mechanical, without the permission of the author (Department of Clinical Genetics, L.A.J. Janssen, Dr. Molewaterplein 50, 3015 GE Rotterdam), or, when appropriate, of the publishers of the publications.

Locus Heterogeneity and the Molecular Basis of Tuberous Sclerosis

(Locus heterogeniteit en
de moleculaire basis van tubereuze sclerosis)

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR AAN DE
ERASMUS UNIVERSITEIT ROTTERDAM OP GEZAG VAN DE
RECTOR MAGNIFICUS
PROF. DR. P.W.C. AKKERMANS M.A.
EN VOLGENS BESLUIT VAN HET COLLEGE VOOR PROMOTIES
DE OPENBARE VERDEDIGING ZAL PLAATSVINDEN OP
WOENSDAG 3 MEI 1995 OM 15.45 UUR

DOOR

LAMBERTUS ANTONIUS JACOBUS JANSSEN
GEBOREN TE HAARLEM

PROMOTIECOMMISSIE

Promotor:	Prof. Dr. D. Lindhout
Co-Promotor:	Dr. D.J.J. Halley
Overige leden:	Prof. Dr. F.T. Bosman Prof. Dr. G.J.B. van Ommen Prof. Dr. M.S. Povey

The research described in this thesis has been performed in the Department of Clinical Genetics of the Academic Hospital Rotterdam and the Faculty of Medicine and Health Sciences of the Erasmus University Rotterdam, The Netherlands.

This study was financially supported by the "Praeventiefonds" (grant nr. 002817230) and the "Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO)" (grant nr. 901-04-121). Katwijk Farma BV, Sanofi Winthrop v.o.f., Yamanouchi Pharma B.V., the Institute of Clinical Genetics Rotterdam and NWO are gratefully acknowledged for financial support to publish this thesis.

Opgedragen aan Helma, Jaap en Nienke,
en aan hen van wie wij onze genen hebben geërfd.

CONTENTS

Chapter 1: Introduction

Chapter 2: Methods for linkage analysis under locus heterogeneity

Chapter 3: Linkage analysis in tuberous sclerosis

CHAPTER 1

INTRODUCTION

1.1

SCOPE OF THIS THESIS

Genetic diseases can be subdivided by their underlying mechanism. A disease may be caused by a single mutation in a specific gene (homogeneity), or a single mutation affecting a member of a group of disease related genes (heterogeneity). Furthermore, a disease may be multifactorial, which indicates that mutations in multiple genes are necessary to obtain the disease, or that environmental factors may act as risk factors in genetically predisposed individuals. It is obvious that the etiology of a single gene disorder (homogeneity) can be unravelled more easily than the etiology of a genetically heterogeneous or multifactorial trait.

The initial aim of the project described in this thesis was to isolate one or more genes responsible for the heterogeneous disorder tuberous sclerosis (TSC). TSC shares with many other genetic diseases that diagnosis and treatment are hampered by a lack of knowledge about the underlying biochemical defect. More insight into the molecular aspects of its etiology may help to overcome this problem and will perhaps lead to the development of a therapy. It is therefore important to identify the responsible genes and to study their function in the cell. On the short term patients can be helped with molecular genetic diagnostics and more accurate genetic counselling, as soon as a responsible gene has been identified. On the long term patients may benefit from clarification of the underlying mechanism by improved therapeutical intervention.

For an increasing number of disorders the trait causing gene is identified by 'positional cloning'. Positional cloning strategies are aimed at finding the responsible gene via positional information. For simplicity, it is often assumed that defects in a single gene are responsible for a certain trait in affected individuals from all families under investigation. In reality, however, locus heterogeneity seems to be the rule, rather than the exception. This can be explained if we consider that many cellular processes are regulated by sequential pathways, requiring several functional proteins. Moreover, many proteins in the cell are multimers, consisting of subunits encoded by different genes. If a pathway or multimeric protein is malfunctioning, this may be due to a mutation at one of several gene loci.

A challenging aspect of the search for the TSC genes, by positional cloning, is the genetic (locus) heterogeneity. This locus heterogeneity prohibited a straightforward approach. It seemed almost inevitable to extend the research towards the more theoretical aspects of heterogeneity analysis, because powerful methods for the analysis of multiple candidate regions were not available. Part of this thesis describes research on the statistical methods (Chapter 2), which also have relevance to other genetically heterogeneous disorders. Throughout all chapters TSC serves as the prime example of a heterogeneous disorder.

Subsequent chapters describe the efforts to localise and isolate the genes responsible for TSC. Two different international consortia were formed aimed at two different stages of the positional cloning process. The first collaboration consisted of eight TSC sclerosis mapping centres: Boston, Cardiff, Durham, Erlangen, Houston, Irvine, London and Rotterdam. The objective of this consortium was to perform a critical evaluation of a series of linkage claims. As demonstrated in Chapter 2.2, the family material collected by the consortium was more than sufficient to study linkage in the presence of locus heterogeneity, provided that a novel approach for heterogeneity analysis was used. Chapter 3 comprises the results of the linkage studies. The last part of this chapter describes how the most favourable linkage results were obtained: 14 families from Cardiff and Rotterdam were used to demonstrate linkage to both chromosomes 9 and 16. The fourth chapter is entirely dedicated to the next steps in the search for the TSC1 gene on chromosome 9: the assembly of YAC and cosmid contigs and the examination of expressed genes. The search for the TSC2 gene on chromosome 16 is described in Chapter 5. This effort required the formation of another consortium, consisting of groups from Cardiff, Leiden, Oxford and Rotterdam resulting in the 'Identification and characterisation of the tuberous sclerosis gene on chromosome 16'.

The molecular work described here led to the mapping and isolation of a large number of genes. One of these genes turned out to be responsible for another heterogeneous disorder: polycystic kidney disease. This and other consequences of the molecular and methodological research, like the prospects for molecular diagnosis and the relevance of our work towards other heterogeneous disorders, are discussed in Chapter 6. Key questions to this discussion are: 'How many TSC genes may exist?', 'Can additional TSC genes be mapped by linkage analysis?', 'Where does the TSC1 gene map?', 'What steps have to be taken towards its isolation?', 'Can the statistical methods described here be used to map other heterogeneous disorders as well?', 'What are the current diagnostic possibilities in TSC?' and 'What is the function of TSC2 and its gene product?'.

In summary, this thesis focuses on the general aspects of positional cloning when complicated by locus heterogeneity. It specifically describes the positional cloning strategies as applied to the heterogeneous disorder tuberous sclerosis: the methodological novelties introduced in the linkage analysis, the results of the linkage analysis, and beyond that the isolation of one of the genes involved in this tragic disease.

1.2

PHENOTYPICAL ASPECTS OF TUBEROUS SCLEROSIS: A DISORDER WITH LOCUS HETEROGENEITY AND VARIABLE EXPRESSION

A few disorders are generally regarded as classical examples of diseases showing locus heterogeneity. Well-known examples are elliptocytosis (Morton, 1956; Ott, 1983), tuberous sclerosis (Edwards, 1990, Evans and Harris 1992) and familial breast cancer (Evans and Harris, 1992; Easton et al., 1993). One of these disorders - tuberous sclerosis - has been the subject of investigations in our group since 1988. The disease is characterized by the growth of benign tumours (hamartomas) and malformations (hamartias) in one or more organs. The brain, skin, heart, eye and kidney are often affected, but other organs may also be involved. The disease was first recognised in 1880 by Désiré-Magloire Bourneville. He described the potato-like sclerotic lesions (tubers) and introduced the name 'sclérose tubéreuse' (tuberous sclerosis) (Gomez, 1988). Today tuberous sclerosis is still the most widely used name. The obsolete name epiloia officially stands for 'epilepsy plus anoia' (mental retardation). Epiloia has also been proposed to be a useful acronym and mnemonic for 'epilepsy, low intelligence and adenoma sebaceum' (McKusick, 1990), but in fact the name is no longer used. Occasionally, especially in French speaking countries, the disorder is designated Bourneville's disease in honour of the French neurologist. Quite confusing is the abbreviation TS for tuberous sclerosis, because it is often used for another genetic disorder affecting the central nervous system: Tourette syndrome. In genetic literature, the disorder is mostly designated as tuberous sclerosis complex (TSC), emphasizing the multiplicity of organs involved.

Tuberous sclerosis is a hereditary disorder with an often severe phenotype

Since the initial description of TSC by Bourneville, many investigators have contributed to an improved characterisation of the disease. Important in this respect was the work done by Pringle, Vogt and Van der Hoeve (reviewed by Gomez (1988)). Perhaps the most significant contribution to this field has been made by Gomez. His book 'tuberous sclerosis' certainly provides the most comprehensive description of the disease. (Gomez, 1988). TSC was first recognised to be a hereditary disorder in 1913 by Berg (Gomez, 1988). Since then a large number of familial cases have been reported, all showing an autosomal dominant pattern of inheritance, with high - although not complete - penetrance (Baraitser and Patton, 1985; Webb and Osborne, 1991).

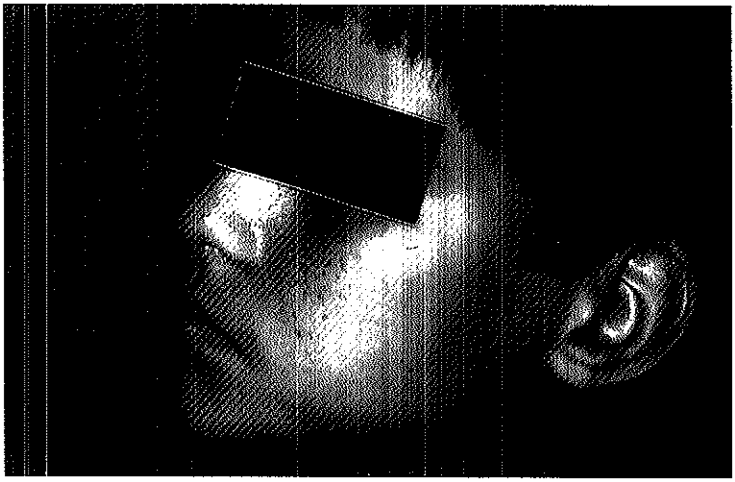


Figure 1.1: Patient with facial angiofibromas. This patient has the unbalanced form of the t(16;22)(p13.3;q11.21) translocation shown in Figure 1.5, which was of crucial importance to the work described in the Chapters 5.1 and 6.2.

Important symptoms of tuberous sclerosis are mental retardation and epilepsy. When present, these symptoms - often associated with autism - strongly affect the patient's life and that of other family members. Already in 1908 Heinrich Vogt associated the occurrence of seizures and mental retardation with a facial symptom, called adenoma sebaceum or facial angiofibroma (Figure 1.1). A study on 300 TSC patients in Rochester revealed that only 6% showed none of the symptoms of Vogt's triad (Gomez, 1988). Mental retardation occurred in 48% of the patients, while 92% had a history of seizures. Other studies reported 52%-68% of the patients to be mentally retarded (Hunt and Lindenbaum, 1984; Osborne et al., 1991; Ahlsén et al., 1994). Recent work by Hunt and Shepherd revealed autistic features among approximately 50% of the Scottish TSC population (Hunt and Shepherd, 1993). All retarded patients had a history of seizures at some age, usually in the first year of life, or somewhat later. Characteristic lesions in the central nervous system (CNS) are cortical tubers, subependymal nodules (Figure 1.2) and giant cell astrocytomas. Histologically, these are all characterised by hypertrophic neurons and bizarre - often enormously enlarged - astrocytes.

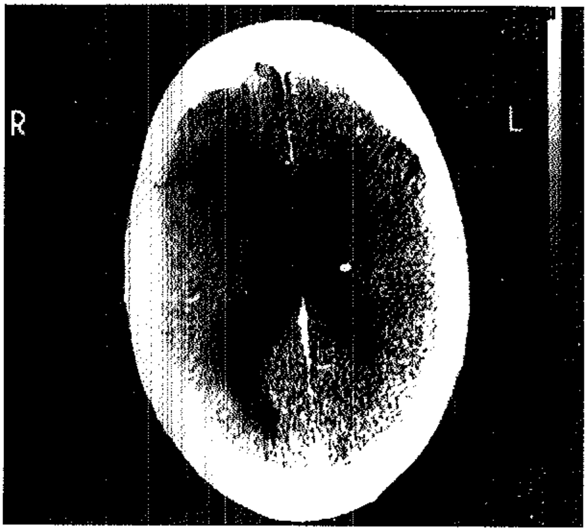


Figure 1.2: CT scan demonstrating a calcified nodule (white spot), characteristic for TSC. The patient is a member of family 4219 and therefore likely to be a TSC1 (chromosome 9-linked) case (Chapter 3.3). There are no significant clinical differences between TSC1 and TSC2 type patients.

A large proportion of TSC patients have kidney complications. 40%-80% of TSC patients show renal cysts and angiomyolipomas, generally occurring bilateral.

Although pathologically distinguishable, the renal cysts resemble those seen with autosomal dominant polycystic kidney disease. Although classified as benign, angiomyolipomas often have to be treated, in order to prevent a fatality from sudden bleeding (Steiner et al., 1993; Fleury, 1989; Van Baal et al., 1994).

Involvement of the kidneys is second to that of the CNS as a cause of mortality in adults and adolescents. In females, involvement of the lungs may also be fatal, while in children cardiac failure due to rhabdomyoma is a frequent cause of death. (Fenoglio et al., 1976; Gomez, 1988; Shepherd and Gomez, 1991). About 60% of all patients have a history of rhabdomyoma, while males have been reported to be more frequently affected with rhabdomyoma than females (Fenoglio et al., 1976).

The skin findings in tuberous sclerosis are often quite obvious. Most patients (90%) show depigmented spots, called hypomelanotic macules, which are most often already visible at birth (Gomez, 1988). Other skin findings are facial angiofibromas (see Figure 1.1), ungual fibromas, shagreen patches and fibrous forehead plaques.

Apart from the organs discussed so far, many other organs may be involved. As shown in Table 1.2, some manifestations, like the ophthalmological findings, are of diagnostic importance, while others are not. Only very few organs have never been shown to be involved in TSC: the skeletal muscles, the spinal roots and ganglia, the peripheral nervous system and probably the spinal cord (Gomez, 1988).

The prevalence of tuberous sclerosis

Studies pertaining to the epidemiology of TSC are few. The resulting prevalence estimates, as listed in Table 1.1, are likely to be underestimates of the real figure, since mild cases of TSC are often not identified. The authors of the studies included in Table 1.1 were aware of the fact that their prevalence figures were underestimates. In the last Oxford study and the Wessex study an increased proportion of mentally retarded children under the age of 5 was seen. This can only be explained by assuming that a large number of non-retarded affecteds were not (yet) referred to a participating clinician (Osborne et al., 1991). Ahlsén reasoned that the real figure could best be approximated by studying a young adolescent population, 11 to 15 years old. This group is regularly surveyed, while symptoms of TSC that are often not present before puberty, such as facial angiofibroma, will be recognisable among people of this age category (Ahlsén et al., 1994). The results

Table 1.1:
The prevalence of TSC.

Year	Population/cohort	prevalence	reference
1935	mentally handicapped	1/30,000	Gunter & Penrose
1968	idem	1/23,000	Zaremba
1956	Northern Ireland	1/150,000	Stevenson and Fisher
1968	Oxford region, England	1/100,000	Nevin and Pearce
1971	Hong Kong	1/170.000	Singer
1984	Oxford (age<65)	1/29,000	
	idem (age<30)	1/20,000	
	idem (age<5)	1/15,000	Hunt and Lindenbaum
1985	Rochester, Minnesota	1/10,000	Wiederholt et al.
1989	West-Scotland	1/27,000	
	idem (age<10)	1/10,000	Sampson et al.
1991	Wessex, England	1/34,000	
	idem (age<5)	1/15,000	Osborne et al.
1994	Western Sweden	1/30,000	
	idem (11<age<15)	1/6,800	Ahlsén et al.

of this study - a prevalence of 1 in 6800 - make it likely that the disease occurs much more often than previously presumed, although this figure will probably also be an underestimate. The prevalence of the TSC trait at birth will undoubtedly be even higher, perhaps 1/5,800 (Osborne et al., 1991).

Although TSC is a hereditary disorder with high penetrance, most affected children have normal parents. It has been estimated that about two-third of all cases occur *de novo* (Hunt and Lindenbaum, 1984; Sampson et al., 1989; Osborne et al., 1991). The phenotypes of these sporadic cases are caused by new mutations. Therefore, for this group of patients, the risk of getting an affected child is equal to that of familial cases: 50%. The high proportion of *de novo* cases reflects the high mutation rate of $2.5 \cdot 10^{-5}$ (Sampson et al., 1989), which is comparable with that of Duchenne muscular dystrophy (van Essen et al., 1992), and only slightly lower than the mutation frequency of $10 \cdot 10^{-5}$ in neurofibromatosis (Öbe, 1984). This places TSC in a group of disorders that are considered to have a remarkably high mutation rate.

Clinical variability in tuberous sclerosis

Until now the diagnosis of TSC has always been based on clinical and morphological findings. The number of symptoms and the severity of the disorder vary from patient to patient, even within families, among people carrying the same mutation in all somatic cells. A large number of very specific (pathognomonic) and less specific signs have been associated with the disease. For diagnostic purposes these signs can be divided in three groups: primary, secondary, and tertiary criteria, as listed in Table 1.2.

Table 1.2:
Criteria for the diagnosis of TSC (Roach et al., 1992; Neumann and Kandt, 1993)

Primary category: definitive (pathognomonic) criteria:	
Skin:	facial angiofibromas (adenoma sebaceum)
	multiple ungual fibromas
CNS:	cortical tubers[h]
	subependymal nodules[r]
	giant cell astrocytoma[h]
Retina:	multiple retinal astrocytoma
Secondary criteria:	
Skin:	shagreen patches
	fibrous forehead plaque
Lungs:	lymphangiomyomatosis[h]
Kidneys:	angiomyolipoma[r,h,u]
	cysts[h]
Heart:	rhabdomyomas[r,h,u]
CNS:	cerebral tubers[r]
	non-calcified subependymal nodules[r]
Retina:	single hamartoma
A certain diagnosis of TSC in a first degree relative[*]	
Tertiary criteria:	
Skin:	hypomelanotic macules
	confetti-like spots
Kidneys:	cysts[r,u]
CNS:	heterotopic white matter[r]
	infantile spasms
Gingiva:	fibromas
Teeth:	enamel pits
Rectum:	polyps[h]
Bones:	cysts[r]
Lungs:	lymphangiomyomatosis[r]
Hamartomas in other organs[h]	

[r]) Radiologic finding; [h]) Histologic finding; [u]) Ultrasound; [*]) The affection status of relatives is not used as a diagnostic criterium in our linkage studies. CNS: Central nervous system

The diagnosis of TSC can be made definitive when one primary criterium, or two secondary criteria, or one secondary and two tertiary criteria are fulfilled. A presumptive diagnosis can be made if one symptom from the secondary and one from the tertiary category, or three symptoms from the tertiary category are seen. Persons showing only one of the secondary symptoms, or persons fulfilling only two tertiary criteria are not more than suspect of having TSC. The secondary criterion of a relative with a certain diagnosis of TSC was not used in our linkage studies.

The disease shows an remarkably variable expression, even within families. The family depicted in Figure 6.4 may serve as an example for this. Most signs listed in Table 1.2 occur in a minority of patients. For diagnostic purposes the occurrence of the butterfly-shaped facial angiofibromas (adenoma sebaceum) is the most important sign, because it occurs in 50-90% of all patients and because it is pathognomonic of TSC. As discussed above, mental retardation and epilepsy are also very frequent symptoms of tuberous sclerosis. These are not included in Table 1.2, because of their limited diagnostic value.

In Chapter 3 we demonstrate that TSC is a genetically heterogeneous disorder with trait-causing loci on human chromosomes 9 and 16, while a third locus on for instance chromosome 11 can not be excluded completely (see Figure 2.3). The question has often been raised whether there is any relationship between the observed clinical variability and locus heterogeneity. In our research group, intensive searches for consistent clinical differences between chromosome 9 linked and other families have not revealed anything significant (S. Verhoef, pers. comm.). Similar studies by other investigators revealed the same negative outcome, with one exception: ungual fibromata have been reported to occur more frequently among chromosome 9 linked families than among unlinked families (Northrup et al., 1992). However, this theory was recently disputed, since it could not be confirmed in the English family data (Povey et al., 1994). The consequence of this is that it is not possible to use clinical features for the assignment of TSC patients to the genetic subgroup of patients with a gene defect on chromosome 9 (TSC1) or to the subgroup with a defect in the TSC2 gene on chromosome 16. Therefore, this assignment relies completely on linkage methods and thus indirectly on family size, and the reliability of the state of affectedness.

TSC associated hamartomas are defined as benign and malignancies resulting from these lesions are rarely observed. It is striking that the TSC-specific hamartoma also occur sporadically in non-TSC patients. Sporadic TSC-like lesions, such as for instance asymptomatic angiomyolipomas, generally occur as unilateral small single hamartomas and are believed to be a forme fruste of TSC if the existence of affected relatives can be demonstrated (Van Baal et al., 1989). In the absence of affected relatives, the hypothesis that both the sporadic isolated hamartomas and the disease TSC are caused by mutations in one and the same gene, is still not only still plausible and very interesting, but also of practical value. The hypothesis implies that the gene is susceptible to mutations and that the altered gene products give rise to uncontrolled cellular proliferation. This suggests that the TSC gene(s) are tumour-suppressor gene(s) or proto-oncogene(s).

References

Ahlsén G, Gilberg IC, Lindblom R, Gilberg C. Tuberous sclerosis in Western Sweden. Arch Neurol 1994;51:76-81.

Baraitser M, Patton MA. Reduced penetrance in tuberous sclerosis. J Med Genet 1985; 22: 29-31.

Easton DF, Bishop DT, Ford D et al. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. Am J Hum Genet 1993;52:678-701.

Edwards JH. The linkage detection problem. Ann Hum Genet 1990;54:253-275.

Evans DGR, Harris R. Heterogeneity in genetic conditions. Q J Med 1992;84:563-5.

Fenoglio JJ, McAllister HA, Ferrans VJ. Cardiac rhabdomyoma: a clinicopathologic and electron microscopic study. Am J Cardiol 1976;38:241-251.

Fleury P. Tubereuze sclerose, veranderingen in de diagnostische en de therapeutische benadering. Tijdschr Kindergeneeskd 1989;57:158-167.

Gomez MR. Tuberous Sclerosis, 2nd edition. 1988 (New York: Raven Press).

Gunther M, Penrose LS. The genetics of epiloia. J Genet 1935;31:413-430.

Hunt A, Lindenbaum RH. Tuberous sclerosis: A new estimate of prevalence within the Oxford region. J Med Genet 1984;21:272-277.

Hunt A, Shepherd C. A prevalence study of autism in tuberous sclerosis. J Autism Dev Disord 1993;23:323-339.

McKusick VA. Mendelian Inheritance in man. 9th edition. 1990 (Baltimore and London: The Johns Hopkins University Press).

Morton NE. The detection and estimation of linkage between the gene for elliptocytosis and the RH blood type. Am J Hum Genet 1956;8:80-96.

Neumann HPH, Kandt RS. Klinik and Genetic der tuberösen Sklerose. Dtsch Med Wschr 1993; 118:1577-1583.

Nevin NC, Pearse WG. Diagnostic and genetic aspects of tuberous sclerosis. J Med Genet 1968;5:273-280.

Northrup H, Kwiatkowski DJ, Roach ES et al. Evidence for genetic heterogeneity in tuberous sclerosis: one gene locus on chromosome 9 and at least one locus elsewhere. Am J Hum Genet 1992;51:709-720.

Obe G. Mutations in man. 1984 (Berlin, Heidelberg, New York, Tokyo: Springer-Verlag).

Osborne JP, Fryer A, Webb D. Epidemiology of tuberous sclerosis. N Y Acad Sci 1991;615:125-127.

Ott J. Linkage analysis and family classification under heterogeneity. Ann Hum Genet 1983;47:311-320.

Povey S, Burley MW, Attwood J et al. Two loci for tuberous sclerosis: one on chromosome 9q34 and one on 16p13. Ann Hum Genet 1994;58:107-127.

Roach ES, Smith M, Huttenlocher P et al. Report of the diagnostic criteria committee of the National Tuberous Sclerosis Association. J Child Neurol 1992;7:221-224.

Sampson JR, Scahill SJ, Stephenson JBP et al. Genetic aspects of tuberous sclerosis in the West of Scotland. J Med Genet 1989;26:28-31.

Shepherd CW, Gomez MR. Mortality in the Mayo Clinic tuberous sclerosis complex study. N Y Acad Sci 1991;615:375-377.

Singer K. Genetic aspects of tuberous sclerosis in a Chinese population. Am J Hum Genet 1971;23:33-40.

Steiner MS, Goldman SM, Fishman EK, Marshall FF. The natural history of renal angiomyolipoma. J Urol 1993;150:1782-1786.

Stevenson AC, Fisher OD. Frequency of epiloia in Northern Ireland. Br J Prev Soc Med;1956;10:134-135.

Van Baal JG, Fleury P, Brummelkamp WH. Tuberous sclerosis and the relation with renal angiomyolipoma. A genetic study on the clinical aspects. Clin Genet 1989;35:167-173.

Van Baal JG, Smits NJ, Keeman JN et al. The evolution of renal angiomyolipomas in patients with tuberous sclerosis. J Urol 1994;152:35-38.

Van Essen AJ, Busch HFM, te Meerman GJ, ten Kate LP. Birth and population prevalence of Duchenne muscular dystrophy in the Netherlands. Hum Genet 1992;88:258-266.

Webb D, Osborne JP. Non-penetrance in tuberous sclerosis. J Med Genet 1991; 28: 417-419.

Wiederholt WC, Gomez MR, Kurland LT. Incidence and prevalence of tuberous sclerosis in Rochester, Minnesota, 1950 through 1982. Neurology 1985;35:600-603.

Zaremba J. Tuberous sclerosis: a clinical and genetic investigation. J Ment Defic Res; 1968;12:63-80.

1.3

POSITIONAL CLONING UNDER LOCUS HETEROGENEITY

The search for disease genes: past, present and future

Already in the nineteenth century, when nothing was known about the concept of DNA, genes were defined as the fundamental unit of inheritance. According to recent estimates, the human genome contains about 80,000 genes (Antequera and Bird, 1993) and the total of the (haploid) genome consists of $3 \cdot 10^9$ DNA bases. Mutations in genes may predispose to genetic disorders. In the past, molecular research on human genetic disorders proceeded largely through isolation and characterisation of proteins, followed by the cloning of corresponding genes. Several disorders have been studied successfully using this approach, examples are the globin-related disorders, like sickle cell anaemia and the thalassaemia syndromes (Orkin and Kazazian, 1984), and familial hypercholesterolemia (Brown and Goldstein, 1986). The approach is based on knowledge of the aminoacid sequence, availability of antibodies, or other data, like knowledge of receptor-ligand binding properties. This so-called 'functional cloning approach' can only be applied if the underlying biochemical defect is known. However, for the majority of human genetic disorders information on the primary protein defect is not available.

When the number of molecular techniques increased a completely new route, which leads to the same endpoint (a cloned disease gene) became feasible. A comparison of the functional cloning approach with the alternative, as depicted in Figure 1.3 and 1.4 respectively, shows that the alternative route has the opposite

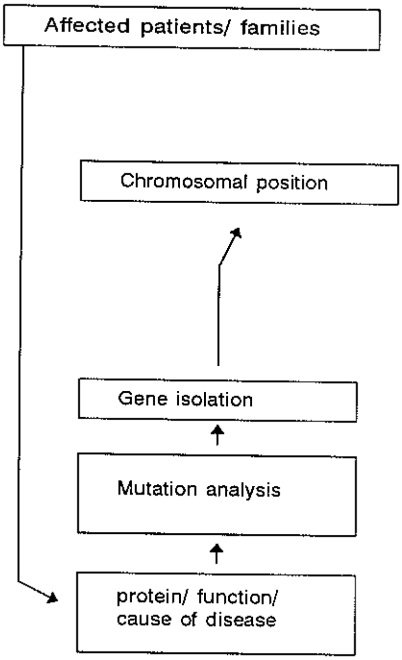

Affected patients/ families
Chromosomal position
Gene isolation
Mutation analysis
protein/ function/ cause of disease

Figure 1.3: A schematic representation of the various steps in a functional cloning project.

direction. This prompted the initial name for this approach: 'reverse genetics' (Ruddle, 1984; Orkin et al., 1986). In the opening article of the very first issue of Nature Genetics, Collins proposed to drop this confusing term and to replace it by the more widely accepted term 'positional cloning' (Collins, 1992; Monaco, 1994). Positional cloning strategies involve the isolation of trait-causing genes on the basis of their chromosomal location, without the need for any prior knowledge of the defective protein.

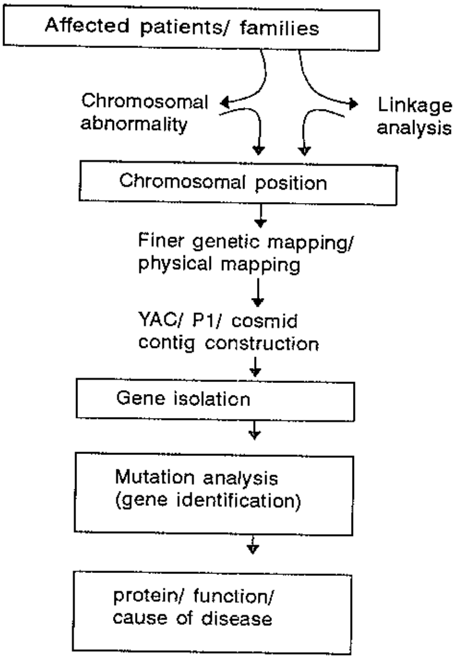

Affected patients/ families
Chromosomal abnormality
Linkage analysis
Chromosomal position
Finer genetic mapping/ physical mapping
YAC/ P1/ cosmid contig construction
Gene isolation
Mutation analysis (gene identification)
protein/ function/ cause of disease

Figure 1.4: A representation of the various steps involved in positional cloning.

As depicted in Figure 1.4, there are two different methods that can be used for the first step in the process: localising the gene on a chromosomal map. A map position can be deduced from chromosomal abnormalities, or linkage analysis, or a combination of these. Linkage analysis is aimed at finding cosegregation of the disease with genetic markers. For each autosomal locus an individual carries two alleles: one inherited from the father and one from the mother. Both alleles can be identical (homozygosity) or different and therefore distinguishable (heterozygosity). If a frequent molecular difference between both alleles exists, the locus can serve as genetic marker, provided that the chromosomal position of the locus is known. In the absence of a deleterious effect on the phenotype, such differences are called polymorphisms. A well-known example is the ABO blood group marker, a protein polymorphism with three different forms (alleles): the codominant alleles A and B and the recessive O allele. The protein is encoded by a recently cloned glycosyl-transferase gene, which maps to chromosome 9q34 (Yamamoto et al., 1990a). At least three human disorders have been shown to be 'linked' with this marker: cosegregation of the marker alleles with the disease alleles of nail patella syndrome (Renwick and Lawler, 1955; Renwick and Schulze, 1965), torsion dystonia (Ozelius et al., 1989) and TSC1 (Connor et al., 1987; Fryer et al., 1987; Chapter 3) has been demonstrated significantly.

In general, linkage analysis is regarded to be laborious when a gene has to be mapped by a genome wide search. If chromosomal abnormalities are available, linkage analysis in combination with a cytogenetic investigation will be more effective. However, most often chromosomal abnormalities associated with the particular disease of interest are not available. This is mainly, but not solely, due to the infrequent occurrence of cytogenetic findings as a cause of a specific monogenic disease. In the ideal situation, cytogeneticists and molecular geneticists would systematically work together on the investigation of all available cytogenetic abnormal material in positional cloning projects. However, the communication is often hampered. The cytogenetic aberrancy depicted in Figure 1.5 - a translocation involving chromosomes 16 and 22 - awaited a molecular follow-up for 12 years. Recently it turned out to be crucial for the identification of the TSC2 and ADPKD1 genes (Chapters 5 and 6.4).

The efficiency of linkage analysis, with or without the aid of cytogenetic data, is determined by the information content of the genetic marker. In the early days, protein polymorphisms, such as the ABO system, have been very important. To date, protein polymorphisms are somewhat outdated, because more useful classes

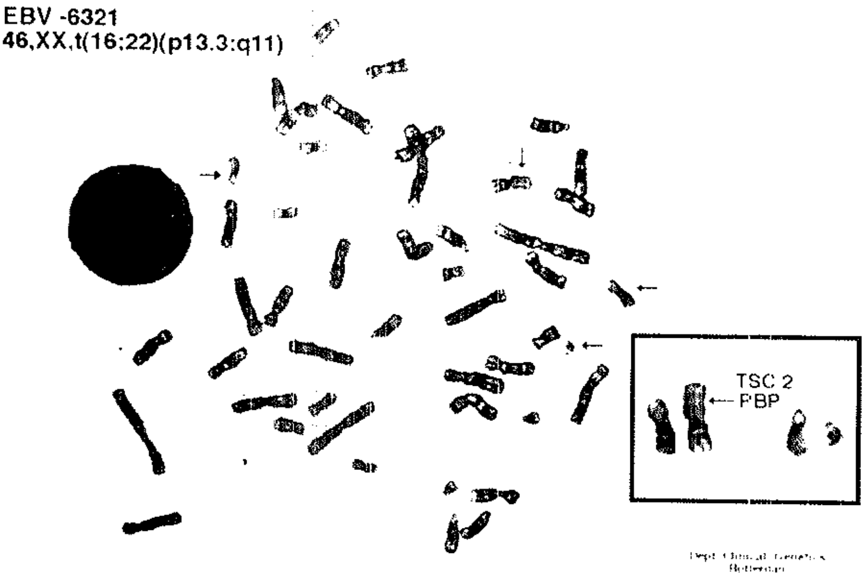


Figure 1.5: Karyogram showing a balanced t(16;22)(p13.3;q11.21) translocation associated with tuberous sclerosis and autosomal dominant polycystic kidney disease. Although the chromosomal abnormality was found in 1980, it was not used in any molecular work until 1992.

of genetic markers and more efficient techniques have become available.

A type of polymorphisms that can be typed at the DNA level are the so-called restriction fragment length polymorphisms (RFLPs) (Botstein et al., 1980). RFLPs are based on the presence or absence of recognition sites for specific endonucleases (restriction enzymes) or on insertion/deletion events that caused differences in distance between these sites. RFLPs are normally visualised by Southern hybridisation (Sambrook et al., 1989). A special type of RFLP is the variable number of tandem repeat (VNTR) marker. The number of repeats present determine the length of the restriction fragment.

To date, it is possible to implement the polymerase chain reaction (PCR) in the study of RFLPs. The DNA fragment containing the polymorphic site is amplified prior to endonuclease digestion. Two different RFLPs in amplifiable fragments of the recently cloned ABO gene have been shown to be the cause of the protein polymorphism mentioned above (Yamamoto et al., 1990b).

Currently, the most modern type of genetic markers are di-, tri- and tetranucleotide markers, also called simple tandem repeat (STR) sequences or microsatellite markers. The markers are often highly polymorphic and require only small amounts of DNA, because their analysis is based on PCR techniques (Weber and May, 1989). To date, thousands of microsatellite markers are available, while most of these markers have been mapped very precisely (Weissenbach et al., 1992).

A meiosis is informative if it occurs in a parent who is heterozygous at both the marker and the disease loci, provided that the contribution of both parents in the child's genotype remains assignable. Therefore, the usefulness of a certain marker in linkage studies depends on the number of alleles and the frequency of the different alleles among the population. Informative markers, that are not linked with the disease locus, will show cosegregation in only 50% of all meioses. The other half of the meioses are said to be 'recombinant'. If loci map close to each other on the same chromosome, recombinations will occur infrequently and the recombination fraction will be considerably lower than 50%. Therefore, the recombination frequency (Θ) is a good measure of genetic distance, and so is its deduced measure, the centimorgan (1 cM ≈ 1% recombinations). For a statistical evaluation of linkage results it is customary to calculate the likelihood for a certain segregation of alleles through a family due to chance (Θ=0.5) and to compare that likelihood with the likelihood for a segregation reflecting linkage (Θ<0.5). Usually we compare the ratio of both likelihoods and calculate the logarithm of the likelihood ratio: the lod score (Z). This procedure can be used to calculate lod scores ($Z_{(\Theta)}$) for each value of Θ between 0 and 0.5. A lod score larger than 3.0 - corresponding to a likelihood ratio of 1000:1 or more - is generally regarded as significant evidence for linkage. Lod scores of -2.0 or lower are accepted as evidence for exclusion of linkage. A detailed overview of human linkage analysis has been given by Ott (Ott, 1991).

As soon as a chromosomal position has been found by means of linkage analysis, by deletion mapping, breakpoint mapping, or by a combination of these methods, refined genetic mapping and physical mapping become feasible (Figure 1.4). Genetic fine mapping involves an intensive study of the most useful recombination events. Flanking markers have to be identified and new markers, mapping closer to the disease locus, may have to be developed. The resolution of fine mapping is usually limited to about 1 cM (Collins, 1992), depending on the number of available informative meioses. Physical mapping aims at the construction of a map encompassing a few megabases (Mb), containing all important map elements in the region, such as markers, breakpoints, cloned genes, CpG islands, restriction sites for rare-cutters, and available genomic clones. This step in the positional cloning process (Figure 1.4) therefore involves techniques like haplotype analysis, pulsed field gel electrophoresis (PFGE) (van Ommen et al., 1986; Burmeister and Ulanovsky, 1992), hybrid mapping (Verkerk et al., 1991) and fluorescent in-situ hybridisation (FISH) (Kievits et al., 1990; Trask, 1991). Haplotype analysis denotes the combined study of data from all genetic markers in the region. It involves the determination of the order of markers and the localisation of the disease locus by mapping recombination events. PFGE is a Southern hybridisation technique, applied on large restriction fragments ranging in size from less than 100 kb up to several megabases. Hybrid mapping involves the mapping of genes and markers relative to somatic cell hybrids containing (segments of) human chromosomes against a rodent background. The FISH technique is based on the cytogenetic study of biotinylated, or dioxigenin labelled probes. The main objective of these techniques is to place all map elements in the correct order and to approximate the physical distances. 1 cM roughly corresponds to approximately 1 Mb, although it should be noted that genetic and physical distances are not linearly related and that this relation depends on the chromosomal region studied.

When a physical map has been established, a systematic search for expressed sequences (genes) may start. While known genes can be examined right away, unknown genes first have to be isolated. This requires the availability of large stretches of genomic sequences. Although only a small percentage of the genomic sequences will be expressed, it is practice to assemble as many genomic clones as possible and to sort out the expressed parts later. Large contigs (groups of overlapping clones) of genomic DNA clones from the candidate area can be assembled through chromosome walking and jumping (Poustka et al., 1987). Various cloning vectors can be used: Yeast artificial chromosomes (YACs) (Burke et al., 1987; Schlessinger, 1990) usually contain inserts of up to 1 or 2 Mb, P1 vectors (Smoller et al., 1991) can be used for cloning inserts of 100 kb, while inserts of about 40 kb are suitable for cloning in cosmids (Evans et al., 1989). The use of phage clones for this purpose is becoming obsolete, because of the small insert size.

There are various ways by which cloned genomic material can be used for the isolation of genes. Although not an absolute requisite, one may start by mapping (possibly) expressed sequences by searching for evolutionary conservation using 'zoo-blots' (Monaco et al., 1986), screening Northern blots with single copy fragments, or by mapping CpG islands (Bird, 1986; Larsen et al., 1992; Valdes et al., 1994). CpG islands are CG-rich stretches of DNA that mark the 5′ end of most housekeeping genes. A more modern starting point is exon amplification, also known as exon trapping. It relies on the presence of splice sites in the genomic DNA. Splice sites are functional sequences, flanking exons, that are required for RNA splicing. If a genomic fragment, cloned behind a splice donor site, contains an exon, this will be spliced properly into mature mRNA (Duyk et al., 1990). Improved variants of the technique have been developed, applying the more sensitive and highly efficient pSPL1 vector system or a modified form of this vector (Buckler et al., 1991; Hamaguchi et al., 1992; Krizman and Berget, 1993; Datson et al., 1994; Church et al., 1994).

An essential step in gene isolation is the screening of cDNA libraries (i.e. libraries of cloned DNA copies of mRNA species) with selected or unselected genomic fragments. It has been demonstrated that it is possible to directly hybridise whole YAC or cosmid inserts to cDNA libraries on filters after suppression of the hybridisation of repeated sequences in the radioactively labelled probe (Elvin et al., 1990). Nevertheless, the best results may be expected if somewhat smaller probes of 1-10 kb are used. A modern alternative is 'cDNA selection' (Parimoo et al., 1991; Lovett et al., 1991; Korn et al., 1992; Morgan et al., 1992). This term denotes a group of techniques, all based on the same principle. A cDNA library is hybridised to (immobilised) genomic fragments, obtained from the region of interest. The hybridising cDNAs are then eluted and amplified by PCR, followed by a few more rounds of hybridisation/amplification, resulting in a highly enriched cDNA sub-library.

The techniques mentioned here exhibit intrinsic differences and may therefore complement each other very well. However, if one fails to isolate the desired gene, despite the use of a combination of these methods, it will be necessary to sequence long stretches of genomic DNA. Although rather laborious, genomic sequencing combined with computer analysis can be a forceful tool in the identification of potential exons (Uberbacher and Mural, 1991; Gish and States, 1993). Other techniques, such as RACE (Rapid Amplification of cDNA Ends) can then be used to obtain the desired cDNA clone (Frohman et al., 1988; Monaco, 1994).

When a gene has been isolated that may serve as a proper candidate, based on its chromosomal position and/or its characteristics, it may still be very time consuming to find convincing evidence confirming its trait-causing identity. The easiest approach is to screen for substantial rearrangements, initially on large scale by pulsed field gel electrophoresis (PFGE). On a smaller scale ordinary Southern hybridisation can be applied directed at finding purely intragenic mutations. If substantial deletions and insertions appear to be absent, more refined techniques have to be employed. An overwhelming number of such techniques is currently available, including direct sequencing, RNAase cleavage, denaturing gradient-gel electrophoresis (DGGE) and related techniques, carbodiimide modification, chemical cleavage of mismatch, single-strand conformation polymorphism (SSCP), heteroduplex analysis, reverse transcriptase PCR (RT-PCR), protein truncation test and transgenic (knock-out) animal models. Most of these techniques have been comprehensively reviewed by Cotton (Cotton, 1993). Formally, this step concludes the positional cloning process.

Finally, when the gene has been identified, the opportunity arises to search and find answers to the questions that existed for a long time. Questions like 'what kind of protein is defect', 'what is the nature of the defect', 'how and at what stage does the defect cause the disease', 'what is the normal function of the protein', 'can a clear genotype-phenotype correlation be determined', 'can molecular diagnostics be offered' and so forth were presumably the main reason to initiate the positional cloning project in the first place. Various types of experiments may help to provide answers to these important questions. This step completes the process depicted in Figure 1.4.

To date positional cloning is the most prominent approach towards the cloning of genes. However, in the near future positional cloning will be rather different. As more informative markers will become available, chromosomal positions will be found more easily. If other genes in the vicinity have already been isolated by positional cloning, one may be able to use the physical maps, YACs and cosmids that resulted as spin-off. Moreover, a first-generation physical map, consisting of 33,000 YAC clones and covering about 87% of the human genome, is already available (Cohen et al., 1993) These resources will speed up the positional cloning process enormously. By the end of 1994 5055 human genes were cloned and included in the genome database (GDB). Together with a large number of randomly sequenced cDNA clones, these genes form the group of so-called expressed sequence tags (ESTs), of which many are polymorphic. The ESTs will greatly facilitate the construction of high-resolution genetic and physical maps of expressed sequences (Takahashi and Ko, 1993). In the future this information may be applied to a different approach that will gradually replace positional cloning.

As more ESTs are mapped to specific chromosome regions, disease loci mapping to the same regions will have increasing numbers of potential candidate genes. This has led to the concept of the positional candidate approach, as proposed by Ballabio (Ballabio, 1993). The method combines essential parts of the positional cloning strategy with a candidate gene approach. The specific features of candidate genes will be compared to the features predicted by the disease symptoms, in order to find the strongest candidate. Functional domains will be matched with a biochemical defect, trinucleotide repeats will be matched with the clinical phenomenon of anticipation and expression patterns will be compared. Ballabio stated "it could be anticipated that the search for disease genes will be performed at the computer and not at the bench!". However, this statement can only be partly true, since it ignores the fact that all hypotheses that arise from a computer study, still have to be proven in a molecular experiment. A more significant issue neglected by the highly optimistic positional candidate outlook, is the growing complexity of linkage problems. While perhaps most of the uncomplicated, purely Mendelian, disorders have been mapped, a large group of complex disorders still awaits chromosomal localisation of the genes involved. An important factor in this context is locus heterogeneity.

Locus heterogeneity

Linkage analysis is often the most time-consuming step in the positional cloning process. Moreover, a number of factors may complicate the linkage analysis and may thus influence the rate-limiting aspect of this step in a very negative way.
The most important complicating factors are: incomplete penetrance (not all gene carriers show the affected phenotype), uncertainty of diagnosis, genomic imprinting (only the maternal or paternal allele is expressed), polygenic inheritance (the interaction between different genes determines the phenotype) and locus heterogeneity (gene defects at different loci cause the same phenotype).
Here, we will focus on the problem of locus heterogeneity.

In a linkage study of a disease with locus heterogeneity, one may expect to find families with positive lod scores at a certain chromosomal location and families with negative lod scores, even at the position of the disease locus. All linked families show a positive lod score, unless family members have been clinically misclassified. Most non-linked families show a negative score, although a few smaller families may show a positive lod score due to cosegregation of disease and marker genotype by chance. At some distance from the disease locus, the situation is even more complicated. Most of the unlinked families show negative lod scores again, but as the distance from the gene increases, linked families may show one or more recombinations, resulting in a reduced lod score. If only linkage data are available, it is impossible to distinguish between linked and unlinked families with 100% certainty. However, if a family is large enough to yield significant lod scores by itself, a reasonably reliable assignment may be possible. Furthermore, if a large set of families is available for analysis, powerful methods that facilitate an effective linkage analysis under heterogeneity may be applied. Some of these methods, namely the admixture test (A-test), B-test, C-test, M-test, predivided sample test and the two-locus lod score method, are discussed in Chapter 2.1.

Due to the problems described above, the positional cloning of heterogeneous disorders is complicated. The more extended scheme of positional cloning in case of heterogeneity is depicted in Figure 1.6, showing two major differences compared with Figure 1.4. The most prominent differences are the additional difficulties in linkage analysis. If no chromosomal abnormality is known, the first step in positional cloning after family ascertainment is a potential obstacle. This can only be avoided if a provisional separation of linked and unlinked families can be achieved, for instance by one of the methods for heterogeneity analysis listed above. Linkage information obtained from the linked families can be used to narrow down the position of the linked locus, whereas the other families may be useful to find the unlinked loci. Finer genetic mapping by means of haplotype analysis can only be performed on large families that yield significant evidence for linkage.

Gene identification - the fifth step in Figure 1.6 - is also different for heterogeneous disorders. The available material will consist of *l* unrelated individuals, who are affected due to a mutation at the locus under investigation, and *n* affecteds with a mutation at an unlinked locus. Therefore only mutation analysis in *l* individuals can yield conclusive evidence for gene identification. If only one candidate gene is under investigation, it is often possible to increase the total number of patients, resulting in a higher number *l+n*. This assures the presence of a minimal number of *l*. If a large number of candidate genes have to be explored, a more complicated strategy, that provides significant exclusion criteria, will have to be applied. Recently, Kwiatkowski and his coworkers investigated the VAV2 gene on chromosome 9q34 as a candidate for TSC1 (Henske et al., 1995). The applied strategy consisted of three steps. First, a number of families was selected that showed good evidence for linkage (P>0.9). Second, RT-PCR was performed on RNA from affected members of these families and the products were tested for the presence of an intragenic polymorphism. Four unrelated individuals were found to be heterozygous at the RNA level, demonstrating that both alleles were transcribed. Finally, the transcripts of the four heterozygous individuals were completely sequenced and no mutations were found, indicating that VAV2 can not be a candidate for TSC1 in these four patients.

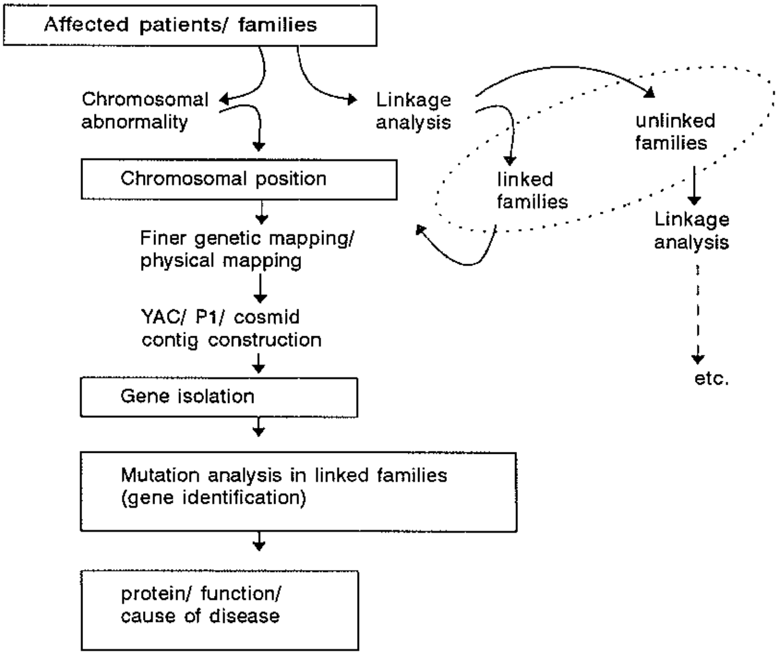

Affected patients/ families
Chromosomal abnormality
Linkage analysis
Chromosomal position
linked families
unlinked families
Linkage analysis
Finer genetic mapping/ physical mapping
YAC/ P1/ cosmid contig construction
etc.
Gene isolation
Mutation analysis in linked families (gene identification)
protein/ function/ cause of disease

Figure 1.6: A schematic representation of a positional cloning strategy under locus heterogeneity.

Examples of disease genes obtained by positional cloning strategies

Today it is almost impossible to provide a complete list of genes obtained by positional cloning. Each month several new genes are added to the list. Genes for complex disorders are, however, poorly represented. Only very few genes responsible for heterogeneous disorders have yet been cloned. Here a few typical examples of genes obtained by positional cloning strategies will be mentioned.

The first human gene ever cloned by positional cloning strategies was the X-CGD gene (Royer-Pokora et al., 1986), responsible for the X-linked form of the phagocytic disorder chronic granulomatous disease. Linkage information and two interstitial deletions, directed the search to Xp21 near the locus for Duchenne's muscular dystrophy (DMD). Phage clones, originating from a DMD cloning study, were used to enrich a cDNA library for Xp21 clones in a subtraction experiment. Two overlapping cDNA clones were shown to encode X-CGD and the results were published in July 1986, several months before the larger DMD gene itself was cloned.

In the years that followed, many disease-causing genes - especially for X-linked disorders - have been cloned by the procedure outlined in Figure 1.4. An illustrative example of such an X-linked disorder is fragile X syndrome (Verkerk et al., 1991). The fragile X syndrome is the most frequent form of inherited mental retardation, with a prevalence of 1/1250 males (Webb et al., 1986; Oostra et al., 1993). A highly characteristic cytogenetic finding is the 'fragile site' near the end of the long arm of the X chromosome. By means of linkage analysis and physical mapping the candidate region on Xq27.3 was narrowed down to 2.0 - 2.5 megabases. In a search for YACs from this region a 475 kb YAC was found, that was shown by FISH to cross the fragile site. The YAC was subcloned into a cosmid mini library. These cosmids were used to screen a human fetal brain cDNA library, which resulted in the isolation of the trait-causing gene, designated FMR1. Significant to the identification of the gene was the finding of a 5.1 kb EcoRI fragment that appeared to be enlarged in patients. This fragment turned out to contain an instable trinucleotide repeat causing fragile X syndrome, when present in an excess of 200 copies (Verkerk et al., 1991).

For both X-CGD and the fragile X syndrome a cytogenetic abnormality could be visualised and used as a guideline in the search for the responsible gene. If no cytogenetic information is available and in the absence of obvious candidate genes, the search may be a long-term effort, as was the case for Huntington's disease (HD). HD is an autosomal dominant progressive neurodegenerative disorder, characterised by motor disturbance, cognitive loss and psychiatric manifestations (Martin and Gusella, 1986). The HD locus was assigned to chromosome 4p in 1983 in one of the first successful linkage analyses using polymorphic DNA markers in humans (Gusella et al., 1983). Despite the early localisation and the formation of a large consortium, it took a complete decade to identify the gene (The HD collaborative research group, 1993). The main obstacle was the conflicting linkage data. This problem was overcome by exploiting a linkage disequilibrium strategy, based on the assumption that only a small number of founder mutations were present. Linkage disequilibrium has indeed been demonstrated, in other words: patients shared alleles at polymorphic sites. Ancient recombination events disturb linkage disequilibrium, hence the closer a marker is to the HD gene, the stronger the linkage disequilibrium. Next, exon trapping was used to find genes in the 500 kb candidate region, defined by linkage disequilibrium studies. The trait-causing gene - IT15 - shares the presence of an instable trinucleotide repeat with FMR1 as the mutational mechanism.

Positional cloning is even more complicated when genes at multiple loci cause an almost or completely identical phenotype (locus heterogeneity). Very recently, the chromosome 17-linked gene for familial breast cancer has been identified (Miki et al., 1994). Due to locus heterogeneity, linkage remained undetected until 1990. In that year, linkage to chromosome 17q21 was reported by Hall, with a lod score of 2.35 and evidence for heterogeneity (Hall et al., 1990). Recently Easton and the Breast Cancer Linkage Consortium reported on the analysis of 214 breast cancer families (Easton et al., 1993). Linkage was confirmed and the proportion of linked families was estimated to be 62% by applying the admixture test. It was shown that the combination of breast cancer with ovarian cancer mainly occurred in families linked to chromosome 17. It has been suggested that the linked and unlinked forms may be associated with a different age at onset (Hall et al., 1990). These observations are of particular interest, since they indicate that the linked and unlinked forms show a phenotypical dissimilarity. In the next chapter we will demonstrate that under these conditions the admixture test is prone to a systematic bias if clinical differences are associated with differences in family size (Chapter 2.3). We re-analyzed the same data as used by Easton, applying the method described in Chapter 2.3, and found only a small systematic bias (Sandkuijl and Janssen, unpublished data). Our results indicate that the number of linked families is slightly overestimated and that the real figure would be approximately 3% lower. The bias in the recombination fraction was almost nihil. The recent identification of the causative gene was achieved through direct screening of cDNA libraries and careful examination of genomic sequence. The so-called BRCA1 gene is composed of 22 coding exons, distributed over about 100 kb of genomic DNA. Several point mutations and small deletion-insertion mutations up to 11 bp were detected. The 7.8 kb transcript encodes a protein of 1863 aminoacids, containing a presumed zinc finger domain. This finding suggests that the BRCA1 product may function as a DNA binding protein. A previously presumed tumour suppressor role of BRCA1 is still subject of discussion (Miki et al., 1994; Futreal PA et al., 1994).

A second example of a heterogeneous disorder is autosomal dominant polycystic kidney disease (ADPKD). Like Fragile X syndrome and familial breast cancer, ADPKD is one of the most frequent monogenic genetic diseases. The main symptom of the disease is the development of cystic kidneys, commonly leading to renal failure in adult life. Positional cloning strategies applied to this disorder are discussed in Chapter 6.4. Briefly, a locus, responsible for ADPKD has been mapped to chromosome 16p13.3 in 1985 (Reeders et al., 1985). This locus, designated PKD1, is responsible for the disease in only 85% of all cases (Peters and Sandkuijl, 1992). Most of the remaining cases show linkage to markers in the PKD2 region on chromosome 4 (Kimberling et al., 1993; Peters et al., 1993). The available linkage data had shown a colocalisation of TSC2 and PKD1 at 16p13.3. The close vicinity of these loci was further demonstrated by a family with a translocation involving 16p and segregating both diseases in different individuals. This positional information led to the cloning of the 3' part of the PKD1 gene in the spring of 1993, but it took almost a year to prove the identity of the novel gene, called 'the PBP gene' for polycystic breakpoint gene (Chapter 6.4).

Positional cloning in tuberous sclerosis

When we started our project in 1988, a TSC locus had been assigned to chromosome 9q34 near ABO and the abelson oncogene (Fryer et al., 1987; Connor et al., 1987a). However, further data provided conflicting evidence. Apparently, some groups were able to confirm the provisional localisation (Connor et al., 1987b; Fahsold and Rott, 1988), while others were not (Smith et al., 1987; Northrup et al., 1987; Renwick, 1987; Kandt et al., 1989). In 1988 Clark described a TSC patient with trisomy for 11q23.3-qter and postulated a candidate locus for TSC in that region (Clark et al., 1988). A linkage study in 15 American families supported a TSC locus at 11q14-q23 (Smith et al., 1990). Although all these data were seemingly conflicting, none of all these studies provided any evidence for locus heterogeneity. In 1989 a single apparently unlinked family among a group of chromosome 9 linked families was reported by Sampson et al., (1989). A few months later the study of nine Dutch families confirmed that locus heterogeneity was the most likely explanation for this finding, (Janssen et al., 1990).

The following chapters describe work performed since the initial reports of locus heterogeneity in TSC (1990-1994). The identification of the TSC2 gene was the final step in that particular part of the positional cloning strategy. Nevertheless, it was the beginning of a broader study at both the DNA and protein level, aimed at determination of the protein function in the normal cell and at clarification of the mechanism causing TSC.

References

Antequera F, Bird A. Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci USA 1993;90:11995-11999.

Ballabio A. The rise and fall of positional cloning? Nature Genet 1993;3:277-279.

Bird AP. CpG-rich islands and the function of DNA methylation. Nature 1986;321:209-213.

Botstein D, While RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphism. Am J Hum Genet 1980;32:314-331.

Brown MS, Goldstein JL. A receptor-mediated pathway for cholesterol homeostasis. Science 1986;232:34-47.

Buckler AJ, Chang DD, Graw SL, et al. Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. Proc Natl Acad Sci USA 1991;88:4005-4009.

Burke DT, Carle GF, Olson MV. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 1987;236:806-812.

Burmeister M, Ulanovsky L. Pulsed-field gel electrophoresis. Protocols, methods and theories. (Totowa, New jesey: Humana Press).

Church DM, Stotler CJ, Rutter JL et al. Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. Nature Genet 1994;6:98-105.

Clark RD, Smith M, Pandolfo M et al. Tuberous sclerosis in a liveborn infant with trisomy due to t(11q23.3;22q11.2) translocation: Is neural cell adhesion molecule a candidate gene for tuberous sclerosis? Am J Hum Genet 1988;43: A44.

Cohen D, Chumakov I, Weissenbach J. A first-generation physical map of the human genome. Nature 1993; 366:698-701.

Collins FS. Positional cloning: let's not call it reverse anymore. Nature Genet 1992;1:3-6.

Connor JM, Yates JRW, Mann L et al. Tuberous sclerosis: analysis of linkage to red cell and plasma protein markers. Cytogenet Cell Genet 1987a;44:63-64.

Connor JM, Pirrit LA, Yates JRW et al. Linkage of the tuberous sclerosis locus to a DNA polymorphism detected by v-abl. J Med Genet 1987b;24:544-546.

Cotton RGH. Current methods of mutation detection. Mutation Res 1993;285:125-144.

Datson NA, Duyk GM, Van Ommen JB, Den Dunnen JT. Specific isolation of 3' terminal exons of human genes by exon trapping. Nucleic Acids Res 1994;22:4148-4153.

Duyk GM, Kim S, Myers RM, Cox DR. Exon trapping: A genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. Proc Natl Acad Sci USA 1990;87:8995-8999.

Easton DF, Bishop DT, Ford D et al. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. Am J Hum Genet 1993;52:678-701.

Elvin P, Slynn G, Black D, et al. Isolation of cDNA clones using yeast artificial chromosome probes. Nucleic Acids Res 1990;18:3913-3917.

Evans GA, Lewis K, Rothenberg BE. High efficiency vectors for cosmid microcloning and genomic analysis. Gene 1989;79:9-20.

Fahsold R, Rott HD. Tuberous sclerosis: linkage to alpha-acid glycoprotein (ORM). Clin Genet 1988;34:394.

Frohman MA, Dush MK, Martin GR. Rapid production of full-length cDANs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc Natl Acad Sci USA 1988;85:8998-9002.

Fryer AE, Chalmers A, Connor JM et al. Evidence that the gene for tuberous sclerosis is on chromosome 9. Lancet 1987 i:659-661.

Futreal PA, Liu Q, Schattuck-Eidens D, et al. BRCA1 mutations in primary breast and ovarian cancers. Science 1994; 266: 120-122.

Gish W, States DJ. Identification of protein coding regions by database similarity search. Nature Genet 1993;3:266-272.

Gusella JF, Wexler NS, Conneally PM et al. A polymorphic DNA marker genetically linked to Huntington disease. Nature 1983;306:234-238.

Hall JM, Friedman L, Guenther C et al. Closing in on a breast cancer gene on chromosome 17q. Am J Hum Genet 1990;50:1235-1242.

Hamaguchi M, Sakamoto H, Tsuruta H et al. Establishment of a highly selective and specific exon-trapping system. Proc Natl Acad Sci USA 1992;89:9779-9783.

Henske EP, Short MP, Jozwiak S et al. Identification of VAV2 on 9q34 and its exclusion as the tuberous sclerosis gene TSC1. Ann Hum Genet 1995; 59: 25-37.

Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 1993;72:971-983.

Janssen LAJ, Sandkuyl LA, Merkens EC et al. Genetic heterogeneity in tuberous sclerosis. Genomics 1990;8:237-242.

Kandt RS, Pericak-Vance MA, Hung W-Y et al. Absence of linkage of ABO blood group locus to familial tuberous sclerosis. Exp Neurol 1989;104:223-228.

Kievits T, Dauwerse JG, Wiegant J, et al. Rapid subchromosomal localization of cosmids by nonradioactive in situ hybridisation. Cytogenet Cell Genet 1990;53:134-136.

Kimberling WJ, Kumar S, Gabow PA et al., Autosomal dominant polycystic kidney disease: localization of the second gene to chromosome 4q13-q23. Genomics 1993;18:467-472.

Krizman DB, Berget SM. Efficient selection of 3' terminal exons from vertebrate DNA. Nucleic Acids Res 1993;21:5198-5202.

Larsen F, Gunderson G, Lopez R, Prydz H. CpG islands as gene markers in the human genome. Genomics 1992;13:1095-1107.

Martin JB, Gusella JF. Huntington's disease: pathogenesis and management. N Engl J Med 1986;315:1267-1276.

Mike Y, Swensen J, Shattuck-Eidens D, et al. A strong cnadidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 1994; 266: 66-71.

Monaco AP, Neve RL, Colletti-Feener C, Bertelson CJ et al. Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. Nature 1986;323:646-650.

Monaco AP. Isolation of genes from cloned DNA. Curr Opin Genet Dev 1994;4:360-365.

Northrup H, Beaudet AL, O'Brien WE et al. Linkage of tuberous sclerosis to ABO blood group. Lancet 1987;2:804-805.

Oostra BA, Willems PJ, Verkerk AJMH. Fragile X syndrome: A growing gene. In G Genome Analysis Vol 6: Genome mapping and neurological disorders. Davies KE , Tilgman SM (eds.) 1993; pp. 45-75. (New York: Cold Spring Harbour Laboratory Press).

Orkin SH, Kazazian HH. Mutation and polymorphism of the human ß-globin gene and its surrounding DNA. Ann Rev Genet 1984;18:131-171.

Orkin SH. Reverse genetics and human disease. Cell 1986;47:845-850.

Ott J. Analysis of human genetic linkage. Revised edition. (Baltimore, London: The Johns Hopkins University Press)

Ozelius L, Kramer PL, Moskowitz CB et al. Human gene for torsion dystonia located on chromosome 9q32-q34. Neuron 1989;2:1427-1434.

Peters DJM, Sandkuijl LA. Genetic heterogeneity of polycystic kidney disease in Europe. Contributions Nephrol. 1992;97:128-139.

Peters DJM, Spruit L, Saris JJ et al. Chromosome 4 localization of a second gene for autosomal dominant polycystic kidney disease. Nature Genet 1993;5:359-362.

Poustka A, Pohl TM, Barlow DP et al. Construction and use of human chromosome jumping libraries from NotI-digested DNA. Nature 1987;325:353-355.

Reeders ST, Breuning MH, Davies KE et al. A highly polymorphic DNA marker linked to adult polycystic kidney disease on chromosome 16. Nature 1985;317:542-544.

Renwick JH, Lawler SD. Genetic linkage between the ABO and nail-patella loci. Ann Hum Genet 1955;19: 312-331.

Renwick JH, Schulze J. Male and female recombination fractions for the nail patella: ABO linkage in man. Ann Hum Genet 1965;28:379-392.

Renwick JH. Tuberous sclerosis and ABO. Lancet 1987;2:1096-1097.

Royer-Pokora B, Kunkel LM, Monaco AP et al. Cloning the gene for an inherited human disorder - chronic granulomatous disease - on the basis of its chromosomal location. Nature 1986;322:32-38.

Ruddle FH. The William Allan memorial award address: reverse genetics and beyond. Am J Hum Genet 1984;36:944-953.

Sambrook J, Fritsch EF, Maniatis T. Molecular cloning, a laboratory manual. 2nd edition. 1989. (New York: Cold Spring Harbour Laboratory Press).

Sampson JR, Yates JRW, Pirrit LA et al. Evidence for genetic heterogeneity in tuberous sclerosis. J Med Genet 1989;26:551-516.

Schlessinger D. Yeast artificial chromosomes: tools for mapping and analysis of complex genomes. Trends Genet 1990;6:248-258.

Smith M, Haines J, Trofatter J et al. Linkage studies in tuberous sclerosis. Cytogenet Cell Genet 1987;46:694-695.

Smith M, Smalley S, Cantor R et al. Mapping of a gene determining tuberous sclerosis to human chromosome 11q14-11q23. Genomics 1990;6:105-114.

Smoller DA, Petrov D, Hartl DL. Characterization of bacteriophage P1 library containing inserts of drosophila DNA of 75-100 kilobase pairs. Chromosoma 1991;8:487-494.

Takahashi N, Ko MSH. The short 3'-end region of complementary DNAs as PCR-based polymorphic markers for and expression map of the mouse genome. Genomics 1993;16:161-168.

Trask BJ. Fluorescence in situ hybrididsation: applications in cytogenetics and gene mapping. Trends Genet 1991;7:149-154.

Uberbacher EC, Mural RJ. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc Natl Acad Sci USA 1991;88:11261-11265.

Valdes M, Tagle DA, Collins FS. Island rescue PCR: A rapid and efficient method for isolating transcribed sequences from yeast artificial chromosomes and cosmids. Proc Natl Acad Sci USA 1994;91:5377-5381.

Van Ommen GJ, Verkerk JM, Hofker MH, et al. A Physical map of 4 million bp around the Duchenne muscular Dystrophy gene on the human X-chromosome. Cell 1986;47:499-504.

Verkerk AJMH, Pieretti M, Sutcliffe JS et al. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 1991;65:905-914.

Webb TB, Bundey SE, Thake AI, Todd J. Population incidence and segregation ratios in the Martin-Bell syndrome. Am J Med Genet 1986;23:573-580.

Weber JL, May PE. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 1989;44:388-396.

Weissenbach J, Gyapay G, Dib C et al. A second generation linkage map of the human genome. Nature 1992;359:794-801.

Yamamoto F-I, Marken J, Tsuji T, et al. Cloning and characterisation of DNA complemantary to human UDP-GalNAc:Fucα1→2Galα1→3GalNAc Transferase (Histo-blood Group A transferase) mRNA. J Biol Chem 1990a;265:1146-1151.

Yamamoto F-i, Hakomori S-i. Sugar-nucleotide donor specificity of Histo-blood group A and B transferases is based on amino acid substitutions. J Biol Chem 1990b;265:19257-19262.

CHAPTER 2

METHODS FOR LINKAGE ANALYSIS UNDER LOCUS HETEROGENEITY

2.1

INTRODUCTION: ANALYSIS OF LOCUS HETEROGENEITY IN PRACTICE

Bart Janssen
Dicky Halley
Dick Lindhout
Lodewijk Sandkuijl

To be Submitted

Abstract

When a proportion of families segregating a certain genetic trait shows linkage to a particular marker, while other families carrying the same trait are unlinked, we are dealing with locus heterogeneity. Many - and perhaps most - human genetic disorders show locus heterogeneity. Usually it is difficult to rule out the possiblility of locus heterogeneity. Therefore, it is a complicating factor that may not be neglected if a linkage study is undertaken. This review provides an overview of the consequences of locus heterogeneity with respect to the various phases of a linkage study: the ascertainment of families, the analysis itself and the interpretation and presentation of the results. The properties of the commonly used statistical methods - the A-test, B-test, M-test and predivided sample test - and two relatively new methods - the C-test and the 'two-locus lod score methods'- are discussed.

Glossary

Alpha: Proportion of linked families (α).

A-test: Admixture test. Heterogeneity test, assuming an admixture of α linked and ($1-\alpha$) unlinked families, based on equation (2) (refer to text for details).

Asymptotic distribution: Statistic distribution obtained with an infinite number of families.

Bilineal families: Families with disease alleles coming in from more than one (grand)parent. (this term is also used for 'multilineal' families)

B-test: Heterogeneity test, based on the assumption that Θ follows a beta distribution (refer to text for details).

Complex disease: A disease that does not show simple Mendelian inheritance. Factors like incomplete penetrance, uncertainty of diagnosis, locus heterogeneity, epistasis and genomic imprinting complicate linkage analysis in these diseases.

C-test: Heterogeneity test that combines the M-test with a simulation (refer to text for details).

df: Degree(s) of freedom.

Exclusion mapping: Lod scores lower than -2.0 are widely accepted as evidence against linkage. Exclusion mapping aims at obtaining such scores for a whole chromosome/ chromosomal area.

High density pedigrees: Pedigrees heavily loaded with affected individuals.

ICA: Imaginary chromosome approach: method for the simultaneous analysis of two or more loci on different chromosomes, applying the A-test.

Location score: Lod score resulting from multipoint analysis, expressed as 2 times the difference in Ln likelihood.

Mixed families: Bilineal (multilineal) families segregating two or more disease loci.

Monogenic disorder: Genetic disorder caused by only one gene.

M-test: Heterogeneity test, based on equation 1 (see to text for details).

Glossary (continued)

Phenocopies:	(Apparently) affected individuals, who do not carry the disease gene. The phenotype is caused by a non-genetic factor.
PIC:	Polymorphism Information Content
Posterior probability:	Conditional probability (of being linked).
Power (statistical):	The probability that H_0 (the hypothesis of non-linkage) will be rejected in case of linkage.
Predivided sample test:	Heterogeneity test (based on the M-test) for the analysis of two or more distinct groups of families (see text for details).
Theta:	Recombination fraction (Θ).
Two-locus method:	'Two-locus lod score method' applying a modified LINKAGE program called TMLINK.
Type I error:	False positive result.
w:	Posterior probability.
$Z_{(\Theta)}$:	Lod score under homogeneity.
$Z_{(\alpha,\Theta)}$:	Lod score under heterogeneity

Introduction

Linkage analysis of traits with unknown biochemical basis is the first step in positional cloning of causative genes. This strategy has proven to be quite successful, provided that only a single gene is mutated. It has been shown however, that many cellular processes involve multimeric proteins. It is also known that proteins often act in sequential pathways. Since these processes require correct products from multiple genes, phenotypically similar effects of these biochemical processes may arise from mutations at different loci. This common phenomenon has been termed 'locus heterogeneity'. Other terms like 'linkage heterogeneity', 'etiological heterogeneity', 'non-allelic heterogeneity' and 'genetic heterogeneity' are often used to refer to the same phenomenon. In this paper we will use the terminology 'locus heterogeneity', since some of the alternatives have a somewhat broader definition. Sometimes allelic heterogeneity and non-allelic (locus) heterogeneity are mentioned together, because they may both be associated with clinical variability and diagnostic difficulties. Allelic heterogeneity occurs if different mutations, affecting the same gene, cause the same disease phenotype. From a linkage point of view this type of heterogeneity is basically different from locus heterogeneity. In this review we will focus on locus heterogeneity only and describe the associated problems and the commonly used methods and strategies that may be helpful in the analysis of locus heterogeneity.

If locus heterogeneity occurs it is - for obvious reasons - always regarded as a complicating factor in linkage analysis. Nevertheless, opinions with respect to locus heterogeneity have changed considerably over the last few years. Half a decade ago locus heterogeneity was regarded to be relatively uncommon and sometimes assumed differences in genetic distance were considered to provide a very realistic alternative explanation for the findings (Hulten, 1988; Goodfellow, 1985). If a linkage study in an independent set of families did not yield positive lod scores in an area where other investigators previously reported a significant score, the prior finding was most often judged to be a false linkage claim, in other words the original claim was classified as a type I error despite the possibility of locus heterogeneity (Kandt et al., 1989; Van Haeringen et al., 1989). Although very powerful tests for heterogeneity were already available, these were not used very frequently. To date, proposed mechanisms that deviate from the one-trait/one-locus concept are widely accepted. In contrast, heterogeneity of genetic distances are believed to be a rare phenomenon, with the exception of male-female differences. Nowadays, negative lod scores obtained in obvious candidate areas (e.g. obtained by previous linkage studies) are most often considered as evidence for locus heterogeneity, rather than evidence for exclusion. It thus seems that we are moving from an overestimate of the number of type I errors to a neglect of their possible occurrence. The number of publications on studies that involve the evaluation of locus heterogeneity has shown a vast increase, as shown in Figure 2.1. The number of reported analyses under heterogeneity increased from nil in 1984 to almost 40 in 1993. Table 2.1 shows an overview of recent heterogeneity analyses and their conclusions.

The recent advances in heterogeneity analysis may be due to a growing lack of simple monogenic disorders that combine a high prevalence with a simple Mendelian mode of inheritance and absence of definitive mapping data. As a result of the identification of the most prominent disease-causing genes from this group, linkage analysis is now being applied to more complex diseases. Many factors, such as incomplete penetrance, uncertain diagnostic criteria, phenocopies,

Table 2.1
Disorders for which locus heterogeneity has been tested (1989-1993)
(Up to 4 most recent/relevant studies are listed)

Trait	Sub-	Mode of	Families		Heterogeneity		Reference
	type	inheritance		Method		Chromosome	
AD Cerebellar Ataxias	SCA2	AD	14	A-test	Y	12q,?	Gispert et al., 1993.
	ADCA1	AD	10	A-test	Y	6p,?	Khati et al., 1993.
AD Polycystic Kidney Disease		AD	328	A-test	Y	16p,?	Peters et al., 1992.
		AD	35	A-test	Y	16p,?	Wright et al., 1993.
	non-16	AD	8	A-test	N	4q	Peters et al., 1993.
AR Muscular Dystrophy		AR	13	A-test, M-test	N	13q	Azibi et al., 1993.
Alport Syndrome		XL	31	A-test, M-test	N	Xq	M'Rad et al., 1992.
Atopic IgE Responses		AD	64	A-test	N	11q	Young et al., 1992.
Charcot-Marie-Tooth Disease	CMT1	AD	5	A-test	N	17p	McAlpine et al., 1990.
	CMT1	AD	7	A-test	Y	1,17p,?	Chance et al., 1990.
	X-linked	XL	3	A-test	Y	Xp,Xq	Ionasescu et al., 1991, 1992.
	CMT2	AD	5	A-test	Y	1p,?	Ben Othmane et al., 1993.
Congenital Stationary Night Blindness		XL	8	A-test	N	Xp	Musarella et al., 1989.
Cutaneous Malignant Melanoma-Dysplastic Nevus		AD	13	M-test	Y	1p,?	Goldstein et al., 1993.
Diabetes							
	MODY	AD	3	A-test, M-test	Y	20q	Bowden et al., 1992.
	MODY	AD	16	A-test	Y	7p,?	Froguel et al., 1992.
	NIDDM	AD or AR	18	A-test	-	?	Elbein et al., 1992.

Table 2.1 continued

Early Onset Periodontitis		var. models	20	Exclusion by A-test	excl. from 4q if AD		Hart et al., 1993.
Eclampsia/ Pre-eclampsia		AR	10	A-, B-, M-test	excl. from chr. 6		Wilton et al., 1990.
Fam. Adenomatous Polyposis		AD	18	A-test	N	5q	Mareni et al., 1993
Fam. Alzheimer Disease	-	AD	32	M-test, A-test (ICA)	-	19 and/or 21	Pericak-Vance et al., 1991.
	early onset	AD	10	A-test	Y	14,21	Mullan et al., 1992.
	-	AD	49	A-test, M-test	Y	14,21,?	Schellenberg et al., 1993
	-	AD	94	A-, C-, M-test, TMLINK	Y	19,21	Several studies, summarised by Wijsman, 1993.
Fam. Amyotrophic Lateral Sclerosis		AD	23	A-test	Y	21q,?	Siddique et al., 1991.
Fam. Benign Hypercalcemia		AD	5	A-test (ICA)	Y	3q,19p	Heath et al., 1993.
Fam. Breast/ Ovarian Cancer		AD	43	A-test	Y	17q,?	Smith et al., 1993.
		AD	19	A-test, M-test	Y	17q,?	Mazoyer et al., 1993.
	Premenopausal Breast C.	AD	35	A-test, M-test	-	17q or 17q,?	Haile et al., 1993.
	-	AD	214	A-test	Y	17q, ?	Easton et al., 1993.
Fam. Dyslexia		AD	21	A-test	-	15,? or ?	Smith et al., 1990.
Fam. Hypertrophic Cardiomyopathy		AD	5	A-test	Y	14q,?	Epstein et al., 1992.
		AD	3	A-test	Y	14q,1q, 15q,?	Thierfelder et al., 1993.
Fam. Mediterranean Fever		AR	23	A-test, M-test	N	16p	Shohat et al., 1992.
		AR	18	A-test, M-test	-	17q,? or 16p	Aksentijevich et al., 1993.

Table 2.1 continued

Fanconi Anemia	AR	34	A-test	Y	20q,?	Mann et al., 1991.
FSHD Muscular Dystrophy	AD	10	A-test	-	4q or 4q,?	Wijmenga et al., 1991
	AD	65	A-test	N	4q	Sarfarazi et al., 1992a.
	AD	7	A-test	Y	4q,?	Gilbert et al., 1993.
Gerstmann-Sträussler-Scheinker Syndrome	AD		A-test	N	20p	Speer et al., 1991.
Gorlin syndrome (NBCCS)	AD	15	A-test	N	9q	Chenevix-Trench et al., 1993.
Hereditary Hydronephrosis	AD	5	A-test	Y	6p,?	Izquierdo et al., 1992.
Hereditary Multiple Exostoses	AD	11	A-test	Y	8q,?	Cook et al., 1993.
Huntington Disease	AD	12	A-test	N	4p	Ajmar et al., 1991.
	AD	63	A-test	N	4p	Conneally et al., 1989.
Idiopathic Torsion Dystonia	AD	27	A-test	Y	9q, ?	Warner et al., 1993.
Juvenile Myoclonic Epilepsy	AD or AR					
		23	A-test	N	6p	Weissbecker et al., 1991.
	idem	25	A-test	Y	6p,? or ?	Whitehouse et al., 1993.
Limb-Girdle Muscular Dystrophy	AR	11	A-test	Y	15q,?	Passos-Bueno et al., 1993.
Leber Hereditary Optic Neuroretinopathy	XL	6	A-test	N	Xp	Vilkki et al., 1991.
Low-voltage EEG	AD	17	A-test	-	20q,?	Anokhin et al., 1992.
	AD	17	A-test	Y	20q,?	Steinlein et al., 1992.

Table 2.1 continued

Malignant Hyperthermia		AD	16	A-test (ICA)	Y	17q,19q,?	Levitt et al., 1992 .
Malignant Melanoma		AD	26	A-test	N	9p	Nancarrow et al., 1993.
Manic Depression		var. models					
			136		N	?	Price 1989.
		AD	6	A-test	Y	11p,?	Lim et al., 1993.
Marfan Syndrome		AD	17	A-test	N	15q	Kainulainen et al., 1991.
		AD	22	A-test	-	15q	Sarfarazi et al., 1992b.
Multiple Endocrine Neoplasia Neoplasia / MTC		AD	34	A-test	N	10	Narod et al., 1992a.
		AD	18	A-test	N	10	Narod et al., 1989.
Neurofibromatosis							
	NF II	AD	12	A-test	N	22	Narod 1992b.
	NF I	AD	22	A-test	N	17q	Upadhyaya et al., 1989.
Neuronal Ceroid Lipofuscinosis							
	CLN1	AR	31	M-test	Y	1p,?	Järvelä, 1991
	-	AR	25	M-test	Y	1p,16p,?	Williams et al., 1993.
Osteogenesis Imperfecta		AD	38	estimation of α	no further het.	7q,17q	Sykes et al., 1990.
Panic Disorder			36	A-test, B-test	excl. from 16q		Crowe et al., 1990.
Retinitis Pigmentosa	X-linked	XL	9	A-test	Y	Xp (2x)	Chen et al., 1989.
	X-linked	XL	20	A-test	Y	idem	Musarella et al., 1990.
	X-linked	XL	62	A-test	Y	X (3x)	Ott et al., 1990.

Table 2.1 continued

Schizophrenia		AD	15	A-, B-, M-test	-	5q or 5q,?	Kennedy et al., 1989.
		AD	29	A-test	-	5q,? or ?	McGuffin et al., 1990.
		AD or AR	9	A-test	-	-	Coon et al., 1993.
Stickler Syndrome		AD	10	A-test	N	12q	Bonaventure et al., 1992.
Spinal Muscular Atrophy	type II, III	AR	13	A-test	N	5q	Brzustowicz et al., 1990.
	I, II, III	AR	16	A-test	Y	5q,?	Gilliam et al., 1990.
	II, III	AR	38	A-test	Y	5q,?	Brzustowicz et al., 1993.
Treacher Collins Syndrome		AD	8	A-test	N	5q	Wang Jabs et al., 1991.
Tuberous Sclerosis		AD	15	A-test	N	11q	Smith et al., 1990.
		AD	22	A-test (ICA)	Y	9q,?	Haines et al., 1991.
		AD	128	A-test (ICA)	Y	9q,?	Chapters 2.2 and 3.2.
		AD	14	A-test	Y	9q,?	Northrup et al., 1992.
Usher Syndrome		AR	10	A-test	Y	1q,?	Smith et al., 1992.
		AR	15	A-test	Y	14q,?	Kaplan et al., 1992.
	type I	AR	36	A-test	-	?	Keats et al., 1992.
Von Hippel-Lindau Disease		AD	46	A-test	N	3p	Maher et al., 1992.
		AD	29	A-test	N	3p	Crossey et al., 1993.
		AD	38	A-test	N	3p	Richards et al., 1993.
Waardenburg Syndrome	type I and II	AD	44	A-, B-, M-test	Y	2q,?	Farrer et al., 1992.
X-linked Ametogenesis Imperfecta		XL	3	A-test	Y	Xp, Xq	Aldred et al., 1992.
X-linked Deafness		XL	7	A-test	Y	Xq (2X)	Reardon et al., 1991.

AD= Autosomal Dominant; AR= Autosomal Recessive; XL= X-linked

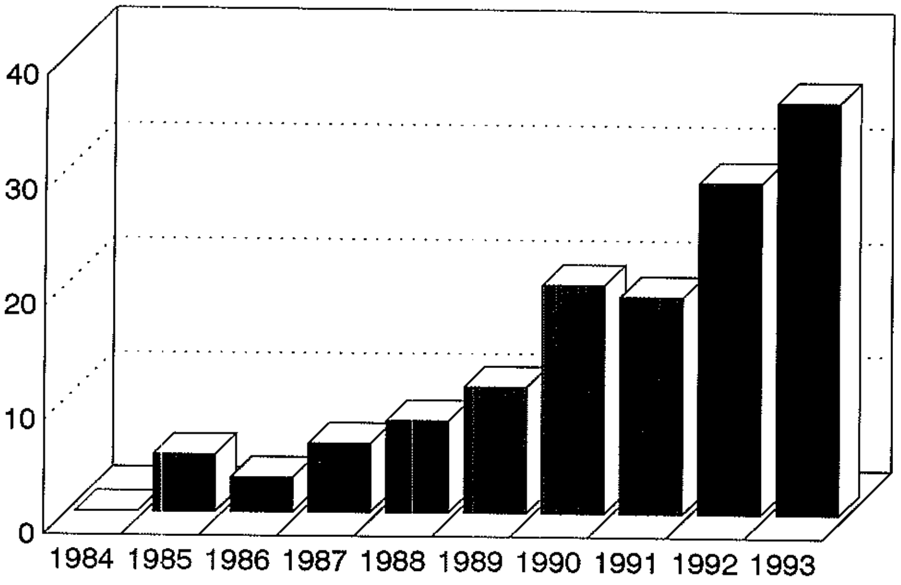


Figure 2.1: Identified number of reported studies that involve statistical testing for locus heterogeneity (1984-1993).

and locus heterogeneity complicate positional cloning. This review addresses the problem of locus heterogeneity, the various steps that have to be taken in planning a linkage study if locus heterogeneity is suspected and the problems that may occur in the interpretation of results. First we will focus on the available methods that can be used for statistical testing of locus heterogeneity.

Available tests for locus heterogeneity

Most statistical analyses on disease traits are carried out by means of likelihood ratio tests. Use of these so-called lod score methods on genetically homogeneous traits has led to the cloning of a large number of trait causing genes. It is therefore not surprising that lod score methods are also preferred in studies complicated by possible locus heterogeneity, especially since homogeneity is often assumed until the alternative has been proven. However, if the specification of a reliable genetic model for a disease seems problematic, for instance because the mode of inheritance is uncertain, one might prefer a parameter free method, like the affected sib pair method or the analysis of allelic association. Goldin and Gershon have demonstrated that for the purpose of detecting linkage in the presence of some unlinked families these methods can be quite powerful. Using the affected sib pair method and a significance threshold corresponding to $p<0.05$, a sample size of 140 informative sib pairs is sufficient to obtain more than 80% power, if at least 25% of the families is linked, provided that the a recombination fraction is 5% or less and that the disease shows a prevalence of 1% or less (Goldin and Gershon 1988). By studying allelic association one can obtain comparable power with a similar sample size and level of locus heterogeneity in the case of a strong founder effect, with at least 40% of the individuals with a linked disease gene carrying a mutation in the same infrequent (<10%) haplotype (Gershon et al., 1989). Although the theoretical possibility exists to perform a statistical test for locus heterogeneity using multipoint affected sib pair data, the non-parametric methods mentioned here are generally not followed by a heterogeneity test. Table 2.1 shows that only three heterogeneity tests are commonly used: the M-test, A-test and B-test. These tests are based on lod score methods and can be applied on pairwise data (lod scores) or multipoint results (location scores).

The M-TEST / predivided sample test

In 1956 Morton devised a simple test for assessing the significance of linkage heterogeneity, which he applied on the elliptocytosis linkage problem. He demonstrated that some of the 14 elliptocytosis families were linked to the rhesus factor while others were not (Morton 1956). Morton's test evaluates the differences between the lod score obtained for each family *i* at its own optimal recombination frequency ($\hat{\Theta}_i$) and the recombination frequency favoured by the overall data set ($\hat{\Theta}$):

$$Z = \sum_{i=1}^{k} 2\ln \frac{L_i(\hat{\theta}_i)}{L_i(\hat{\theta})} \qquad \textbf{(1)}$$

The test can be used in its original form, where the hypothesis of k different recombination fractions (heterogeneity) is tested against the null-hypothesis of only one Θ (homogeneity) at (k-1) df. Alternatively, one can divide the family material in k groups, rather than k families, based on clinical or geographical criteria and test for the presence of k different thetas. The latter variant of the M-test is generally known as the 'predivided sample test'. Calculations of the significance of findings, assuming an underlying chi-square distribution, are not entirely satisfactory. Rish has shown that the M-test is quite conservative under most circumstances (Risch, 1988). However, for medium theta values the test turned out to be more liberal. Moreover, it looses statistical power if the recombination fraction exceeds 0.3 (Risch, 1988; Ott, 1991).

The A-TEST

In 1963 Smith introduced a method for heterogeneity testing, currently known as 'the admixture test' or 'A-test' (Smith, 1963). A lod score under heterogeneity is defined ($Z_{(\alpha,\Theta)}$) as a function of the proportion of linked families (α) and the recombination fraction (Θ):

$$Z_{(\alpha,\theta)} = \log(\alpha \cdot 10^{Z_{(\theta)}} + (1-\alpha)) \qquad (2)$$

The A-test has been incorporated in the HOMOG programs (Ott, 1991). These programs optimise the $Z_{(\alpha,\Theta)}$ by testing a grid of α values ranging from 0 - 1. The asymptotic distribution is quite complicated, but can be approximated by a chi-square distribution (Faraway, 1993). Despite this, it is obvious that the A-test may be called the most popular method for heterogeneity analysis (Table 2.1). In a large series of studies the statistical power of the test has been evaluated and found to be substantial (Cavalli-Sforza and King, 1986; Levinson, 1993, Martinez and Goldin 1989; Chapter 2.2; Le Boyer et al., 1990; Chen et al., 1992). Using theoretical calculations (Cavalli-Sforza and King, 1986) or simulations (Levinson, 1993) the number of families segregating a dominant trait with two affected offspring required to reach $Z_{(\alpha,\Theta)}$=3.0 was estimated to be 83-95 at Θ=0.05 and α=0.5, depending on the informativeness of the marker. However, especially for dominant traits, evidence for heterogeneity can only be obtained if larger families or multipoint data are being used. 110 families with 3 affected offspring have been shown to be sufficient to demonstrate linkage and heterogeneity for a locus in between two PIC=0.37 markers, spaced 10cM, if α=0.5 (Martinez and Goldin, 1989). For a recessive model this was achieved with 96 families of the same type.

An advantage of the A-test is its broad applicability. Possible alternative applications to address specific problems are largely unexplored. An example of a relatively new application is the combined analysis of two separate chromosomal regions. In 1990 this application was introduced as 'Imaginary Chromosome Approach' (ICA) (Janssen et al., 1990). The ICA involves the analysis of positive evidence for linkage to one chromosomal region synchronously with (positive or negative) linkage information for one or more other regions. A complete analysis of two or three regions can therefore be performed by one single HOMOG2 or HOMOG3 analysis. In practice it means that HOMOG input files from different chromosomal regions are combined; location scores X_{a1} to X_{an} from the first chromosomal region are followed by X_{b1} to X_{bm} locations from a second region etc. The procedure is legitimate because the HOMOG programs are not designed to interpolate lod scores in between locations. The program just selects the combination of locations that provides the highest $Z_{(\alpha,\Theta)}$. This can easily be checked by rearranging the positions of locations X_1, X_2, X_3, X_4, X_5 in the input file into for instance X_2, X_4, X_5, X_1, X_3. Both analyses will reveal the same optimum. The synchronous analysis of multipoint data from multiple candidate regions in a combined analysis (ICA) renders significant extra power, compared with conventional analyses (Chapter 2.2). Only if α exceeds 60% conventional analyses provide more power, due to the lower significance thresholds that may be applied, as discussed later. ICA-like approaches have been used successfully in the analysis of familial Alzheimer disease, familial benign hypercalcemia, hyperthermia and tuberous sclerosis (Table 2.1).

Another example of an alternative application of the A-test is its use in exclusion mapping. While the results of other methods are confined to evidence for heterogeneity, the A-test also provides α and Θ estimates. This can be utilised by calculating a $Z_{(\alpha,\Theta)}$ for a series of α values at each map position. Loci can be excluded for a certain α range if $Z_{(\alpha,\Theta)}<2.0$. Using this approach, Hart recently excluded a gene for early onset periodontitis from a candidate area on chromosome 4q for all values of α between 40% and 100% (Hart et al., 1993).

The B-TEST

The youngest of the three most popular tests is the B-test, introduced by Risch in 1988 (Risch, 1988). For each family, linkage analysis will result in a lod score and a theta value. However, for small families only few theta values are possible. Resulting theta values follow a certain prior distribution, with one mass point (homogeneity) or multiple mass points (heterogeneity). Rish assumed that this may be represented as a beta distribution with two parameters. In a test procedure, called the B-test (B for Beta), the log likelihoods for this model and the null-hypothesis are evaluated. The B-test is somewhat more powerful than the A-test if there is at least one tightly linked locus (Risch 1988).

Less common tests

Another powerful test is the so-called C-test as proposed by MacLean (MacLean et al., 1992). It combines Morton's M-test with a computer simulation. The test is very reliable, since the simulation results provide all information on the probability distribution under the null hypothesis. Unfortunately, this step also means that the method demands a lot of computing time. Perhaps this explains why the method has not gained wide popularity. MacLean demonstrated that the test reaches or equals Neyman Pearson limits for optimal power and thereby suggested that the test would be more powerful than other tests. This suggestion was recently refuted by Faraway, who showed that there is very little difference in power between the A-test and the C-test (Faraway , 1994).

All methods discussed so far involve two steps: the calculation of lod scores per family, followed by a heterogeneity test. However, heterogeneity analysis within one single program instead of two separate programs is intuitively appealing, since not only inter-familial, but also intra-familial heterogeneity will be allowed. Recently, the LINKAGE program has been modified to allow for the direct analysis of a two-locus disease model (Schork et al., 1993). The so-called 'two-locus lod score' methods (TMLINK) are not only useful for the analysis of intra-familial heterogeneity, but also seem to be the only way to allow for differences in penetrance, phenocopy rate or mode of inheritance between two disease loci. Hereafter, we will use the expression 'mixed families' for families segregating disease alleles at two or more disease loci. The advantage of the 'two-locus' method is proportionate to the degree of mixed families in the data set. Epistasis - a situation in which disease alleles at both loci are required to obtain the phenotype - can be considered as an extreme example of heterogeneity with 100% mixed families and is thereby a typical problem for 'two-locus' analysis. Recently Tienari reported the use of this method in linkage analysis on multiple sclerosis. Under the assumption of epistasis, linkage to chromosome 6 and 18 was proven simultaneously (Tienari et al., 1994). Under comparable conditions (only one linked locus under investigation, low degree of mixed families, equal penetrance, etc.) the power of the 'two-locus' method has been shown to be about equal to the A-test, even if a certain proportion of mixed families is present (Goldin, 1992; Durner et al., 1992; Schork, 1993). A considerable disadvantage of the method is the substantial computational burden (Schork et al., 1993; Tienari et al., 1994).

Parameter estimates

Some of the methods described above operate via very useful parameters, while others do not. The parameters of the A-test - the recombination fraction Θ and the proportion of linked families α - are indispensable to positional cloning efforts and precise risk calculations. On the other hand the parameters used by the B-test are not really meant to be applied in practical gene mapping studies or risk calculations. The TMLINK-based 'two-locus' method can also be used to calculate Θ values, whereas the α can be approximated by optimising the gene frequencies and calculating the occurrence of the first disease allele relative to all disease alleles. However, it should be noted that the relative gene frequency is not by definition equal to the proportion of linked families (α). We recently demonstrated that considerable differences may occur - especially for recessive traits - due to differences in the likelihood for the homozygous state if gene frequencies are unequal (Chapter 2.3). This means that α represents a family type specific parameter. α values calculated for nuclear families can not be used for risk calculations in consanguineous pedigrees. The inequality of α and the relative gene frequency also implies that it is not allowed to transform the latter into the former if a recessive trait is studied by 'two-locus' (TMLINK) methods and that a biased α is to be expected if a dominant trait is analyzed this way.

Although the A-test is based on the assumption that all families are either linked or unlinked, it has been shown that a small proportion of mixed families can be handled. This leads, however, to an overestimation of Θ. In the presence of many mixed families, a slightly more accurate estimate of Θ may be obtained using the 'two-locus' lod score method (TMLINK) (Durner et al., 1992). However, the Θ estimates obtained by this procedure are very dependent on the correctness of each of the gene frequencies and can therefore be largely biased if wrong frequencies have been specified (Goldin, 1992).

Analytical evaluation of the performance of the A-test procedure with respect to the accuracy of the α and Θ estimates revealed that the estimates are most often quite reliable (Chapter 2.3). However, both estimators seem to be sensitive to dissimilarities in family size and structure between linked and unlinked types of families. Biases in Θ were shown to be minimal if multipoint analysis with flanking markers was performed. The ICA approach combines highly reliable Θ and α estimates if all loci can be captured in between flanking markers.

Planning a linkage study: Sampling strategies

It is virtually impossible to collect family material sufficient for detecting linkage, without - consciously or unconsciously - implementing a certain selection strategy. It is tempting to extend pedigrees as far as possible and to include as many promising families as possible, especially 'high density' pedigrees, with a high number of affecteds per family. On the other hand investigators may also use a negative selection, by excluding small families, bilineal families (with disease alleles coming in from more than one grandparent), families with only mildly affected patients, or families clearly displaying a certain mode of inheritance, while another mode was expected.

Bilineal pedigrees are often excluded from linkage studies, even under the assumption of homogeneity. By using simulated heterogeneous models, Durner has demonstrated that on average bilineal pedigrees have little effect on the maximum lod score and no measurable effect on the Θ estimate (Hodge, 1992; Durner et al., 1992). It is propitious that the A-test and 'two-locus' method are not seriously hampered by a limited number of mixed families, since mixed families will inevitably be present among bilineal families in case of locus heterogeneity. However, especially for the A-test, it is inconceivable that an analysis can be successful when the sample consists of a single bilineal family or a few of these families only. This is of considerable relevance, since many investigators have proposed that heterogeneity can be avoided by selecting only one or just a few extended pedigrees with many affected members. This policy involves mainly 'high density' pedigrees and therefore results in an increased proportion of mixed families. This can be overcome by using a single large pedigree from a genetic isolate. A disadvantage of this is that disorders found in an isolate may be rare and of no epidemiological significance.

If the linked form shows a recessive mode of inheritance, while the unlinked form is dominant, selecting 'high density' pedigrees will result in selection against the recessive form, at the cost of not finding linkage for the recessive form (Durner et al., 1992). Uncertainties about the true mode of inheritance may be a reason for selecting dominant-appearing or recessive-appearing families followed by an analysis under the corresponding model. According to Cox this selection only affects the power and does not lead to false evidence for locus heterogeneity if the results are compared (Cox et al., 1988). It should be noted however, that their highest lod scores were always obtained if no selection was made at all, regardless of the genetic model used in the analysis.

Selection procedures can affect the outcome of a heterogeneity test if the linked form is associated with a different phenotype or family size. Under these circumstances, negative selections based on family size will at least lead to a biased α estimate as discussed above (Chapter 2.3). Geographical selection procedures are another possible cause of α biases.

In summary, not much can be gained by applying negative selection criteria and combinations of these criteria may constitute real pitfalls. In contrast, positive actions, like extending the size and number of pedigrees, will increase the power and improve the accuracy of the parameters, provided that all types of families obtain equal attention.

Planning a linkage study: Strategies for analysis

When a linkage study has been planned thoughtfully, families have been ascertained, and markers have been tested, it all comes down to the actual analysis of data. If locus heterogeneity is already suspected, one only has to choose between the methods listed above. If it is unknown whether the disease may be caused by genes at one or more loci, the investigator also has to choose between analysis under homogeneity or analysis allowing for heterogeneity. Unless heterogeneity can be adequately excluded, as in the case of the single family from a genetic isolate, heterogeneity analysis is recommended, despite the extra computational effort that has to be made. A large series of studies have pointed out that disease loci can easily be incorrectly excluded from a chromosomal location under the assumption of homogeneity, while analysis under the assumption of heterogeneity does not result in a significant loss of power, even if the homogeneity model happens to be the correct model (Rich, 1989; Martinez and Goldin 1991; Chapter 2.2; Faraway 1993). If the proportion of linked families within the selected group of families under evaluation was between 40% and 100%, not much difference in power has been observed between analysis under homogeneity ($Z_{(\Theta)}$>3.0) and analysis under heterogeneity (evaluating $Z_{(\alpha,\Theta)}$) (Risch et al., 1989; Chapter 2.2; Faraway, 1993). The power of analysis under homogeneity is always poor if less than 40% of the families shows linkage. If multipoint linkage analysis is performed on heterogeneous data, the chances of finding a locus under the assumption of homogeneity are even worse (Martinez and Goldin, 1991). In 1986 Lander and Botstein postulated that a locus that accounts for 60% of all cases might still be statistically excluded under the false assumption of homogeneity. With an informative marker within 1 cM of the trait-related locus, linkage might still be excluded from a region of about 20 cM around the marker (Lander and Botstein, 1986). Three years later a tuberous sclerosis linkage study provided exclusion evidence for a region of 20 cM around the marker ABO (Kandt et al., 1989). Another four years later, at the Second International Chromosome 9 Workshop, a gene for tuberous sclerosis was assigned to an area of 1 or 2 cM around the same marker with an α of approximately 50%! (Kwiatkowski et al., 1993). The surprising parallel between theory and reality in this example emphasizes the importance of heterogeneity analysis. Several research groups, such as the Usher Syndrome Consortium (Bronya et al., 1992), have already made the A-test an integral part of their genome-wide search.

We have listed several robust methods for analysis under locus heterogeneity. The characteristics of the disease entity, the availability of fast computers and the number of markers to be tested, will help to make a choice. The predivided sample test will be a logical choice if a relationship between clinical differences and locus heterogeneity can be hypothesized, while the A-test seems a prudent choice if positional information is required. Table 2.1 demonstrates that many investigators apply more than one method.

Although multipoint analysis is absolutely not recommendable for the analysis of (probably) heterogeneous data under homogeneity, it may add to the robustness and power of an analysis under heterogeneity. We recently showed that clinical misclassifications of family members do not necessarily cause a significant bias if multipoint analysis is used in combination with the A-test procedure (Chapter 2.3). It has also been shown that the sensitivity of multipoint analysis towards misspecifications of the model can be solved by using inflated disease allele frequencies (Risch and Giuffra, 1992). Unfortunately this will lead to a biased α estimate if the A-test is used. To date our own policy is to include as much information in a heterogeneity test as possible, by - preferably - using multipoint data and if possible combining multipoint results from multiple candidate regions into one ICA analysis. In terms of statistical power, the latter appears to be ideal in comparison with any other strategy mentioned in this paper (Chapter 2.3).

Interpretation and presentation of results

Significance thresholds

Lod scores derived under heterogeneous models ($Z_{(\alpha,\Theta)}$) have an additional degree

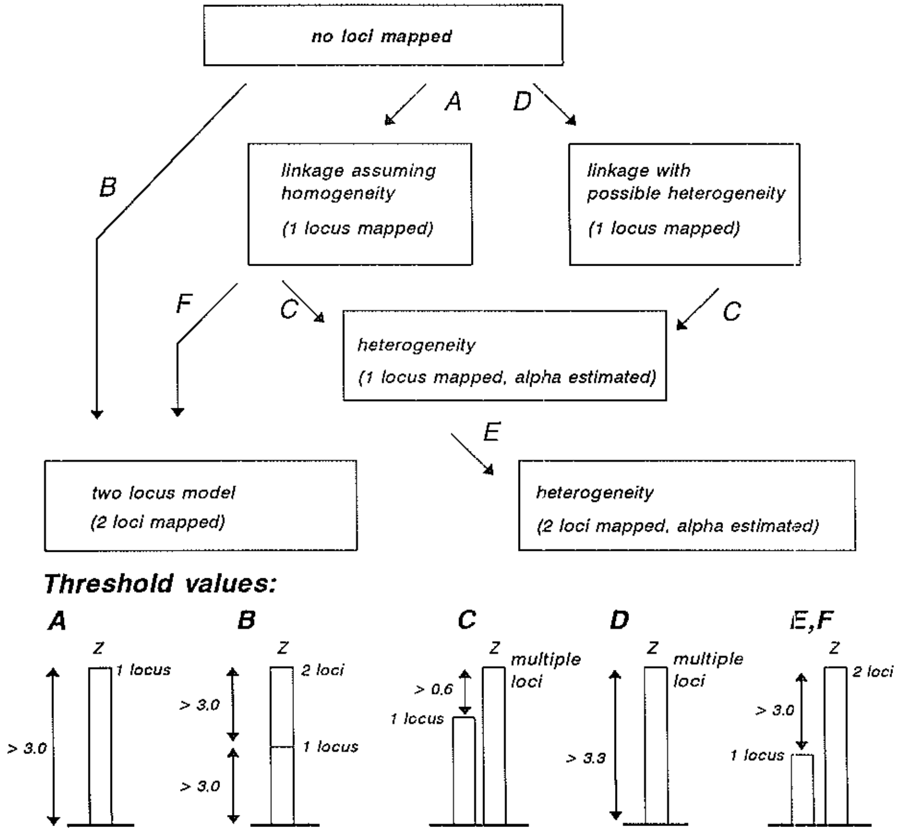


Figure 2.2: Mapping genes under heterogeneity is often a multi step process. Sometimes heterogeneity can only be demonstrated after mapping a locus under the assumption of homogeneity (A,C), while other studies involve heterogeneity analysis from the start (D). Using the ICA the number of mapped loci under a heterogeneous model can be increased (E). Alternatively, the 'two-locus' methods can be applied to find two loci without estimating alpha (B,F). The significance thesholds (expressed as lod scores) associated with these various routes are indicated below.

of freedom ($\alpha \neq 1.0$). Therefore if $Z_{(\alpha,\Theta)}=3.0$ (transformed via equation 2) the result is suggestive, but not significant. If a threshold of 3.0 is not sufficient, what level is? Assuming the A-test to have an asymptotic chi-square distribution, threshold values of 3.7 or 3.8 have been proposed (Risch, 1988, Chapter 2.2). After careful re-evaluation of the statistic Faraway concluded that a threshold of 3.3 (lnL=7.55) would be appropriate (Faraway, 1993). If two loci are being mapped simultaneously a lod score increase of 3.0 units compared with the single locus analysis should be demonstrated. For the 'two-locus' methods this means that 3.0+3.0=6.0 is sufficient to prove 'double linkage' and heterogeneity. Such a requirement seems reasonable, because the gene frequencies are fixed values and do not represent an extra degree of freedom, provided that no attempts are being made to optimise these parameters. The same criteria can be applied to ICA-like approaches. This means that 3.3 + 3.0 = 6.3 is the minimal requirement. If the first locus was associated with a lod score of 5.0, a total of 8.0 is sufficient evidence for 'double linkage' and heterogeneity. On the other hand, if the first locus was associated with a lod score of for instance 1.0, a total two-locus ICA score of 8.0 is only sufficient evidence for linkage of the second locus and heterogeneity. These criteria are outlined in Figure 2.2.

False claims for linkage under heterogeneity

The current policy to declare heterogeneity as soon as one fails to confirm linkage with a independent set of families or a single new family has a major disadvantage. This 'locus-exclusion-per-family strategy' neglects the possibility that the original claim for linkage might have been a type I error. The advancement of positional cloning efforts largely depends on mechanisms for detecting erroneous locus assignments as soon as possible. Strategies that lack such a mechanism, such as 'locus-exclusion-per-family', are therefore not favourable. It is much more sensible to re-evaluate an assignment using a large set of families and a method that involves a statistical test for locus heterogeneity. In the ideal situation the data set also contains the families used in the original linkage-assignment study. In 1990 an assignment of a schizophrenia locus to chromosome 5q - supported by a lod score of 6.49 (Sherrington et al., 1988) - was re-evaluated by McGuffin and coworkers (McGuffin et al., 1990). They combined data from 5 independent studies, including the original data supporting a locus on chromosome 5q. Using the A-test the lod score dropped from 6.49 to 3.36. In other words: the results contained only marginal evidence for a locus on 5q. Even more conclusive were the chromosome 11 exclusion data obtained by a consortium of tuberous sclerosis research groups (Chapter 3.2). Although a subset of the data originally gave rise to a lod score of 3.26 at D11S144 (Smith et al., 1990), the combined data provide strong evidence for exclusion for all values of α between 40% and 100% (Figure 2.3). The evidence for linkage was reduced to $Z_{(\alpha,\Theta)\max}$=0.41.

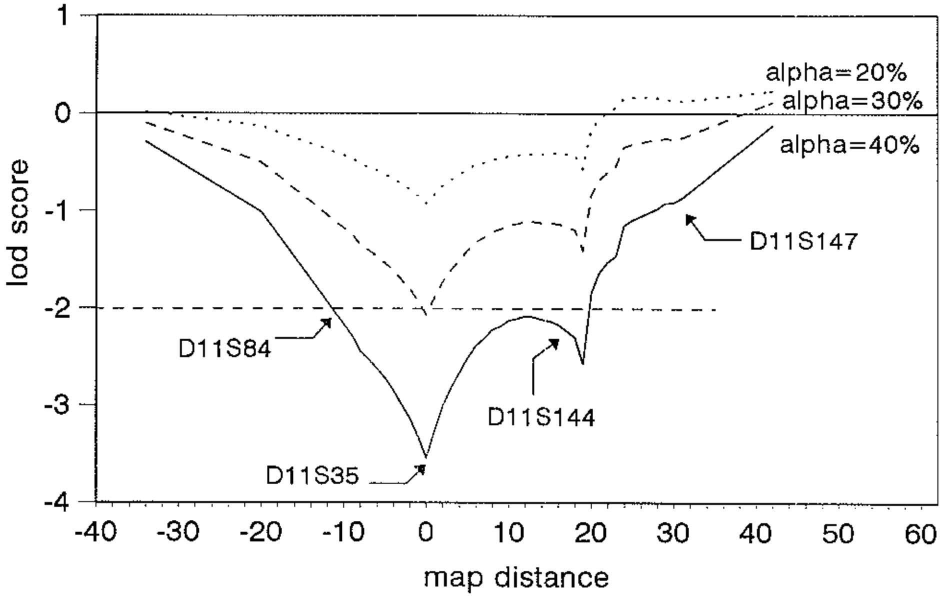


Figure 2.3: Tuberous sclerosis exclusion map of chromosome 11q21-23. Data presented in Chapter 3.2 were reanalysed. The three lines indicate $Z_{(\alpha,\Theta)}$ going from centromere to telomere through the 11q21-23 region with α=20% (top), α=30% and α=40% (bottom) (map distance in cM). This kind of maps may be helpful in the analysis of any heterogeneous disorder, because they clearly indicate where possible loci of minor importance still may reside and which markers have to be tested to study this possiblility.

According to Faraway, $Z_{(\Theta)}$>3.0 and $Z_{(\alpha,\Theta)}$>3.3 both correspond to a significance level of 1 error per 5000 tests. It is however unknown how many actual linkage analyses are being performed worldwide per year. This number might be quite high, especially if we include modifications of models and optimisations of parameters. Since mainly positive results will be published, we may expect that chance effects are an important source of incorrect claims for linkage and heterogeneity. Heterogeneity may also be falsely concluded if for instance DNA samples have been switched within a family or a certain group of families. Other possible causes of incorrectly deduced heterogeneity are germline mosaicism (Arveiler et al., 1990) and clustered occurrence of phenocopies, for instance due to environmental factors.

Presentation of results

There is considerable variation in the way the outcome of heterogeneity analyses is presented. With respect to the A-test the most common form of presentation is a table containing all elements of the chi-square test. Although quite fundamental, this form of presentation seems incompatible with a direct evaluation of $Z_{(\alpha,\Theta)}$, as recommended here. Furthermore, there is a growing tendency to present results as

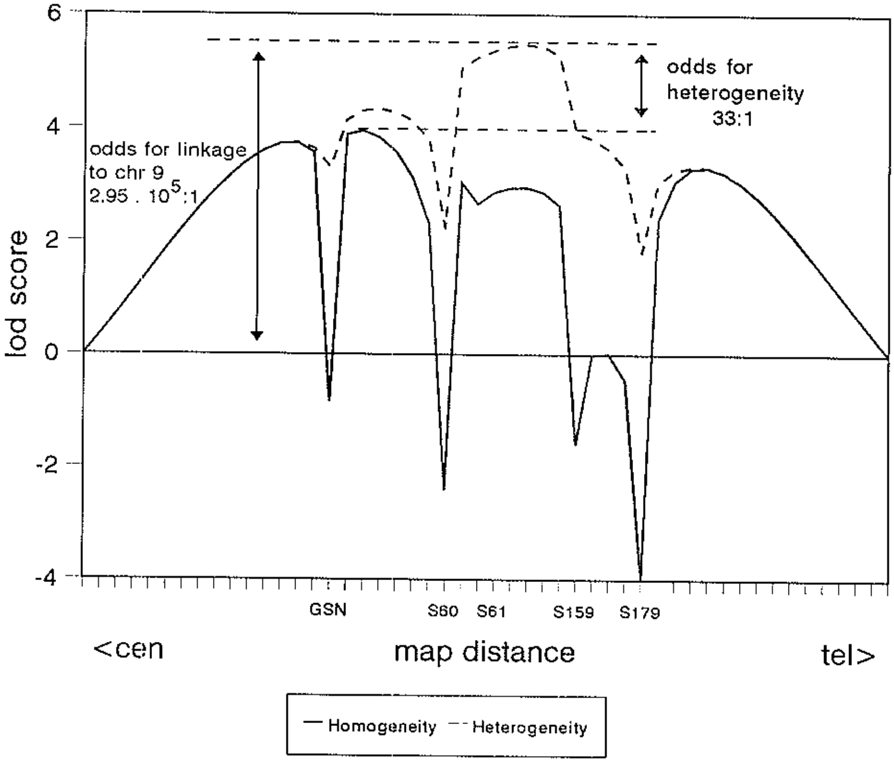


Figure 2.4: Example of rather elegant way of presenting positive results in case of heterogeneity: hereditary haemorrhagic telangiectasia. The thick line indicates the lod score ($Z_{(\Theta)}$) under the assumption of homogeneity, whereas the heterogeneity score ($Z_{(\alpha,\Theta)}$) is depicted by the broken line (map distance in cM). This single picture combines all relevant data: a) the evidence for locus heterogeneity, b) the most likely position of the linked locus (in between D9S61 and D9S159), c) the odds for linkage to chromosome 9 and d) the position of the locus under the assumption of homogeneity (near GSN;$Z_{(\Theta)}$=3.95). Data and figure were obtained from a linkage report by Heutink (Heutink et al., 1994). Methodological procedure: extended HOMOG output contains lnL values for each combination of α and Θ. For each map position (Θ) a maximum lnL was found by optimising α. This was converted into likelihoods, the depicted likelihood ratios and the depicted log likelihoods.

location scores (lod scores for each position on a map of markers), especially under homogeneity. There are two main classes of results: positive evidence for linkage and evidence for exclusion. For both classes we will indicate how results derived under heterogeneous models can be presented as location scores.

By using the HOMOG programs positive evidence for linkage can easily be transformed into location scores (see legend to Figure 2.4). By comparing the $Z_{(\alpha,\Theta)}$ with the $Z_{(\Theta)}$ (α=100%) likelihood ratios for linkage and heterogeneity can be obtained that serve as a good alternative for the components of a chi-square test. An example of such a presentation is given in Figure 2.4. If desired, α values can be depicted and a 90% confidence interval can be indicated by subtracting 1 unit from the $Z_{(\alpha,\Theta)\max}$. In the past a different approach has been used by several authors who presented location scores obtained from a multipoint/HOMOG analysis, followed by a selection step and another multipoint analysis on the linked families only. However, in this procedure information is used twice. This often leads to inflated lod scores (Chapter 3.3) and underestimates of the recombination fraction (Ott, 1983).

In principle, negative evidence can be presented in the same way as positive evidence, with the single limitation that it is not possible to estimate α. Even if the data set contains a large number of large families, it will still be possible that a locus - responsible for the disorder in a small minority of families - resides in the area. If α approaches 0.0 the $Z_{(\alpha,\Theta)}$ in equation (2) also approaches 0.0. This means that the usual threshold of -2.0 can never be reached for small α values. The best approach is to present location scores for several values of α. An example of such an exclusion map has been presented by Hart (Hart et al., 1993). Another example, derived according to the same principle, is shown in Figure 2.3.

Classification of families

Once linkage analysis has been performed and heterogeneity has been proven, a logic next step would be to determine which families are linked and which are not. A good classification of families is important for risk estimations in genetic counselling and may also lead to new impulses in clinical studies, since seemingly trivial phenotypic differences may turn out to be typical for the linked or unlinked class. Moreover differences in the course of the disease may exist, leading to a differentiated prognosis. As discussed above, family classifications aimed at further analysis of linkage data is not recommendable.

The HOMOG programs apply a classification method proposed by Ott, providing an estimated posterior probability w for linkage (Ott, 1983; Chapter 6.5). It is however unclear whether one should use w>0.5 or w>α as a classification rule (Ott, 1983; Ott, 1991). Although the former seems intuitively appealing, many investigators use the latter rule, which is equal to classifying all families with positive lod scores in the linked category. The posterior probability w can be used for risk calculations in genetic counselling as outlined by Ott (Ott, 1989). However, if differences in family size or phenotypic dissimilarities between the linked and unlinked forms exist, the procedure is not adequate (Chapter 2.3).

Currently it is common to calculate w by means of the A-test as described above. Nevertheless alternatives, such as the C-test, are available and robust (Maclean, 1994).

Conclusions

To date three methods for heterogeneity testing are frequently used: the predivided sample test (M-test), the A-test and the B-test. The predivided sample test is most often applied to test for locus heterogeneity in case two or more groups can be distinguished prior to performing linkage analysis. The test is very suitable for this purpose. The B-test is specially designed to analyze two groups that show a subtle difference in the recombination fraction and is slightly more powerful than the M-test. The A-test also performs well, especially in the case of loose linkage (Risch 1988). These tests thus complement each other nicely. The A-test is the most popular test and is commonly applied to a spectrum of different problems. We have therefore focused part of our overview on the A-test and highlighted the usefulness of the lod score derived via equation (2) ($Z_{(\alpha,\Theta)}$).

If the results of a heterogeneity analysis is presented as in Figure 2.3 or 2.4, it is immediately clear why linkage and locus heterogeneity have been concluded or to what extent linkage has been excluded. In many psychiatric disorders almost 100% of the genome has been excluded assuming homogeneity. This in itself is already an indication for locus heterogeneity. If exclusion maps derived under heterogeneous models become available for the whole genome, it will be possible to consider which parts of the genome might probably harbour a gene locus responsible for the disease in a minority of cases.

More in general we may conclude that the existence of many as yet unmapped genes for complex disorders, the growing interest in heterogeneity analysis and the availability of powerful methods comprise all necessary elements for the successful mapping of many of these genes within the near future.

References

Ajmar F, Mandich P, Bellone E, Abbruzzese G. Genetic analysis of Huntington disease in Italy. Am J Med Genet 1991;39:211-214.

Aksentijevich I, Gruberg L, Pras E, et al. Evidence for linkage of the gene causing familial Mediterranean fever to chromosome 17q in non-Ashkenazi Jewish families: second locus or type I error? Hum genet 1993;91:527-534.

Aldred MJ, Crawford PJM, Roberts E, et al. Genetic heterogeneity in X-linked amelogenesis imperfecta. Genomics 1992;14:567-573.

Anokhin A, Steinlein O, Fischer C, et al. A genetic study of the human low-voltage electroencephalogram. Hum genet 1992;90:99-112.

Arveiler B, de Saint-Basile G, Fischer A et al. Germ-line mosaicism simulates genetic heterogeneity in Wiskott-Aldrich syndrome. Am J Hum Genet 1990;46:906-911.

Azibi K, Bachner L, Beckmann JS, et al. Severe childhood autosomal recessive muscular dystrophy with the deficiency of the 50 kDa dystrophin-associated glycoprotein maps to chromosome 13q12. Hum Mol Genet 1993;2:1423-1428.

Ben Othmane K, Middleton LT, Lorraine J, et al. Localization of a gene (CMT2A) for autosomal dominant Charcot-Marie-Tooth disease type 2 to chromosome 1p and evidence of genetic heterogeneity. Genomics 1993;17:370-375.

Bonaventure J, Philippe C, Plessis G, et al. Linkage study in a large pedigree with Stickler syndrome: exclusion of COL2A1 as the mutant gene. Hum Genet 1992;90:164-168.

Bowden DW, Akots G, Rothschild CB, et al. Linkage analysis of maturity-onset diabetes of the young (MODY): genetic heterogeneity and nonpenetrance. Am J Hum Genet 1992;50:607-618.

Brzustowicz LM, Lehner T, Castilla LH, et al. Genetic mapping of chronic childhood-onset spinal muscular atrophy to chromosome 5q11.2-13.3. Nature 1990;344:540-541.

Brzustowicz LM, Mérette C, Kleyn PW, et al. Assessment of nonallelic genetic heterogeneity of chronic (type II and III) spinal muscular atrophy. Hum Hered 1993;43:380-387.

Cavalli-Sforza LL, King M-C. Detecting Linkage for genetically heterogeneous diseases and detecting heterogeneity with linkage data. Am J Hum Genet 1986;38:599-616

Chance PF, Bird TD, O'Connell P, et al. Genetic linkage and heterogeneity in type 1 Charcot-Marie-Tooth disease (hereditary motor and sensory neuropathy type 1). Am J Hum Genet 1990;47:915-925.

Chen J-D, Halliday F, Keith G, et al. Linkage heterogeneity between X-linked retinitis pigmentosa and a map of 10 RFLP loci. Am J Hum Genet 1989;45:401-411.

Chen WJ, Faraone SV, Tsuang MT. Linkage studies of schizophrenia: a simulation study of statistical power. Genet Epidemiol 1992;9:123-39.

Chenevix-Trench G, Wicking C, Berkman J, et al. Further localization of the gene for nevoid basl cell carcinoma syndrome (NBCCS) in 15 Australasian families: linkage and loss of heterozygosity. Am J Hum Genet 1993;53:760-767.

Conneally PM, Haines JL, Tanzi RE, et al. Huntington disease: no evidence for locus heterogeneity. Genomics 1989;5:304-308.

Cook A, Raskind W, Halloran Blanton S, et al. Genetic heterogeneity in families with hereditary multiple exostoses. Am J Hum Genet 1993;53:71-79.

Coon H, Byerley W, Holik J, et al. Linkage analysis of schizophrenia with five dopamine receptor genes in nine pedigrees. Am J Hum Genet 1993;52:327-334.

Cox NJ, Hodge SE, Marazita M et al. Some effects of selection strategies on linkage analysis. Genet Epidemiol 1988;5:289-97.

Crossey PA, Maher ER, Jones MH, et al. Genetic linkage between Von Hippel-Lindau disease and three microsatellite polymorphisms refines the localisation of the VHL locus. Hum Mol Genet 1993;2:279-282.

Crowe RR, Noyes R, Samuelson S, et al. Close linkage between panic disorder and α-haptoglobin excluded in 10 families. Arch Gen Psychiatry 1990;47:377-380.

Durner M, Greenberg DA, Hodge SE. Inter- and intrafamilial heterogeneity: effective sampling strategies and comparison of analysis methods. Am J Hum Genet 1992;51:859-70.

Easton DF, Bishop DT, Ford D, et al. Genetic analysis in familial breast and ovarian cancer: results from 214 families. Am J Hum Genet 1993;52:678-701.

Elbein SC, Hoffman MD, Matsutani A, Permutt MA. linkage analysis of GLUT1 (HepG2) and GLUT2 (liver/islet) genes in familial NIDDM. Diabetes 1992;41:1660-1667.

Epstein ND, Fananapazir L, Lin HJ, et al. Evidence of genetic heterogeneity in five kindreds with familial hypertrophic cardiomyopathy. Circulation 1992;85:635-647.

Faraway JJ. Distribution of the admixture test for the detection of linkage under heterogeneity. Genet Epidemiol 1993;10:75-83.

Faraway JJ. Testing for linkage under heterogeneity: A test versus C test. Am J Hum Genet 1994;54:563-564.

Farrer LA, Grundfast, KM, Amos J, et al. Waardenburg syndrome (WS) type I is caused by defects at multiple loci, one of which is near ALPP on chromosome 2: first report of the WS consortium. Am J Hum Genet 1992;50:902-913.

Froguel Ph, Vaxillaire M, Sun F, et al. Close linkage of glucokinase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus. Nature 1992;356:162-164.

Gershon ES, Martinez M, Goldin L, Gelernter J, Silver J. Detection of marker associations with a dominant disease gene in genetically complex and heterogeneous deiseases. Am J Hum Genet 1989;45:578-85.

Gilliam TC, Brzustowicz LM, Castilla LH, et al. Genetic homogeneity between acute and chronic forms of spinal muscular dystrophy. Nature 1990;345:823-825.

Gispert S, Twells R, Orozco G, et al. Chromosomal assignment of the second locus for autosomal dominant cerebellar ataxia (SCA2) to chromosome 12q23-24.1. Nat Genet 1993;4:295-299.

Goldin LR, Gershon ES. Power of the affected-sib-pair method for heterogeneous disorders. Genet Epidemiol 1988;5:35-42.

Goldin LR. Detection of linkage under heterogeneity: comparison of the two-locus vs. admixture models. Genet Epidemiol 1992;9:61-6.

Goldstein AM, Dracopoli NC, Ho EC, et al. Further evidence for a locus for cutaneous malignant melanoma-dysplastic nevus (CMM/DN) on chromosome 1p, and evidence for genetic heterogeneity. Am J Hum Genet 1993;52:537-550.

Goodfellow PJ, White BN, Holden JJA et al. Linkage analysis of a DNA marker localized to 20p12 and multiple endocrine neoplasia type 2A. Am J Hum Genet 1985;37:890-7.

Haile RW, Cortessis VK, Millikan R, et al. A linkage analysis of D17S74 (CMM86) in thirty-five famiies with premenopausal bilateral breast cancer. Cancer Res 1993;53:212-214.

Haines JL, Short MP, Kwiatkowski DJ, et al. Localization of one gene for tuberous sclerosis within 9q32-9q34, and further evidence for heterogeneity. Am J Hum Genet 1991;49:764-772.

Hart TC, Marazita ML, McCanna KM, et al., Reevaluation of the chromosome 4q candidate region for early onset periodontitis. Hum Genet 1993;91:416-22.

Heath H, Jackson CE, Otterud B, Leppert MF. Genetic linkage analysis in familial benign (hypocalciuric) hypercalcemia: evidence for locus heterogeneity. Am J Hum Genet 1993;53:193-200.

Heutink P, Haitjema T, Breedveld GJ et al. Linkage of hereditary haemorrhagic telangiectasia to chromosome 9q34 and evidence for locus heterogeneity. J Med Genet 1994;31:933-936.

Hodge SE. Do bilineal pedigrees represent a problem for linkage analysis? Basic principles and simulation results for single-gene diseases with no heterogeneity. Genet Epidemiol 1992;9:191-206.

Hulten MAJ. Linkage heterogeneity in autosomal dominant polycystic kidney disease. Lancet 1988;ii:451-2.

Ionasescu VV, Trofatter J, Haines JL, et al. Heterogeneity in X-linked recessive Charcot-Marie-Tooth neuropathy. Am J Hum Genet 1991;48:1075-1083.

Ionasescu VV, Trofatter J, Haines JL, et al. X-linked recessive Charcot-Marie-Tooth neuropathy: clinical and genetic study. Muscle Nerve 1992;15:368-373.

Izquierdo L, Porteous M, Paramo PG, Connor JM. Evidence for genetic heterogeneity in hereditary hydronephrosis caused by pelvi-ureteric junction obstruction, with one locus assigned to chromosome 6p. Hum Genet 1992;89:557-560.

Janssen LAJ, Sandkuyl LA, Merkens EC et al. Genetic heterogeneity in tuberous sclerosis. Genomics 1990;8:237-242

Järvelä I. Infantile neuronal ceroid lipofuscinosis (CLN1): linkage disequilibrium in the Finnish population and evidence that variant late infantile form (variant CLN2) represents a nonallelic locus. Genomics 1991;10:333-337.

Kainulainen K, Steinmann B, Collins F, et al. Marfan syndrome: no evidence for heterogeneity in different populations, and more precise mapping of the gene. Am J Hum Genet 1991;49:662-667.

Kandt RS, Pericak-Vance MA, Hung WY et al. Absence of linkage of ABO blood group locus to familial tuberous sclerosis. Exp Neurol 1989;104:223-8.

Kaplan J, Gerber S, Bonneau D, et al. A gene for Usher syndrome type 1 (USH1A) maps to chromosome 14q. Genomics 1992;14:979-987.

Keats BJB, Todorov AA, Atwood LD, et al. Linkage studies of Usher syndrome type 1: exclusion results from the Usher syndrome consortium. Genomics 1992;14:707-714.

Kennedy JL, Giuffra LA, Moises HW, et al. Molecular genetic studies in schizophrenia. Schizophr Bull 1989;15:383-391.

Khati C, Stevanin G, Durr A, et al. Genetic heterogeneity of autosomal dominant cerebellar ataxia type 1. Neurology 1993;43:1131-1137.

Kwiatkowski DJ, Armour J, Bale AE et al. Report of the Second International Workshop on Human Chromosome 9. Cytogenet Cell Genet 1993;64:94-106.

Lander ES, Botstein D. Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc Natl Acad Sci USA 1986;83:7353-57.

LeBoyer M, Babron MC, Clerget-Darpoux XX. Sampling strategy in linkage studies of affective disorders. Psychol Med 1990;20:573-9.

Levinson DF. Power to detect linkage with heterogeneity in samples of small nuclear families. Am J Med Genet 1993;48:94-102

Levitt RC, Olckers A, Meyers S, et al. Evidence for the localization of a malignant hyperthermia susceptibility locus (MHS2) to human chromosome 17q. Genomics 1992;14:562-566.

Lim LCC, Gurling H, Curtis D, et al. Linkage between tyrosine hydroxylase gene and affective disorder cannot be excluded in two of six pedigrees. Am J Med Genet 1993;48:223-228.

MacLean CH, Ploughman LM, Diehl SR, Kendler KS. A new test for linkage in the presence of locus heterogeneity. Am J Hum Genet 1992;50:1259-66.

Maclean CJ, Sham PC, Ploughman LM et al. Reply to Faraway (letter). Am J Hum Genet 1994;54:564-567.

Maher Er, Bentley E, Payne SJ, et al. Presymptomatic diagnosis of von Hippel-Lindau disease with flanking DNA markers. J Med Genet 1992;29:902-905.

Mann WR, Venkatraj VS, Allen RG, et al. Fanconi anemia: evidence for linkage heterogeneity on chromosome 20q. Genomics 1991;9:329-337.

Mareni C, Stella A, Origone P, et al. Linkage studies in Italian families with familial adenomatous polyposis. Hum Genet 1993;90:545-550.

Martinez MM, Goldin LR. The detection of linkage and heterogeneity in nuclear families for complex disorders: One versus two marker loci. Am J Hum Genet 1989;44:552-9

Martinez M, Goldin LR. Detection of linkage for heterogeneous disorders by using multipoint linkage analysis. Am J Hum Genet 1991;49:1300-5.

Mazoyer S, Lalle P, Narod SA, et al. Linkage analysis of 19 French breast cancer families with five chromosome 17q markers. Am J Hum Genet 1993;52:754-760.

McAlpine PJ, Feasby TE, Hahn AF, et al. Localization of a locus for Charcot-Marie-Tooth neuropathy type 1a (CMT1A) to chromosome 17. Genomics 1990;7:408-415.

McGuffin P, Sargeant M, Hetti G, et al. Exclusion of a schizophrenia susceptibility gene from the chromosome 5q11-q13 region: new data and a reanalysis of previous reports. Am J Hum Genet 1990;47:524-535.

Morton NE. The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. Am J Hum Genet 1956;8:80-96.

M'Rad R, Sanak M, Deschenes G, et al. Alport syndrome: a genetic study of 31 families. Hum Genet 1992;90:420-426.

Mullan M, Houlden H, Windelspecht M, et al. A locus for familial early-onset Alzheimer's disease on the long arm of chromosome 14, proximal to the α1-antichymotrypsin gene. Nat Genet 1992;2: 340-342.

Musarella MA, Weleber RG, Murphey WH, et al. Assignment of the gene for complete X-linked congenital stationary night blindness (CSNB1) to Xp11.3. Genomics 1989;5:727-737.

Musarella MA, Anson-Cartwright L, Leal SM, et al. Multipoint linkage analysis and heterogeneity testing in 20 X-linked retinitis pigmentosa families. Genomics 1990;8:286-296.

Nancarrow DJ, Mann GJ, Holland EA, et al. Confirmation of chromosome 9p linkage in familial melanoma. Am J Hum Genet 1993;53:936-942.

Narod SA, Sobol H, Nakamura Y, et al. Linkage analysis of hereditary thyroid carcinoma with and without pheochromocytoma. Hum Genet 1989;83:353-358.

Narod SA, Lavoué M-F, Morgan K, et al. Genetic analysis of 24 French families with multiple endocrine neoplasia type 2A. Am J Hum Genet 1992a;51:469-477.

Narod SA, Parry DM, Parboosingh J, et al. Neurofibromatosis type 2 appears to be a genetically homogeneous disease. Am J Hum Genet 1992b;51:486-496.

Northrup H, Kwiatkowski DJ, Roach ES, et al. Evidence for genetic heterogeneity in tuberous sclerosis: one locus on chromosome 9 and at least one locus elsewhere. Am J Hum Genet 1992;51:709-720.

Ott J. Linkage analysis and family classification under heterogeneity. Ann Hum Genet 1983;47:311-320.

Ott J, Bhattacharya S, Chen JD, et al. Localizing multiple X chromosome-linked retinitis pigmentosa loci using multilocus homogeneity tests. Proc Natl Acad Sci USA 1990;87:701-704.

Ott J. Analysis of human genetic linkage. Revised edition. Baltimore: Johns Hopkins University Press, 1991.

Passos-Bueno MR, Richard I, Vainzof M, et al. Evidence of genetic heterogeneity in the autosomal recessive adult forms of limb-girdle muscular dystrophy following linkage analysis with 15q probes in Brazilian families. J Med Genet 1993;30:385-387.

Pericak-Vance MA, Bebout JL, Gaskell PC, et al. Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. Am J Hum Genet 1991;48:1034-1050.

Peters DJM, Sandkuijl LA. Genetic heterogeneity of polycystic kidney disease in Europe. Contributions Nephrol 1992;97:128-139.

Peters DJM, Spruit L, Saris JJ, et al. Chromosome 4 localization of a second gene for autosomal dominant polycystic kidney disease. Nat Genet 1993;5:359-362.

Price RA. Affective disorder not linked to HLA. Genet Epidemiol 1989;6:299-304.

Reardon W, Middleton-Price HR, Sandkuijl L, et al. A multipedigree linkage study of X-linked deafness: linkage to Xq13-q21 and evidence for genetic heterogeneity. Genomics 1991;11:885-894.

Richards FM, Maher ER, Latif F, et al. Detailed genetic mapping of the von Hippel-Lindau disease tumour suppressor gene. J Med Genet 1993;30:104-107.

Risch N. A new statistical test for linkage heterogeneity. Am J Hum Genet 1988;42:353-64.

Risch N. Linkage detection tests under heterogeneity. Genet Epidemiol 1989;6:473-80.

Risch N, Giuffra L. Model misspecification and multipoint linkage analysis. Hum Hered. 1992;42:77-92.

Sarfarazi M, Wijmenga C, Upadhyaya M, et al. Regional mapping of facioscapulohumeral muscular dystrophy gene on 4q35: combined analysis of an international consortium. Am J Hum Genet 1992a;51:396-403.

Sarfarazi M, Tsipouras P, Del Mastro R, et al. A linkage map of 10 loci flanking the Marfan syndrome locus on 15q: results of an International Consortium Study. J Med Genet 1992b;29:75-80.

Schellenberg GD, Payami H, Wijsman EM. Chromosome 14 and late-onset familial Alzheimer disease (FAD). Am J Hum Genet 1993;53:619-628.

Schork NJ, Boehnke M, Terwilliger JD, Ott J. Two-trait-locus linkage analysis: A powerful strategy for mapping complex genetic traits. Am J Hum Genet 1993;53:1127-36.

Sherrington R, Brynjolfsson J, Petursson H et al. Localization of a susceptibility locus for schizophrenia on chromosome 5. Nature 1988;336:164-167.

Shohat M, Bu X, Shohat T, et al. The gene for familial Mediterranean fever in both Armenians and non-Ashkenazi Jews is linked to the α-globin complex on 16p: evidence for locus homogeneity. Am J Hum Genet 1992;51:1349-1354.

Siddique T, Figlewicz DA, Pericak-Vance MA, et al. Linkage of a gene causing familial amyotrophic lateral sclerosis to chromosome 21 and evidence of genetic-locus heterogeneity. N Engl J Med 1991;324:1381-1384.

Smith CAB. Testing for heterogeneity of recombination fraction values in human genetics. Ann Hum Genet 1963;27:175-82.

Smith M, Smalley S, Cantor R, et al. Mapping of a gene determining tuberous sclerosis to human chromosome 11q14-11q23. Genomics 1990;6:105-114.

Smith RJH, Pelias MZ, Daiger SP, et al. Clinical variability and genetic heterogeneity within the Acadian Usher population. Am J Med Genet 1992;43:964-969.

Smith SA, Easton DF, Ford D, et al. Genetic heterogeneity and localization of a familial breast-ovarian cancer gene on chromosome 17q12-q21. Am J Hum genet 1993;52:767-776.

Smith SD, Pennington BF, Kimberling WJ, Ing PS. Familial dyslexia: use of genetic linkage data to define subtypes. J Am Acad Child Adolesc Psychiatry 1990;29:204-213.

Speer MC, Goldgaber D, Goldfarb LG, et al. Support of linkage of Gerstmann-Sträussler-Scheinker syndrome to the prion protein gene on chromosome 20p12-pter. Genomics 1991;9:366-368.

Steinlein O, Anokhin A, Yping M, et al. Localization of a gene for the human low-voltage EEG on 20q and genetic heterogeneity. Genomics 1992;12:69-73.

Sykes B, Ogilvie D, Wordsworth P, et al. Consistent linkage of dominantly inherited osteogenesis imperfecta to the type 1 collagen loci: COL1A1 and COL1A2. Am J Hum Genet 1990;46:293-307.

Tienari PJ, Terwilliger JD, Ott J et al. Two-locus linkage analysis in multiple sclerosis (MS). Genomics 1994;19:320-325.

Thierfelder L, MacRae C, Watkins H, et al. A familial hypertrophic cardiomyopathy locus maps to chromosome 15q2. Proc Natl Acad Sci USA 1993;90:6270-6274.

Upadhyaya M, Sarfarazi M, Huson SM, et al. Close flanking markers for neurofibromatosis type I (NF1). Am J Hum Genet 1989;44:41-47.

Van Haeringen A, Bergman W, Nelen MR et al., Exclusion of the dysplastic nevus syndrome (DNS) locus from the short arm of chromosome 1 by linkage studies in Dutch families. Genomics 1989;5:61-4.

Vilkki J, Ott J, Savontaus M-L, et al. Optic atrophy in Leber hereditary optic neuroretinopathy in probably determined by an X-chromosomal gene closely linked to DXS7. Am J Hum Genet 1991;48:486-491.

Wang Jabs E, Li X, Coss CA, et al. Mapping the Treacher Collins syndrome locus to 5q31.3→q33.3. Genomics 1991;11:193-198.

Warner TT, Fletcher NA, Davis MB, et al. Linkage analysis in British and French families with idiopathic torsion dystonia. Brain 1993;116:739-744.

Weeks DE, Ott J. Risk calculations under heterogeneity. Am J Hum Genet 1989;45:819-821.

Weissbecker KA, Durner M, Janz D, et al. Confirmation of linkage between juvenile myoclonic epilepsy locus and the HLA region of chromosome 6. Am J Med Genet 1991;38:31-36.

Whitehouse WP, Rees M, Curtis D, et al. Linkage analysis of idiopathic generalized epilepsy (IGE) and marker loci on chromosome 6p in families of patients with juvenile myoclonic epelepsy: no evidence for an epilepsy locus in the HLA region. Am J Hum Genet 1993;53:652-662.

Wijmenga C, Padberg GW, Moerer P, et al. Mapping of facioscapulohumeral muscular dystrophy gene to chromosome 4q35-qter by multipoint linkage analysis and in situ hybridization. Genomics 1991;9:570-575.

Wijsman EM. Genetic analysis of Alzheimer's disease: a summary of contributions to GAW8. Genet Epidemiol 1993;10:349-360.

Williams R, Vesa J, Järvelä I, et al. Genetic heterogeneity in neuronal ceroid lipofuscinosis (NCL): evidence that the late-infantile subtype (Jansky-Bielschowsky disease; CLN2) is not an allelic form of the juvenile or infantile subtypes. Am J Hum Genet 1993;53:931-935.

Wilton AN, Cooper DW, Brennecke SP, et et. Absence of close linkage between maternal genes for susceptibility to pre-eclampsia/eclampsia and HLA DRß. Lancet 1990;336:635-637.

Wright AF, Teague PW, Pound SE, et al. A study of genetic linkage heterogeneity in 35 adult-onset polycystic kidney disease families. Hum Genet 1993;90:569-571.

Young RP, Sharp PA, Lynch JR, et al. Confirmation of genetic linkage between atopic IgE responses and chromosome 11q13. J Med Genet 1992;29:236-238.

COMPUTER SIMULATION OF LINKAGE AND HETEROGENEITY IN TUBEROUS SCLEROSIS: A CRITICAL EVALUATION OF THE COLLABORATIVE FAMILY DATA

LAJ Janssen
LA Sandkuijl
JR Sampson
DJJ Halley

J Med Genet 1992; 29: 867-874.

Abstract

The existence of locus heterogeneity for a genetic disease may complicate linkage studies considerably, especially when very few large disease families are available. In this situation a modest collection of families is unlikely to be sufficient for successful localisation of one or more disease genes. Recently, eight research groups working on tuberous sclerosis (TSC) brought together linkage data pertaining to the candidate chromosomes 9, 11 and 12, for a large group of families. In a series of simulation studies we determined the probability of detecting linkage and linkage heterogeneity in this set of families.

On average the TSC families are very small, in most cases there are less than 2 informative meioses. The size distribution of chromosome 9 linked families was similar to that of non-linked families. This indicates that a dramatic difference in the clinical severity of major genetic forms of TSC is unlikely.

The results of our simulation studies show that this set of families can generate highly significant evidence for linkage and heterogeneity. When two TSC genes are equally common, the strongest evidence for linkage and heterogeneity could be obtained using a method based on the incorporation of multiple candidate regions in a single analysis, with an average lod score of 24.27.

Introduction

During the past decade, linkage analysis has been successfully applied to localise the genes responsible for many different inherited diseases (McKusick, 1991). The availability of a large series of highly polymorphic markers throughout the genome has facilitated linkage mapping considerably. For a number of diseases, however, linkage results have been reported that have not been confirmed by further studies, as was initially the case for Charcot Marie Tooth disease, and is still the case for a number of psychiatric disorders including schizophrenia and manic depression (St Clair et al., 1989; Kelsoe et al., 1989). Locus heterogeneity has been suggested as a possible explanation for such differing results. While it is commonly recognised that locus heterogeneity may complicate linkage studies, it has not precluded the accurate localisation of major genes for polycystic kidney disease and, eventually, Charcot Marie Tooth disease. For mapping under locus heterogeneity, it is necessary to identify the subset of families that show linkage to a given chromosomal region. For large families, that can generate significant evidence for linkage when studied individually, this can be done directly. When the average family size is small, however, this distinction cannot be made in a straightforward manner, and it is intuitively clear that a large number of families will be required for detection of one or several of the responsible genes. This has been confirmed in theoretical studies using computer simulation and analytical methods (Narod, 1991; Martinez and Goldin, 1990).

Various statistical methods have become available to make optimal use of the linkage information in a series of (relatively) small families. The most commonly applied test is the admixture test (Smith, 1963; Ott, 1991), which involves simultaneous estimation of the location(s) of the responsible genes and of the proportion of families segregating for each of those genes. This method has been instrumental in detecting several loci causing retinitis pigmentosa on the X chromosome (Ott et al., 1990).

For tuberous sclerosis, a neuro-cutaneous disorder characterised by widespread hamartosis, linkage studies have yielded conflicting results. While some studies provided evidence for a locus close to the ABO blood group gene on chromosome 9 (Connor et al., 1987; Fryer et al., 1987), other studies could not confirm this linkage, and indicated chromosomes 11 and 12 as possible alternative locations (Smith et al., 1990; Fahsold et al., 1991). Considerable efforts have been made to combine linkage data from the groups participating in the TSC-consortium, and to analyze that data with a variety of statistical methods. Unfortunately, these studies have not entirely resolved the controversy, although consistent evidence emerged for a TSC1 gene on the long arm of chromosome 9 and for the existence of locus heterogeneity (Smith et al., 1991; Povey et al., 1991; Chapter 3.1).

For the simultaneous evaluation of the different chromosomal regions that may be involved in TSC, we have developed an extension of the admixture approach (Chapter 2.1; Janssen et al., 1990), which we have previously called the Imaginary Chromosome (IC) approach. In a single analysis, this method will evaluate linkage results for all relevant chromosomal regions. For a given family, this allows simultaneous evaluation of positive evidence for one of the regions and negative linkage information for the other regions. While this approach has the advantage of making maximal use of all available information, it shares with some other statistical methods the disadvantage of not being transparent. Thus, it is not immediately apparent how much evidence a given data set can potentially yield in an analysis of linkage and heterogeneity which includes several chromosomal regions simultaneously. Also, it is not entirely clear what level of evidence is to be regarded as 'significant', the criterion that researchers are most interested in. Ott (1991) suggested that the likelihood ratio favouring linkage or heterogeneity should always be reported, leaving the decision whether a certain likelihood ratio is to be regarded as significant to the individual researchers. Indeed, uniform guidelines are hard to define, as the statistical behaviour of combined tests of linkage and heterogeneity in multipoint linkage analysis has not been adequately investigated.

Recently, we have analyzed the collaborative TSC data set, after rigorous checking for data errors, and after reassessment of diagnoses following uniform criteria. The results, presented in Chapter 3.2, indicate existence of a major locus (TSC1) on chromosome 9 and at least one other locus elsewhere in the genome. Chromosome 12 might harbour another TSC gene of minor importance. To help interpret these results, we carried out extensive computer simulations, employing both the

conventional admixture test and its multilocus extension. In the current study, we will address the following questions:
- How much linkage information is potentially present in the combined TSC families, and how is that information distributed over the families ?
- Is there any difference in effective size between the families that show linkage to the chromosome 9 markers and those that yield negative lod scores ?
- How much evidence for linkage and heterogeneity may one expect to find under various degrees of locus heterogeneity in a data set of this size ?

The answers to these questions can provide both information about the power and size of the collaborative TSC data set and insight into the performance of the admixture test when using multiple candidate regions.

Materials and methods

Family material

In our studies we used 128 families from 8 different centres: Irvine, Boston, Houston, Durham (USA), Cardiff, London (UK), Erlangen (Germany) and Rotterdam (The Netherlands). This set of families is identical to that used in the collaborative linkage study (Chapter 3.2). Almost all families used in this study have been described before (Smith et al., 1990; Fahsold et al., 1991; Janssen et al., 1990; Kandt et al., 1991; Sampson et al., 1989; Northrup et al., 1992; Haines et al., 1991a,b; Povey, 1994), although phenotypic data have been fully reviewed with amendment of every individual's affection status for these studies (Chapter 3.2).

For our simulation studies we classified all individuals for whom actual marker tests had been carried out as "available", others were classified as "unavailable" for DNA analysis. Apart from family structure, diagnostic assessment and availability no other family data was used.

Computer simulation

In our simulation studies of linkage and heterogeneity, three separate steps can be distinguished:

i) Preparation of hypothetical marker data for all persons for whom DNA samples were available. These marker data were generated using the computer program SLINK (Ott et al., 1989), which takes the disease status of all family members into account. Penetrance and gene frequency were as used in the collaborative linkage study (Chapter 3.2). We assumed that a hypothetical marker was located at 5 % recombination from each of the TSC genes. For each family, 100 distinct replicates were made, each with new hypothetical marker data. Three levels of informativeness were simulated: eight alleles (PIC value 0.86), four alleles (PIC value 0.7) and two alleles (PIC value 0.375). In a separate series of simulations, we generated data for markers with similar informativeness, but unlinked to any of the TSC genes. From the simulated data for each individual family, simulated versions of the entire data set were created by selecting for each family a linked or unlinked replicate with probabilities α and 1-α respectively (where α denotes the assumed proportion of families linked to a particular region).

ii) The lod score calculations, by regular analysis of linkage using the generated marker data. Lod scores were calculated for each replicate, varying the recombination frequency from 0.0 to 0.5 in steps of 0.01. Lod score calculations on these replicates were performed batch-wise, using the MLINK option of the LINKAGE package (version 5.03). In these calculations, we applied the same gene frequency as in the collaborative linkage analysis ($1 \cdot 10^{-4}$) (Chapter 3.2), while no allowance was made for new mutations.

iii) Heterogeneity analysis using single locus and multilocus versions of the admixture test. All admixture tests that we applied are based on the assumption that disease genes may exist at two or more locations in the genome, but that penetrance and mutation frequency are equal for all genes involved. Lod scores calculated in the previous step were transferred to a slightly modified version of the HOMOG program (Ott, 1991). For simultaneous evaluation of several chromosomal regions the HOMOG2 program was used. The necessary input was created by appending a list of lod scores for an unlinked replicate of a given family to a list of lod scores for a linked replicate of the same family or vice versa (probabilities α and 1-α respectively). Thereby we created an input file containing two chromosomal regions, each represented by one marker. Each family was linked to either the first or the second region. Together the regions formed a so-called "imaginary chromosome".

Information content of families

In order to get a more precise measure of the usefulness of each family for linkage studies, we calculated the mean lod score (expected lod score) at 5% recombination over the 100 replicates obtained in simulations of an 8-allele marker. The resulting mean $Z_{(\Theta=0.05)}$ was divided by 0.215 (the expected lod score at Θ=0.05 for one completely informative meiosis). The quotient was termed the "effective number of informative meioses" (EFNIM), since it enables us to compare families with completely different structures. This measure is closely related to Edwards' "equivalent observations" (Edwards, 1976) (see Appendix I).

Results

Power of TSC families to detect linkage under homogeneity

We estimated the power to detect linkage under homogeneity, by simulating markers with low and high informativeness. We specified a recombination frequency of 5 % for the following reasons: i) the chromosome-9 markers most closely linked to the TSC1 locus show approximately 5 % recombination (Chapter 3.2), and ii) in a random genome search it is frequently attempted to test markers that divide the genome into 10 cM intervals, which implies a frequency of recombination between the disease gene and the closest marker of at most 5 %.

By simulating a two-allele marker we learned what values of the lod score (Z_{max}) might be expected if only regular RFLPs (PIC 0.37) were used for linkage mapping. Over 100 distinct simulations of the entire data set the two allele marker gave a mean overall lod score of 16.8. The highest lod score value was 24.8, while Z_{max} exceeded 10.5 in 95% of all replicates.

Simulations of an 8-allele marker gave us insight into the lod scores that could be obtained with dinucleotide repeats and other highly informative markers. A similar amount of informativeness can be expected from a map of 2, 4 and 5 allele markers in a multipoint analysis as performed on the actual data of the TSC collaborative group (Chapter 3.2). The highest lod score obtained with the 8-allele marker was 56.7, the mean Z_{max} was 41.8 and at least 95% of all replicates revealed a Z_{max} of 33.6 or higher.

Rather than comparing individual families via their mean lod scores, we decided to describe the families in terms of their "effective number of informative meioses" (EFNIM). The EFNIM of a family represents the number of informative meioses one may expect to score on average in that family for a marker system with given informativeness. The $Z_{(\Theta=0.05)}$ values and EFNIM values for all families are presented in Table 2.2a and 2.2b.

Table 2.2a:
The power of the chromosome 9-linked (TSC1) families[1)]

		Simulated: 8 allele marker at Θ=5%			2 allele marker at 5%	Unlinked 8 allele marker (Θ=50%)	
Family Nr.	Original Fam. Nr.	Max. $Z_{(max)}$	Mean $Z_{(\Theta=5\%)}$	EFNIM $Z_{(\Theta=5\%)}$	Mean $Z_{(\Theta=5\%)}$	Mean	Power to exclude >5cM
1	B 1	1.35	0.41	1.92	0.18	-0.46	0%
2	B 2	0.00	0.00	0.00	0.00	0.00	0%
5	B 5	0.70	0.29	1.36	0.12	-0.28	0%
8	B 8	1.38	0.42	1.96	0.15	-0.49	0%
9	B 9	0.70	0.19	0.89	0.10	-0.25	0%
12	B 12	0.73	0.20	0.95	0.07	-0.29	0%
14	B 14	0.00	0.00	0.00	0.00	0.00	0%
20	B 20	2.22	0.70	3.26	0.31	-0.73	0%
21	B 21	0.36	0.10	0.45	0.04	-0.08	0%
22	B 22	0.23	0.04	0.18	0.02	-0.05	0%
1006	C 6	0.25	0.09	0.42	0.03	-0.05	0%
1009	C 9	0.25	0.09	0.40	0.02	-0.08	0%
1010	C 10	0.25	0.12	0.57	0 03	-0.07	0%
1012	C 12	0.77	0.14	0.67	0.04	-0.08	0%
1013	C 13	2.89	1.55	7.20	0.59	-1.53	46%
1015	C 15	0.48	0.10	0.49	0.03	-0.17	0%
2005	L 5	0.25	0.10	0.48	0.04	-0.09	0%
2008	L 8	2.24	0.95	4.40	0.32	-1.07	15%
2009	L 9	0.57	0.22	1.01	0.11	-0.26	0%
2010	L 10	0.25	0.08	0.38	0.02	-0.09	0%
2011	L 11	0.57	0.22	1.04	0.11	-0.28	0%
2014	L 14	0.44	0.20	0.92	0.04	-0.21	0%
2015	L 15	0.06	0.00	0.00	0.00	-0.00	0%
2016	L 16	0.95	0.21	0.97	0.09	-0.35	0%
2018	L 18	1.43	0.74	3.42	0.28	-0.83	4%
2023	L 23	0.00	0.00	0.00	0.00	0.00	0%
2026	L 26	0.70	0.24	1.10	0.08	-0.29	0%
3001	I 1	1.58	0.82	3.80	0.31	-0.89	15%
3011	I 11	1.16	0.62	2.88	0.21	-0.63	6%
3015	I 15	3.23	1.34	6.25	0.47	-1.27	25%
3019	I 19	0.24	0.08	0.37	0.03	-0.05	0%
3021	I 21	0.24	0.03	0.15	0.01	-0.07	0%
3026	I 26	1.60	0.54	2.52	0.10	-0.88	15%
3028	I 28	0.25	0.11	0.53	0.05	-0.08	0%
3029	I 29	0.00	0.00	0.00	0.00	0.00	0%
3033	I 33	0.25	0.09	0.40	0.01	-0.10	0%
3101	I 101	0.86	0.34	1.60	0.12	-0.36	0%

Table 2.2a (continued)

		Simulated: 8 allele marker at Θ=5%			2 allele marker at 5%	Unlinked 8 allele marker (Θ=50%)	
Family Nr.	Original Fam. Nr.	Max. $Z_{(max)}$	Mean $Z_{(\Theta=5\%)}$	EFNIM $Z_{(\Theta=5\%)}$	Mean $Z_{(\Theta=5\%)}$	Mean	Power to exclude >5cM
4046	R 2046	0.00	0.00	0.00	0.00	0.00	0%
4067	R 2067	0.25	0.06	0.29	0.02	-0.12	0%
4068	R 2068	0.00	0.00	0.00	0.00	0.00	0%
4077	R 2077	1.27	0.50	2.31	0.17	-0.55	0%
4219	R 1219	2.12	0.78	3.61	0.24	-0.73	5%
4221	R 1221	0.83	0.15	0.71	0.04	-0.05	0%
4222	R 1222	1.00	0.08	0.36	0.02	-0.01	0%
4264	R 1264	0.99	0.27	1.27	0.13	-0.25	0%
4467	R 1467	0.04	0.00	0.00	0.00	-0.01	0%
5080	E 2080	0.02	0.00	0.00	0.00	-0.00	0%
5159	E 3159	0.46	0.07	0.35	0.03	-0.03	0%
5240	E 2240	2.03	1.07	4.97	0.37	-0.95	20%
5452	E 1452	0.00	0.00	0.00	0.00	0.00	0%
5733	E 3733	0.00	0.00	0.00	0.00	0.00	0%
7430	D 430	0.88	0.31	1.44	0.11	-0.32	0%
7474	D 474	1.45	0.82	3.81	0.40	-0.77	14%
7479	D 479	0.99	0.31	1.45	0.08	-0.37	0%
7781	D 781	0.25	0.04	0.20	0.02	-0.07	0%
7873	D 873	1.57	0.81	3.78	0.37	-0.85	9%
7981	D 981	1.99	0.64	2.99	0.24	-0.83	11%
9004	H 4	3.83	1.91	8.89	0.84	-2.02	59%
9005	H 5	0.57	0.23	1.07	0.05	-0.24	0%
9006	H 6	0.06	0.00	0.02	0.00	-0.00	0%
9007	H 7	0.83	0.25	1.17	0.09	-0.32	0%
9008	H 8	0.25	0.07	0.32	0.02	-0.10	0%
9010	H 10	0.07	0.00	0.01	0.00	-0.00	0%

1) Results of simulation studies presented as lod scores, EFNIM values and percentages of all replicates. B=Boston, C=Cardiff, L=London, I=Irvine, R=Rotterdam, E=Erlangen, D=Durham, H=Houston, $Z_{(max)}$=lod score maximized over Θ, $Z_{(\Theta=5\%)}$=lod score at a recombination fraction of 0.05, EFNIM=Effective number of informative meioses (see text).

Table 2.2b:
The power of the non-9 linked (TSC1) families[2)]

		Simulated: 8 allele marker at Θ=5%			2 allele marker at 5%	Unlinked 8 allele marker (Θ=50%)	
Family	Original	Max.	Mean	EFNIM	Mean	Mean	Power to
Nr.	Fam. Nr.	$Z_{(max)}$	$Z_{(\Theta=5\%)}$	$Z_{(\Theta=5\%)}$	$Z_{(\Theta=5\%)}$		exclude >5cM
3	B 3	0.36	0.07	0.34	0.02	-0.09	0%
6	B 6	0.05	0.00	0.00	0.00	-0.00	0%
7	B 7	0.00	0.00	0.00	0.00	0.00	0%
10	B 10	0.87	0.24	1.10	0.11	-0.25	0%
15	B 15	0.25	0.14	0.63	0.03	-0.11	0%
16	B 16	0.04	0.00	0.00	0.00	-0.00	0%
18	B 18	0.48	0.11	0.51	0.04	-0.13	0%
19	B 19	0.04	0.00	0.01	0.00	0.00	0%
1003	C 3	0.73	0.35	1.65	0.11	-0.36	0%
1004	C 4	1.58	0.87	4.07	0.27	-0.84	4%
1005	C 5	0.36	0.05	0.25	0.01	-0.10	0%
1007	C 7	0.56	0.09	0.41	0.03	-0.13	0%
1011	C 11	0.48	0.15	0.70	0.08	-0.08	0%
2001	L 1	6.74	3.17	14.76	1.30	-3.75	90%
2002	L 2	0.04	0.00	0.00	0.00	0.00	0%
2003	L 3	0.57	0.26	1.22	0.09	-0.28	0%
2004	L 4	0.36	0.04	0.20	0.00	-0.10	0%
2006	L 6	0.36	0.10	0.46	0.04	-0.15	0%
2007	L 7	2.63	1.34	6.23	0.66	-1.93	55%
2012	L 12	0.68	0.23	1.06	0.09	-0.13	0%
2013	L 13	0.48	0.15	0.70	0.06	-0.12	0%
2017	L 17	0.25	0.06	0.29	0.01	-0.11	0%
2019	L 19	0.57	0.28	1.30	0.12	-0.26	0%
2020	L 20	2.03	1.03	4.82	0.42	-1.08	18%
2021	L 21	0.25	0.05	0.23	0.01	-0.12	0%
2022	L 22	0.46	0.06	0.26	0.02	-0.09	0%
2024	L 24	0.70	0.32	1.49	0.14	-0.27	0%
2025	L 25	1.08	0.34	1.59	0.16	-0.29	0%
3003	I 3	1.56	0.74	3.46	0.31	-0.89	11%
3004	I 4	0.70	0.26	1.19	0.08	-0.17	0%
3008	I 8	0.67	0.15	0.69	0.05	-0.09	0%
3016	I 16	1.28	0.63	2.95	0.24	-0.67	7%
3020	I 20	0.42	0.08	0.36	0.03	-0.05	0%
3024	I 24	0.46	0.04	0.18	-0.01	-0.15	0%
3034	I 34	0.06	0.00	0.01	0.00	-0.00	0%
4079	R 207	0.87	0.44	2.04	0.15	-0.39	0%

Table 2.2b (continued):

		Simulated: 8 allele marker at Θ=5%			2 allele marker at 5%	Unlinked 8 allele marker (Θ=50%)	
Family Nr.	Original Fam. Nr.	Max. $Z_{(max)}$	Mean $Z_{(\Theta=5\%)}$	EFNIM $Z_{(\Theta=5\%)}$	Mean $Z_{(\Theta=5\%)}$	Mean	Power to exclude >5cM
4197	R 1197	0.04	0.00	0.00	0.00	-0.00	0%
4223	R 1223	0.25	0.11	0.49	0.04	-0.06	0%
5024	E 2024	0.70	0.24	1.10	0.07	-0.42	0%
5246	E 2246	0.25	0.06	0.29	0.02	-0.15	0%
5383	E 3383	0.49	0.16	0.73	0.07	-0.03	0%
5449	E 1449	0.00	0.00	0.00	0.00	0.00	0%
5459	E 1459	0.00	0.00	0.00	0.00	0.00	0%
5462	E 3462	0.00	0.00	0.00	0.00	0.00	0%
5465	E 1465	2.19	0.84	3.92	0.30	-1.01	22%
5468	E 3468	0.00	0.00	0.00	0.00	0.00	0%
5865	E 865	0.64	0.07	0.32	0.05	-0.06	0%
5866	E 4865	0.57	0.23	1.05	0.03	-0.31	0%
7085	D 1085	0.02	0.00	0.00	0.00	-0.00	0%
7111	D 1111	2.29	1.12	5.19	0.44	-1.15	23%
7188	D 1188	0.82	0.33	1.51	0.10	-0.35	0%
7427	D 427	1.11	0.40	1.86	0.22	-0.37	0%
7431	D 431	0.83	0.12	0.57	0.04	-0.06	0%
7437	D 347	3.72	1.97	9.16	0.68	-2.37	64%
7473	D 473	0.48	0.08	0.35	0.01	-0.16	0%
7482	D 482	0.09	0.00	0.01	0.00	0.00	0%
7507	D 507	1.52	0.60	2.80	0.27	-0.77	11%
7821	D 821	0.00	0.00	0.00	0.00	0.00	0%
9001	H 1	0.68	0.11	0.50	0.05	-0.11	0%
9003	H 3	0.98	0.46	2.14	0.19	-0.42	0%
9011	H 11	2.75	1.34	6.24	0.53	-1.44	36%
9012	H 12	0.07	0.00	0.02	0.00	-0.00	0%
9013	H 13	1.70	0.80	3.73	0.32	-1.00	22%
9014	H 14	2.00	1.06	4.95	0.38	-1.00	21%
9015	H 15	0.57	0.26	1.21	0.12	-0.18	0%

2) As Table 2.2a.

EFNIM values for the chromosome 9-linked group of TSC1 families were compared with those obtained for non-9 linked families. Assignment of families to either of these two groups was made according to the posterior probabilities for linkage to chromosome 9 as determined in the actual linkage study (Chapter 3.2) using real linkage data for chromosome 9 markers. The distribution of the families over 11 EFNIM categories of increasing informativeness is shown in Figure 2.5. There are no marked differences in size (as summarised by EFNIM) between the two groups of TSC families. Another obvious finding is the paucity of large families in both groups. Most families are in the 0 - 1 EFNIM category. This is illustrated by the mean EFNIM values for both family groups: 1.41 for the 9-linked families and 1.46 for the non 9-linked group.

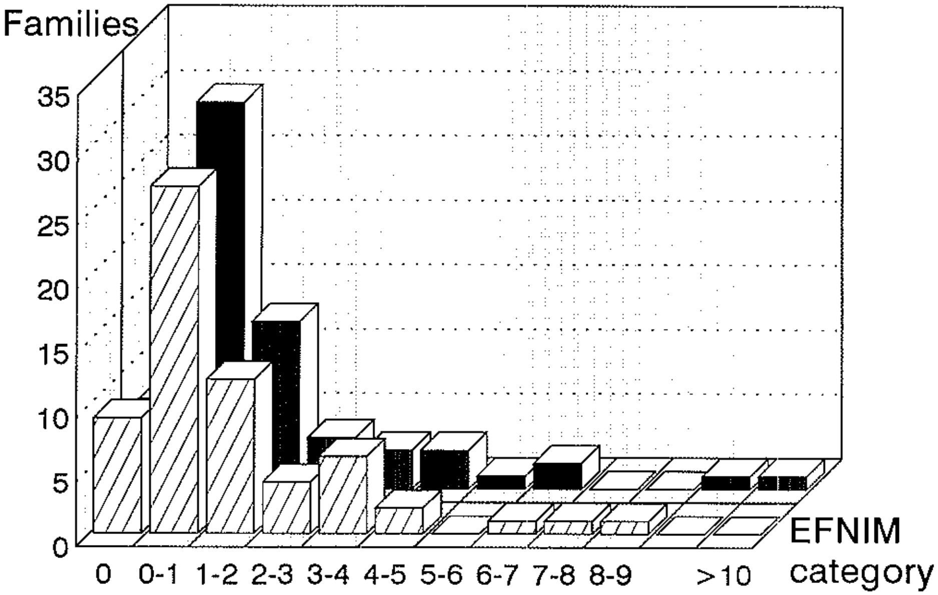


Figure 2.5: Family distribution by Effective number of informative meioses (EFNIM). The distribution of all 63 TSC1 families is shown in front (hatched), the 65 non-9 linked families are shown behind (solid).

Exclusion power of TSC families

In general, exclusion studies are only valid when the mode of inheritance is specified correctly in the analysis. In a real exclusion study for TSC one would therefore have to take into account the locus heterogeneity. In our simulation study we evaluated the power of the families for an uncomplicated exclusion study (i.e. under linkage homogeneity), with the sole purpose of comparing the families. As expected, families that contained much information for linkage detection also contributed most information for exclusion (Table 2.2).

Together, the TSC1 families had the power to exclude linkage over large genetic distances. When an 8-allele marker was examined, 95% of the replicates excluded 31 cM or more on either side of the marker. With a 2-allele marker the exclusion distance decreased to 12 cM.
The non 9-linked families were able to exclude similar areas: at least 29 cM for an 8-allele marker and at least 11 cM for a 2-allele marker in 95% of the replicates.

Power of TSC families to detect heterogeneity
We tested the performance of the families in heterogeneity analyses. We analyzed HOMOG input files (containing lod scores from linked and unlinked families) and HOMOG2 input files (containing lod scores from two regions, with each family linked to only one of these). Thus we tried to answer two questions:
1. What is the power of this data set for detection of heterogeneity?
2. What is gained by including information from the alternative region in a two-locus heterogeneity analysis?

The interpretation of results of linkage analysis in terms of statistical significance is not always simple. Under homogeneity a lod score of 3.0 or more is accepted as 'significant' evidence for linkage. A lod score of 3.0 corresponds to an odds ratio of 1000 : 1. This high threshold has a statistical basis (usually, $p<0.05$ is already regarded as significant evidence): when two loci are selected at random the chances that they will show linkage are small. It has been estimated that the prior probability of two loci showing true linkage is only 1 in 50. If the theoretical odds ratio of 1000 : 1 is corrected for the low (1 in 50) prior expectation of finding linkage, the resulting frequency of false positive linkages will be 0.05.

When one is examining the possibility of locus heterogeneity, however, no prior expectations can be formulated about whether a disease will show locus heterogeneity or not. Therefore lod scores of 0.834, approximately corresponding to a chi-square of 3.841 ($p=0.05$ at 1 df) have been conventionally accepted as 'significant'. The situation changes again, when examining the possibility that the presumed second locus is also located within the region for which markers were tested. The prior expectation for the second locus being in the tested region is low. Therefore, a threshold lod score difference of 3.0 seems a prudent choice. Accordingly, we formulated the following tests of various hypotheses:

H0. No locus exists in the area (null-hypothesis. The lod score is 0.0 by definition).

H1. One locus maps in the area tested and there is no heterogeneity. A lod score of at least 3.0 is required.

$H2_{(1\ mapped)}$. One locus maps in the area, heterogeneity exists, but it is assumed that the second locus is not located in the tested region. Both the null hypothesis and the H1 hypothesis should be rejected in favour of this alternative when the lod score exceeds that for H1 by at least 0.834 and the lod score exceeds that for H0 by 3.834 (=3.0 + 0.834).

$H2_{(2\ mapped)}$. Two loci exist in the tested area: There are three requirements for significance: There must be a lod score difference with $H2_{(1\ mapped)}$ of 3.0, with H1 of 3.834 and with H0 of 6.834.

We tested these hypotheses in heterogeneity analyses using simulated data sets containing a certain proportion of linked families (α's of 10%, 30%, 50%, 70% and 90%). These analyses made use of data for a four-allele marker system, since this closely resembles the combined informativeness of true marker maps in the collaborative data (when expected and observed lod scores were compared). Each series consisted of 100 replicates and therefore involved 100 runs of the HOMOG program. By combining the simulated data for a linked and an unlinked marker into an imaginary chromosome, we also created a typical HOMOG2 problem, with 2 loci to be mapped within the tested area.

For α=0.5, a mean lod score favouring $H2_{(2\ mapped)}$ over H0 of 24.27 was obtained. These results indicate that the family material is highly suitable for any type of linkage or heterogeneity analysis (Figure 2.6). When at least 50% of the families were assumed to be linked, it was even possible to detect linkage under the (false) assumption of locus homogeneity in 94/100 attempts. For most α values the imaginary chromosome approach was more powerful than conventional HOMOG analysis.

We also studied the precision of the obtained estimates for the recombination fraction and the proportion of linked families. It was counted how often the correct values for α (± 10 %) and Θ (between 2 and 8 %) were found in replicates that yielded significant evidence for heterogeneity and/or linkage. The α and Θ values obtained from the 1-locus and 2-locus heterogeneity tests were quite accurate (Figures 2.7 and 2.8). However, for very low true alphas the precision of the tested methods was found to be insufficient.

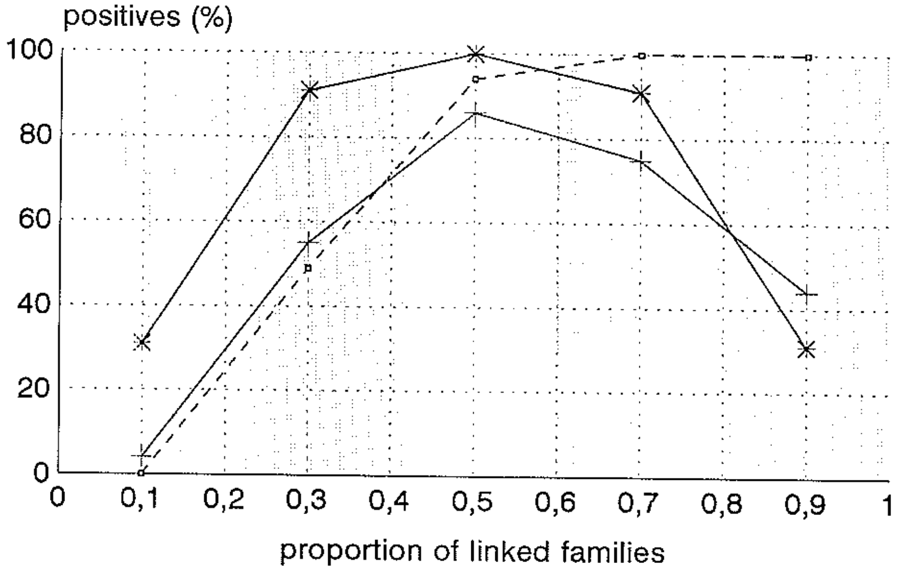


*Figure 2.6: Power to detect linkage and heterogeneity. Broken line (H1) = power to detect linkage under the assumption of homogeneity (Z>3.0), +-+ ($H2_{(1\ mapped)}$) = power to detect linkage and heterogeneity if only one locus is known (see text), *-* ($H2_{(2\ mapped)}$) = power to detect linkage and heterogeneity if two loci are known, applying the imaginary chromosome approach (see text).*

Discussion

Suitability of family material

Our family material was found to be very suitable for heterogeneity analysis. The family structures provide sufficient information both for detection and exclusion of linkage. Furthermore family structure is very similar in the chromosome 9 linked and unlinked groups. The suitability of our family set has been confirmed by heterogeneity analyses of the collaborative data set (Chapter 3.2). For major genes, detection of linkage and heterogeneity will not be problematic. The precision of the resulting values for α and Θ are acceptable. However, mapping genes responsible for TSC in a small minority of families will be difficult, particularly when the α value approaches 10% or less. Under these circumstances the additional power offered by the imaginary chromosome approach is very useful, although still only 31/100 analyses reached significance. This may be important with respect to the non-significant chromosome 12 findings of the collaborative linkage study (Chapter

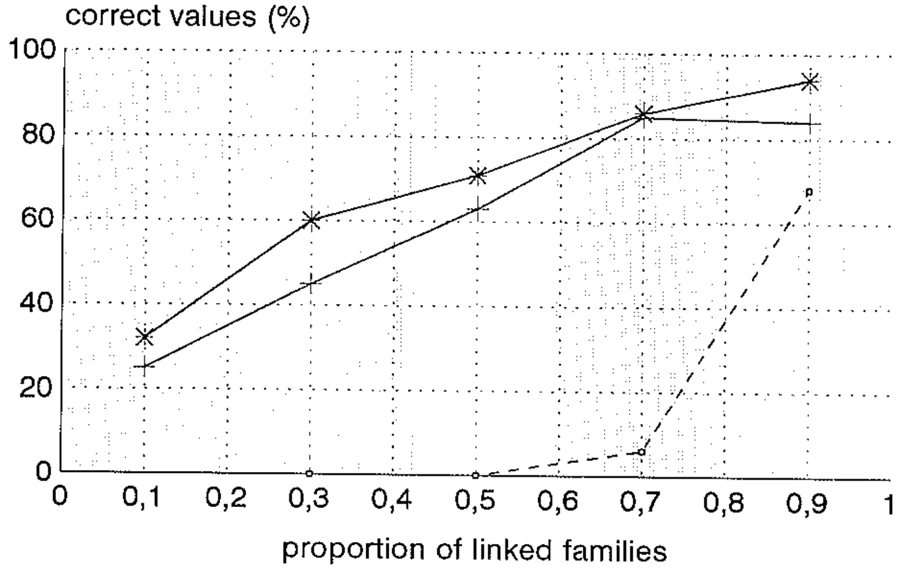


Figure 2.7: Precision of Θ estimates (Θ obtained = true Θ ±3%) under heterogeneity. Only results supported by significant lod scores were evaluated. H(1): broken line, $H2_{(1\ mapped)}$:+-+, $H2_{(2\ mapped)}$:-*.*

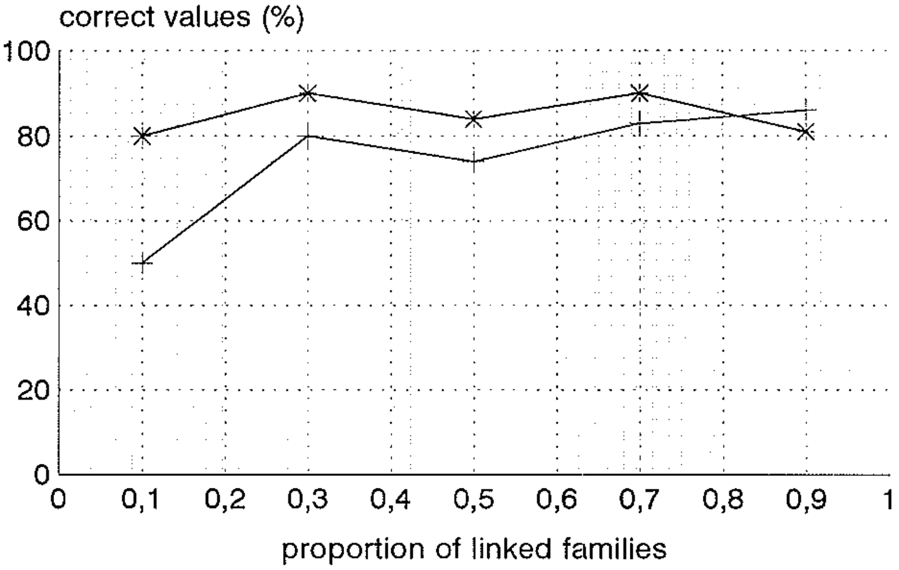


Figure 2.8: Precision of α estimates (α obtained= true α ±10%) under heterogeneity. Only results supported by significant lod scores were evaluated. H(1): broken line, $H2_{(1\ mapped)}$:+-+, $H2_{(2\ mapped)}$:-*.*

3.2). Our findings indicate that existence of a minor locus may only rarely yield significant evidence against homogeneity. As long as reasonable Θ values emerge, it may be wise to continue the study of a putative minor locus. Linkage methods, however, can contribute little to such a study. Other avenues, such as the t(3;12) translocation in the case of TSC3 (Fahsold et al., 1991), will have to be explored.

Power of the methodological approach

Overall the imaginary chromosome approach has been shown to be a powerful method for mapping loci when locus heterogeneity occurs. If multiple candidate regions have been identified it seems sensible to analyze these simultaneously. Only when the α exceeds 0.8 does the conventional HOMOG approach ($H2_{(1\ \mathrm{mapped})}$) perform better than the imaginary chromosome approach. Since a high α implies that a small number of families are assigned to the alternative locus, we may presume that the power of these families is insufficient to meet the required lod score difference of 3.0 between $H2_{(2\ \mathrm{mapped})}$ and $H2_{(1\ \mathrm{mapped})}$, as defined above.

Although we found the HOMOG analyses on our data to be quite satisfactory, we disagree with the thresholds normally accepted for significance levels. Older versions of the HOMOG programs used to indicate results as significant ($p<0.05$) when the difference in lod score for hypothesis $H2_{(2\ \mathrm{mapped})}$ versus the H1 is 1.000 or more; in more recent publications (Ott et al., 1990) odds of 50:1 or 100:1 in favour of $H2_{(2\ \mathrm{mapped})}$ have been regarded as significant. We feel that a lod difference of 3.834 is more appropriate (evidence for a second localisation should be at least as convincing as evidence required for a first localisation). Our revision of the lod score difference required for localisation of new TSC genes is relevant to the previous apparent support for the putative chromosome 11 TSC locus (Janssen et al., 1990). The results obtained using conventional criteria for significance are apparently misleading. In contrast the new results of the collaborative study do not provide any support for a TSC2 locus on chromosome 11 (Chapter 3.2).

Description of families using the "effective number of informative meioses"

We have used the "effective number of informative meioses" for describing family size. The EFNIM value gives a much better description of the family size than does the number of generations, the number of affecteds, or the number of relatives. In contrast to the simulated lod score itself, it provides us with an instant image of the family size. We propose the use of EFNIM values to describe families in cases where it is not feasible to show the pedigrees themselves.

The EFNIM distribution of the TSC1 (chromosome 9-linked) and non-TSC1 families showed that there is no obvious difference in family size. This implies that there is no marked difference in the biological fitness associated with the different genetic types of tuberous sclerosis.

Acknowledgements

We would like to thank the TSC collaborative group, particularly Moyra Smith, Jonathan Haines, Jean Amos, David Kwiatkowski, Priscilla Short, Hope Northrup, Susan Blanton, Sue Povey, Mari Wyn Burley, J Michael Connor, Mark Nellist, Phillip Brook-Carter, Paul Fleury, Arjenne Hesseling-Janssen, Senno Verhoef, Ray Kandt, Raimund Fahsold, Hans-Dieter Rott, Margaret Pericak-Vance, Corinne Merkens, Anneke Maat-Kievit and Dick Lindhout, for their part in the enormous effort to eslish this very useful set of families.

This study was funded by the Netherlands Praevention Fund (Grant 28-1723) and the National Tuberous Sclerosis Association of the USA (Grant 91-03).

Appendix I: Informativeness of families for linkage.

In linkage studies, the analysis of simple phase-known families is straightforward: one can count recombinants and non-recombinants; no complicated statistical analysis is required. When some family members are not available for analysis, or when phase is not known, the calculations can become extremely involved: the evidence for linkage is then summarized via the best estimate of the recombination frequency and the corresponding maximum lod score. To ease the interpretation of lod scores, Edwards (1976) calculated the so called 'equivalent observations' (n), that is, the number of recombinants and nonrecombinants which would give the same lod score.

$$n = \frac{Z_{\max}}{\log 2 \ + \ \hat{\theta}\log\hat{\theta} \ + \ (1-\hat{\theta})\log(1-\hat{\theta})} \qquad if \ \ \hat{\theta} > 0 \tag{1}$$

$$n = \frac{Z_{\max}}{\log 2} \qquad if \ \ \hat{\theta} = 0 \tag{2}$$

According to this equation, a data set yielding a lod score of 0.419 at Θ=0.2, for instance, is equivalent to five meioses, four of which are non-recombinants. It should be noted that this measure is calculated for an entire data set; it is only relevant when in the calculations the best estimate of the recombination frequency obtained in that same data set was used. To illustrate this point: a data set containing five completely informative meioses (four non-recombinants) will yield exactly five equivalent observations when a recombination frequency of 0.2 is assumed, but it will yield 25 equivalent observations when analyzed under Θ=0.4, or only 2 when Θ=0.1. This also implies that equivalent observations are only additive over data sets when obtained for the same value of Θ.

While the number of equivalent observations is a very convenient measure to summarize a given data set, it cannot be used directly to compare families with respect to their informativeness for disease mapping for several reasons. Firstly, the calculation of the number of equivalent observations is based on the best estimate of Θ in the entire data set. When the ratio non-recombinants : recombinants in an individual family deviates from the overall Θ, the number of equivalent observations from that family will not accurately reflect its information content, as became apparent in the previous example. Secondly, the calculated number of equivalent observations depends on the actual marker segregation in the family: when several parents are by chance homozygous for the marker, the information content of that family may seem small. For a fair comparison between families, the effects of chance marker segregation have to be eliminated.

In our TSC families, we simulated marker segregation for markers with various degrees of informativeness, linked to a putative TSC locus with 5 % recombination. For the calculation of EFNIM values, lod scores were calculated in these replicates for that same value of Θ. Frequently, the maximum lod score in a particular replicate occurred at some other value of Θ; the mean lod score over all replicates, however, peaked at a value of 5 %. That mean lod score was used to compare families via the calculated EFNIM (effective number of informative meioses). The EFNIM was calculated via

$$EFNIM = \frac{\bar{Z}_\theta}{\log 2 + \theta \log\theta + (1-\theta)\log(1-\theta)} \quad \textit{with}\ \theta = 0.05 \qquad (3)$$

This equation is closely related to Edward's formula, but here the assumed recombination frequency is used in the calculations rather than the best estimate of the recombination frequency. Also, the calculations are not carried out for each simulated replicate separately, but rather using the mean lod score over all replicates. Because EFNIM values are based on a large number of simulations (in this case 100), chance fluctuations of marker informativeness reduce to the average informativeness of the marker as characterised by the PIC value. The EFNIM values calculated here were obtained for a marker with eight alleles (PIC 0.86). EFNIM values for a less informative marker (with a PIC value of, say, 0.375) can be approximated via

$$EFNIM_{(PIC_1)} \approx EFNIM_{(PIC_2)} \times \frac{PIC_1}{PIC_2} \tag{4}$$

How accurate this approximation is depends on the actual family structure: when many persons have not been tested, reconstruction of their marker genotypes is only possible for highly polymorphic markers. For such families, less polymorphic markers will yield smaller EFNIM values than calculated via equation 4.

As two examples, consider families 1 and 1013 in Table 2.2a.

Family 1 yielded an EFNIM of 1.92 for a marker with PIC=0.86. Using (4), a marker with PIC=0.375 (2 allele with equal frequencies) should yield an EFNIM of 0.84. When calculated from the mean lod score for the marker with two alleles the actual EFNIM is indeed 0.84.

For family 1013, however, the predicted EFNIM for the two allele marker is 3.14, while the observed EFNIM is 2.74. In this large family, analysis with a marker with eight alleles allows reconstruction of missing genotypes which is frequently not possible using a two-allele system.

References

Connor JM, Yates JRW, Mann L, Aitken DA, Stephenson JBP. Tuberous sclerosis: Analysis of linkage to red cell and plasma protein markers. Cyt Cell Genet 1987;44:63-64.

Edwards JH. The interpretation of lods in linkage analysis. Cyt Cell Genet 1976; 16: 289-293.

Fahsold R, Rott H-D, Lorenz P. A third gene locus for tuberous sclerosis is closely linked to the phenylalanine hydroxylase gene locus. Hum Genet 1991;88:85-90.

Fryer AE, Chalmers A, Connor JM, Fraser I, Povey S, Yates AD, Yates JRW, Osborne JP. Evidence that the gene for tuberous sclerosis is on chromosome 9. Lancet 1987;i:659-661.

Haines J, Amos J, Attwood J, Bech-Hansen NT, et al. Genetic heterogeneity in tuberous sclerosis: study of a large collaborative dataset. Ann N Y Acad Sci 1991a;615:256-264.

Haines JL, Short MP, Kwiatkowski DJ, Jewell A, Andermann E, Bejjani B, Yang C-H, Gusella JF, Amos JE. Localization of one gene for tuberous sclerosis within 9q32-9q34, and further evidence for heterogeneity. Am J Hum Genet 1991b;49:764-772.

Janssen LAJ, Sandkuijl LA, Merkens EC, Maat-Kievit JA, Sampson JR, Fleury P, Hennekam RCM, Grosveld GC, Lindhout D, Halley DJJ. Genetic heterogeneity in tuberous sclerosis. Genomics 1990;8:237-242.

Kandt RS, Pericak-Vance MA, Hung W-Y, Gardner RJM, Crossen PE, Nellist MD, Speer MC, Roses AD. Linkage studies in tuberous sclerosis: chromosome 9?, 11? or maybe 14!. Ann N Y Acad Sci 1991;615:284-297.

Kelsoe JR, Ginns EI, Egeland JA, Gerhard DS, et al. Re-evaluation of the linkage relationship between chromosome 11p loci and the gene for bipolar affective disorder in the Old Order Amish. Nature 1989;342:238-243.

Martinez M, Goldin LR. Power of the linkage test for a heterogeneous disorder due to two independent inherited causes: a simulation study. Genet Epidemiol 1990;7:219-230

McKusick VA. Current trends in mapping human genes. FASEB J 1991;5:12-20

Narod SA. Power of the admixture test to detect genetic heterogeneity. Genet Epidemiol 1991;8:209-216

Northrup H, Kwiatkowski DJ, Roach ES, Dobyns WB, Lewis, RA, Herman GE, Rodriguez E, Daiger SP, Blanton SH. Am J Hum Genet 1992; in press.

Ott J. Computer-simulation methods in human linkage analysis. Proc. Natl Acad Sci USA 1989;86:4175-4178.

Ott J. Bhattacharya S, Chen JD, Denton MJ, et al. Localizing multiple X chromosome-linked retinitis pigmentosa loci using multilocus homogeneity tests. Proc Natl Acad Sci USA 1990; 87:701-704.

Ott J. Analysis of human genetic linkage (revised edition). The Johns Hopkins University Press, Baltimore MD, 1991.

Povey S, Attwood J, Janssen LAJ, Burley M et al. An attempt to map two genes for tuberous sclerosis using novel two-point methods. Ann N Y Acad Sci 1991;615:298-305.

Povey S, Burley MW, Attwood J, et al. Two loci for tuberous sclerosis: one on 9q34 and one on 16p13. Ann Hum Genet 1994;58:107-127.

Sampson JR, Yates JRW, Pirrit LA, Fleury P, Winship I, Beighton P, Connor JM. Evidence for genetic heterogeneity in tuberous sclerosis. J Med Genet 1989;26:511-516.

Smith CAB. Testing for heterogeneity of recombination fraction values in Human Genetics. Ann Hum Genet 1963;27:175-182

Smith M, Smalley S, Cantor R, Pandolfo M, Gomez MI, Baumann R, Flodman P, Yoshiyama K, Nakamura Y, Julier C, Dumars K, Haines J, Trofatter J, Spence MA, Weeks D, Conneally M. Mapping of a gene determining tuberous sclerosis to human chromosome 11q14-q23. Genomics 1990:6:105-114.

St. Clair D, Blackwood D, Muir W, Baillie D, Hubbard A, Wright A, Evans HJ. No linkage of chromosome 5q11-q13 markers to schizophrenia in Scottish families. Nature 1989; 339:305-309.

LINKAGE ANALYSIS UNDER LOCUS HETEROGENEITY: BIASED PARAMETER ESTIMATES USING THE ADMIXTURE APPROACH

Bart Janssen
Dicky Halley
Lodewijk Sandkuijl

Submitted for publication

Summary

The admixture test is a popular method for the analysis of linkage data when locus heterogeneity is suspected. It can be applied on pairwise linkage data, multipoint data and even for the simultaneous analysis of data from multiple dispersed candidate regions. However, very little is known about the conditions for the use of the method under these divergent circumstances. By performing analytical evaluations, we demonstrate that a systematic bias in estimates of the recombination fraction (Θ) and the proportion of linked families (α) may occur if the actual frequency of linked families is not identical among small and large families. We reason that the admixture test should be used with caution if recessive diseases or diseases with presumed phenocopies are under investigation. The phenotype and the probability of developing the phenotype at a certain age should be equal for family members of linked and unlinked families. If dissimilarities in family size can not be ruled out the extent of bias should be

considered and size specific alpha values should be used in risk calculations.

Introduction

Linkage analysis is a common first step towards the positional cloning of disease genes. For simple Mendelian disorders its statistical methodology is straightforward, but locus heterogeneity may present challenging complications. A widely applied statistical method for analysis of heterogeneity, the admixture test (A-test) introduced by Smith (Smith, 1961; Ott, 1991), assumes that a linkage sample consists of a mixture of disease families, some of which show linkage to one area of the genome while others are unlinked to that region, presumably showing linkage to another location. The A-test can be applied to map one single locus in the presence of unlinked families, two loci on one chromosome, or even for the combined analysis of two or more dispersed chromosomal regions. The latter application has been introduced as 'Imaginary Chromosome Approach' (ICA) Janssen et al., 1990). The support for linkage and heterogeneity in a given family is usually calculated via the following equation, which was implemented in the HOMOG programs by Ott (Ott, 1991):

$$Z_{i(\alpha,\theta)} = \log(\alpha . 10^{Z_{i(\theta)}} + (1-\alpha)) \qquad (1)$$

where Θ is the frequency of recombination between disease locus 1 and the marker studied, α represents the proportion of disease families segregating for disease locus 1, and Z_i indicates the lod score in family i. This statistical method is based on the assumption that families never segregate for mutations at multiple loci simultaneously. While this condition is not too restrictive for infrequent genetic diseases, it may be violated for common diseases.

For the simultaneous analysis of two regions (ICA) equation (1) can easily be extended to:

$$Z_{i(\alpha,X_1,X_2)} = \log(\alpha \cdot 10^{Z_{i(X_1)}} + (1-\alpha) \cdot 10^{Z_{i(X_2)}}) \qquad (2)$$

Where X_1 and X_2 denote the locations of gene locus 1 and 2.

Equation (1) is derived from the more complete formulation:

$$Z_{i(\alpha,\theta)}=\log(\frac{\alpha.P_{(D_i,M_i|q_1,f_1,\theta)}+(1-\alpha).P_{(D_i,M_i|q_2,f_2,0.5)}}{\alpha.P_{(D_i,M_i|q_1,f_1,0.5)}+(1-\alpha).P_{(D_i,M_i|q_2,f_2,0.5)}}) \tag{3}$$

where $P_{(D,M)}$ is the probability of the observed disease and marker phenotypes in a given family, q_1 and q_2 are the disease gene frequencies at the linked and unlinked loci, and f_1 and f_2 are the vectors of penetrances for the various genotypes at the two loci. In the latter formulation, it becomes clear that equation 3 reduces to equation 1 only when the gene frequencies and penetrances are identical for both disease genes. Usually, the gene frequencies are kept identical for all loci throughout the analysis and differences in gene frequencies are parametrized via α, the proportion of linked families. Possible difficulties in the interpretation of α stem from its ambiguous definition: the study unit in heterogeneity analysis is a 'family', which is insufficiently defined in terms of size and structure unless a specific sampling strategy has been adopted.

In general, simplification of equation 3 into equation 1 is only adequate when the frequency of recombination between disease gene and marker is the only feature that distinguishes between the two family types. As soon as there is any additional dissimilarity that may further discriminate between the two loci, equation 1 will be incomplete. Possible causes of dissimilarity are differences in age at onset, number of offspring of gene carriers, family structure and fitness in relation to severity.

We have evaluated the role of these factors as sources of bias in the estimation of α and Θ. Apart from their obvious relevance for positional cloning, accurate estimates of α and Θ are essential for risk calculations in genetic counseling. Our evaluations involved both theoretical, analytical comparisons and investigations using published data, with the standard or more advanced applications of the A-test, with multipoint instead of pairwise data or multiple chromosomes (ICA) instead of a single chromosome.

Methods

Analytical method

A computerised model was designed and implemented to evaluate possible bias in estimates of α and Θ, using several variants of the A-test, including ICA. Our model made the simplifying assumption that all disease families can be described in terms of a discrete number of completely informative phase-known meioses. We took the number of informative meioses per family as a measure of the family size. All calculations were carried out for the following map situation:

$$---M_1\overset{\theta_1}{-}-X_1\overset{\theta_2}{-}-M_2---//---M_3\overset{\theta_3}{-}-X_2\overset{\theta_4}{-}-M_4---$$

where X_1 and X_2 again represent locations of two genes that independently affect the same phenotype. The map of markers contains multipoint linkage information for both disease loci, but this information can be reduced to two-locus linkage information by fixing the recombination frequency for one of the markers flanking each disease locus to 0.5 (e.g. $\Theta_2=\Theta_4=0.5$). A further reduction leads to the classical admixture situation, where some families show linkage to a marker, while others are unlinked (e.g. $\Theta_2=\Theta_3=\Theta_4=0.5$).

For a family type with a given number of informative meioses, all possible combinations (c) of recombination or non-recombination over all intervals were enumerated. The probability of occurrence of each configuration (P_c) was calculated for given (true) values of Θ_1 through Θ_4, and given (true) alpha. Initially, it was assumed that the probability for a family to be segregating for X_1 or X_2 was independent of family size. Later calculations allowed for different values of alpha depending on the size of the family. Once the probabilities for all family types were calculated, best estimates of alpha and theta(s) were obtained analogous to the procedure followed in HOMOG. Briefly, an exhaustive search was carried out over a wide range of values for theta and alpha for the combination - Θ_{out} and α_{out} - that leads to the best overall lod score. For each possible configuration a lod score $Z_{c(\alpha,X1,X2)}$ was calculated. An expected lod score for each family type e was obtained by:

$$\bar{Z}_{e(\alpha,X_1,X_2)}=\sum P_c \cdot Z_{c(\alpha,X_1,X_2)} \tag{4}$$

where $Z_{c(\alpha,X1,X2)}$ was obtained as in equation (2).

Weights were assigned to various family types according to their frequency of occurrence; this approach corresponds to having a infinitely large sample of families. The maximum likelihood values of alpha and theta obtained in this manner were compared to those specified in the first step, the generation of family configurations.

Description of families (EFNIM)

The computerised model described above was appropriate for the analytical evaluation of hypothetical family data, but for the analytical study of real data the actual family structure had to be converted. The information content of each family had to be defined as a whole number of informative meioses.

In 1992 we introduced the 'effective number of informative meioses' (EFNIM) for the specification of family size in our large set of tuberous sclerosis (TSC) families (Chapter 2.2). The properties and advantages of the EFNIM have been described in that paper and its appendix. Apart from the family size, the EFNIM also depends on the actual theta between locus and marker and on the informativeness of the marker under investigation.

In brief we estimated the EFNIM of existing families at Θ=0.05 and PIC=0.86 by simulating 100 replicates, using the SLINK program (Ott, 1989). The mean lod score was calculated for each replicate and used to derive a mean EFNIM by division by the expected Z_Θ for one fully informative meiosis. For Θ=0.05 the expected Z_Θ=0.215. For the evaluation of the tuberous sclerosis data we previously calculated EFNIM values as described and rounded these to whole numbers for the purpose of this analysis.

Results

Relationship between true alpha, true theta and their estimates

First of all we tested whether the use of more advanced applications of the A-test - multipoint analysis and ICA - influences the accuracy of the estimators. Various α_{true} values were tested in simple families with two informative meioses (Table 2.3). The obtained α_{out} was always equal to the α_{true}. We also did not find deviant results for Θ_{out}, even when the locus was positioned extremely to the right or left side of an interval (multipoint analysis), or when one locus was positioned close to a marker on the first chromosome and another far removed from all markers on the second chromosome (ICA) (data not shown).

Table 2.3
Alpha and theta values obtained from pairwise and multipoint heterogeneity analysis

EFNIM linked/ unlinked	True			Obtained			Max. lod score obtained with 10 families	Lod score at α_{true} and Θ_{true} (10 families)
	α	Θ1	Θ2	α	Θ1	Θ2		
2 / 2	50%	10%	50%*	50%	10%	50%	0.791	0.791
	30%	10%	5%	30%	10%	5%	0.424	0.424
2 / 4	50%	10%	50%*	42%	14%	50%	0.563	0.510
	50%	10%	10%	37%	10%	10%	0.684	0.614
	50%	10%	5%	38%	10%	5%	0.767	0.697
	30%	10%	50%*	30%	24%	50%	0.173	0.102
	30%	10%	10%	17%	10%	10%	0.183	0.096
	30%	10%	5%	17%	10%	5%	0.201	0.108
4 / 2	50%	10%	50%*	63%	11%	50%	2.156	2.111
	50%	10%	10%	60%	10%	10%	2.964	2.909
	30%	10%	10%	41%	10%	10%	1.399	1.329
7 / 9	30%	10%	50%*	27%	9%	50%	1.377	1.372
	30%	10%	10%	29%	10%	10%	2.270	2.270
	30%	10%	5%	29%	10%	5%	2.686	2.685

Notes: 'EFNIM linked/unlinked'=2/4 means that all linked families were of the EFNIM=2 type (2 informative meioses), while the unlinked families were of the EFNIM=4 type. 'Theta1': the true or obtained distance (Θ_{true} or Θ_{out}) between the first marker (M1) and the locus (M1 - (Θ_1) - Locus - (Θ_2) - M2). 'Theta2'=0.5 denotes the pairwise analysis (M1 - (Θ_1) - Locus). 'Lod score' refers to $Z_{(\alpha,X1)}$. Analyses marked with (*) are pairwise analyses.

Family size

The influence of family size was tested by assigning different numbers of informative meioses (NIM) to the group of linked and unlinked families. By investigating families (NIM=2) linked to a marker at Θ=0.1 together with an equal number of unlinked families of the NIM=4 type we actually studied a hypothetical situation where the NIM specific $\alpha_{(NIM=2)}$=1.0 and $\alpha_{(NIM=4)}$=0.0, while the overall α=0.5 according to the definition of α. The resulting α_{out} and Θ_{out} estimates, which were optimised by the A-test differed considerably from the respective true values. A Θ_{out}=0.14 was obtained, which differs 0.04 from the Θ_{true}. The α_{out} (0.42) also differed substantially from reality (input value: α_{true}=0.5) (Table 2.3). Figures 2.9 and 2.10 show α_{out} and Θ_{out} in relation to α_{true} for several different family distributions (Figure 2.9) and α_{out} and Θ_{out} in relation to Θ_{true} for another set of quite extreme family distributions (Figure 2.10). The figures clearly show that theta estimates may deviate considerably when unlinked families are much larger than linked families, for small α_{true} values and also when Θ_{true} is small. Dissimilar family size also causes an pronounced bias in the alpha estimate.

When we evaluated multipoint information instead of pairwise data, the accuracy of the obtained Θ_{out} improved drastically (Table 2.3). When we extended the example mentioned above with one additional marker, placed at Θ_2=0.1 opposite the first marker (Θ_1=0.1), we obtained unbiased estimates for Θ_{out}, while the bias in α_{out} increased (α_{out}=0.37). All our single-chromosome multipoint evaluations revealed Θ_{out} values identical to Θ_{true}. However, the accuracy of the A-test with regard to the resulting α_{out} decreased for small family sizes. Under some circumstances the estimated α was shown to be almost half its actual value (NIM linked=2; NIM unlinked =4). This effect was associated with a major increase of the lod score (Table 2.3).

We also tested the performance of the A-test when the imaginary chromosome approach (ICA) was applied. The combined analysis of two loci instead of one in between markers at 10% recombination (Θ_1=Θ_2=Θ_3=Θ_4=0.1, α=0.5)) again revealed unbiased Θ estimates, while the α_{out} estimate showed an important improvement (α_{out}=0.43) compared with the pairwise and multipoint analyses. More in general the results - as shown in Table 2.4 - confirm that an important improvement of the outcome can be achieved by using a two-locus ICA application. The results show a fairly accurate Θ_{out} , while the accuracy of α_{out} indicates a major improvement compared with the single-chromosome multipoint test and occasionally even shows a better estimate than α_{out} obtained from the pairwise analysis (Table 2.3).

Not only the accuracy of the estimators, but also the lod score yield improves substantially if markers closely linked to both loci can be analyzed.

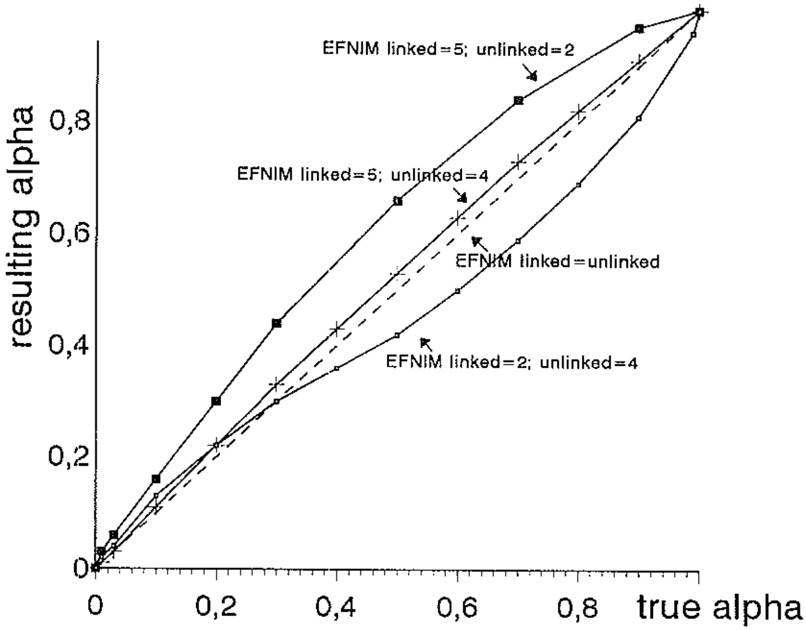


Figure 2.9a: Alpha values obtained with A-test when true alpha varies, while true theta=0.1.

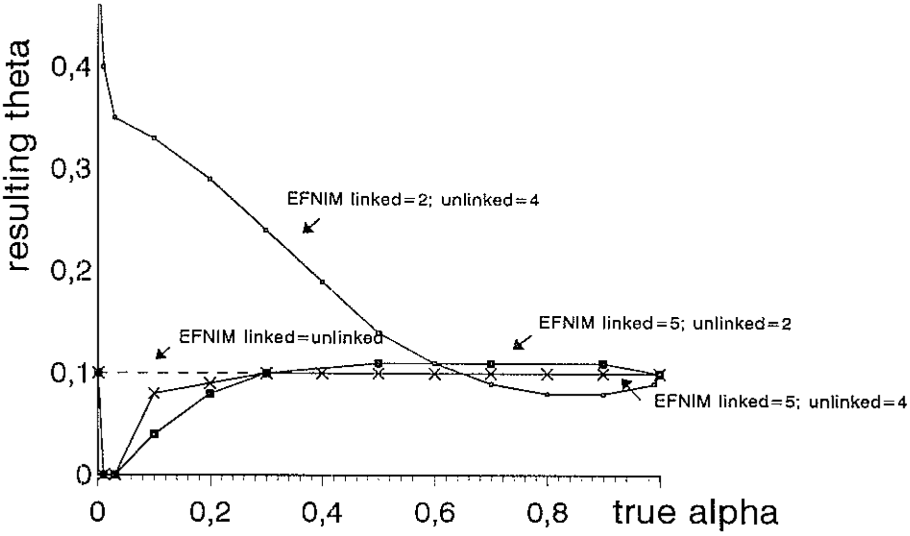


Figure 2.9b: Theta values obtained with A-test when true alpha varies, while true theta=0.1.

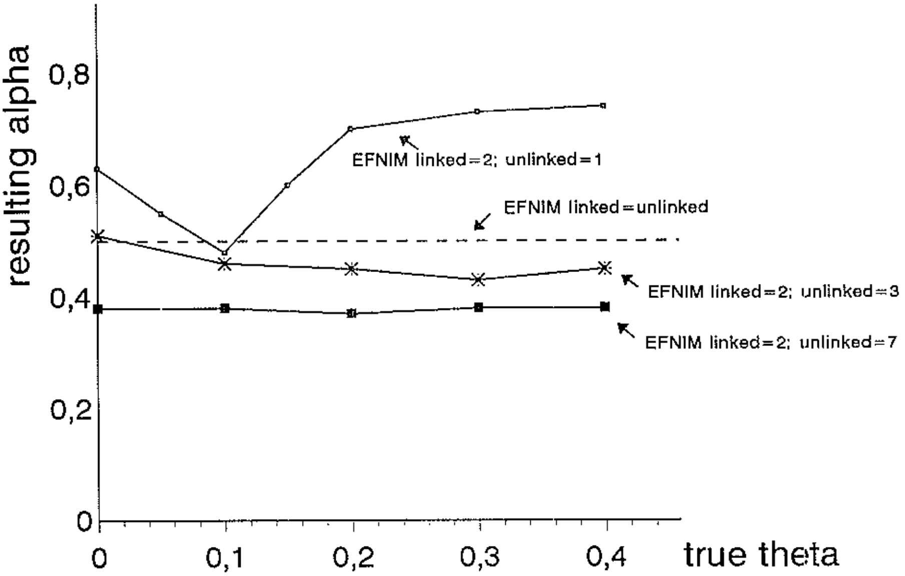


Figure 2.10a: Alpha values resulting from A-test when true theta varies, while true alpha=0.5.

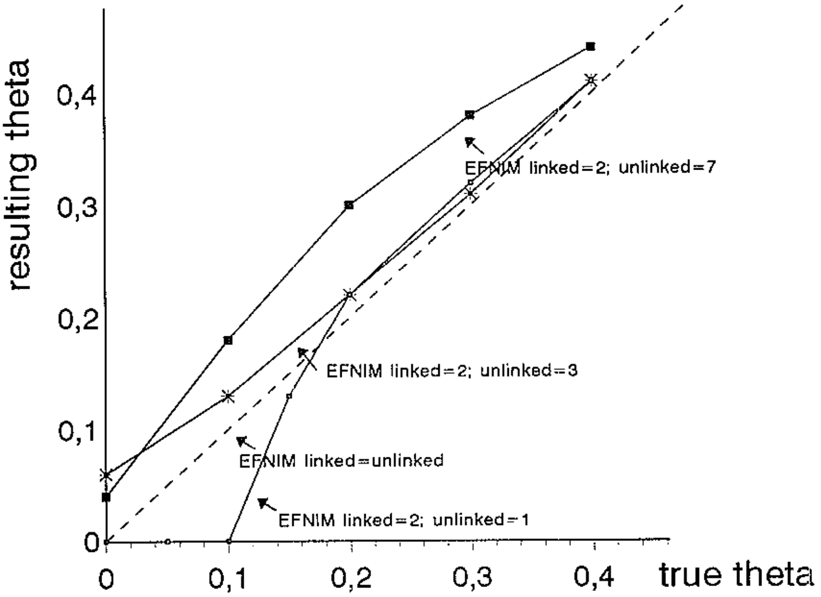


Figure 2.10b: Theta values resulting from A-test when true theta varies, while true alpha=0.5.

Table 2.4
Alpha and theta values obtained from simultaneous analysis of two chromosomal regions (ICA).

EFNIM linked /un-linked	True					Obtained					Lod score obtained with 10 families
		Chromosome 1:		Chromosome 2:			Chromosome 1:		Chromosome 2:		
	α	Θ1	Θ2	Θ1	Θ2	α	Θ1	Θ2	Θ1	Θ2	
2 / 4	50%	10%	10%	10%	50%	42%	10%	10%	10%	50%	3.440
		10%	30%	10%	50%	40%	9%	31%	11%	50%	3.190
		10%	10%	10%	10%	43%	10%	10%	10%	10%	4.523
		10%	10%	50%	50%	37%	10%	10%	50%	50%	0.684
2 / 4	30%	10%	10%	10%	50%	22%	10%	10%	11%	50%	4.162
		10%	10%	10%	30%	22%	10%	10%	10%	30%	4.477
		10%	10%	10%	10%	24%	10%	10%	10%	10%	5.729
		10%	10%	50%	50%	17%	10%	10%	50%	50%	0.183

Notes: the table is similar to Table 2.3. 'Lod score' refers to $Z_{(\alpha,X1,X2)}$, α *denotes the alpha of locus 1 on chromosome 1. The last line in each block denotes a single chromosome analysis (only one region is known).*

Informativeness of markers

We also investigated the influence of differences in informativeness of markers on the outcome of the A-test. A biased outcome of an ICA study because of differences in informativeness in both regions is very well conceivable considering direct relationship between the NIM and the PIC of the markers. Using a modified version of the software we studied two groups of families, both of the same NIM category. In this part of our study we did not regard the NIM of a certain family as a constant, but varied the NIM depending on the chromosome under investigation. This way we created a test model involving two chromosomal regions, one marker on each chromosome, which allows for different informativeness of the markers. We varied the α_{true} and Θ_{true}, but we did not find any deviations in the outcome (data not shown).

Misclassification

Clinical misclassification is perhaps the most common source of errors. We therefore investigated the effect of misclassification analytically. As expected, our results indicated that - for pairwise analysis on small families - misclassification caused an increase of Θ_{out} and α_{out} (Table 2.5). As long as map positions in between flanking markers remained favoured above positions 'outside the map' the results of multipoint heterogeneity analyses showed no deviations of Θ_{out}, while α_{out} became smaller than α_{true}. Multipoint ICA analysis also turned out to be more resistant to errors. No deviations of α_{out} were observed, while deviations of Θ_{out} were again mild or absent.

Table 2.5
The effect of clinical misclassification.
(EFNIM linked=2; EFNIM unlinked=2)

Misclas. rate	True					Obtained				
	Chromosome 1:			Chromosome 2:		Chromosome 1:			Chromosome 2:	
	α	$\Theta1$	$\Theta2$	$\Theta1$	$\Theta2$	α	$\Theta1$	$\Theta2$	$\Theta1$	$\Theta2$
10%	50%	10%	50%	50%	50%	55%	20%	50%	50%	50%
	50%	10%	10%	50%	50%	40%	10%	10%	50%	50%
	50%	10%	20%	50%	50%	40%	10%	20%	50%	50%
20%	50%	10%	50%	50%	50%	50%	25%	50%	50%	50%
10%	50%	10%	10%	10%	10%	50%	10%	10%	10%	10%
	50%	10%	20%	10%	20%	50%	12%	18%	12%	18%

Notes: The first and fourth line of data represent pairwise analyses; the second and third represent multipoint tests; the last two lines represent ICA analyses.

Analysis of actual data

If the Effective number of informative meioses (EFNIM) - the equivalent of NIM in actual data - of each family under investigation is known, the possible effect of dissimilar family size on a heterogeneity analysis carried out in practice can be diagnosed. As an example we re-analyzed data from studies on chromosome 9-linked tuberous sclerosis (TSC). The size distribution of the tuberous sclerosis family data had already been reported to be roughly similar for the linked and unlinked group (Chapter 2.2), but no statistical evidence for this had been provided. Figure 2.11 shows the EFNIM distribution of this set of 128 families. We tested the family data for a possible bias due to differences in the EFNIM distribution, using the reported distance from the flanking marker on the proximal side of TSC1 (ABL), which is 5% (Chapter 3.2). From the same study we obtained the distance (7%) to the distal flanking marker (D9S14). Our testAtest multipoint analysis revealed a bias of +1% in α_{out} and no bias at all in the theta values.

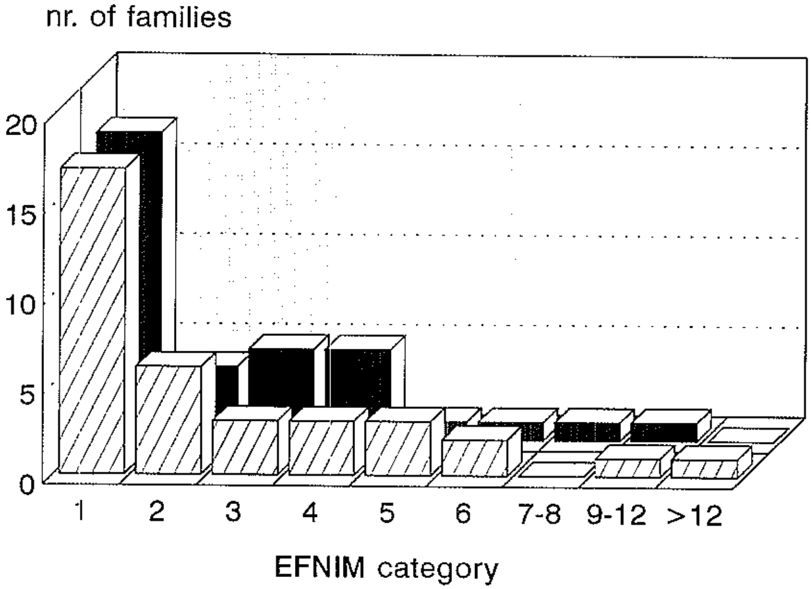


Figure 2.11: Tuberous sclerosis family distribution (family set used in collaborative analysis) by 'effective number of informative meioses' (EFNIM). The distribution of unlinked families is shown in front (hatched), the chromosome 9 linked families are shown behind (solid). Linked and unlinked TSC families show an almost equal size distribution.

Discussion

The origin and extent of bias

The outcome of our analyses demonstrates that the estimators of the A-test may be biased. The extent of the bias is method dependent, but the variants of the method themselves are not the cause of the inconsistency. We determined the characteristics of the linkage problem under investigation as the cause of bias. Linkage problems consist of two parts: on one hand the family data, which may contain a certain ascertainment bias and an amount of misclassification, and on the other hand the disease entity itself with its specific characteristics, such as penetrance, severity, age at onset, phenotype, mode of inheritance, mutation rate and phenocopy rate. By selecting patients from a subset of the population an investigator can provoke a pure selection bias of α. If an ascertainment strategy aims at extending linked families more than unlinked families another kind of bias will be induced. This bias affects both α and Θ and has to be termed a systematic bias or an inconsistency. Misclassification errors have also been shown to be a biasing factor for both α and Θ, although these errors have a clinical origin. With respect to the disease related features, one can tentatively conclude that biases occur if the genotype-phenotype correlations are not identical for all disease loci. It will however be difficult to draw a definiteve conclusion that can be applied on all these features.

The extent of bias has been shown to be method-dependent. Our study indicates a few possibilities to minimise the bias in actual linkage analysis. The most precise estimates for α were obtained in pairwise tests, provided that linked families were not larger than unlinked families. The best estimates for Θ were always obtained in multipoint analyses. Therefore it seems wise to perform both types of analysis and to compare the outcome carefully. If one or more candidate regions for the second locus are known, the imaginary chromosome approach (ICA) is recommendable, since this method generally revealed reasonable precise estimates for both α and Θ, even if a few persons have been clinically misclassified. In our study we have shown that it is possible to compute the systematic bias. If a certain data set provokes a bias due to dissimilarities in family size, it is feasible to correct the results afterwards, that is after LINKAGE and HOMOG analysis. In the case of TSC this would mean that we subtract 1% from the original α_{out} (51%) (Chapter 3.2), resulting in a corrected α estimate of 50%.

Clinical implications

Apart from geographical differences, it has become a custom to talk about one single alpha per disease locus. We now propose a change. Instead of determining one single alpha, it is more realistic to determine the relationship between alpha and family size, unless linked and unlinked families can be proven to have similar sizes. In the case of TSC, 128 linked and unlinked families showed an almost equal size distribution. The slight dissimilarity was not sufficient to induce a significant systematic α bias.

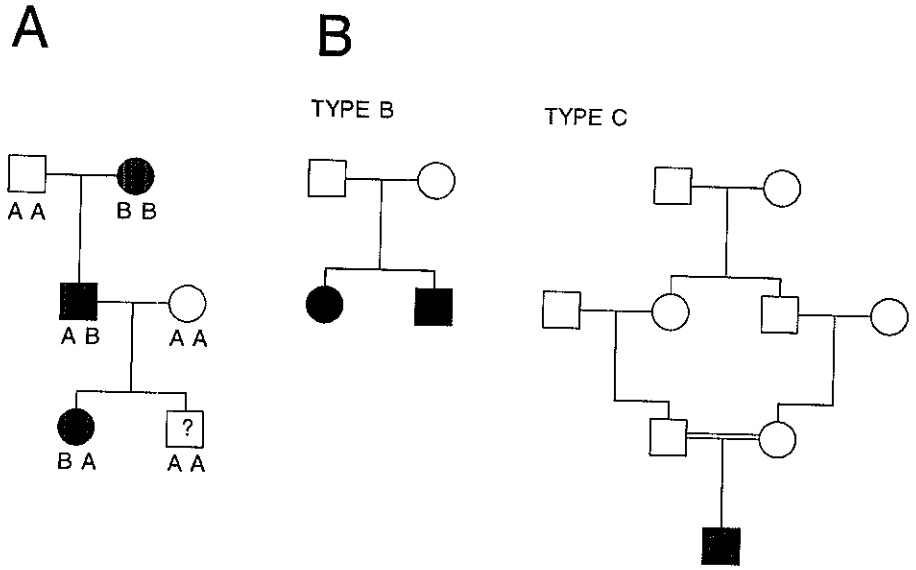


Figure 2.12:
A: Example of family segregating for a dominant disorder, closely linked to the depicted marker (Θ=2%). The risk for the youngest child of developing the disease is discussed in the text.
B,C: Examples pedigrees segregating a recessive disorder. In type C the α is almost equal to the relative gene frequency (q'), whereas no strict correlation exists between α and q' in a set of type B families (see text).

It is however not difficult to imagine a data set with an overall α=80%, while the size specific alpha for the EFNIM=1 type of families is not more than 60%. We have mentioned several potential causes for such a dissimilarity. This will result in a biased overall estimate of α_{out}, but even more striking than the bias in α_{out} is the consequence of the dissimilar size distribution for risk calculations. The effect can be well illustrated when we consider the risk to be affected of an individual in an EFNIM=1 type of family, as depicted in Figure 2.12 (A), who inherited the low-risk haplotype. When this individual had been given a risk of 7.4 % assuming an a-priori risk (α) for being linked of 80%, a dramatically increased risk of 14.2% can be calculated by applying the size-specific $\alpha_{EFNIM=1}$ of 60%. In practice the effect will be even worse, because a biased α_{out} will be used instead of the true overall α of 80%, due to the systematic bias.

Does α represent the relative gene frequency?
In our study we have referred to various different types of alpha. By assuming

different size-specific alphas, we demonstrated that α_{true} and the empirically derived α_{out} are essentially different. One may question whether all of these different types of α properly represent the relative gene frequency ($q_1/(q_1+q_2)$). The definition of α (the proportion of linked families) predicts a very strong relationship. However, this comparison leads to unexpected complications as illustrated in the following example.

Consider an autosomal recessive disease which may result from mutations at any of two disease genes at different locations. Let us assume that mutant alleles at locus 1 occur with frequency 0.02, while the mutant alleles at locus 2 have a frequency of 0.01. Further assume that penetrance is complete in homozygous individuals and absent in heterozygotes. Simple probability calculations (e.g. by MLINK) on the pedigrees shown in Figure 2.12 yield the following result: of all families of type B 80 % is segregating for a mutation at locus 1. The discrepancy between the α of 80% and the relative gene frequency of 67% is considerable. No discrepancy between q and α can be demonstrated for family type C, where 67% of all families segregate for a mutation at locus 1.

This problem is due to differences in the frequencies for the homozygous states (q_1^2 and q_2^2) and are therefore determined by the family structure. In the case of a recessive mode of inheritance this problem occurs when gene frequencies are unequal. The example clearly shows that parametrization of differences in gene frequencies via α is not completely satisfactory. In other words: none of the various types of alpha mentioned above gives a proper representation of the relative gene frequency. Although less pronounced this problem may also occur in the case of dominant inheritance when gene frequencies are unequal and one of the gene frequencies is high. This might for instance be the case for some psychiatric disorders and for predisposing factors for cardiovascular disease.

We may conclude that α does not represent the relative gene frequency. We have also shown that expressing the proportion of linked families by means of α is somewhat unsatisfactory because of the inconsistency and the ambiguous definition of 'a family'. If the use of α for these purposes is inappropriate, what does α represent? Although we did not directly investigate this question, we propose that α most of all represents the relative information content present for the mapping of each locus.

Significance tresholds

The biases in α_{out} and Θ_{out} are associated with higher lod scores at α_{out} and Θ_{out} compared with α_{true} and Θ_{true}. In most cases this 'inflation' is moderate. However, in some of our more extreme examples we demonstrated resulting lod scores ($Z_{(\alpha,X1,X2)}$) that were almost twice as high as the lod score at α_{true} / Θ_{true} (Table 2.4). This means that results of heterogeneity tests have to be interpreted with care if the significance level approximates the threshold. If one uses more conservative threshold values, as proposed by several authors (Chapter 2.2; Rish, 1989), this will in general not cause any problem.

Conditions and recommendations

Our analytical evaluations of the A-test show that - for uncomplicated dominant traits - the procedure is quite reliable. We also demonstrated that more recent variants, basically the analysis of multipoint data and the ICA, are correct applications. Under more extreme circumstances systematic biases may occur. We can not confirm the conclusions of Maclean et al. (1994) that "both α and Θ are much too large" and "estimates of α bear little relation to true α". We realise that for some investigators an unbiased estimate of α and Θ is desired, but not essential, while others, like genetic counsellors, really need to rely on these parameters. If more reliable estimates of the parameters are desired, the following conditions and recommendations for the use of the A-test may help to minimise the bias:

I) There is only one mutation per family allowed: As discussed above the A-test demands a single mutated locus per family and an equal gene frequency. For dominant disorders this can be overcome by stating that only one mutated allele per family is allowed. In practice this means that by reducing the frequency of the mutated gene substantially the probability of the homozygous state can be reduced effectively. Analyses on recessive disorders should be performed and interpreted with care, because α is a poorly defined parameter.

II) Sizes of linked and unlinked families may not differ: This study was based on the principle that the proportion of linked families (α) is not necessarily the same for all types of families. However, the A-test does assume identical α's. In practice such a dissimilarity might reflect a clinical difference, such as a difference in severity, number of offspring, or penetrance. If this condition can not be met, one should be aware of the possiblility that a systematic bias of α and Θ might occur. If a considerable bias is to be expected, other methods like TMLINK (Lathrop and

Ott, 1990). should be applied.

III) Ascertainment strategies should be planned with care: Differences in ascertainment influence the α directly. Geographical and phenotypical selection biases should be avoided. Age at onset and penetrance differences also affect the α directly, because these factors determine whether a certain family will come to the attention of a clinician. Furthermore these three factors may have an effect on the family size and provoke a systematic bias as mentioned above (II).

IV) Phenocopies are not allowed throughout the calculations: All affecteds are assumed to have inherited the disease gene.

V) The amount of information in the analysis should be maximized: Some applications of the A-test are more error-prone than others. When a correct Θ estimate is desired, multipoint analysis can be recommended, especially if misclassifications or size differences are present. When multiple candidate regions are known, we strongly recommend the ICA.

Software availability

The testAtest program is available on request from the first author.

Acknowledgements

We thank Dr. D. Lindhout and Dr. J. Ott for helpful comments on the manuscript and on the analysis. We also like to thank Dr. Peters and Dr. Sampson for their support. This study was funded by the Netherlands Organisation for Scientific Research (NWO). A pilot study preceding this project was funded by the National Tuberous Sclerosis Association (USA).

References

Smith CAB: Homogeneity tests for linkage data. Proc Sec Int Congr Hum Genet 1961;1:212-213

Ott J: Analysis of human genetic linkage (revised edition), Johns Hopkins, Baltimore. 1991.

Janssen LAJ, Sandkuyl LA, Merkens EC, et al. Genetic heterogeneity in tuberous sclerosis. Genomics 1990;8:237-242

Ott J: Computer-simulation methods in human linkage analysis. Proc Natl Acad Sci USA 1989;86:4175-4178.

Rish N: Linkage detection tests under heterogeneity. Genet Epidemiol 1989;6:473-480.

MacLean CJ, Sham PC, Ploughman LM, et al. (Letter). Am J Hum Genet 1994;54:564-567.

Lathrop GM, Ott J: Analysis of complex diseases under oligogenic models and intrafamilial heterogeneity by the LINKAGE programs. Am J Hum Genet 1990;47:A188.

CHAPTER 3

LINKAGE ANALYSIS IN TUBEROUS SCLEROSIS

3.1

A COMPARATIVE STUDY ON GENETIC HETEROGENEITY IN TUBEROUS SCLEROSIS: EVIDENCE FOR ONE GENE ON 9Q34 AND A SECOND GENE ON 11Q22-23

L.A.J. Janssen
S. Povey
J. Attwood
L.A. Sandkuyl
D. Lindhout
P. Flodman
M. Smith
J.R. Sampson
J.L. Haines
E.C. Merkens
P. Fleury
P. Short
J. Amos
D.J.J. Halley

Ann NY Acad Sci 1991; 615: 306-315.

Linkage studies in tuberous sclerosis

Tuberous sclerosis (TSC) is a dominantly inherited disorder. Although the variability in expression is high, complete penetrance may be assumed. Following a suggestion of positive lod scores between TSC and ABO (Connor et al., 1987a), Fryer and coworkers reported linkage between TSC and the ABO blood group locus with a peak cumulative lod score of 3.85 at $\Theta = 0$ (Fryer et al., 1987). This was supported by Connor et al., who found linkage to the abl oncogene with a lod score of 3.18 at $\Theta = 0$. (Connor et al., 1987b). These findings enabled Connor and coworkers to perform a first-trimester exclusion of TSC in a pregnancy at risk. The abl oncogene maps approximately 10cM proximal of ABO on 9q24 (Figure 3.1). Subsequent studies mapped the TSC locus between MCT136 and MCOA12 (Sampson et al., 1989).

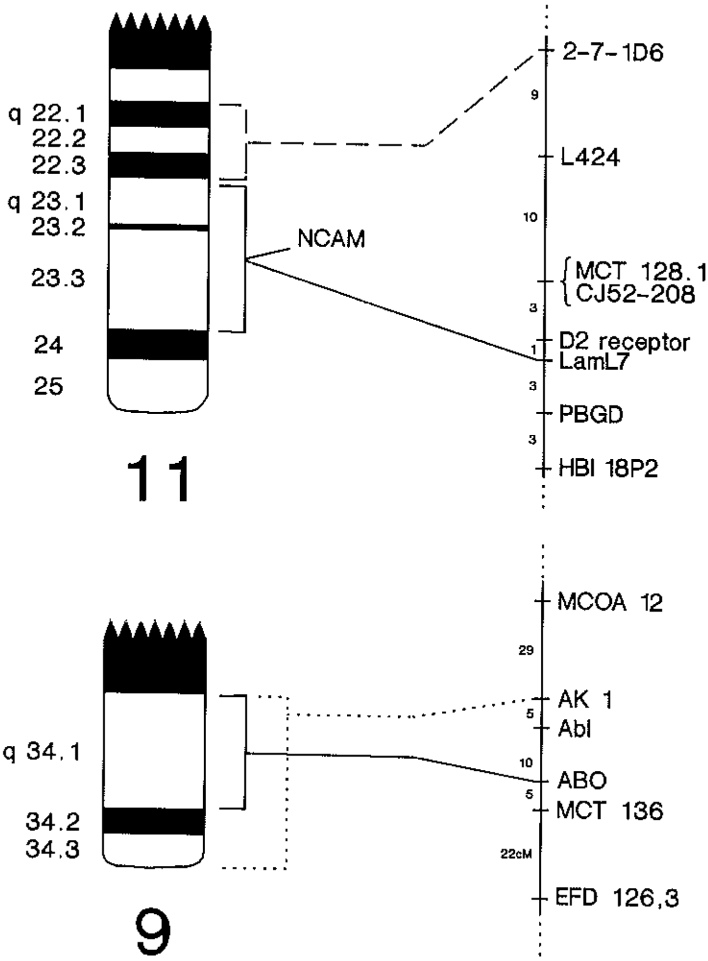


Figure 3.1. Genetic maps of chromosome 9 (top) and 11 (bottom) with physical locations of selected markers indicated on the karyogram. The positions of the markers used in our study (MCOA12, AK1, abl, ABO, MCT128.1, CJ52.208, LamL7 and PBGD) are indicated as are their positions towards other markers of interest. The order of ABO and MCT136 is in doubt (J.L. Haines, pers. communication). Scale is in centimorgans (cM).

In contrast with these reports, a substantial number of families were described with free recombination between TSC and the chromosome 9 markers (Northrup et al., 1987; Renwick, 1987). Recently Kandt et al. excluded TSC from an area of 20 cM encompassing the ABO locus (Kandt et al., 1989). In 1988 Clark et al described a TSC patient with trisomy for 11q23.3-qter (Clark et al., 1988). The mother was reported to have a balanced t(11q23;22q11.2) rearrangement, but there was no evidence of TSC in either parent. The gene for neural cell adhesion molecule, which is located on 11q23, was postulated as a candidate locus. Linkage studies in 15 North American families using four chromosome 9q markers and six chromosome 11q markers revealed evidence for linkage to chromosome 11 (Smith et al., 1990). A maximum two-point lod score of 3.26 at $\Theta = 0.08$ was obtained with the marker MCT128.1.

The present and other (Haines et al., 1991) collaborative efforts aim at ending the controversy by combining data from several studies, including those in which linkage to either chromosome 9 or 11 has been established. Data were combined without selection for any type of families. We compared different statistical strategies to investigate the possibility of locus heterogeneity.

Locus heterogeneity

In the last decade reverse genetics has become an important tool to localize and identify disease genes. This approach is based on the search for co-segregation of a disease trait and genetic markers of know location. The growing number of mapped infomative polymorphic markers enhanced the success of this approach. Several genes causing diseases have been mapped and identified this way. This approach is based, however, on the assumption that data from informative meioses in different families can be combined. This assumption would be invalid if the disease-causing locus is not identical in all families, that is, if locus heterogeneity exists. Given the complexity of biochemical pathways, it is expected that locus heterogeneity will prove to be a feature of a significant proportion of genetic diseases. Of interest, then, is the analysis of diseases that have been mapped in some families to a locus, although in other phenotypically similar families a mutation at the mapped locus cannot account for the observed data.

Tuberous sclerosis may very well be a member of this small group of diseases (Sampson et al., 1989; Janssen et al., 1990; Haines et al., 1989; Haines et al., 1991), showing locus heterogeneity. Other examples of diseases belonging to this group are elliptocytosis, X-linked retinitis pigmentosa, Charcot-Marie-Tooth disease type 1, adult polycystic kidney disease, and manic depression. Only for Charcot-Marie-Tooth and X-linked retinitis pigmentosa has more than one locus been mapped

(Griffiths et al., 1989; Vance et al., 1989; Ott et al., 1990).

In some diseases the assignment of families to a certain locus is facilitated by the type of segregation. This is the case when one or more loci are X-linked, while other loci are autosomal. Diseases like limb girdle muscular dystrophy and manic depression are phenotypically indistinguishable from their X-linked counterparts, Becker dystrophy and X-linked manic depression. Most inherited diseases show autosomal inheritance, however. For autosomal diseases with locus heterogeneity, arguments for classification apart from linkage data are hard to obtain. Lack of linkage data may explain why the diseases have not yet been recognized as being heterogeneous. Even when linkage data are available, a locus may be overlooked, if the statistical analysis does not allow for locus heterogeneity. Lander and Botstein estimated that a trait-causing locus that accounts for 60% of all cases may be missed completely in a linkage study under the false assumption of homogeneity. If the tested marker lies within 1 cM of the disease locus, linkage may still be excluded from a region of about 20 cM encompassing the marker. This implies that the recent exclusion data of Kandt et al. do not undermine the validity of the previous assignment of a TSC locus to 9q34. The power of reverse genetics is clearly reduced in the case of locus heterogeneity. This complicated problem requires the development and use of powerful statistical approaches.

Locus heterogeneity in tuberous sclerosis

For most heterogeneous traits only one candidate region is known. Since there are two candidate regions for TSC, TSC1 (9q) and TSC2 (11q), our problem is quite exceptional. We compared various approaches that allow for locus heterogeneity. These methods may be divided into two groups: The first group contains approaches that start with a separation of families, followed by conventional two-point analysis or multipoint analysis, followed by weighted family assignment to the putative disease loci. Results obtained this way are presented in this paper.

A well-known test for heterogeneity is the "classic" admixture test (Smith et al., 1963). It has been used in the analysis of almost all genetically heterogeneous diseases mentioned above. For each map location (X) and for each proportion of linked families (α) a lod score under heterogeneity $Z_{i(\alpha,x)}$ is computed.

$$Z_{i(\alpha,X)} = \log(\alpha . 10^{Z_{i(X)}} + (1-\alpha))$$

The combination of α and X for which $Z_{i(\alpha,x)}$ is maximal is evaluated in a Chi square test, using the maximum value of $Z_{i(\alpha,x)}$, assuming homogeneity (α=1) as a reference. The admixture tests were performed by using the HOMOG computer program (Ott et al., 1985).

An alternative approach is also based on the admixture test. Instead of one position (X), two positions (X_1 and X_2 are evaluated. The equation

$$Z_{i(\alpha,X_1,X_2)} = \log(\alpha \cdot 10^{Z_{i(X_1)}} + (1-\alpha) \cdot 10^{Z_{i(X_2)}})$$

is similar to the equation given above, where X_2 is at Θ = 0.5. This alternative approach can be taken to test for two distinct loci on one chromosome. Such a test has been performed successfully for X-linked retinitis pigmentosa (Ott et al., 1990). If both loci are not on the same chromosome, a so-called "imaginary chromosome" has to be constructed, as described previously (Janssen et al., 1990). We composed the imaginary chromosome by combining the results of two multipoint analyses (one for each candidate region) in a head-to-tail orientation. The most distal marker on chromosome 9 and the most proximal marker on chromosome 11 were flanking the junction. The recombination fraction between these markers and the junction was set at 0.5. By the use of this imaginary construct, the TSC linkage problem to markers on separate chromosomes was reduced to a linkage problem with two regions of interest on a single (imaginary) chromosome. The imaginary chromosome was analyzed with the programs HOMOG2 and POINT4 (Ott et al., 1985). Both programs are part of the HOMOG package. This method is only valid in the absence of further heterogeneity in the data set. Therefore a final check using the HOMOG3 program was performed. This program allows for a third locus, which may be unlinked to either candidate region.

Seventy-three families were analyzed, the majority of which have been included in previous publications (Sampson et al., 1989; Smith et al., 1990; Janssen et al., 1990). For two-point and multipoint analyses the various programs of the LINKAGE package were utilized.

Results obtained with the "classic" admixture tests

The simple application of the "classic" admixture test is the analysis of two-point linkage data using HOMOG. Two of these tests were performed, one for the ABO locus on chromosome 9 and one for the anonymous chromosome 11 marker MCT128.1. The results of these tests are shown in Table 3.1. Significant support for heterogeneity could only be obtained with the ABO linkage data. A similar test using MCT128.1 revealed only insignificant support for heterogeneity.

A more advanced application of the admixture test uses a map of markers. The markers MCOA12 (D9S28, *Msp*I), AK1 (*Taq*I or protein polymorphism), abl (*Pst*I or

Table 3.1:
Results of the various admixture tests.

Analysis: Marker(s)/ programs	Components of Chi-square test for H_2 versus H_2:			Most likely location assuming heterogeneity
	X^2	p value	α	
ABO/ MLINK-HOMOG	4.593	0.0161 (1 df)	TSC1: 38%	TSC1: at ABO
MCT128.1/ MLINK-HOMOG	1.057	0.1520 (1 df)	TSC2: 47%	TSC2: 35 cM from MCT128.1
Map of chromosome 9 markers/ LINKMAP-HOMOG	10.172	0.0007 (1 df)	TSC1: 45%	TSC1: 6 cM prox. of ABO
Map of chromosome 11 markers/ LINKMAP-HOMOG	0.505	0.2386 (1 df)	TSC2: 30%	TSC2: 35 cM dist. of PBGD
Combined map (imaginary chromosome) LINKMAP-HOMOG	15.448	0.0002 (2 df)	TSC1: 48% TSC2: 52%	TSC1: 6 cM prox. of ABO TSC2: 35 cM dist. of PBGD

Note: H_1 = hypothesis of linkage under homogeneity; H_2 = hypothesis of linkage under heterogeneity; α = proportion of families linked with indicated locus.

*Taq*I), and ABO were used as chromosome 9 markers in a five-point analysis. Another five-point analysis was carried out using the chromosome 11 markers MCT128.1 (D11S144, *Msp*I), CJ52.208 (D11S351, *Msp*I), LamL7 (D11S29, *Taq*I and PBGD (PBGD, *Msp*I or *Pst*I). All inter-marker distances are given in Figure 3.1. The results obtained with HOMOG are presented in table 3.1. The use of a map of four chromosome 9 markers instead of one resulted in an increased significance when heterogeneity was tested versus homogeneity. However, a similar study on the chromosome 11 region showed no increased significance.

Results obtained with the imaginary chromosome approach

By combining the results of both five-point analyses into one larger structure an imaginary chromosome was constructed. The results obtained with this approach are shown in Tables 3.1, 3.2 and 3.3. The outcome of the analysis showed strong support for locus heterogeneity, with one putative TSC locus 6 cM proximal to ABO (TSC1) and other putative TSC locus 35 cM distal to PBGD (TSC2). Additional tests revealed that other possible orders could not always be significantly excluded (Table 3.3). The most likely location of both putative TSC loci was confirmed by a HOMOG3 analysis. These results revealed no evidence for a third locus in our data set.

The validity of the model

Although the quality of the model we tested permitted locus heterogeneity to be proven, we may still question whether we tested the optimal model. The model can be regarded to be less optimal if we used false inter-marker distances or orders, if we did not test markers from the real TSC regions, if further locus heterogeneity is the case, or if the penetrance (100%) is not correct. Testing a less optimal model might results in reduced significance an incorrect putative locations. The putative TSC2 locus is not flanked by a distal marker. The additional tests on the chromosome 11 region revealed an alternative TSC2 locus 30 cM proximal to MCT128.1. This alternative position is almost as likely as the position 35 cM distal PBGD (Table 3.3). The validity of the tested region on chromosome 11 seems questionable.

Table 3.2:
The 73 TSC families and their assignment to chromosome 9 or 11

Family	Assignment to chromosome by method			Lod scores at putative TSC loci	
	ICA	MLSM	NLSM	TSC1	TSC2
Rot 2079	9 (0.80)	9	9	0.41	-0.24
Rot 2046	9 (0.61)	-	9	0.23	0.00
Rot 2067	11 (0.56)	-	-	0.00	0.07
Rot 2068	11 (0.95)	11	11	-1.03	0.24
Rot 2077	-	11	-	0.24	0.21
Rot 1222	11 (0.89)	11	11	-0.67	0.22
Rot 1219	9 (0.87)	9	-	0.79	-0.06
Rot 1264	9 (0.68)	9	-	0.38	0.02
Car 0001	9 (0.98)	9	9	1.07	-0.63
Car 0002	-	-	-	0.00	-0.03
Car 0003	11 (1.00)	-	11	-3.64	-0.01
Car 0004	11 (0.58)	-	11	-0.14	-0.03
Car 0005	9 (0.66)	9	9	0.21	-0.11
Car 0006	9 (0.87)	-	-	0.84	-0.01
Car 0007	9 (1.00)	9	9	2.79	-0.59
Car 0008	11 (0.52)	9	-	0.00	0.00
Irv 0004	11 (0.90)	-	11	-0.92	0.00
Irv 0008	11 (1.00)	11	11	-3.10	0.15
Irv 0011	9 (0.60)	9	9	0.00	-0.21
Irv 0015	11 (1.00)	11	11	-3.09	0.44
Irv 0016	11 (0.61)	11	11	-0.26	-0.10
Irv 0019	11 (0.52)	-	-	0.00	0.00
Irv 0020	11 (0.57)	-	-	-0.14	-0.05
Irv 0021	9 (0.75)	9	9	0.49	-0.03
Irv 0023	11 (0.52)	-	-	0.00	0.00
Irv 0024	11 (0.80)	11	11	-0.41	0.15
Irv 0026	11 (0.71)	11	-	0.01	0.38
Irv 0028	11 (0.55)	-	-	0.00	0.06
Irv 0029	9 (0.55)	-	9	0.00	-0.12
Irv 0033	11 (0.56)	-	9	0.00	0.07
Irv 0101	11 (0.53)	-	-	0.23	0.25
Lon 5400	11 (1.00)	11	11	-2.05	0.26
Lon 5348	11 (0.95)	11	11	-1.05	0.20
Lon 5431	11 (0.73)	-	9 & 11	-0.49	-0.08

Table 3.2 (continued)

Family	Assignment to chromosome by method			Lod scores at putative TSC loci	
	ICA	MLSM	NLSM	TSC1	TSC2
Lon 5406	11 (0.89)	11	11	-0.86	0.00
Lon 5244	11 (0.86)	11	11	-0.76	0.00
Lon 5384	11 (0.87)	-	11	-0.58	0.21
Lon 5214	9 (0.60)	9	9	0.00	-0.20
Lon 5372	9 (0.75)	-	-	0.52	0.00
Lon 5386	9 (0.79)	9	9	0.53	-0.08
Lon 5272	9 (0.72)	-	-	0.52	0.07
Lon 5301	11 (0.79)	-	9 & 11	-0.59	-0.07
Lon 5275	9 (0.63)	9	9	0.00	-0.27
Lon 5349	9 (0.65)	9	9	0.24	-0.07
Lon 5235	9 (0.66)	-	9	0.25	-0.06
Lon 5477	9 (0.52)	9	9	0.00	-0.08
Lon 5379	9 (0.53)	9	9	0.00	-0.08
Lon 5252	9 (0.97)	9	9	1.19	-0.29
Lon 5385	11 (0.75)	-	9 & 11	-0.52	-0.06
Lon 5241	9 (0.68)	-	-	0.52	0.15
Lon 5274	9 (0.71)	-	11	0.42	0.00
Lon 5350	11 (0.88)	11	11	-0.66	0.15
Lon 5388	11 (0.81)	11	9 & 11	-0.66	-0.08
Lon 5441	9 (0.53)	-	-	0.09	0.00
Lon 5404	9 (0.78)	-	9	0.52	-0.07
Lon 5412	9 (0.66)	-	9	0.25	-0.06
Lon 5298	9 (0.62)	-	-	0.25	0.00
Bos 1	9 (0.81)	-	-	0.69	0.04
Bos 2	9 (0.57)	-	-	0.15	0.00
Bos 3	11 (0.52)	-	-	0.00	0.00
Bos 4	11 (0.81)	-	9 & 11	-0.66	-0.07
Bos 5	11 (0.60)	-	-	0.00	0.14
Bos 6	9 (0.79)	9	-	0.51	-0.10
Bos 7	11 (0.81)	-	9 & 11	-0.64	-0.04
Bos 8	9 (0.68)	-	9 & 11	0.00	-0.36
Bos 9	9 (0.91)	9	9	1.07	0.01
Bos 10	9 (0.52)	-	9	0.00	-0.07
Bos 11	11 (0.71)	-	9	-0.41	-0.05

Table 3.2 (continued)

Family	Assignment to chromosome by method				Lod scores at putative TSC loci	
	ICA		MLSM	NLSM	TSC1	TSC2
Bos 12	9	(0.54)	-	9 & 11	0.00	-0.10
Bos 13	11	(0.54)	-	-	0.00	0.03
Bos 14	9	(0.67)	-	11	0.41	0.06
Bos 15	9	(0.62)	-	-	0.25	0.00
Bos 16	11	(0.83)	-	-	-0.66	0.00
Bos 17	11	(0.70)	-	-	-0.33	0.00

Note: ICA = imaginary chromosome approach; MLSM = family assignment by maximum lod score method; NLSM = possible family assignments by negative lod score method. (Since families can be selected twice by using the NLSM, all possible assignments are indicated instead of a definitive assignment.) Next to the family assignment by the ICA, the posterior probability of the assignment of each family (by ICA) is given in parentheses. The last two columns show the lod scores at the putative TSC1 and TSC2 loci, as calculated using the imaginary chromosome approach. Rot = Rotterdam, Car = Cardiff, Irv = Irvine, Lon = London and Bos = Boston.

Table 3.3:
Odds against possible orders when compared to the most likely order*,#

Locus order	Odds against locus order
MCOA12 - AK1 - abl - TSC1 - ABO	1
TSC1 - MCOA12 - AK1 - abl - ABO	$4.8 \cdot 10^4$
MCOA12 - TSC1 - AK1 - abl - ABO	199
MCOA12 - AK1 - TSC1 - abl - ABO	1.32
MCOA12 - AK1 - abl - ABO - TSC1	2.95
MCT128 - CJ52 - L7 - PBGD - TSC2	1
TSC2 - MCT128 - CJ52 - L7 - PBGD	1.22
MCT128 - TSC2 - CJ52 - L7 - PBGD	$9.3 \cdot 10^9$
MCT128 - CJ52 - TSC2 - L7 - PBGD	$1.3 \cdot 10^8$
MCT128 - CJ52 - L7 - TSC2 - PBGD	$2.5 \cdot 10^3$

**) Given at the top of each list.*

#) The odds are calculated using the imaginary chromosome approach.

Exclusion of relatives without symptoms in additional analyses did not alter the outcome. Therefore the assumed full penetrance did not influence the model to a large extent (results not shown). Furthermore, the maps of chromosome 9 and chromosome 11 markers are rigorously tested by several groups (Lathrop et al., 1988; Charmley et al., 1990). Therefore analysis of the data under the assumption of further locus heterogeneity or the analysis of more distant chromosome 11 markers are the best candidate strategies for improving the model.

Comparison with results obtained using alternative methods

Although locus heterogeneity may not be an uncommon feature, only a few examples are known. So far it has not been possible to test methods for heterogeneity analysis on a large scale. Therefore, it is wise to use multiple methods in each study, in order to avoid biases. Povey and coworkers used two alternative methods in the analysis of the same family material (Povey et al., 1991). Both methods start with a separation of families, followed by a conventional two-point analysis.

The first approach is called the Negative Lod Score Method (NLSM) (Povey et al., 1991). The method only uses linkage information from a candidate region if the alternative region is associated with negative lod scores (Edwards, 1990). This analysis is followed by conventional two-point analysis on the selected families. This way the problem is simplified into two segments for which homogeneity is assumed. No linkage data are used twice.

The second method is called the Maximum Lod Score Method (MLSM) (Povey et al., 1991) It uses the maximum lod scores from both candidate regions to classify the families (N. Morton, personal communication).

Part of the results obtained by these methods are included in Table 3.2. The results on the chromosome 11 localization support a TSC2 locus proximal of MCT128 (Povey et al., 1991). A position between 2-7-1D6 and L424 seems likely. There seems to be an inconsistency within the results on the localization of the TSC1 locus. The NLSM supports a locus between ABO and EFD126.3 which is in contrast with the position near abl as supported by the MLSM (Povey et al., 1991) and ICA. This discrepancy is not more than a seeming discrepancy since none of the resulting locus orders are inconsistent with the odds against these orders as given in Table 3.3.

All approaches used have shown their value. The imaginary chromosome approach has increased power in heterogeneity analysis when compared to the "classic" admixture tests. The imaginary chromosome approach seems to be more refined and less biased, since data from all of the families are used to identify the two most likely locations, without prior selection. However, misleading results could be generated by using an incorrect model. For instance, if one of the putative locations was inaccurate, then the probabilities of linkage assigned to each family would be incorrect, and the associated lod scores would be low. Because of these potential problems, the use of an alternative method such as the NLSM or the MLSM to verify the results is also recommended.

Concluding remarks

Our studies revealed evidence for a model with two different loci independently causing TSC. The first locus (TSC1) maps on chromosome 9 between the abl oncogene and the ABO blood group locus. Several methods confirmed a position about 6 cM proximal to ABO. The second locus TSC2 maps on chromosome 11q22-23. The exact position of the TSC2 locus remains unclear. Our current studies aim at finding closer linked markers to both TSC1 and TSC2 loci. On chromosome 11 a set of informative markers that span a wider range is under investigation. On chromosome 9 our studies focus on new markers, closely linked to abl and ABO.

Acknowledgements

We wish to thank J.A. Maat-Kievit, R.C.M. Hennekam, G.C. Grosveld, H. Stroink, A.P. Oranje, M. Wildervanck de Blecourt-Devilee, B. Prahl-Andersen, H.L.J. Tanghe, N. Morton, J.H. Edwards, J. Yates, M.W. Burley, J. Osborne, R.F. Mueller, L.J. Al-Gazali, M. Super, A.E. Fryer, M. Kuit, M.F. Niermeijer and H. Galjaard (Director of the Clinical Genetics Foundation, Rotterdam) for their assistance and support.

References

Charmley P, Foroud T, Wei S, et al. A primary linkage map of the human chromosome 11q22-23 region. Genomics 1990;6:316-323.

Clark RD, Smith M, Pandolfo M, et al. Tuberous sclerosis in a lifeborn infant with trisomy due to t(11q23.3;22q11.2) translocation: Is neural cell adhesion molecule a candidate gene for tuberous sclerosis? Am J Hum Genet 1988;43:44a.

Connor JM, Yates JRW, Mann L, et al. Tuberous sclerosis: Analysis of linkage to red cell and plasma protein markers. Cyt Cell Genet 1987;44:63-64.

Connor JM, Pirrit LA, Yates JRW, et al. Linkage of the tuberous sclerosis locus to a DNA polymorphism detected by v-abl. J Med Genet 1987;24:544-549.

Connor JM, Loughlin SAR, Whittle MJ. First trimester prenatal exclusion of tuberous sclerosis. Lancet 1987;i:1269.

Edwards JH. The linkage detection problem. Ann Hum Genet 1990;54:253-275.

Fryer AE, Chalmers A, Connor JM, et al. Evidence that the gene for tuberous sclerosis is on chromosome 9. Lancet 1987;i:659-661.

Griffiths LR, Zwi MB, McLeod JG, et al. Heterogeneity evidence and linkage studies on Charcot-Marie-Tooth disease. Neurology 1989;39:280-281.

Haines JL, Amos J, Attwood J, et al. Linkage heterogeneity in tuberous sclerosis. Cyt Cell Genet 1989;51:1010.

Haines JL, Amos J, Attwood J, et al. Genetic heterogeneity in tuberous sclerosis. Study of a large collaborative dataset. Ann NY Acad Sci 1991; 615:256-264.

Janssen LAJ, Sandkuyl LA, Merkens EC, et al. Genetic heterogeneity in tuberous sclerosis. Genomics 1990;8:237-242.

Lander ES, Bottstein D. Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc Natl Acad Sci USA 1986;83:7353-7357.

Kandt RS, Pericak-Vance MA, Hung W-Y, et al. Absence of linkage of ABO blood group locus to familial tuberous sclerosis. Exp Neurol 1989;104:223-228.

Lathrop GM, Lalouel JM, Julier C, Ott J. Strategies for multilocus analysis in humans. Proc Natl Acad Sci USA. 1984;81:3443-3446.

Lathrop GM, Nakamura Y, O'Connell P, et al. A mapped set of genetic markers for human chromosome 9. Genomics 1988;3:361-366.

Northrup H, Beaudet AL, O'Brien WE, et al. Linkage of tuberous sclerosis to ABO blood group. Lancet 1987;ii:804-805.

Ott J, Mensink EJBM, Thompson A, et al. Heterogeneity in the map distance between X-linked agammaglobulinemia and a map of nine RFLP loci. Hum Genet 1986;27:280-283.

Ott JS, Bhattacharya S, Chen JD, et al. Localizing multiple X chromosome-linked retinitis pigmentosa loci using multilocus homogeneity tests. Proc Natl Acad Sci USA 1990;87:701-704.

Povey S, Attwood J, Janssen LAJ, et al. An attempt to map two genes for tuberous sclerosis using novel two-point methods. Ann NY Acad Sci 1991;615:298-305.

Renwick JH. Tuberous sclerosis and ABO. 1987;ii:1096-1097.

Sampson JR, Yates JRW, Pirrit LA, et al. Evidence for genetic heterogeneity in tuberous sclerosis. J Med Genet 1989;26:511-516.

Smith CAB. Testing for heterogeneity of recombination fraction values in human genetics. Ann Hum Genet 1963;27:175-182.

Smith M, Smalley S, Cantor R, et al. Mapping of a gene determining tuberous sclerosis to human chromosome 11q14-11q23. Genomics 1990;6:105-114.

Vance JM, Nicholson GA, Yamaoka LH, et al. Linkage of Charcot-Marie-Tooth type 1a to chromosome 17. Exp Neurol 1989;104:186-189.

Note added in 1994:
Since the submission for publication of this paper in 1990, evidence for linkage to chromosome 11 has weakened gradually. To date the findings are most often interpreted as a type I error (false evidence for linkage).

3.2

LINKAGE INVESTIGATION OF THREE PUTATIVE TUBEROUS SCLEROSIS DETERMINING LOCI ON CHROMOSOMES 9q, 11q and 12q

J.R. Sampson
L.A.J. Janssen
L.A. Sandkuijl
and the
Tuberous Sclerosis Collaborative group

J Med Genet 1992; 29: 861-866.

Abstract

Previous linkage studies in tuberous sclerosis have implicated three disease determining loci, at 9q, 11q and 12q. We have collated phenotypic and genotypic data on 1622 members of 128 families with TSC in order to simultaneously evaluate the evidence for these putative loci. Affection status in the family members has been reassessed using uniform diagnostic criteria and genotypic data extensively checked prior to analysis under alternative models of locus heterogeneity. One TSC determining locus, accounting for approximately 50% of the families studied, has been found to map in the region of D9S10 on 9q34 but no evidence has been found to support the existence of major loci on 11q or 12q. A locus, or loci, elsewhere in the genome is likely to account for TSC in most non-9 linked families.

Introduction

Genetic linkage studies in tuberous sclerosis (TSC) have been hampered by a relative paucity of large families, by variable expression and occasional non-penetrance of the causative genes and, in particular, by locus heterogeneity. A combination of these factors is likely to account for the apparently discrepant conclusions reached by a number of previous linkage investigations (Fryer et al., 1987; Smith et al., 1990; Kandt et al., 1989; Fahsold et al., 1991a).

Linkage between TSC and the ABO blood group locus on distal 9q was first indicated by a study of 19 UK families generating a lod score of 3.85 at $\Theta = 0$ (Fryer et al., 1987). In one family uninformative for ABO a lod score of 1.2 at $\Theta = 0$ was obtained with AK1, which maps to the same region. Further support for this assignment came from study of an RFLP at the Abelson oncogene locus (ABL) in a subset of these families. A lod score of 3.18 at $\Theta = 0$ was observed (Connor et al., 1987). In contrast, two later multifamily studies obtained negative scores between TSC and markers from 9q34 (Kandt et al., 1989; Smih et al., 1987). The authors of the latter studies suggested non-linkage or locus heterogeneity as possible explanations for the discrepancy between their own findings and those of the earlier studies, but did not obtain significant evidence for locus heterogeneity. Soon two family types, those linked and those not linked to ABL, were clearly documented and D9S16 and D9S10 were identified as probable flanking markers for the ABL linked locus, TSC1 (Sampson et al., 1989). Assignment of the non-ABL linked TSC gene to 11q14-23 was proposed following a study of 15 families which generated positive lod scores with 5 markers from 11q (Smith et al., 1990). Assuming 90% penetrance, lods of 3.26 ($\Theta = 0.08$) and 2.88 ($\Theta = 0$) were obtained with MCT128.1 (D11S144) and TYR respectively. Analysis with the USERM9 subprogram from MENDEL (Lange and Boehnke, 1983) revealed no evidence of locus heterogeneity and the lods were calculated assuming homogeneity.

Janssen et al. (1990) used an alternative approach (the "imaginary chromosome method") for synchronous evaluation of the putative TSC loci on 9q and 11q. Nine families were studied with multiple markers from both chromosomal regions. Two TSC loci, one in each region, gave a significantly better fit than one locus alone ($P=0.044$) and maximum likelihood estimates for the positions of TSC1 and TSC2 were just proximal to ABL and at L7 (D11S29) respectively. No evidence for a family type unlinked to either region was found.

Following the 10th International Human Gene Mapping Conference a consortium of groups undertaking linkage studies in TSC was formed. Linkage data for markers on 9q and 11q were shared and analysed using alternative applications of the HOMOG programs (Haines et al., 1991a; Chapter 3.1) and novel two-point approaches (Povey et al., 1991). All analyses reached similar conclusions; a TSC locus could clearly be demonstrated on 9q and the evidence for locus heterogeneity was overwhelming. Positioning of TSC1 in relation to 9q34 markers was imprecise reflecting uncertainty in classification of families as TSC1 or TSC2 type. Localisation of TSC1 between ORM and D9S14 was consistent with the findings of the alternative analytical approaches. Evidence for a locus on chromosome 11 was considered to be strong, although flanking markers could not be identified.

A more recent investigation of 22 families with multiple markers from 9q and 11q confirmed locus heterogeneity with one locus (TSC1) on distal 9q (Haines et al., 1991b). Recombination events in families likely to be 9-linked suggested localisation of TSC1 to the interval ASS - D9S7. No evidence for a locus on 11q could be established in the data as a whole or in a subset of families unlikely to be 9-linked.

The report of a third putative TSC locus on 12q, identified through a disease associated translocation (Fahsold et al., 1991b) and linked to PAH (Fahsold et al., 1991a) was therefore of considerable interest. In 15 families a lod score of 4.33 at $\Theta=0$ was observed between TSC and PAH. HOMOG analysis using multipoint lod scores derived with D12S7, S8 and PAH revealed no evidence of heterogeneity and cumulative lods observed with markers from 9q and 11q were substantially negative.

Previous linkage studies have therefore implicated three loci in TSC, but have not evaluated these together. We set out to assess the evidence for the putative loci using a large and rigorously checked data set. This was collected from the centres constituting the Tuberous Sclerosis Collaborative Group and included the data subsets on which the previously suggested assignments were based.

Materials and methods

Both genotypic and phenotypic data were available for 1622 members of 128 multiplex TSC families. Much of the family material has been documented in previous publications and it is considered in more detail in Chapter 2.2.

Phenotypic Data

Clinical and radiological information which could be used to classify affection status was collected on all family members. Modified linkage format pedigree files were used for this purpose. Age at assessment and presence, absence or uncertainty with regard to the following diagnostic criteria were recorded: adenoma sebaceum, peri-ungual fibromata, shagreen patches, hypomelanotic macules, fibrous forehead plaques, retinal phakomata, typical peri-ventricular calcification on brain CT scan or MRI scan, cortical tubers on MRI scan or at post mortem, and renal angiomyolipomata or cysts detected radiologically or at PM. In addition, information on seizures, mental retardation, brain tumours and any unusual features was recorded.

The data was reviewed by one of us (JRS) and used to classify each individual into one of four liability classes: AFFECTED, where a definitive diagnosis could be made according to the criteria of Gomez, UNKNOWN 1 where no information was available, where incomplete but normal clinical information was available or where findings of uncertain clinical significance had been made (eg solitary renal cyst, seizures or mental retardation but no other positive signs), UNKNOWN 2 where complete clinical examination proved normal but where radiological workup was incomplete and NORMAL where full clinical and radiological workup including Wood's light examination, ophthalmoscopy, brain CT or MRI scan and renal CT or ultrasound scan revealed no evidence of TSC in a family member over 16 years of age.

Liability classes corresponding to a penetrance of 95%, a phenocopy rate of 2% and a conservative estimate of clinical gene expression (ie disease apparent without radiological investigation) of 67% were used in the analysis. The liability classes were selected to minimise mapping errors consequent on misinterpretation of clinical or radiological signs.

Genotypic Data

A modified version of MLINK (provided by Sandkuijl) was used to identify definite and probable double recombinants by examining all possible combinations of 3 markers in the same chromosomal region in each family. Those with probable (>10:1) double recombinants were retyped (in each case with resolution to a zero or single recombinant solution) or excluded from the analysis. Recombination fractions were calculated for pairs of markers in each chromosomal region for each

group and these were found to be consistent with published data and did not show significant intergroup variation (data not shown).
A database of available marker data was compiled and marker loci were selected for inclusion in the analysis on the basis of three criteria: adequate coverage of the proposed candidate regions, comprehensive typing by the contributing groups and high informativity. Where a family was typed for two or more polymorphic systems at a single marker locus the most informative system in that family was selected.

The marker loci used in the analysis and their map positions, taken from published data (Lathrop et al., 1988; Kwiatkowski et al., 1992; Povey et al., 1992; Julier et al., 1990; Sanal et al., 1990; Junien et al., 1991; O'Connell et al., 1987; Craig and McBride, 1991) are shown in Figure 3.2.

```
Chromosome 9:
Cen-GSN-------AK1-----ASS-ABL------S10------S14---------S7*-Tel
        0.09       0.05  0.01  0.06      0.06       0.14

Chromosome 11:
Cen-S84-------S35-----------S144-S351---S29----PBGD----S147-Tel
        0.09         0.15        0     0.04  0.04     0.04

Chromosome 12
Cen-S8---------------S7---------------PAH-Tel
            0.19             0.16
```

Figure 3.2: Genetic maps of the chromosomal regions analysed. Genetic distances expressed as sex averaged Θ.
**) D9S7 data were omitted from the analysis.*

Allele frequencies were calculated (from founder members and spouses only) for the data of each group and for the data as a whole. They did not differ significantly from published frequencies (Williamson et al., 1991) which were used in linkage analysis. Dinucleotide repeat markers were recoded to 4/5 alleles without loss of linkage information.

Linkage Analysis
For each family MLINK was used to calculate multipoint lod scores at specified

positions across maps of the three candidate regions. There were 45, 40 and 43 positions at 1 cM intervals between the markers on chromosomes 9, 11 and 12 respectively. For each family these positions were defined by calculating the distances to and from the nearest informative proximal and distal flanking markers using Haldane's mapping function (Haldane, 1919). Therefore, markers uninformative in a given family were not included in the analysis. Lod scores were also calculated at four "off map" positions at $\Theta = 0.1$, 0.2, 0.3 and 0.4 either side of each region. Interference was assumed to be absent and male and female recombination frequencies equal. The mutation rate was set at 2.5×10^{-5} and the disease gene frequency at 1×10^{-4}.

In summary, 1 multipoint run was therefore used to calculate the lod score at 1 position in 1 family. Arrays of lod scores calculated at identical positions independent of the pattern of informativity formed the input files for heterogeneity tests using the HOMOG programs.

Because the large number of haplotypes generated in highly informative families exceeded the capacity of PCs lod scores were computed, where necessary, on a DEC 5830 and on Sun workstations using several overlapping multipoint analyses as described by Ott (Ott, 1991).

Heterogeneity Analysis

Sets of lod scores for different chromosomal regions can be analysed synchronously using the HOMOG programs (Ott, 1983), by their combination into a single array. This method has been termed the "imaginary chromosome approach" (Janssen et al., 1990; Chapter 2.1)) and is suitable for simultaneous mapping of 2 or more loci.

The programs HOMOG, HOMOG2 and HOMOG 3 were used to evaluate the evidence for locus heterogeneity and for 1,2 or 3 TSC loci in the tested regions. Analyses were performed for all combinations of chromosomal regions for the data as a whole and the subsets of data contributed by each centre were also submitted to a more restricted series of analyses.

The HOMOG programs assume equal penetrance, mutation rate and gene frequency for the different disease determining loci.

Results

Informativity in the Tested Regions

The number of families informative at each locus analysed and the number of families informative for each chromosomal region as a whole are indicated in Table 3.4.

Evidence for Heterogeneity and for TSC Loci on Chromosomes 9, 11 and 12

The relative likelihoods for models of heterogeneity including TSC loci in all possible combinations of the chromosome 9, 11 and 12 regions are summarised in Figure 3.3.

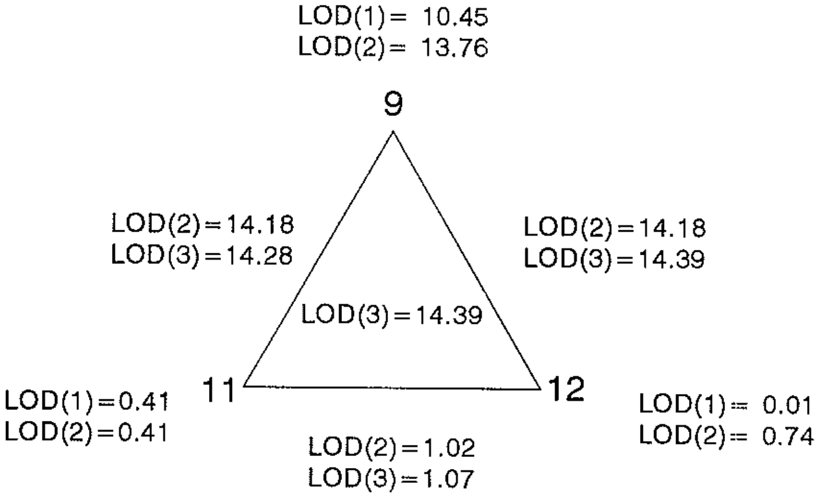


Figure 3.3: Scheme showing likelihoods for models of heterogeneity incorporating TSC loci on all combinations of chromosomes 9, 11, and 12. Refer to text for details.

Lod scores at the points of the triangle were calculated using data for the indicated single chromosomal region alone. At the points of the triangle Lod(1) indicates the logarithm of the odds computed under homogeneity and Lod(2) that computed under heterogeneity using the HOMOG program (i.e. assuming that there is a proportion of unlinked families in which the disease locus maps to an unspecified alternative region). There is significant evidence for a TSC locus on 9, but not on 11 or 12. The chromosome 9 data clearly supports heterogeneity (9 and non-9 only model), Lod(2) being 3.31 lod units above Lod(1).

Table 3.4:
Number of informative families at each marker locus and in each chromosomal region (totals)

Chromosome 9		Chromosome 11		Chromosome 12	
Marker	Families	Marker	Families	Marker	Families
GSN	38	S84	38	S8	34
AK1	28	S35	60	S7	62
ASS	103	S144	73	PAH	49
ABL	72	S351	62		
S10	79	S29	33		
S14	21	PBGD	47		
		S147	41		
Total	106		101		62

Lod scores shown along the sides of the triangle were computed using the data for the corresponding 2 chromosomal regions. These were calculated under heterogeneity allowing for 2 and 3 loci (Lod(2) and Lod(3)) using the HOMOG2 and HOMOG3 programs respectively. HOMOG2 allows for the presence of two loci on the map. Alternatively one of the loci may be placed "off map" (that is one of the loci is unlinked to either region) and here the analysis will yield results identical to those of the HOMOG program. Only lods calculated from data including chromosome 9 are significant and the data for 11 and 12 do not significantly improve the lod score over that for the 9 and non-9 only model. Minor differences in lod scores between analyses not reaching significance may reflect variation in Θ values associated with peak lod scores in unlinked families.

Analysing all three chromosomal regions synchronously using HOMOG3 the lod (indicated as Lod(3) inside the triangle) is not significantly higher than that for the 9 non-9 only model.

In no analysis was a locus positioned on the chromosome 11 map. Analyses of the chromosome 12 data alone, the combined chromosome 11 and 12 data, and the HOMOG3 analysis of the 9, 11 and 12 data, while not reaching siginificance over the 9 non-9 only model did place one locus between D12S7 and PAH. The corresponding α values were 10 - 20%.

Evaluation of the data from the individual contributing centres (analyses not shown) showed that the data from all except Erlangen were consistent with the conclusions drawn from the data set as a whole. Interestingly, data sets previously found to support a locus on chromosome 11 did not do so in this analysis, but did support the chromosome 9 locus. This possibly reflects gain of information from recent marker studies, or loss of information due to our restrictive diagnostic classification. On analysis of all families the main evidence for a locus on chromosome 12 (which did not reach significance) came from the Erlangen families. When analysed separately the Erlangen data favoured a locus on chromosome 12 (at D12S7) in 95% of families (lod score = 2.88). When analysed with the locations and α values obtained in the combined analysis, we observed that the Erlangen families contributed only 0.88 to the overall Lod(3) score.

In summary, there is significant evidence for a TSC locus in the tested region of chromosome 9 (TSC1), but not in the tested regions of 11 or 12. Although locus heterogeneity exists there is no evidence that a major non-9 locus maps to the other regions studied.

The Proportion of Families Segregating for a TSC1 Mutation

In the simplest model of heterogeneity reaching significance, 9 and non-9 only, the maximum likelihood estimate of α (proportion of 9-linked families) is 51%. In the other models of heterogeneity reaching significance the maximum likelihood estimates of α varied between 45% and 55%

Localisation of TSC1

All HOMOG and HOMOG2 analyses incorporating the chromosome 9 data identify the maximum likelihood position for TSC1 as 1cM proximal to D9S10. Two of three HOMOG3 analyses also identify this position and one (9 and 11 data only) identifies D9S10.

The POINT4 program was used to estimate the support region for TSC1 based on a difference of 1 unit of $\log_{10}$likelihood from maximum under all sets of parameter values (α_1, α_2, and disease gene locations, x_1, x_2 and x_3). This identified a region extending from 1cM distal to ABL to 4 cM distal to D9S10 (Figure 3.2).

Under analysis of locus heterogeneity only one combination of α and map location

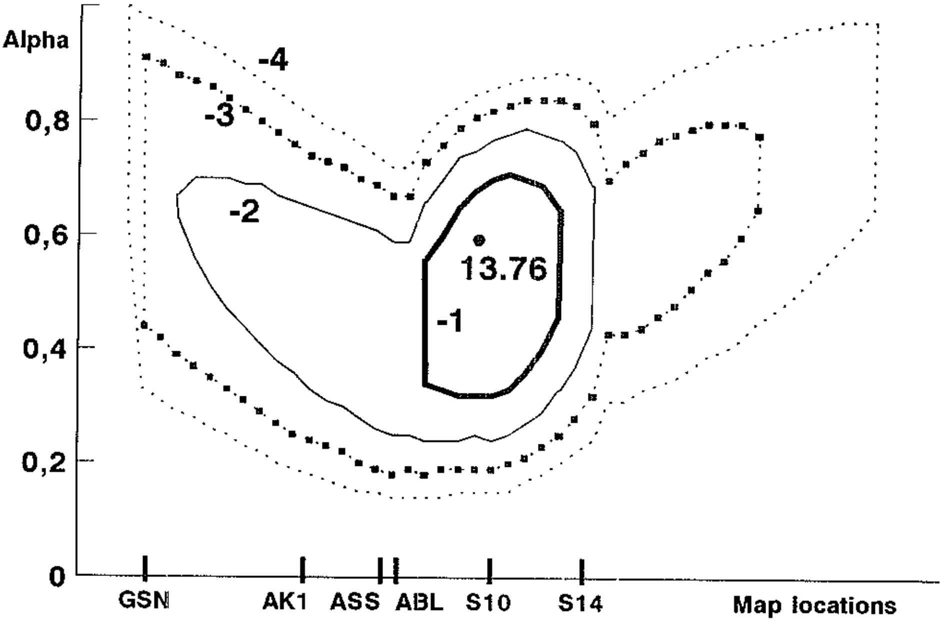


Figure 3.4: Possible alpha values (proportion of chromosome 9 linked families) and map locations for TSC1 corresponding to Z_{max} -1, -2, -3, and -4 lod units.

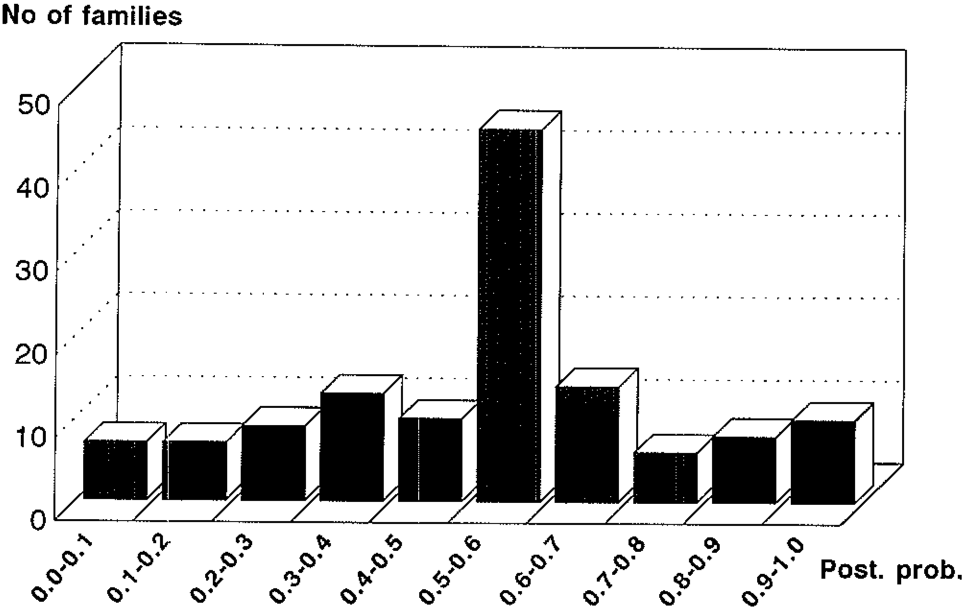


Figure 3.5: Distribution of posterior probabilities for families being TSC1 type.

(or Θ) will typically be associated with the maximum likelihood solution. However, various combinations of α and map location will be associated with equal, but lower, likelihoods. Figure 3.4 shows the point of maximum likelihood (Z_{max} = 13.76) and contours corresponding to (Z_{max} - 1), ((Z_{max}- 2) and (Z_{max}-3) lod units obtained on analysis of the chromosome 9 data for all families.

Posterior Probabilities of 9-Linkage, by Family

Figure 3.5 shows the distribution of posterior probabilities for families segregating a TSC1 mutation as determined by HOMOG analysis of the data for chromosome 9; that is with TSC1 positioned 1 cM proximal to D9S10.

Discussion

Linkage analysis has proved to be a successful strategy for mapping disease genes in situations where the majority, or even all individuals with the disease phenotype have mutations at a single locus (Rommens et al., 1989; Brook et al., 1992). However, many genetic diseases are likely to reflect dysfunction 2of heterodimeric proteins, acceptor-receptor interaction or interruption of a metabolic pathway at one of a number of intermediary steps. Even mutations affecting entirely separate cellular or physiological systems may generate similar phenotypes. Locus heterogeneity, and its implications for linkage approaches, should therefore be anticipated. When large families are available for 1 or more of the contributing loci (Chance et al., 1990), or when mutations at one of the loci are responsible for the majority of familial cases (Peters and Sandkuijl, 1992), it will generally be possible to accurately map such a locus. In contrast, when only small families are available, or when several loci exist, each being responsible for a small proportion of families, localisation of the respective loci via linkage analysis will be virtually impossible. In Chapter 3.2 we evaluate in more detail the average family size in the current data set, and its suitability for localisation of multiple TSC loci.

We have used an approach enabling evaluation of three putative TSC determining loci, on 9q, 11q and 12q, in a single analysis. The possible existence of loci in these regions was supported by previously published investigations obtaining lod scores reaching conventional levels for significance (Fryer et al., 1987; Smith et al., 1990; Fahsold et al., 1991a). Larger collaborative analyses had apparently confirmed the existence of the chromosome 9 locus (TSC1) and provided strong but not incontrovertible evidence for a second locus on 11q (Haines et al., 1991a; Povey et al., 1991; Chapter 3.1). We chose to reevaluate all clinical information and ignore all previously ascribed affection statuses. This resulted in changes in many families. We also sought to minimise typing errors by assessing marker-marker estimates of Θ in the data of each contributing group, and by identifying and eliminating probable double recombinants.

Our analysis confirms the assignment of one gene determining TSC to 9q34, and indicates its most probable location as just proximal to D9S10. Mutations in this gene account for an estimated 45-55% of the 128 families studied. There is minimal support for the putative locus on 12q and none for the putative locus on 11q. While the possibility of very rare loci in any region can never be ruled out, it is clear that a locus outside the tested chromosomal regions accounts for the majority of non- 9 linked families. Even the subsets of data which previously implicated loci at 11q and 12q did not give significant evidence to support their existence in this analysis. None of the analyses placed a TSC gene on the chromosome 11 map in any proportion of families and this situation was not changed by analysis of a more proximal subset of markers, TYR-D11S84 (Janssen, unpublished results). Some analyses mapped a small proportion of families to a position between D12S7 and PAH, but the likelihoods obtained were not significant over an "off-map" localisation for all non-9 families. The possibility of a rare chromosome 12 locus might usefully be clarified by improved clinical evaluation of the very few families hinting at linkage in this data set.

Localisation of the major non-9 TSC gene will assist with linkage resolution of TSC1. Haplotype analysis in the few large 9-linked families available and the identification of key recombinant individuals should be particualrly helpful in this regard. Even with the mapping of TSC1 and TSC2 to intervals of a few cM, the combination of locus heterogeneity and small mean family size will severely limit the application of linkage approaches in the diagnostic setting. For the majority of families this must await the isolation and characterisation of the TSC genes.

Footnote

It is not possible to provide lod scores for all families individually in a paper of this kind, but those interested may approach J.R. Sampson or L.A.J. Janssen for a compter disk containing these.

Acknowledgements

Finanical support for this study was provided by: The Tuberous Sclerosis Association (GB), The Prevention Fund (Netherlands), The National Tuberous Sclerosis Association (USA), Action Research, the MRC, the Welcome Trust, the American Cancer Society (Grant PDT-435), NIH, the National Institute of Childhood Diseases (Grant HD 23745) and NINDS (5P01 NS26630-34).

We are also grateful to E Andermann, J Attwood, K Handa, V Berdoukas and members of the Tuberous Sclerosis Working Group (Rotterdam).

References

Brook JD, McCurrach ME, Harley HG, et al. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. Cell 1992;88:799-808

Chance PF, Bird TD, O'Connell P, et al. Genetic linkage and heterogeneity in type 1 Charcot-Marie-Tooth disease (hereditary motor and sensory neuropathy type 1). Am J Hum Genet 1990;47:915-925

Connor JM, Pirrit LA, Yates JRW, et al. Linkage of the tuberous sclerosis locus to a DNA polymorphism detected by v-abl. J Med Genet 1987;24:544-546

Craig IW and McBride OW. Report on the committee on the genetic constitution of chromosome 12. Cytogenet Cell Genet 1991;58:555-579

Fahsold R, Rott H-D, Lorenz P. A third gene locus for tuberous sclerosis is closely linked to the phenylalanine hydroxylase locus. Hum Genet 1991a;88:85-90

Fahsold R, Rott H-D, Claussen U, Schmalenberger B. Tuberous sclerosis in a child with de-novo translocation t(3;12) (p26.3;q23.3). Clin Genet 1991b;40:326-328

Fryer AE, Chalmers A, Connor JM, et al. Evidence that the gene for tuberous sclerosis is on chromosome 9. Lancet 1987;i:659-661

Haines J, Amos J, Attwood J, et al. Genetic Heterogeneity in tuberous sclerosis: study of a large collaborative dataset. Ann N Y Acad Sci 1991a;615:256-264

Haines JL, Short MP, Kwiatkowski DK, Localisation of one gene for Tuberous Sclerosis within 9q32-9q34, and further evidence for heterogeneity. Am.J.Hum.Genet 1991b;49:764-722.

Haldane JBS. The combination of linkage values and calculation of distances between the loci of linked factors. J Genet 1919;8:299-309

Janssen LAJ, Sandkuyl LA, Merkens EC, et al. Genetic heterogeneity in Tuberous Sclerosis. Genomics 1990;8:237-242

Julier C, Nakamura Y, Lathrop M, et al. A detailed genetic map of the long arm of chromosome 11. Genomics 1990;7:335-345

Junien C and van Heyningen V. Report of the committee on the genetic constitution of chromosome 11. Cytogenet Cell Genet 1991;58:459-554

Kandt RS, Pericak-Vance MA, Hung W-Y, et al. Absence of linkage of ABO blood group locus to familial tuberous sclerosis. Exp Neurol 1989;104:223-228

Kwiatkowski DJ, Henske EP, Weimer K, et al. Construction of a GT polymorphism Map of Human 9q. Genomics 1992;12:229-240

Lange K and Boehnke M. Extensions to pedigree analysis. V. Optimal calculations of Mendelian likelihoods. Hum Hered 1983;33:291-301

Lathrop GM, Nakamura Y, O'Connell P, et al. A mapped set of genetic markers for human chromosome 9. Genomics 1988;3:361-366

O'Connell P, Lathrop GM, Law M, et al. A primary linkage map for human chromosome 12. Genomics 1987;1:93-102

Ott J. Linkage analysis and family classification under heterogeneity. Ann Hum Genet 1983;47:311-320

Ott J. Analysis of human genetic linkage (Revised Edition). The Johns Hopkins University Press, Baltimore and London, 1991.

Peters DJM and Sandkuijl LA. Genetic heterogeneity of polycystic kidney disease in Europe. Contrib Nephrol 1992;97:128-139

Povey S, Attwood J, Janssen LAJ, et al. An attempt to map two genes for tuberous sclerosis using novel two-point methods. Ann N Y Acad Sci 1991;615:298-305

Povey S, Smith M, Haines J, et al. Report on the First International Workshop on Chromosome 9, held at Girton College Cambridge, UK, March 22-24, 1992. Ann Hum Genet 1992;56:167-221.

Rommens J, Ianuzzi M, Kerem B, et al. Identification of the cystic fibrosis gene: chromosome walking and jumping. Science 1989;245:1059-1066

Sampson JR, Yates YRW, Pirrit LA, et al. Evidence for genetic heterogeneity in tuberous sclerosis. J Med Genet 1989;26:551-516

Sanal O, Wei S, Foroud T, et al. Further mapping of an ataxia-telangiectasia locus to the chromosome 11q23 region. Am J Hum Genet 1990:47:860-866

Smith M, Haines J, Trofatter J, et al. Linkage studies in tuberous sclerosis. Cytogenet Cell Genet 1987;46:694-695

Smith M, Smalley S, Cantor R, et al. Mapping of a gene determining tuberous sclerosis to human chromosome 11q14- q23. Genomics 1990;6:105-114

Williamson R, Bowcock A, Kidd K, et al. Report of the DNA committee and catalogues of cloned and mapped genes, markers formatted for PCR and DNA polymorphisms. Cytogenet Cell Genet 1991;58:1190-1832

REFINED LOCALIZATION OF TSC1 BY COMBINED ANALYSIS OF 9Q34 AND 16P13 DATA IN 14 TUBEROUS SCLEROSIS FAMILIES

Bart Janssen
Julian Sampson
Mieke van der Est
Wout Deelen
Senno Verhoef
Ian Daniels
Arjenne Hesseling
Phillip Brook-Carter
Mark Nellist
Dick Lindhout
Lodewijk Sandkuijl
Dicky Halley

Hum Genet 1994; 94: 437-440.

Abstract

Tuberous sclerosis (TSC) is a heterogeneous trait. Since 1990, linkage studies have yielded putative TSC loci on chromosomes 9, 11, 12 and 16. Our current analysis, performed on 14 Dutch and British families, reveals only evidence for loci on chromosome 9q34 (TSC1) and chromosome 16p13 (TSC2). We have not found any indication for a third locus for TSC, linked or unlinked to either of these chromosomal regions. The majority of our families shows linkage to chromosome 9. We refined the candidate region for TSC1 to a region of approximately 5 cM between ABL and ABO.

Introduction

Tuberous sclerosis (TSC) is an autosomal dominant disorder, characterised by hamartomas that may affect numerous organ systems (for a review see Gomez (1991)). Positional cloning has been hampered by locus heterogeneity. Linkage was initially reported to markers on chromosome 9q34 (Fryer et al., 1987). However, these findings were subsequently disputed, until significance for locus heterogeneity was demonstrated and the existence of a chromosome 9 linked locus (TSC1) was confirmed (Haines et al., 1991a,b; Chapter 3.1; Northrup et al., 1992).

In the period 1990-1992 linkage studies in TSC were dominated by two main strategies: the development of more efficient methods for heterogeneity analysis, and the search for other loci responsible for the disease. Several methods for linkage analysis under heterogeneity have been used. The HOMOG programs (Ott, 1991), implementing the admixture test (A-test), have been used successfully by several research groups (Haines et al., 1991a,b; Northrup et al., 1992). A more advanced application of the A-test, which we designated the "Imaginary Chromosome Approach" (ICA) (Janssen et al., 1990), involves the synchronous analysis of multipoint linkage data from multiple candidate regions. The A-test based approaches have proven quite powerful and provide an efficient tool for the assignment of the gene defect in each family to one of the various loci (Chapter 2.2). Alternative, quite transparent, techniques that were not based on the A-test, supported the findings obtained with A-test based approaches (Povey et al., 1991).

Additional chromosomal locations for TSC have been sought since the demonstration of locus heterogeneity. A second locus was provisionally assigned to chromosome 11 (Smith et al., 1990), but subsequent heterogeneity analyses failed to position a locus within the predicted region on 11q14-23 (Haines et al., 1991a,b; Chapter 3.1; Povey et al., 1991) or attained marginal significance levels (Janssen et al., 1990). A de novo translocation t(3;12) in a TSC patient led to the provisional assignment of a third locus to chromosome 12, through linkage analysis in 15 German families (Fahsold et al., 1991). However, a large collaborative heterogeneity analysis, involving data from all three regions 9q34, 11q14-23 and 12q23.3 did not provide evidence for either a locus on chromosome 11 or a locus on chromosome 12 (Chapter 3.2). In 1992 Kandt et al. reported linkage to markers on the tip of chromosome 16p in 5 large non-9 linked TSC families. Apart from the incontrovertible evidence for a locus on 16p13.3, the authors also showed a clear lack of evidence for loci on chromosome 11 or 12 in these families. However, since only non-9 linked families were studied, no comment on the importance, or even the existence, of a locus on 9q34 could be made.

We planned a heterogeneity analysis, methodologically similar to previous studies

(Chapter 3.2), utilizing data from 9q34 and 16p13.3, with the aims of gaining better insight into the TSC heterogeneity problem and of achieving more precise map positions for each locus.

Materials and methods

We selected 14 large families from a mixed group of 24 nuclear families and extended pedigrees from Cardiff and Rotterdam. Previously, simulation studies had been performed in order to determine the power of each family (Chapter 2.2). Families that never showed two or more informative meioses were left out of the present analysis, because this type of families can not contribute significantly, unless an extremely large number of families is available. 13 of the 14 selected families had already been included in previous analyses. The only changes with regard to family structures or affection status was extension of family 1013 and clarification of previously unknown statuses in this family. One new 5-generation family (family 4210) has been added. We used the same genetic parameters (corresponding to a penetrance of 95% and a phenocopy rate of 2%) as in our previous study (Chapter 3.2). On chromosome 9 we typed markers at ABL (dinucleotide repeat), D9S64 (dinucleotide repeat), ABO and D9S10 (MCT136). Inter marker distances (2, 3 and 2 cM. respectively) were taken from the report of the second chromosome 9 workshop (Kwiatkowski et al., 1993). At ABO the serological typing was utilized, unless molecular typing of the O-allele had been performed. In family 1013 D9S10 was uninformative and therefore replaced by the dinucleotide repeat D9S66, which maps on the same cosmid. On chromosome 16 we typed the markers D16S85 (3'HVR), D16S259 (pGGG1) and the dinucleotide repeats D16S291 (16AC2.5) and D16S283 (SM7). Inter marker distances (6, 1 and 1 cM respectively) were taken from Germino et al. (1992), or inferred from physical distances. On both chromosomes lod scores were calculated at intervals of 1 cM. Outside the inter marker regions lod scores were calculated at several positions from 0% to 50% recombination. Heterogeneity was studied on chromosome 9 and 16 data synchronously by ICA, allowing for two linked loci on the combined chromosomes, or three loci, one of which is unlinked. The data were analyzed by the programs HOMOG2 and POINT4 from the HOMOG package (Ott, 1991).

Results and discussion

Assuming homogeneity, we obtained a maximum cumulative lod score $Z_{(\alpha=1)MAX}$ of 4.14 at 10% recombination (Θ=0.1) from ABL (Table 3.5, Figure 3.6). At each map position (x1) we calculated the maximum lod score $Z_{(\alpha,x1,x2)MAX}$ by optimizing α (proportion of linked families) and the position of the other locus (x2). On chromosome 9 the lod score peaked at D9S64 ($Z_{(\alpha,x1,x2)MAX}$=8.92, α=0.65). By comparison with $Z_{(\alpha=1)MAX}$ we found an odds ratio of $6.0 \cdot 10^4$:1 in favour of heterogeneity. Still assuming heterogeneity we calculated the lod score for TSC1 being unlinked to chromosome 9. This position is associated with a $Z_{(\alpha,x1=\infty,x2)MAX}$=0.73, with α=0.72 and the remainder located at D16S291. ($Z_{(\alpha,x1=\infty,x2)}$ may be unequal to 0 if $\alpha \neq 1$.) Comparison of the two lod scores under heterogeneity revealed substantial evidence for TSC1 being chromosome 9 linked (Z=8.2) with an odds ratio of $1.5 \cdot 10^8$:1.

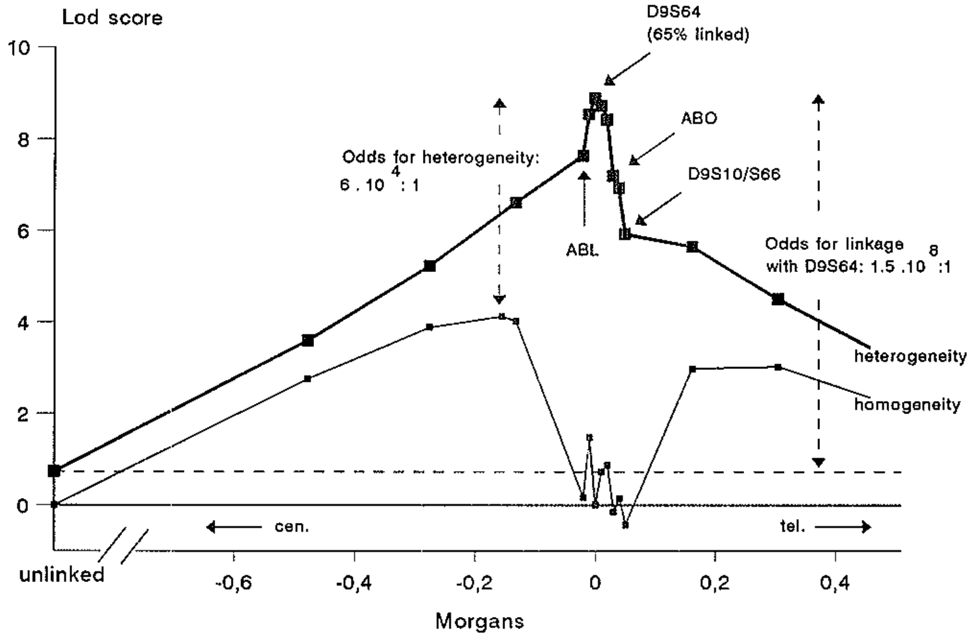


Figure 3.6: Results of multipoint analysis with markers on chromosome 9. The thin line indicates the cumulative lod score under homogeneity. The thick line indicates the lod score under heterogeneity (Z(α,x1,x2)), as derived from the POINT4 program. At each position the Z(α,x1,x2) was calculated by optimizing α and the position of TSC2. Odds for heterogeneity and linkage are depicted with broken arrows (see text).

To further delineate the area a 90% confidence interval was constructed, including all locations with a lod score exceeding $Z_{(\alpha,x1,x2)MAX}$-1. ABL and ABO do not lie within this confidence interval and are therefore likely to be flanking markers encompassing a TSC1 region of 5 cM.

On chromosome 16 (under the assumption of homogeneity) we obtained a maximum lod score $Z_{(\alpha=1)MAX}$ of 0.52. Under heterogeneity the locus mapped at D16S291, with α=0.35 and TSC1 placed at D9S64 (Figure 3.7) ($Z_{(\alpha,x1,x2)MAX}$=8.92 as discussed above). The unlinked position was associated with a $Z_{(\alpha,x1,x2=\infty)MAX}$ of 6.52 at α=0.43 and TSC1 placed at D9S64. This results in an odds ratio in favour of linkage of "TSC2" to chromosome 16 of $2.5{\cdot}10^2$:1 (Z=2.4). This result, although not independently significant according to the very stringent guidelines that we proposed in Chapter 2.1 and 2.2, provides a clear confirmation of Kandt's finding (Kandt et al., 1992), since non of our families were in his study. Due to the limited amount of information in the chromosome 16 linked families, we could not define

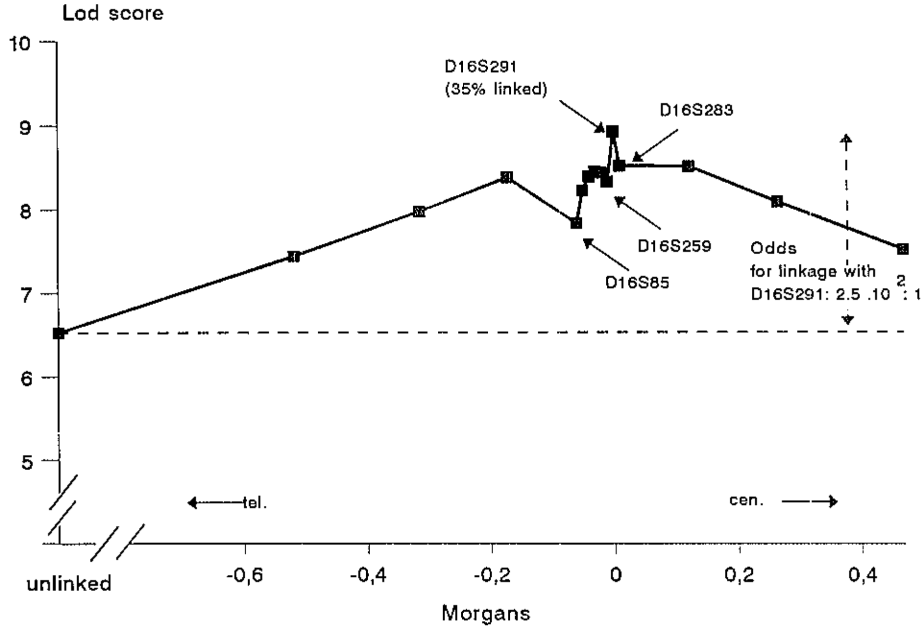


Figure 3.7: Results of multipoint analysis with chromosome 16 markers (as Figure 3.6). Since the lod score under homogeneity peaked at 0.52, these lod scores are not shown.

Table 3.5:
Lod scores and posterior probabilities (HOMOG2) for individual families:

	Max. lod score		Lod score at		Posterior probability of
Fam.	Chr.9	Chr.16	D9S64	D16S291	being 9-linked
4079	0.20	0.86	-1.20	0.86	0.015
4210	0.00	1.14	-3.79	1.13	0.000
4077	1.27	0.26	1.27	-0.41	0.989
4219	1.81	0.00	1.81	-1.74	1.000
4221	0.83	0.05	0.83	-0.27	0.959
4222	0.00	0.00	-0.11	-0.11	0.649
4264	0.72	0.03	0.72	-0.69	0.979
1003	0.00	0.14	-0.59	0.04	0.300
1004	0.00	0.88	-3.32	0.88	0.000
1007	0.00	0.00	-0.15	-0.12	0.636
1011	0.00	0.05	-0.22	-0.41	0.745
1012	0.81	0.00	0.80	-0.26	0.955
1015	0.48	0.00	0.05	-0.52	0.876
1013	3.90	0.00	3.90	-5.43	1.000
total			0.01	-7.05	

a narrow confidence interval for TSC2. The ratio of large and small families was about the same for both the TSC1 and the TSC2 group of families.

In order to examine the possibility of a third trait causing locus we repeated the analysis. Instead of two, three α-values were introduced, where α_1 and α_2 are like before and where α_3 denotes the proportion of families linked to neither region. A maximum lod score was obtained if α_3=0%, indicating that this data set yields no evidence for a third locus at all.

Our results demonstrate the usefulness of ICA. In order to avoid systematic biases we have chosen to evaluate the lod scores under heterogeneity directly, rather than selecting families by their posterior probability of being linked to one of the regions, followed by evaluation of their $Z_{(\alpha=1)}$ values, as other authors have done (Haines et al., 1992a,b; Northrup et al., 1992). If we had used such a method the (inflated) lod score in favour of linkage to 16p13.3 would have been 2.9 instead of 2.4.

During the course of this study the TSC2 gene on chromosome 16 has been isolated (Chapter 5). The gene maps less than 200 kb distal of D16S291. In one of the families (family 4079; posterior probability of being chromosome 16-linked = 0.985) a reduced level of TSC2 transcript was demonstrated in all affected family members (European Chromosome 16 TSC Consortium, 1993), which confirmed the involvement of TSC2 in this family. The excellent agreement between the statistical and the molecular localization of TSC2 is encouraging for a targeted search for TSC1 in the currently favoured area on chromosome 9. It is likely that molecular diagnosis will soon be feasible in large families.

Acknowledgements

We thank all persons involved in the TSC projects in Cardiff and Rotterdam as well as our funding organisations: The Netherlands Organization for Scientific research, Action Research and the Prevention Fund. We are especially thankful to the patients and their organisations: The Tuberous Sclerosis Association (England and Wales) and the Stichting Tubereuze Sclerosis Nederland.

References

Fahsold R, Rott H-D, Lorenz P. A third gene locus for tuberous sclerosis is closely linked to the phenylalanine hydroxylase locus. Hum Genet 1991;88:85-90.

Fryer AE, Chalmers A, Connor JM, et al. Evidence that the gene for tuberous sclerosis is on chromosome 9. Lancet 1987; I: 659-661.

Germino GG, Weinstat-Saslow D, Himmelbrauer H, et al. The gene for autosomal dominant polycystic kidney disease lies in a 750-kb CpG-rich region. Genomics 1992; 13: 144-151.

Gomez MR. Phenotypes of the tuberous sclerosis complex with a revision of diagnostic criteria. Ann NY Acad Sci 1991; 615: 1-7.

Haines J, Amos J, Attwood J, et al. Genetic heterogeneity in tuberous sclerosis: study of a large collaborative dataset. Ann NY Acad Sci 1991a; 615: 256-264.

Haines JL, Short MP, Kwiatkowski DJ, et al. Localization of one gene for tuberous sclerosis within 9q32-9q34, and further evidence for heterogeneity. Am J Hum Genet 1991b; 49: 764-772.

Janssen LAJ, Sandkuyl LA, Merkens EC, et al. Genetic heterogeneity in tuberous sclerosis. Genomics 1990; 8: 237-242.

Kandt RS, Haines JL, Smith M, et al. Linkage of an important gene locus for tuberous sclerosis to a chromosome 16 marker for polycystic kidney disease. Nature Genetics 1992; 2: 37-41.

Kwiatkowski DJ, Armour J, Bale AE, et al. Report on the second international workshop on human chromosome 9. Cytogenet Cell Genet 1993; 64: 94-106.

Northrup H, Kwiatkowski DJ, Roach ES, et al. Evidence for genetic heterogeneity in tuberous sclerosis: one locus on chromosome 9 and at least one locus elsewhere. Am J Hum Genet 1992; 51: 709-720.

Ott J. Analysis of human genetic linkage, rev. edn. 1991. Johns Hopkins University Press, Baltimore London.

Povey S, Attwood J, Janssen LAJ, et al. An attempt to map two genes for tuberous sclerosis using novel two-point methods. Ann NY Acad Sci 1991; 615: 298-305.

Smith M, Smalley S, Cantor R, et al. Mapping of a gene determining tuberous sclerosis to human chromosome 11q14-q23. Genomics 1990; 6: 105-114.

CHAPTER 4

THE SEARCH FOR THE TSC1 GENE

COSMID CONTIGS FROM THE TUBEROUS SCLEROSIS CANDIDATE REGION ON CHROMOSOME 9Q34

M van Slegtenhorst
B Janssen
M Nellist
S Ramlakhan
C Hermans
A Hesseling
A v.d. Ouweland
D Kwiatkowski
B Eussen
J Sampson
P de Jong
D Halley

Eur J Hum Genet (in press)

Abstract

Tuberous Sclerosis (TSC) is a heterogeneous multisystem disorder, with loci on 9q34 (TSC1) and 16p13.3 (TSC2). The TSC2 gene has recently been isolated, while the TSC1 gene has been mapped to a 5 cM region between the markers D9S149 and D9S114. In our effort to localise and clone TSC1, we have obtained three adjacent cosmid contigs that cover the core of the candidate region. The 3 contigs comprise approximately 600 kb and include 80 cosmids, 2 P1 clones, 1 YAC, 5 anonymous markers and 4 STSs. The ABO blood group locus, the Surfeit gene cluster, the dopamine ß hydroxylase gene (DBH) and VAV2, a homologue of the vav oncogene, have all been mapped within the contigs. Exon trapping and mutation screening experiments, aimed at identifying the TSC1 gene, are currently in progess.

Introduction

Tuberous sclerosis (TSC) is an autosomal dominant multisystem disorder. The brain, skin, heart and kidneys are often affected and almost all other tissues and organs may be involved, except muscle syncytia (Gomez, 1988). The disease shows a high penetrance with variable expression and is known for its locus heterogeneity, with one locus mapping to chromosome 9q34 (TSC1) and another to chromosome 16p13.3 (TSC2) (Kandt et al., 1992). The number of families linked to each locus is approximately equal and there is no significant evidence for a third locus (Chapter 3.3). The TSC2 gene has been isolated (Chapter 5) and both genes may act as growth- or tumour- suppressors, since loss of heterozygosity (LOH) has been demonstrated on 9q34 (Carbonara et al., 1994; Green et al., 1994a,b) or 16p13 (Green et al., 1994b) in various hamartomatous tissues from patients with TSC.

The chromosome 9 locus for tuberous sclerosis, TSC1, is tightly linked to the ABO blood group locus (Fryer et al., 1987) and maps in a gene-rich region on chromosome 9q34. Since the initial linkage report by Fryer et al. (1987), the TSC1 region has been refined to a region of 5 cM between D9S149 and D9S114 (Connor et al., 1987; Haines et al., 1991; Northrup et al., 1992; Povey et al., 1994; Janssen et al., 1990; Chapter 3.2; Chapter 3.3). However, there is no consensus on the exact position of TSC1 within this interval, since some groups have found recombinants in favour of a position proximal to ABO and the dopamine ß hydroxylase gene (DBH), while other groups have presented data supporting a location distal of these markers (Sampson and Harris, 1994). The conflicting observations have several possible causes, including misclassification of individuals with only minor clinical findings or non-linkage of one or more families.

Several genes have been mapped within the TSC1 candidate region, including ABO, DBH, the Surfeit gene cluster and VAV2 (Henske et al., 1995; Smith et al., 1994; Woodward et al., 1994)), while other disorders genetically linked to ABO include torsion dystonia (Ozelius et al., 1989) and nail patella syndrome (Renwick and Lawler, 1955; Renwick and Schulze, 1965).

In this paper we present the results of a contig assembly and gene mapping effort, focused on part of the TSC1 candidate region around ABO and DBH. Our detailed map spans 600 kb, corresponding to more than 2 cM of the TSC1 critical region. Eight genes and several known and novel genetic markers have been precisely positioned on a genomic EcoRI map between D9S149 and D9S114.

Materials and Methods

Libraries

The ICI YAC library (Anand et al., 1990) was accessed through the UK Human Genome Mapping Resource Centre and sets of primary, secondary and tertiary pools for PCR screening were provided by R. Elaswarapu. Primary pools from the St. Louis YAC library (Burke et al., 1987) were supplied by J. den Dunnen from the Department of Human Genetics in Leiden (in the framework of the Leiden YAC Screening Center, supported by NWO (NL) and the EC). The P1 library was made from human foreskin fibroblast DNA (Pierce et al., 1992). The library was gridded into 125 96-well plates with approximately 12 different P1 clones per well and pools were made for PCR screening. The chromosome 9-specific cosmid library LL09NC01"P" was constructed at the Biomedical Sciences Division, LLNL, Livermore, CA 94550, USA under the auspices of the National Laboratory Gene Library Projects sponsored by the US Department of Energy. The library was replicated on gridded filters as described (Bentley et al., 1992) at the (NWO/EC) YAC screening core of the Department of Human Genetics in Leiden. Two sets of membranes were used to make pools for PCR screening (Kwiatkowski et al., 1993). The nomenclature of the cosmids in the contigs is the same as the nomenclature of the library source from which they were obtained. Cosmid ABO17 was provided by J. Wolfe.

Cosmid library screening

Hybridization probes were generated by inter Alu PCR (Nelson et al., 1989) using primers CL1 and CL2 (Lengauer et al., 1992) or by isolating end fragments from cosmids in low melting agarose. Probes were randomly labeled, competed with total human DNA, hybridized to nylon filters and washed using standard procedures (Sambrook et al.,1989). Cosmid library screening by PCR was performed by screening two dimensional pools of clones as described by Green and Olson (1990).

YACs, P1 and cosmid clones

Cosmid and P1 DNA was prepared, isolated and fingerprinted using standard techniques (Maniatis et al., 1989). YACs, P1 and cosmid clones were mapped back to 9q34 by FISH (Breen et al., 1992).

STSs

STSs were developed by YAC end rescue inverse PCR or direct sequencing of cosmids. YAC end rescue was performed as described by Silverman et al. (1991) and the products were sequenced directly. Sequence was derived from the cosmid clones by cycle sequencing (Pharmacia) with the appropriate vector primers.

Results and Discussion

Strategy

We aimed to isolate a significant part of the TSC1 critical region between the markers D9S149 and D9S114 on 9q34. Several additional markers are known to map between these, but have not been convincingly associated with genuine recombination events. The initial strategy was to isolate the region in YAC, P1 and cosmid clones. However, attempts to obtain YACs were hampered by underrepresentation of the region in the available libraries. This prompted us to follow an alternative strategy which consisted of cosmid walking complemented by screening P1 libraries.

YAC library screening

Two YACs from the ICI library, 4DD1 (120kb) and 25DG9 (320kb), were identified with primers specific for the ABO locus. STS mapping using primer pairs from both ends of the YACs indicated that the left ends of both inserts overlapped, however inter Alu PCR in combination with hybridization experiments suggested that the region of overlap was small. FISH analysis confirmed the localization of both YACs to chromosome 9q34, however 25DG9 gave an additional signal on chromosome 18 indicating chimerism. This clone was not investigated further. No additional YACs were identified in the ICI library using the end clone STSs from 4DD1 or 25DG9, or with additional markers from the TSC1 candidate region (D9S10, D9S66, DBH). An STS derived from the left arm of YAC 4DD1 was used to screen the St. Louis YAC library and identified two duplicate clones, 51A7 and 61A10 (200kb). FISH analysis mapped 51A7 to 9qter, however STS mapping experiments using primers derived from the right arm of this clone suggested that it contained a large deletion (data not shown) and it was not investigated further. It is interesting to note that the TSC2 locus on chromosome 16 was also found to be underrepresented in YAC libraries (unpublished results).

Contig assembly

Starting points for cosmid contig assembly were ABO, DBH and D9S10 (Figure 4.1). Cosmids were identified with both the left and right end clones of YAC 4DD1

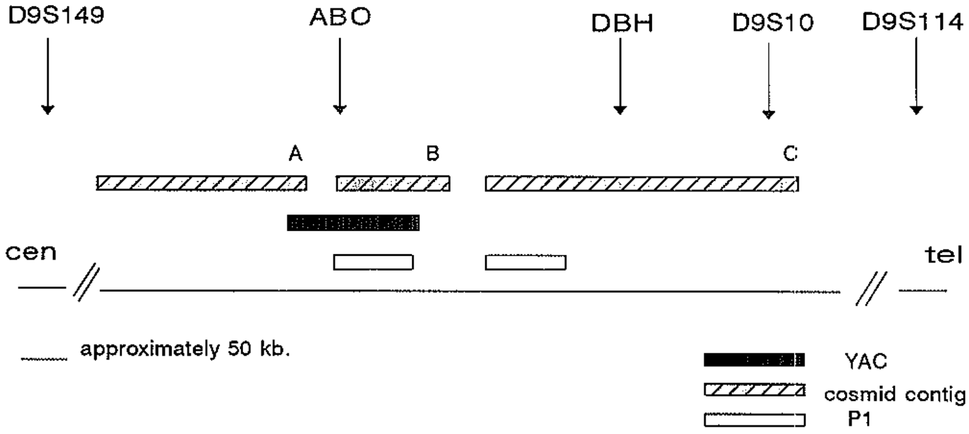
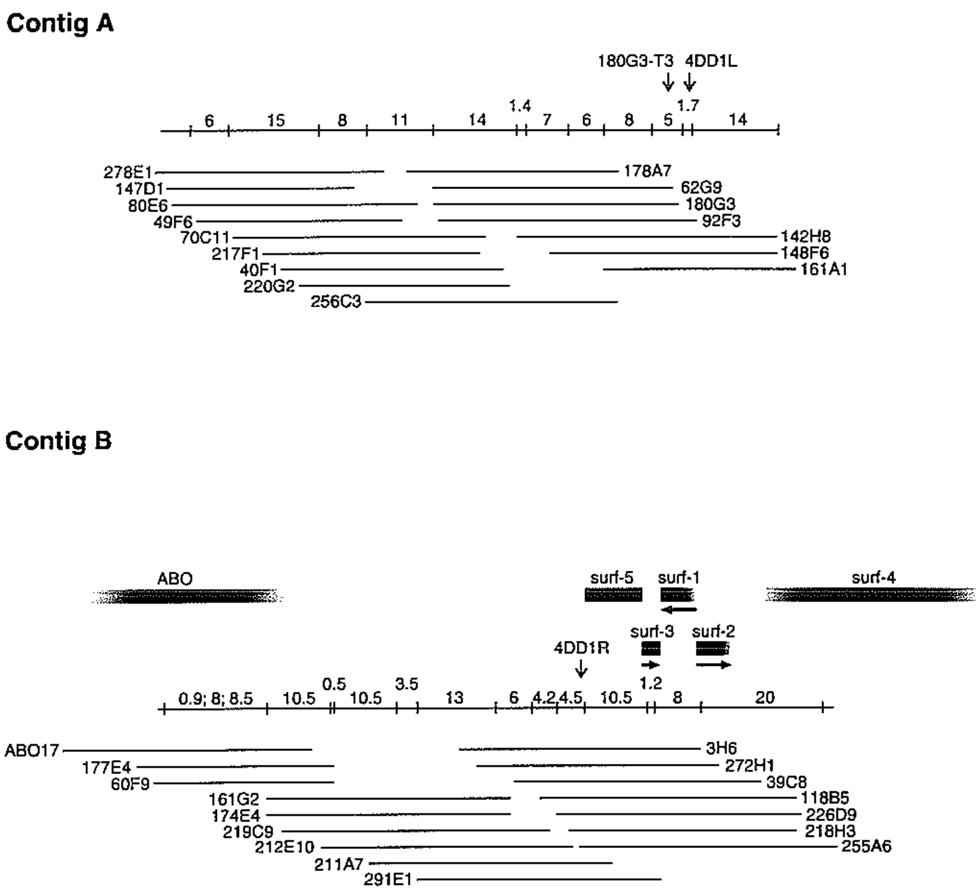


Figure 4.1. Schematic representation of the TSC1 region. The starting points for YAC and cosmid walking are indicated, together with the YAC 4DD1, the cosmid contigs and P1 clones.

and two contigs were constructed of 110kb and 130kb respectively (Figure 4.2, contig A and B). The orientation of the cosmid contigs was consistent with resultsfrom YAC inter Alu PCR screening of the cosmid library and with the YAC STS mapping experiments. No cosmids could be identified distal of cosmid 255A6 (contig B). Only a single non-rearranged cosmid and a single P1 clone were detected at the ABO locus, and no clone could be detected linking the two contigs. However, from the size of the 4DD1 YAC and direct visual hybridization (DIRVISH) experiments of stretched DNA (Wiegant et al., 1991) (unpublished results), we estimated that the gap is approximately 30kb.

Cosmids were identified with the DBH cDNA and probe pMCT136 from the D9S10 locus. DBH and D9S10 map 1 and 2 cM distal of ABO respectively and were linked by chromosome walking, covering a physical distance of 150 kb (contig C). This indicates that the genetic versus physical distance ratio in this region is large. The contig was extended proximal of DBH by 125 kb, but could not be extended further towards ABO. We did isolate a P1 clone with an STS from the proximal end of 251C9, but could not bridge the gap. The distance between clone 251C9 (contig B) and 255A6 (contig C) could not be resolved by interphase FISH, indicating that the gap between contig B and C is small. DIRVISH DNA mapping experiments are in progress to estimate the size of the gap.


Contig A
180G3-T3
4DD1L
6
15
8
11
14
1.4
7
6
8
5
1.7
14
278E1
147D1
80E6
49F6
70C11
217F1
40F1
220G2
256C3
178A7
62G9
180G3
92F3
142H8
148F6
161A1
Contig B
ABO
surf-5
surf-1
surf-4
surf-3
surf-2
4DD1R
0.9; 8; 8.5
10.5
0.5
10.5
3.5
13
6
4.2
4.5
10.5
1.2
8
20
ABO17
177E4
60F9
161G2
174E4
219C9
212E10
211A7
291E1
3H6
272H1
39C8
118B5
226D9
218H3
255A6

Contig C

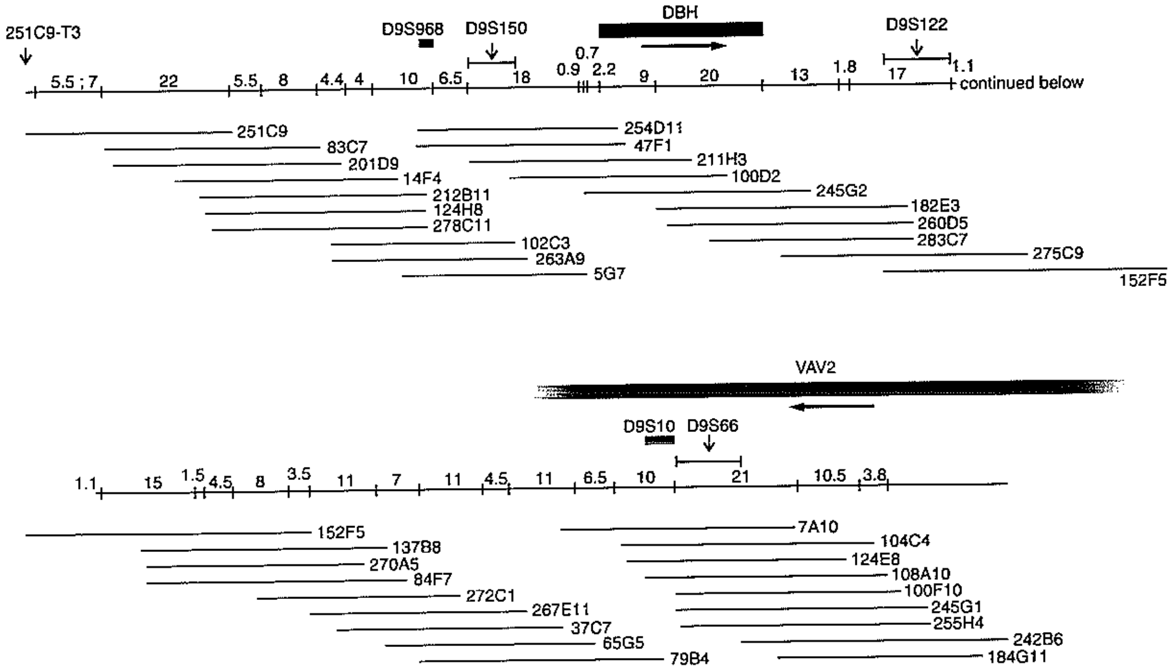


Figure 4.2: Detailed EcoRI restriction map of the three contigs described in this paper. Cosmids are shown below the EcoRI map. Thin bars represent RFLP markers and vertical arrows indicate STSs en microsatellites. Genes are shown above the restriction map as thick bars. The size of the bars indicates the maximal genomic extent. The direction of transcription is indicated by arrowheads. For DBH, Surf-1, Surf-2, Surf-3 and VAV2, the gene structure was studied by Nahmias et al. (in press), Yon et al. (1993) and Henske et al. (1995). The position and orientation of the genes in the cosmid contigs were deduced from our experiments and previously published maps (Nahmias et al., in press; Yon et al., 1993).

Table 4.1:
List of RFLPs in the region

Locus	Enzyme	Probe	fragment sizes	heterozygosity
D9S10	PstI* HindIII (These RFLPs show linkage disequilibrium)	MCT136 MCT136	2.5 and 2.3 kb. 2.2 and 2.0 kb.	50% 50% (200 chrom.)
D9S968	HindIII	RD560	4.5 and 2.6 kb.	14% (115 chrom.)
DBH	(several RFLPs* and $(CA)_n$, all listed in GDB)			
VAV2	PstI	5′ VAV2 (base 1-865)	5, 4.2 and 2.2 kb.	48% (>100 chrom.)

Note: all RFLPs marked with and asterisk are already listed in GDB. The heterozygosity percentages of the new RFLPs (without asterisk) have been determined in at least 100 chromosomes from Caucasians. The map postion of each locus is indicated in Figures 4.1 and 4.2. The VAV2 RFLP maps within the VAV2 gene, distal to the end of the cosmid contig.

Table 4.2:
List of STSes in the region

STS	Primer sequences	Product length	Map position
180G3-T3	5′ GGTGT GGTTC TCCCA AGGG 3′ GAGAG AGGCT TCCTG CTTGC	128 bp	distal part of contig A
4DD1L	5′ CCAAG GGAAG CTGGA GAAGT 3′ CCAGA CCCAG CCTAC ATTTC	97 bp	left arm of YAC 4DD1
4DD1R	5′ CATGC TGTTG GCACT GTTGTA 3′ TTTCT CTTTG GCTTC CCTCTT	135 bp	right arm of YAC 4DD1
251C9-T3	5′ GGAAA GAGGA GCGAG GAAG 3′ CACAA TCTCA CAGTG AATGCC	152 bp	proximal end of contig C

Note: a number of polymorphic STSes at ABO, DBH, VAV2, D9S149, D9S150, D9S122, D9S66 and D9S114 have been described previously and are therefore not included in the list.

In regions of overlap the contigs presented here were consistent with the cosmid contigs constructed by HinfI fingerprinting as described by Nahmias et al. (Nahmias et al., in press). They need at least 50% overlap between cosmids before the clones are joined in a contig. Our data are more detailed and detect smaller overlaps. Additional cosmids have been isolated from the flanking locus D9S149. Chromosome walking experiments are currently focussed on closing the gap between D9S149 and the most proximal ABO contig (contig A).

Mapping of markers and genes in the contigs

RFLPs and unique sequence tagged sites (STSs) are listed in Tables 4.1 and 4.2. The STSs 180G3-T3 and 4DD1L map to adjacent EcoRI fragments in contig A. Two additional STSs, 4DD1R and 251C9-T3 were mapped to contigs B and C respectively. Existing minisatellite repeats (D9S122 and D9S150) (Povey et al., 1994) were precisely positioned within this large contig (Figure 4.2, contig C) and a HindIII polymorphism (D9S968) was detected immediately proximal of DBH.

The position and orientation, where known, of genes identified within the contigs are indicated in Figure 4.2. The role and expression pattern of the ABO blood group transferase indicate that it is not a good candidate for TSC1. The Surfeit gene cluster had been previously mapped by in situ hybridization telomeric to the c-abl and can genes on 9q34 (Yon et al., 1993). A oligonucleotide derived from the Surf-3 cDNA sequence detected a 1.2 kb EcoRI fragment in several cosmids, slightly distal to ABO in contig B. In the mouse this cluster consists of 6 house keeping genes, which are unrelated by sequence homology (Yon et al., 1993). To date the Surfeit genes form the tightest gene cluster known in mammals. Since these genes are in the critical region of TSC1 and not much is known about their function, mutation analysis in TSC patients must be considered.

Our EcoRI mapping data from the DBH locus is consistent with that of Kobayahi et al. (1989). The direction of transcription is towards the telomere. The role of DBH in the conversion of dopamine to noradrenaline and the neurological manifestations of TSC led to the proposal that DBH could be a candidate for the TSC1 gene (Janssen et al., 1993). However, more recent results suggest that TSC1 maps either distal or proximal of DBH and consequently DBH is not such an attractive candidate.

Exon trapping (Buckler et al., 1991) efforts using cosmids from the D9S10 locus identified a gene homologous to the vav oncogene (Smith et al., 1994). This gene, designated VAV2, was considered a good candidate for the TSC1 gene. However, intensive screening failed to identify any mutations, and VAV2 was consequently excluded as a candidate for the TSC1 gene (Henske et al., 1995).

Eight different genes could be placed on the map. The region is gene dense and although some genes map extremely close to each other, we can not exclude the presence of other, as yet unidentified, expressed sequences in the same region. Experiments to identify and characterise additional genes from the TSC1 candidate region are in progress.

Further efforts are directed towards extending the contigs and screening TSC patients for mutations by pulsed field gel electrophoresis (PFGE) using novel probes derived from our cloned material. The identification of large deletions at the TSC2 locus made a significant contribution to the rapid isolation of the TSC2 gene (Chapter 5).

In conclusion we have identified 80 cosmids, 2 P1 clones and a single non-rearranged YAC from the TSC1 candidate region on 9q34. We have constructed a detailed restriction map of three adjacent cosmid contigs and oriented the maps with respect to known and previously unidentified genes and DNA markers. We have shown that DBH and D9S10, previously estimated to be 1 cM apart, are separated by less than 300 kb, and estimate that the physical distance between ABO and DBH is less than 300 kb.

In conjunction with the article by Nahmias and coworkers (Nahmias et al., in press) we have shown that cosmid walking, using a large chromosome specific cosmid library can provide almost complete coverage of a large genomic region. This minimises the need to search for non-chimeric non-rearranged YAC clones, which have been difficult to obtain from the TSC1 region. Moreover, our contigs and the associated maps provide a good tool for generating novel markers and cloning additional genes from this region. It would be of great help to get more excluding data on the recombinants within the region, so that the search for TSC1 can be restricted to a smaller area. LOH studies in tumors of patients and the development of new polymorphic CA repeats in the area, especially between ABO and D9S149, could help to reduce the critical region. Ultimately it is hoped that this work will lead to the identification of the TSC1 gene.

Acknowledgements

We are grateful to Professor H. Galjaard and Professor D. Lindhout for their continuous support. We like to thank Dr. S. Povey, Dr. J. Wolfe, Joseph Nahmias and co-workers for sharing unpublished data and their cooperation. This work was funded by the Dutch Organisation for Scientific Research (NWO).

References

Anand R, Riley JH, Butler R, et al. A 3.5 genome equivalent multi access YAC library: construction, characterisation, screening and storage. Nucleic Acids Res 1990;18:1951-1956.

Bentley DR, Todd C, Collins J, et al. The development and application of automated gridding for efficient screening of yeast and bacterial ordered libraries. Genomics 1992;12:534-541.

Breen M, Arveiler B, Murray I, et al. YAC mapping by FISH using Alu-PCR-generated probes. Genomics 1992;13:726-730.

Buckler AJ, Chang DD, Graw SL, et al. Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. Proc Natl Acad Sci USA 1991;88:4005-4009.

Burke DT, Carle GF, Olson MV. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 1987;236:806-812.

Carbonara C, Longa L, Grosso L, et al. 9q34 loss of heterozygosity in a tuberous sclerosis astrocytoma suggests a growth suppressor-like activity also for the TSC1 gene. Hum Mol Genet 1994;3:1829-1832.

Connor JM, Pirrit LA, Yates JRW, et al. Linkage of the tuberous sclerosis locus to a DNA polymorphism detected by v-abl. J Med Genet 1987;24:544-546.

Fryer AE, Chalmers A, Connor JM, et al. Evidence that the gene for tuberous sclerosis is on chromosome 9. Lancet i 1987:659-661.

Green AJ, Johnson PH, Yates JRW. The tuberous sclerosis gene on chromosome 9q34 acts as a growth suppressor. Hum Mol Genet 1994a;3:1833-1834.

Green AJ, Smith M, and Yates JRW. Loss of heterozygosity on chromosome 16p13.3 in hamartomas from tuberous sclerosis patients. Nature Genet 1994b;6:193-196.

Green ED, Olson MV. Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. Proc Natl Acad Sci USA 1990;87:1213-1217.

Gomez MR: Tuberous sclerosis 2nd edition. New York, Raven Press 1988.

Haines JL, Short MP, Kwiatkowski DJ, et al. Localization of one gene for tuberous sclerosis within 9q32-9q34, and further evidence for heterogeneity. Am J Hum Genet 1991;49:764-772.

Henske EP, Short MP, Jozwiak S, et al. Identification of VAV2 on 9q34 and its exclusion as the tuberous sclerosis gene TSC1. Ann Hum Genet 1995; 59: 25-37.

Janssen LAJ, Sandkuyl LA, Merkens EC, et al. Genetic heterogeneity in tuberous sclerosis. Genomics 1990;8:237-242.

Janssen LAJ, Nellist M, Eussen BE, et al. The map position of three candidate genes for tuberous sclerosis 1: XPAC, DBH and TAN1. Cyt Cell Genet 1993;64:115.

Kandt RS, Haines JL, Smith M, et al. Linkage of an important gene loccus for tuberous sclerosis to a chromosome 16 marker for polycystic kidney disease. Nature Genet 1992, 2:37-41.

Kobayashi K, Kurosawa Y, Fujita K, Nagatsu T. Human dopamine beta hydroxylase gene: two mRNA types having different 3' terminal regions are produced through alternative polyadenylation. Nucleic Acids Res 1989;17:1089-1102.

Kwiatkowski DJ, Zoghbi HY, Ledbetter SA, et al. Rapid identification of yeast artificial chromosome clones by matrix pooling and crude lysate PCR. Nucleic Acids Res 1993;18:7197-7203.

Lengauer C, Riethman HC, Speicher MR, et al. Metaphase and interphase cytogenetics with Alu-PCR-amplified yeast artificial chromosome clones containing the BCR gene and the protoonocogene c-raf-1, c-fms and c-erbB-2. Cancer Res 1992;52:2590-2596.

Sambrook J, Fritsch EF, Maniatis TM. Molecular cloning: A Laboratory Manual. New York, Cold Spring Harbour Press, 1989.

Nahmias J, Hornigold N, Fitzgibbon J, et al. Cosmid contigs spanning 9q34 including the TSC1 candidate region. eur j Hum genet (in press).

Nelson DL, Ledbetter SA, Corbo L, et al. Alu polymerase chain reaction: a method for rapid isolation of human-specific sequences from complex DNA sources. Proc Natl Acad Sci USA 1989;86:6686-6690.

Northrup H, Kwiatkowski DJ, Roach ES, et al. Evidence for genetic heterogeneity in tuberous sclerosis: one locus on chromosome 9 and at least one locus elsewhere. Am J Hum Genet 1992;51:709-720.

Ozelius L, Kramer PL, Moskowitz CB, et al.: Human gene for torsion dystonia located on chromosome 9q32-q34. Neuron 1989;2:1427-1434.

Pierce JC, Sauer B, Sternberg N. A positive selection vector for cloning high molecular weight DNA by the bacteriophage P1 system: Improved cloning efficiency. Proc Natl Acad Sci USA 1992;89:2056-2060.

Povey S, Armour J, Farndon P, et al. Report on the Third International Workshop on Chromosome 9. Ann Hum Genet 1994;58:177-250.

Renwick JH, Lawler SD. Genetical linkage between the ABO and nail-patella loci. Ann Hum Genet 1955;19:312-331.

Renwick JH, Schulze J. Male and female recombination fractions for the nail patella: ABO linkage in man. Ann Hum Genet 1965;28:379-392.

Sampson JR, Harris PC. The molecular genetics of tuberous sclerosis. Hum Mol Genet 1994;3:1477-1480.

Silverman GA, Jockel JI, Domer PH, et al. Yeast artificial chromosome cloning of a two-megabase-size contig within chromosomal band 18q21 establishes physical linkage between BCL2 and plasminogen activator inhibitor type 2. Genomics 1991;9:219-228.

Smith M, Brickey W, Handa K, Gargus JJ. Sequence analysis of MCT 136 (locus D9S10) in the TSC gene region reveals amino acid domains with sequence homology to RAS activating signal molecules VAV and HSOS. Ann Hum Genet 1994;58:235-236.

Wiegant J, Kalle W, Mullenders L, et al. High resolution in situ hybridisation using DNA halo preparations. Hum Mol Genet 1992;1:587-591.

Woodward KJ, Nahmias J, Hornigold N, et al. Mapping chromosome 9q34 by FISH using metaphase chromosomes with specific translocation breakpoints. Ann Hum Genet 1994;58:241-242.

Yon J, Jones T, Garson K, et al. The organization and conservation of the human Surfeit gene cluster and its localization telomeric to the c-abl and can proto-oncogenes at chromosome band 9q34.1. Hum Mol Genet 1993;2:237-240.

Note:

The original version of this paper will appear in the European Journal of Human Genetics. Permission for including this paper in this thesis has been granted by S. Karger AG, Basel, 9 feb. 1995.

CHAPTER 5

TSC2 GENE IDENTIFICATION

5.1

IDENTIFICATION AND CHARACTERIZATION OF THE TUBEROUS SCLEROSIS GENE ON CHROMOSOME 16

The European Chromosome 16 Tuberous Sclerosis Consortium

Cell 1993; 75: 1305-1315.

Summary

Tuberous sclerosis (TSC) is an autosomal dominant multisystem disorder with loci assigned to chromosomes 9 and 16. Using pulsed field gel electrophoresis (PFGE) we identified five TSC associated deletions at 16p13.3. These were mapped to a 120kb region which was cloned in cosmids and from which four genes were isolated. One gene, designated TSC2, was interrupted by all five PFGE deletions and closer examination revealed several intragenic mutations, including one de novo deletion. In this case Northern blot analysis identified a shortened transcript, while reduced expression was observed in another TSC family, confirming TSC2 as the chromosome 16 TSC gene. The 5.5 kb TSC2 transcript is widely expressed and its protein product, tuberin, has a region of homology to the GTPase-activating protein GAP3.

Introduction

Tuberous sclerosis (TSC) is an autosomal dominant disorder, classified as a phakomatosis (van der Hoeve, 1933) and characterised by the widespread development of growths, usually described as hamartomata, in many tissues and organs. The unpredictable distribution of these lesions, particularly within the brain, eyes, skin, kidneys, heart, lungs and skeleton results in a wide variety of signs, symptoms and complications (Gomez, 1988). Therefore the frequency of diagnosed cases is likely to under-represent true prevalence which may be as high as 1 in 5,800 (Osborne et al., 1991). The pathogenesis of TSC is poorly understood and efforts to establish the primary underlying defect have focused on positional cloning of the causative gene(s).

Linkage studies have established locus heterogeneity (Sampson et al., 1989a; Haines et al., 1991a&b; Chapter 3.1 & 3.2; Povey et al.,1991; Northrup et al., 1992) with disease determining genes on chromosomes 9 (Fryer et al., 1987) and 16 (Kandt et al., 1992) leading to apparently indistinguishable phenotypes. In most, if not all, affected multigeneration families the disease can be accounted for by the gene at one or other of these loci (Kwiatkowski et al., 1993). The Genome Database Nomenclature Committee recently agreed that the loci on chromosomes 9 and 16 should be termed TSC1 and TSC2 respectively. Analysis of meiotic recombination events in TSC families has refined the positions of TSC1 and TSC2 to small regions in the telomeric chromosomal bands 9q34.3 and 16p13.3. The candidate region at 16p13.3 extends between the markers MS205.2 (D16S309) and 16AC2.5 (D16S291) (Kwiatkowski et al., 1993), representing an estimated 1.5 megabases of DNA.

Loss of heterozygosity for alleles at 16p has been observed in hamartomata from TSC patients (Green and Yates, 1993; Smith et al., 1993), indicating that a second somatic mutation may be required to produce the TSC phenotype at a cellular level. This observation is consistent with the chromosome 16 TSC gene acting as a tumour suppressor, a feature shared by genes causing the other phakomatoses, neurofibromatosis type 1 (NF1) (Legius et al., 1993) and type 2 (NF2) (Trofatter et al., 1993) and von Hippel-Lindau disease (VHL) (Latif et al., 1993). If a two-hit mechanism, as proposed by Knudson (1971), does apply to TSC, then inactivating constitutional mutations would be anticipated. TSC has not been noted in individuals with the α-thalassaemia/mental retardation (ATR-16) phenotype and terminal deletions of 16p which extend into the distal part of the candidate region (Wilkie et al., 1990). Therefore we investigated the proximal part of the candidate region for deletions.

Some 60% of TSC cases appear to represent new mutations (Sampson et al., 1989b) and we reasoned that a proportion of these might be large deletions. Such deletions, detectable by pulsed field gel electrophoresis (PFGE), would greatly facilitate identification of the gene, as has been demonstrated in NF1, NF2 and VHL (Viskochil et al., 1990; Trofatter et al., 1993; Latif et al., 1993). We have now identified 5 TSC associated constitutional interstitial deletions of between 30 and 75 kb in the proximal part of the candidate region. These have been mapped to a 120kb segment from which we have identified a number of genes, one of which was disrupted by all the deletions. Mutation analysis and expression studies provide strong evidence that this gene, which we term TSC2, is the chromosome 16 tuberous sclerosis determining gene.

Results

Deletions in The TSC Candidate Region

An ATR-16 patient (BO), with a constitutional deletion at 16p (Wilkie et al., 1990) which extends into the TSC candidate region (Figure 5.1), was specifically reassessed for signs of TSC (by clinical evaluation, renal ultrasound and cranial CT imaging) but with negative results. This led us to focus our search for TSC associated deletions on the more proximal part of the candidate region, most of which is spanned by a ClaI restriction fragment of approximately 340kb (Germino et al., 1992, Harris et al., 1990). Using pulsed field gel electrophoresis (PFGE) this fragment was assayed in 255 unrelated TSC patients with SM6, a single copy probe isolated from cosmid cSMII (Figure 5.1). The patients were ascertained during ongoing clinical studies in Cardiff and Rotterdam, and all fulfilled definitive diagnostic criteria as defined by Gomez (1988). Aberrant smaller fragments consistent with constitutional interstitial deletions were observed in 5 cases. As these changes were likely to involve the TSC gene, the region containing the deletions was further characterised.

Mapping of PFGE deletions and genomic cloning

Cosmid walking was initiated from the previously defined loci JCII and N54 (Germino et al., 1990; Himmelbauer et al., 1991). The proximally directed walk established a series of overlapping clones spanning 200kb across the area of the TSC associated deletions, while the distally directed walk was hampered by a duplicated region homologous to sequences more proximal on 16p (Germino et al., 1992). A long range restriction map was constructed in genomic and cloned DNA which was consistent in size with that produced by Germino et al. (1992), although additional

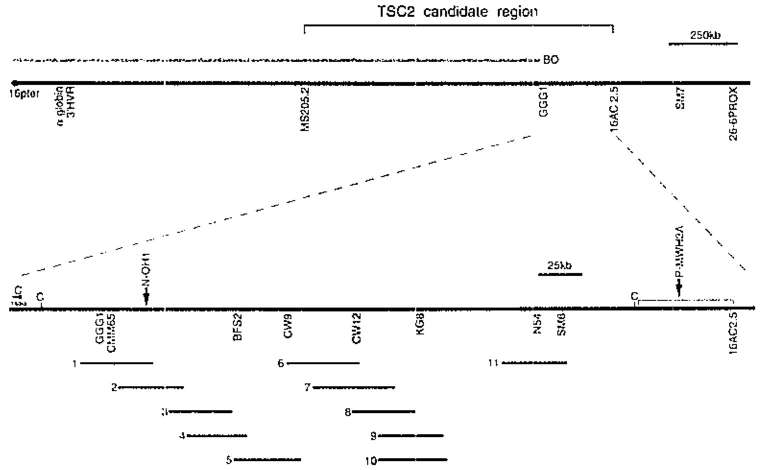


Figure 5.1: A map of the terminal region of chromosome 16p showing the TSC candidate region determined by linkage analysis between MS205.2 (D16S309) and 16AC2.5 (D16S291). The size of the terminal deletion found in ATR-16 patient BO is shown (top) (Summarised from Harris et al., 1990; Harris et al., 1991; Germino et al., 1992; Rack et al., 1993; Wilkie et al., 1990; Germino et al., 1993 and Kwiatkowski et al., 1993). Expanded below is a detailed map of the proximal TSC candidate region showing the ClaI (C) sites, the breakpoints of the two somatic cell hybrids N-OH1 and P-MWH2A and the positions of existing and selected new DNA probes. The positions of cosmids within the contig are shown below the map: 1, JCI; 2, JCII; 3, CC1; 4, CC1-2; 5, CBFS1; 6, CW9D; 7, LADS4; 8, CW12I; 9, JH1K; 10, ZDS5; 11, SMII.

sites for NruI and MluI were identified. Mapping of SacII and BssHII sites enabled the unmethylated CpG islands to be located (Figure 5.5). The area was precisely mapped with EcoRI and other restriction enzymes and many fragments were subcloned (Figure 5.5 and Experimental Procedures for details).

The sizes and positions of the five TSC deletions were more accurately determined by analysing MluI (Figure 5.2) and NruI digested DNA. Successive hybridisations enabled fragments flanking or containing the deletion breakpoints to be identified (Figure 5.3). When suitable material was available a breakpoint fragment was identified in EcoRI, BamHI and/or HindIII digests with probes immediately flanking the deletion, confirming the nature of the rearrangement (Figure 5.4). The precise positions of each of the TSC deletions is summarised in Figure 5.5. Two deletions estimated at 32kb and 46kb, and two of at least 70kb were positioned distally and overlapped one another extensively. A fifth deletion of approximately 35kb was more proximally situated and was shown to be non-overlapping with at least 3 of the distal deletions (Figure 5.5). As each of these deletions was likely to involve part of the chromosome 16 TSC gene, a candidate gene that mapped into all of them was sought.

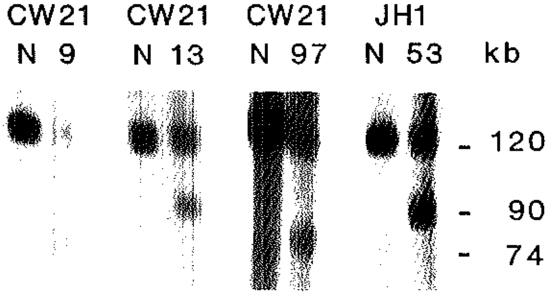


Figure 5.2. PFGE analysis of deletions in TSC individuals.
PFGE of MluI digested DNA from TSC patients and controls probed with the clones CW21 (WS-9, -13, -97) and JH1 (WS-53), which detect an ≈ 120kb fragment in normal individuals and additional smaller fragments in the patients. CW21 is deleted in patient WS-53 so does not recognise the aberrant ≈ 90kb fragment. The WS-97 deletion removes ≈ 75kb including the distal Mlu I site producing an ≈ 74kb junction fragment (see Figure 5.5).

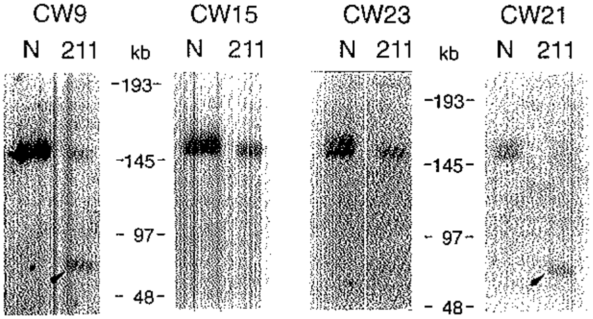


Figure 5.3. PFGE analysis of deletions in TSC individuals.
PFGE of NruI digested DNA of a normal control (N) and WS-211 (211) hybridised with probes flanking the breakpoint at the distal end (CW9 and CW15) and at the proximal end (CW23 and CW21) of the deletion. As well as the normal ≈ 150kb fragment, the same ≈ 80kb breakpoint fragment (arrowed) is seen with the two markers outside of the deletion (CW9 and CW21). CW15 is completely deleted (no breakpoint fragment) while CW23 is mostly deleted although a faint ≈ 80kb fragment can be seen in the WS- 211 track.

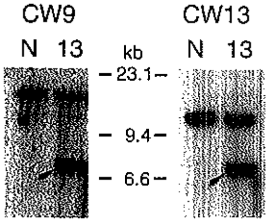


Figure 5.4: Deletion mapping. EcoRI digested DNA of normal control (N) and WS-13 (13) separated on a conventional gel and hybridised with probes, CW9 and CW13, which flank the deletion (see Figure 5.5). The same 7kb breakpoint fragment (arrowed) is seen with both markers consistent with a deletion of 32kb ending within the EcoRI fragments seen by these probes.

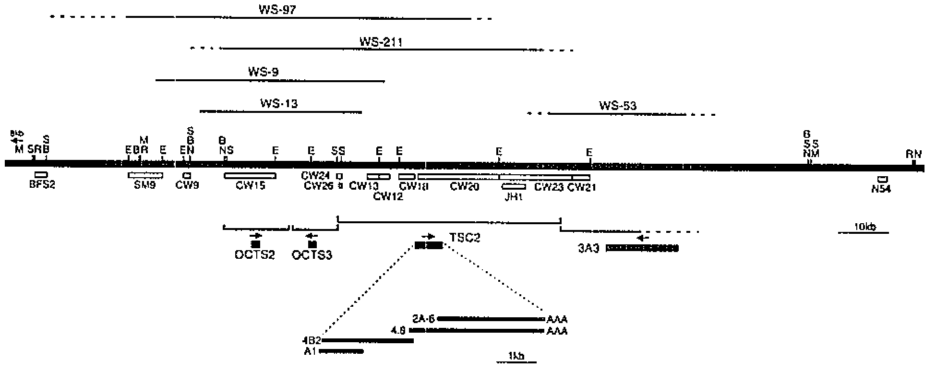


Figure 5.5. A detailed map of the TSC area of chromosome 16. Genomic sites for the enzymes BssHII, B; MluI, M; NotI, N; NruI, R; SacII, S and a partial map of EcoRI, E, sites are shown. The open boxes indicate the size and location of genomic probes (see Experimental Procedures for details). The solid boxes show the sizes of transcripts and their orientations on the chromosome are marked with arrows. The genomic extent of each gene is indicated with brackets. The full proximal extent of 3A3 is unknown. cDNA clones comprising the TSC2 gene are shown enlarged below. The size and location of TSC associated deletions are shown above the map with dashed lines indicating regions of uncertainty. The WS-13 deletion is 32 kb and flanked by CW13 and CW9. A 7 kb EcoRI breakpoint fragment is seen with these two probes (Figure 5.4). WS-9 is a 46 kb deletion with the breakpoints in SM9 and CW12. An 8kb EcoRI breakpoint fragment is seen with these probes. The WS-211 deletion is ≈ 75 kb and the breakpoints lie between CW9 and CW15 distally, and between CW23 and CW21 proximally (Figure 5.3). The distal breakpoint of WS-97 is between BFS2 and SM9 and proximally within CW20, with a region of approximately 75kb deleted. The WS-53 deletion is ≈ 35kb and the distal breakpoint lies within CW23, proximal to JH1. The proximal 0.6kb of TSC2 is deleted. The exact location of the proximal breakpoint of WS-53 is unknown.

Genes in the region harbouring pulsed field deletions
Subcloned probes and fragments from cosmids spanning the region of the TSC associated deletions were used to screen human fetal brain and human kidney cDNA libraries. The mapping of positive clones to the target area was confirmed by hybridisation to panels of somatic cell hybrids, containing derivative 16 chromosomes with breakpoints flanking this region; N-OH1, distal, and P-MWH2A, proximal (Figure 5.1), and a radiation hybrid Hy145.19 which contains this area, as a positive control. Northern blot analysis using RNAs from various human cell lines indicated that the clones derived from four apparently unrelated genes. Hybridisation of the cDNA clones to digests of cosmid, genomic and hybrid DNA indicated the genomic distributions of the genes. Sequence analysis identified the polyA tail of each gene and established their transcriptional orientations.

A gene, termed OCTS2, with a transcript of 1.7kb (cDNA clones OCTS2C and RCTS2) and a second gene termed OCTS3 with a 1kb transcript (cDNA clone OCTS3C) mapped entirely within the four distal deletions, but did not extend as far as the proximal deletion in patient WS-53 (Figure 5.5). A 15kb transcript was recognised by two cDNA clones, 3A3 and AH4, and was termed 3A3. It mapped partly within the WS-53 deletion. Since the distal clone AH4 contained the polyA tail, the gene is transcribed from centromere to telomere and does not extend towards the distal deletions (Figure 5.5).

The cDNA clones 2A6 and 4.9 detect an ≈ 5.5 kb transcript and were identified using an 18kb EcoRI fragment from cosmid ZDS5 (corresponding to the region subcloned in CW23 and CW21). A transcript of the same size was detected by CW26, a genomic probe which maps at a CpG island located within the four distal deletions (Figure 5.5). By means of a cDNA walk the 2A6 and 4.9 clones were connected to clones 4B2 and A1 which mapped to the CW26 region confirming that this single gene is disrupted by all five PFGE deletions. This gene was therefore the best candidate for the chromosome 16 TSC gene. It was designated TSC2 and characterised in detail.

Northern blot analysis indicates that TSC2 is widely expressed with the 5.5 kb transcript seen in all cell lines tested, including those derived from brain, kidney, skin, liver, adrenal gland, colon and white blood cells (Figure 5.6). Expression has also been seen in all tissues tested, including liver, kidney and heart, and in lymphocytes, fibroblasts and biliary epithelium. The high level of TSC2 expression in fibroblasts made it possible to compare the level of transcription in fibroblasts derived from normal controls and TSC patients. In one family in which TSC has been shown to co-segregate with chromosome 16p13.3 markers, but in which the mutation has not been identified, the affected members showed clearly reduced levels of TSC2 transcript (Figure 5.7). Transcripts from adjacent genes showed unaltered levels of expression (data not shown).

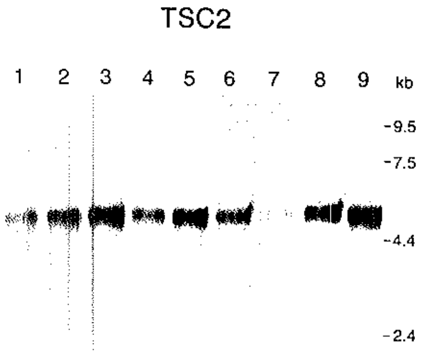


Figure 5.6: The TSC2 cDNA clone 2A6 hybridised to a Northern blot containing mRNA from human tissue specific cell lines: 1, MJ, normal EBV transformed lymphocytes; 2, K562, erythroleukaemia; 3, FS1, normal fibroblasts; 4, HeLa, cervical carcinoma; 5, G401, renal Wilm's tumour; 6, Hep3B, hepatoma; 7, HT29, colonic adenocarcinoma; 8, SW13, adrenal carcinoma; and 9, G-CCM, astrocytoma. The level of expression of the 5.5kb transcript seen here is consistent with other results which show the highest levels of expression in fibroblasts, and cell lines of renal and brain origin. The lowest level of expression is in the colonic adenocarcinoma and EBV transformed lymphocytes. Approximately 1μg of mRNA was loaded per lane.

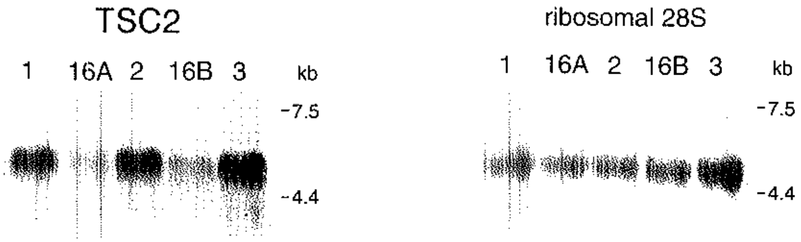


Figure 5.7: Northern blot of total RNA isolated from fibroblasts of a normal control (1), two unrelated TSC individuals (2 and 3), and two sibs from the same chromosome 16-linked TSC family (16A and 16B) hybridised with the TSC2 cDNA clone 2A6. A reduced level of expression is seen in the two family members. The affected mother from this family also showed reduced TSC2 expression (data not shown). Non-specific hybridisation to the ribosomal 28S band under non stringent conditions is shown to indicate that the lanes were equally loaded. Each lane contains approximately 20μg of total RNA.

The combination of non-overlapping PFGE deletions affecting TSC2 and the reduced expression of the TSC2 transcript in TSC patients strongly suggests that the deletions inactivate the structural TSC determining gene rather than a regulatory element for a remote gene. To confirm that TSC2 is indeed a TSC determining gene we sought independent intragenic mutations.

Intragenic mutations affecting TSC2

DNA samples from 260 unrelated TSC patients were screened for confirmatory rearrangements within TSC2 using cDNA sub-clones as hybridisation probes. All patients tested fulfilled the definitive diagnostic criteria of Gomez (1988) and included many of those previously studied by PFGE. In addition to those cases in which PFGE abnormalities had been found, aberrant bands were noted with multiple restriction enzymes in a further 5 patients. Southern analysis using a combination of genomic clones and small cDNA fragments as hybridisation probes demonstrated that these changes represented small deletions. The position of each deleted segment was confirmed relative to the genomic map of EcoRI, HindIII and BamHI sites (Figure 5.8).

The most 5′ deletion found in patient WS-210, was not entirely intragenic as it also involved the OCTS3 related gene. The deletion spans 5-6kb and removes the genomic probe CW26 which contains TSC2 coding sequence. All four other deletions were shown to be entirely within TSC2. A deletion of approximately 1kb in patient 5773 was shown to remove an intronic HindIII site (Figure 5.9). In this case the mutation was also detected in the affected parent. In two further cases (WS-80 and 1737) deletions of approximately 3kb and 5kb respectively were identified. The parents of these cases were thought to be unaffected but were not available for analysis, making it impossible to confirm that the changes represented de novo mutations. In contrast, both clinically unaffected parents of patient WS-11 were available for analysis and the ≈5kb deletion (which was not seen on PFGE) was shown to represent a de novo mutation (Figure 5.10). The deletion removes an intronic HindIII site and the upstream intronic EcoRI site. The genomic probe CW18, which lies between these sites and detects the TSC2 transcript, was shown to be deleted. Leucocyte polyA RNA prepared from this patient showed an abnormal TSC2 transcript of ≈ 4.5kb on Northern analysis (Figure 5.11).

Together these findings confirm that TSC2 is the chromosome 16 tuberous sclerosis determining gene.

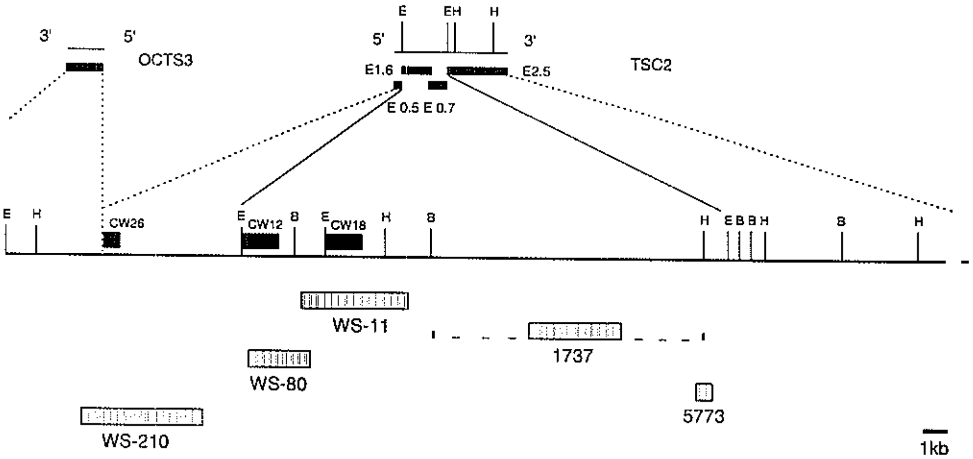


Figure 5.8: Restriction map to show the genomic distribution of TSC2. Genomic probes (CW26, CW12, CW18) and cDNA probes (E0.5, E1.6, E0.7, E2.5) are represented by solid bars, and the position of 5 small deletions (hatched bars) affecting the gene are shown. Exonic EcoRI sites and the 5' and 3' ends of the gene are linked to the genomic map by the diagonal lines.

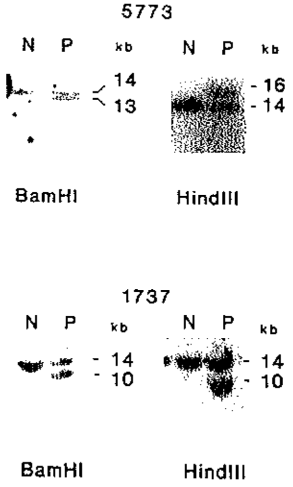


Figure 5.9: Southern blot analysis of deletions in cases 5773 and 1737. HindIII and BamHI digested DNA from the patients (P) and an unrelated control (N) was hybridised with cDNA probe E0.7. This probe detects adjacent HindIII fragments of ≈ 14kb and 2.5kb and a single BamHI fragment of ≈ 14kb. In case 5773 a deletion of ≈ 1kb within the BamHI fragment removes a HindIII site to produce a junction fragment of ≈ 16kb. The ≈ 4kb deletion in case 1737 produces novel HindIII and BamHI fragments of ≈ 10kb. Adjacent fragments were normal.

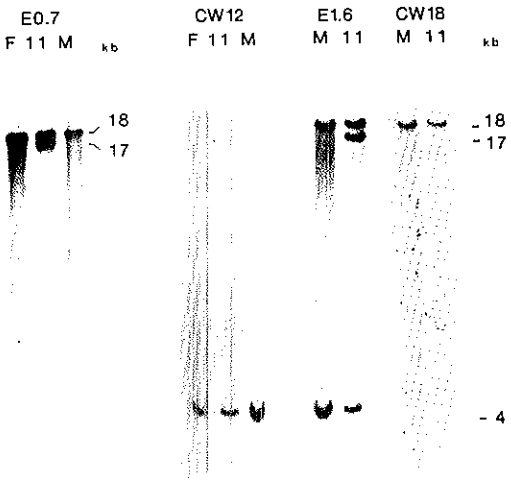


Figure 5.10: Southern blot analysis of the de novo deletion in case WS- 11. EcoRI digested DNA from the patient (11), father (F) and mother (M) was hybridised with probes E0.7, CW12, E1.6 and CW18. E0.7 detects the normal 18kb fragment in WS-11 and both parents and an additional 17kb fragment in WS-11 alone. CW12 detects the normal 4kb fragment in WS-11 and both parents, and the additional 17kb fragment in WS-11 alone, demonstrating that the 17kb fragment is a junction fragment. E1.6 spans the EcoRI site which is deleted in formation of the junction fragment, and so detects both normal fragments of 4kb and 18kb and the 17kb junction fragment. CW18 is deleted on the mutant chromosome so fails to detect the junction fragment. A HindIII junction fragment and novel small BamHI fragment were also seen on Southern analysis and probes recognising a number of VNTR polymorphisms were used to confirm biological parentage (data not shown).

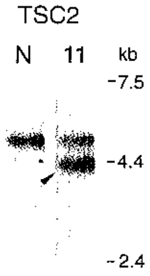


Figure 5.11: The TSC2 cDNA clone 2A6 hybridised to a Northern blot containing 1µg of lymphocyte mRNA from a normal control (N) and TSC patient WS-11 (11) who has an intragenic genomic deletion (see c). An additional abnormal transcript (arrowed) ≈ 1kb smaller than normal is seen in WS-11.

Characterisation of TSC2

To further characterise the TSC2 gene, evolutionary conservation was studied and sequence analysis was performed. A 'zoo-blot' containing genomic DNA from various animal species revealed that the TSC2 gene was conserved throughout the higher vertebrates. Strong signals were obtained from primates and signals indicating lower homology were obtained from several other vertebrates, including rodents, marsupial and reptile. No signal was obtained from fish or non-vertebrate species.

The TSC2 transcript was sequenced completely in both strands. All sequence was confirmed in at least two independent cDNA clones. The coding sequence obtained extends 5474bp (Figure 5.12). Despite repeated cDNA library rescreening, no clones extending further 5′ could be identified. The available sequence approximates to the transcript size determined by Northern blot analysis.

The cDNA contains an open reading frame (ORF) extending from nucleotide 1 through 5370. The second-best ORF is no more than 402bp. At nucleotide position 19 we found an in-frame start codon, matching the Kozak consensus (Kozak, 1987). At the 3′ end we noted two partially overlapping polyadenylation signals (AATAAATAAA) at nucleotide 5425. The occurrence of this doublet may cause differential polyadenylation, since we found polyadenylation sites which differ by up to 15bp in four different cDNA clones.

The total length of the predicted protein is 1784 amino acids with a calculated molecular mass of 198 Kd. There is no apparent signal peptide or signal peptidase cleavage site. Using the method described by Eisenberg et al. (1984), we identified hydrophobic domains, four of which may represent membrane spanning regions. Within a predicted alpha-helical structure, a stretch of 22 amino acids, surrounded by a repeated motif of 9 amino acids, complied with the leucine zipper consensus (Landschulz et al., 1988).

A search for sequence homologies at protein level revealed a region of similarity between the predicted product of TSC2 and the GTPase activating protein GAP3 (or rap1GAP) (Rubinfeld et al., 1991). The region extends over 58 amino acids and the level of residue identity fulfils the criteria of Sander and Schneider (1991) for structural homology. Of the first 39 amino acids, 14 are identical with murine GAP and human GAP3. A core stretch of 17 residues contains identical or similar amino acids with only one mismatch (Figure 5.13).

Figure 5.12

THE TSC2 SEQUENCE:

```
GGTGCGTCCTGGTCCACCATGGCCAAACCAACAAGCAAAGATTCAGGCTTGAAGGAGAAG  60
                  M  A  K  P  T  S  K  D  S  G  L  K  E  K  14

TTTAAGATTCTGTTGGGACTGGGAACACCGAGGCCAAATCCCAGGTCTGCAGAGGGTAAA 120
 F  K  I  L  L  G  L  G  T  P  R  P  N  P  R  S  A  E  G  K  34

CAGACGGAGTTTATCATCACCGCGGAAATACTGAGAGAACTGAGCATGGAATGTGGCCTC 180
 Q  T  E  F  I  I  T  A  E  I  L  R  E  L  S  M  E  C  G  L  54

AACAATCGCATCCGGATGATAGGGCAGATTTGTGAAGTCGCAAAAACCAAGAAATTTGAA 240
 N  N  R  I  R  M  I  G  Q  I  C  E  V  A  K  T  K  K  F  E  74

GAGCACGCAGTGGAAGCACTCTGGAAGGCGGTCGCGGATCTGTTGCAGCCGGAGCGGACG 300
 E  H  A  V  E  A  L  W  K  A  V  A  D  L  L  Q  P  E  R  T  94

CTGGAGGCCCGGCACGCGGTGCT GCTCTGCTGAAGGCCATCGTGCAGGGGCAGGGCGAG 360
 L  E  A  R  H  A  V  L  A  L  L  K  A  I  V  Q  G  Q  G  E 114

CGTTTGGGGGTCCTCAGAGCCCTCTTCTTTAAGGTCATCAAGGATTACCCTTCCAACGAA 420
 R  L  G  V  L  R  A  L  F  F  K  V  I  K  D  Y  F  S  N  E 134

GACCTTCACGAAAGGCTGGAGGTTTTCAAGGCCCTCACAGACAATGGGAGACACATCACC 480
 D  L  H  E  R  L  E  V  F  K  A  L  T  D  N  G  R  H  I  T 154

TACTTGGAGGAAGAGCTGGCTGACTTTGTCCTGCAGTGGATGGATGTTGGCTTGTCCTCG 540
 Y  L  E  E  E  L  A  D  F  V  L  Q  W  M  D  V  G  L  S  S 174

GAATTCCTTCTGGTGCTGGTGAACTTGGTCAAATTCAATAGCTGTTACCTCGACGAGTAC 600
 E  F  L  L  V  L  V  N  L  V  K  F  N  S  C  Y  L  D  E  Y 194

ATCGCAAGGATGGTTCAGATGATCTGTCTGCTGTGCGTCCGGACCGCGTCCTCTGTGGAC 660
 I  A  R  M  V  Q  M  I  C  L  L  C  V  R  T  A  S  S  V  D 214

ATAGAGGTCTCCCTGCAGGTGCTGGACGCCGTGGTCTGCTACAACTGCCTGCCGGCTGAG 720
 I  E  V  S  L  Q  V  L  D  A  V  V  C  Y  N  C  L  P  A  E 234

AGCCTCCCGCTGTTCATCGTTACCCTCTGTCGCACCATCAACGTCAAGGAGCTCTGCGAG 780
 S  L  P  L  F  I  V  T  L  C  R  T  I  N  V  K  E  L  C  E 254

CCTTGCTGGAAGCTGATGCGGAACCTCCTTGGCACCCACCTGGGCCACAGCGCCATCTAC 840
 P  C  W  K  L  M  R  N  L  L  G  T  H  L  G  H  S  A  I  Y 274

AACATGTGCCACCTCATGGAGGACAGAGCCTACATGGAGGACGCGCCCCTGCTGAGAGGA 900
 N  M  C  H  L  M  E  D  R  A  Y  M  E  D  A  P  L  L  R  G 294

GCCGTGTTTTTTGTGGGCATGGCTCTCTGGGGAGCCCACCGGCTCTATTCTCTCAGGAAC 960
 A  V  F  F  V  G  M  A  L  W  G  A  H  R  L  Y  S  L  R  N 314

TCGCCGACATCTGTGTTTCCATCATTTTACCAGGCCATGGCATGTCCGAACGAGGTGGTG 1020
 S  P  T  S  V  F  P  S  F  Y  Q  A  M  A  C  P  N  E  V  V 334

TCCTATGAGATCGTCCTGTCCATCACCAGGCTCATCAAGAAGTATAGGAAGGAGCTCCAG 1080
 S  Y  E  I  V  L  S  I  T  R  L  I  K  K  Y  R  K  E  L  Q 354

GTGGTGGCGTGGGACATTCTGCTGAACATCATCGAACGGCTCCTTCAACAGCTCCAGACC 1140
 V  V  A  W  D  I  L  L  N  I  I  E  R  L  L  Q  Q  L  Q  T 374

TTGGACAGCCCGGAGCTCAGGACCATCGTCCATGACCTGTTGACCACGGTGGAGGAGCTG 1200
 L  D  S  P  E  L  R  T  I  V  H  D  L  L  T  T  V  E  E  L 394

TGTGACCAGAACGAGTTCCACGGGTCTCAGGAGAGATACTTTGAACTGGTGGAGAGATGT 1260
 C  D  Q  N  E  F  H  G  S  Q  E  R  Y  F  E  L  V  E  R  C 414
```

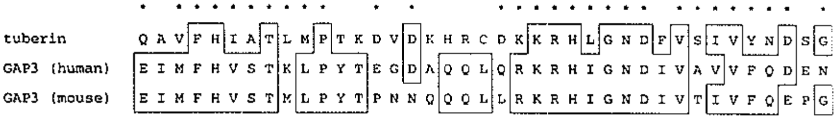
```
GCGGACCAGAGGCCTGAGTCCTCCCTCCTGAACCTGATCTCCTATAGAGCGCAGTCCATC 1320
 A  D  Q  R  P  E  S  S  L  L  N  L  I  S  Y  R  A  Q  S  I 434

CACCCGGCCAAGGACGGCTGGATTCAGAACCTGCAGGCGCTGATGGAGAGATTCTTCAGG 1380
 H  P  A  K  D  G  W  I  Q  N  L  Q  A  L  M  E  R  F  F  R 454

AGCGAGTCCCGAGGCGCCGTGCGCATCAAGGTGCTGGACGTGCTGTCCTTTGTGCTGCTC 1440
 S  E  S  R  G  A  V  R  I  K  V  L  D  V  L  S  F  V  L  L 474

ATCAACAGGCAGTTCTATGAGGAGGAGCTGATTAACTCAGTGGTCATCTCGCAGCTCTCC 1500
 I  N  R  Q  F  Y  E  E  E  L  I  N  S  V  V  I  S  Q  L  S 494

CACATCCCCGAGGATAAAGACCACCAGGTCCGAAAGCTGGCCACCCAGTTGCTGGTGGAC 1560
 H  I  P  E  D  K  D  H  Q  V  R  K  L  A  T  Q  L  L  V  D 514

CTGGCAGAGGGCTGCCACACACACCACTTCAACAGCCTGCTGGACATCATCGAGAAGGTG 1620
 L  A  E  G  C  H  T  H  H  F  N  S  L  L  D  I  I  E  K  V 534

ATGGCCCGCTCCCTCTCCCCACCCCCGGAGCTGGAAGAAAGGGATGTGGCCGCATACTCG 1680
 M  A  R  S  L  S  P  P  P  E  L  E  E  R  D  V  A  A  Y  S 554

GCCTCCTTGGAGGATGTGAAGACAGCCGTCCTGGGGCTTCTGGTCATCCTTCAGACCAAG 1740
 A  S  L  E  D  V  K  T  A  V  L  G  L  L  V  I  L  Q  T  K 574

CTGTACACCCTGCCTGCAAGCCACGCCACGCGTGTGTATGAGATGCTGGTCAGCCACATT 1800
 L  Y  T  L  P  A  S  H  A  T  R  V  Y  E  M  L  V  S  H  I 594

CAGCTCCACTACAAGCACAGCTACACCCTGCCAATCGCGAGCAGCATCCGGCTGCAGGCC 1860
 Q  L  H  Y  K  H  S  Y  T  L  P  I  A  S  S  I  R  L  Q  A 614

TTTGACTTCCTGTTTCTGCTGCGGGCCGACTCACTGCACCGCCTGGGCCTGCCCAACAAG 1920
 F  D  F  L  F  L  L  R  A  D  S  L  H  R  L  G  L  P  N  K 634

GATGGAGTCGTGCGGTTCAGCCCCTACTGCGTCTGCGACTACATGGAGCCAGAGAGAGGC 1980
 D  G  V  V  R  F  S  P  Y  C  V  C  D  Y  M  E  P  E  R  G 654

TCTGAGAAGAAGACCAGCGGCCCCCTTTCTCCTCCCACAGGGCCTCCTGGCCCGGCGCCT 2040
 S  E  K  K  T  S  G  P  L  S  P  P  T  G  P  P  G  P  A  P 674

GCAGGCCCCGCCGTGCGGCTGGGGTCCGTGCCCTACTCCCTGCTCTTCCGCGTCCTGCTG 2100
 A  G  P  A  V  R  L  G  S  V  P  Y  S  L  L  F  R  V  L  L 694

CAGTGCTTGAAGCAGGAGTCTGACTGGAAGGTGCTGAAGCTGGTTCTGGGCAGGCTGCCT 2160
 Q  C  L  K  Q  E  S  D  W  K  V  L  K  L  V  L  G  R  L  P 714

GAGTCCCTGCGCTATAAAGTGCTCATCTTTACTTCCCCTTGCAGTGTGGACCAGCTGTGC 2220
 E  S  L  R  Y  K  V  L  I  F  T  S  P  C  S  V  D  Q  L  C 734

TCTGCTCTCTGCTCCATGCTTTCAGGCCCAAAGACACTGGAGCGGCTCCGAGGCGCCCCA 2280
 S  A  L  C  S  M  L  S  G  P  K  T  L  E  R  L  R  G  A  P 754

GAAGGCTTCTCCAGAACTGACTTGCACCTGGCCGTGGTTCCAGTGCTGACAGCATTAATC 2340
 E  G  F  S  R  T  D  L  H  L  A  V  V  P  V  L  T  A  L  I 774

TCTTACCATAACTACCTGGACAAAACCAAACAGCGCGAGATGGTCTACTGCCTGGAGCAG 2400
 S  Y  H  N  Y  L  D  K  T  K  Q  R  E  M  V  Y  C  L  E  Q 794

GGCCTCATCCACCGCTGTGCCAGACAGTGCGTCGTGGCCTTGTCCATCTGCAGCGTGGAG 2460
 G  L  I  H  R  C  A  R  Q  C  V  V  A  L  S  I  C  S  V  E 814

ATGCCTGACATCATCATCAAGGCGCTGCCTGTTCTGGTGGTGAAGCTCACGCACATCTCA 2520
 M  P  D  I  I  I  K  A  L  P  V  L  V  V  K  L  T  H  I  S 834
```

```
GCCACAGCCAGCATGGCCGTCCCACTGCTGGAGTTCCTGTCCACTCTGGCCAGGCTGCCG 2580
 A  T  A  S  M  A  V  P  L  L  E  F  L  S  T  L  A  R  L  P  854

CACCTCTACAGGAACTTTGCCGCGGAGCAGTATGCCAGTGTGTTCGCCATCTCCCTGCCG 2640
 H  L  Y  R  N  F  A  A  E  Q  Y  A  S  V  F  A  I  S  L  P  874

TACACCAACCCCTCCAAGTTTAATCAGTACATCGTGTGTCTGGCCCATCACGTCATAGCC 2700
 Y  T  N  P  S  K  F  N  Q  Y  I  V  C  L  A  H  H  V  I  A  894

ATGTGGTTCATCAGGTGCCGCCTGCCCTTCCGGAAGGATTTTGTCCCTTTCATCACTAAG 2760
 M  W  F  I  R  C  R  L  P  F  R  K  D  F  V  P  F  I  T  K  914

GGCCTGCGGTCCAATGTCCTCTTGTCTTTTGATGACACCCCCGAGAAGGACAGCTTCAGG 2820
 G  L  R  S  N  V  L  L  S  F  D  D  T  P  E  K  D  S  F  R  934
                         ▸                          ▾

GCCCGGAGTACTAGTCTCAACCAGAGACCCAAGAGTCTGAGGATAGCCAGACCCCCCAAA 2880
 A  R  S  T  S  L  N  Q  R  P  K  S  L  R  I  A  R  P  P  K  954
             ▸                    ▾

CAAGGCTTGAATAACTCTCCACCCGTGAAAGAATTCAAGGAGAGCTCTGCAGCCGAGGCC 2940
 Q  G  L  N  N  S  P  P  V  K  E  F  K  E  S  S  A  A  E  A  974
                                              ▸

TTCCGGTGCCGCAGCATCAGTCTGTCTGAACATGTGGTCCGCAGCAGGATACAGACGTCC 3000
 F  R  C  R  S  I  S  V  S  E  H  V  V  R  S  R  I  Q  T  S  994
                   ▸

CTCACCAGTGCCAGCTTGGGGTCTGCAGATGAGAACTCCGTGGCCCAGGCTGACGATAGC 3060
 L  T  S  A  S  L  G  S  A  D  E  N  S  V  A  Q  A  D  D  S  1014
                      ▸                                   ▾

CTGAAAAACCTCCACCTGGAGCTCACCGAAACCTGTCTGGACATGATGGCTCGATACGTC 3120
 L  K  N  L  H  L  E  L  T  E  T  C  L  D  M  M  A  R  Y  V  1034
                               ▸

TTCTCCAACTTCACGGCTGTCCCGAAGAGGTCTCCTGTGGGCGAGTTCCTCCTAGCGGGT 3180
 F  S  N  F  T  A  V  P  K  R  S  P  V  G  E  F  L  L  A  G  1054
       θ

GGCAGGACCAAAACCTGGCTGGTTGGGAACAAGCTTGTCACTGTGACGACAAGCGTGGGA 3240
 G  R  T  K  T  W  L  V  G  N  K  L  V  T  V  T  T  S  V  G  1074

ACCGGGACCCGGTCGTTACTACGCCTGGACTCGGGGGAGCTGCAGTCCGGCCCGGAGTCG 3300
 T  G  T  R  S  L  L  G  L  D  S  G  E  L  Q  S  G  P  E  S  1094
                                              ▸

AGCTCCAGCCCCGGGGTGCATGTGAGACAGACCAAGGAGGCGCCGGCCAAGCTGGAGTCC 3360
 S  S  S  P  G  V  H  V  R  Q  T  K  E  A  P  A  K  L  E  S  1114

CAGGCTGGGCAGCAGGTGTCCCGTGGGGCCCGGGATCGGGTCCGTTCCATGTCGGGGGGC 3420
 Q  A  G  Q  Q  V  S  R  G  A  R  D  R  V  R  S  M  S  G  G  1134

CATGGTCTTCGAGTTGGCGCCCTGGACGTGCCGGCCTCCCAGTTCCTGGGCAGTGCCACT 3480
 H  G  L  R  V  G  A  L  D  V  P  A  S  Q  F  L  G  S  A  T  1154

TCTCCAGGACCACGGACTGCACCAGCCGCGAAACCTGAGAAGGCCTCAGCTGGCACCCGG 3540
 S  P  G  P  R  T  A  P  A  A  K  P  E  K  A  S  A  G  T  R  1174

GTTCCTGTGCAGGAGAAGACGAACCTGGCGGCCTATGTGCCCCTGCTGACCCAGGGCTGG 3600
 V  P  V  Q  E  K  T  N  L  A  A  Y  V  P  L  L  T  Q  G  W  1194

GCGGAGATCCTGGTCCGGAGGCCCACAGGGAACACCAGCTGGCTGATGAGCCTGGAGAAC 3660
 A  E  I  L  V  R  R  P  T  G  N  T  S  W  L  M  S  L  E  N  1214
                         ▴     θ

CCGCTCAGCCCCTTCTCCTCGGACATCAACAACATGCCCCTGCAGGAGCTGTCTAACGCC 3720
 P  L  S  P  F  S  S  D  I  N  N  M  P  L  Q  E  L  S  N  A  1234

CTCATGGCGGCTGAGCGCTTCAAGGAGCACCGGGACACAGCCCTGTACAAGTCACTGTCG 3780
 L  M  A  A  E  F  F  K  E  H  R  D  T  A  L  Y  K  S  L  S  1254
                                              ♦

GTGCCGGCAGCCAGCACGGCCAAACCCCCTCCTCTGCCTCGCTCCAACACAGACTCCGCC 3840
 V  P  A  A  S  T  A  K  P  P  P  L  P  R  S  N  T  D  S  A  1274
                ▾                          ▸

GTGGTCATGGAGGAGGGAAGTCCGGGCGAGGTTCCTGTGCTGGTGGAGCCCCCAGGGCTG 3900
 V  V  M  E  E  C  S  P  G  E  V  P  V  L  V  E  P  P  G  L  1294
                   ▸
```

```
GAGGACGTTGAGGCAGCGCTAGGCATGGACAGGCGCACGGATGCCTACAGCAGGTCGTCC 3960
 E  D  V  E  A  A  L  G  M  D  R  R  T  D  A  Y  S  R  S  S  1314

TCAGTCTCCAGCCAGGAGGAGAAGTCGCTCCACGCGGAGGAGCTGGTTGGCAGGGGCATC 4020
 S  V  S  S  Q  E  E  K  S  L  H  A  E  E  L  V  G  R  G  I  1334
       ▸  ▸

CCCATCGAGCGAGTCGTCTCCTCGGAGGGTGGCCGGCCCTCTGTGGACCTCTCCTTCCAG 4080
 P  I  E  R  V  V  S  S  E  G  G  R  P  S  V  D  L  S  F  Q  1354

CCCTCGCAGCCCCTGAGCAAGTCCAGCTCCTCTCCCGAGCTGCAGACTCTGCAGGACATC 4140
 P  S  Q  P  L  S  K  S  S  S  S  P  E  L  Q  T  L  Q  D  I  1374
                            ▸                 ▸

CTCGGGGACCCTGGGGACAAGGCCGACGTGGGCCGGCTGAGCCCTGAGGTTAAGGCCCGG 4200
 L  G  D  P  G  D  K  A  D  V  G  R  L  S  P  E  V  K  A  R  1394

TCACAGTCAGGGACCCTGGACGGGGAAAGTGCTGCCTGGTCGGCCTCGGGCGAAGACAGT 4260
 S  Q  S  G  T  L  D  G  E  S  A  A  W  S  A  S  G  E  D  S  1414
                                              ▸

CGGGGCCAGCCCGAGGGTCCCTTGCCTTCCAGCTCCCCCCGCTCGCCCAGTGGCCTCCGG 4320
 R  G  Q  P  E  G  P  L  P  S  S  S  P  R  S  P  S  G  L  R  1434
                                  ▾

CCCCGAGGTTACACCATCTCCGACTCGGCCCCATCACGCAGGGGCAAGAGAGTAGAGAGG 4380
 P  R  G  Y  T  I  S  D  S  A  P  S  R  R  G  K  R  V  E  R  1454
             ▸                    ▾

GACGCCTTAAAGAGCAGAGCCACAGCCTCCAATGCAGAGAAAGTGCCAGGCATCAACCCC 4440
 D  A  L  K  S  R  A  T  A  S  N  A  E  K  V  P  G  I  N  P  1474
                            ▸

AGTTTCGTGTTCCTGCAGCTCTACCATTCCCCCTTCTTTGGCGACGAGTCAAACAAGCCA 4500
 S  F  V  F  L  Q  L  Y  H  S  P  F  F  G  D  E  S  N  K  P  1494
                                                 ▾

ATCCTGCTGCCCAATGAGTCACAGTCCTTTGAGCGGTCGGTGCAGCTCCTCGACCAGATC 4560
 I  L  L  P  N  E  S  Q  S  F  E  R  S  V  Q  L  L  D  Q  I  1514
             θ

CCATCATACGACACCCACAAGATCGCCGTCCTGTATGTTGGAGAAGGCCAGAGCAACAGC 4620
 P  S  Y  D  T  H  K  I  A  V  L  Y  V  G  E  G  Q  S  N  S  1534
             ▾                                      ▸

GAGCTCGCCATCCTGTCCAATGAGCATGGCTCCTACAGGTACACGGAGTTCCTGACGGGC 4680
 E  L  A  I  L  S  N  E  H  G  S  Y  R  Y  T  E  F  L  T  G  1554
                               ♦

CTGGGCCGGCTCATCGAGCTGAAGGACTGCCAGCCGGACAAGGTGTACCTGGGAGGCCTG 4740
 L  G  R  L  I  E  L  K  D  C  Q  P  D  K  V  Y  L  G  G  L  1574

GACGTGTGTGGTGAGGACGGCCAGTTCACCTACTGCTGGCACGATGACATCATGCAAGCC 4800
 D  V  C  G  E  D  G  Q  F  T  Y  C  W  H  D  D  I  M  Q  A  1594
                                                     xxxxxx

GTCTTCCACATCGCCACCCTGATGCCCACCAAGGACGTGGACAAGCACCGCTGCGACAAG 4860
 V  F  H  I  A  T  L  M  P  T  K  D  V  D  K  H  R  C  D  K  1614
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

AAGCGCCACCTGGGCAACGACTTTGTGTCCATTGTCTACAATGACTCCGGTGAGGACTTC 4920
 K  R  H  L  G  N  D  F  V  S  I  V  Y  N  D  S  G  E  D  F  1634
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxθxxxxx▸xxxx

AAGCTTGGCACCATCAAGGGCCAGTTCAACTTTGTCCACGTGATCGTCACCCCGCTGGAC 4980
 K  L  G  T  I  K  G  Q  F  N  F  V  H  V  I  V  T  P  L  D  1654
          ▾                                      ▸

TACGAGTGCAACCTGGTGTCCCTGCAGTGCAGGAAAGACATGGAGGGCCTTGTGGACACC 5040
 Y  E  C  N  L  V  S  L  Q  C  R  K  D  M  E  G  L  V  D  T  1674

AGCGTGGCCAAGATCGTGTCTGACCGCAACCTGCCCTTCGTGGCCCGCCAGATGGCCCTG 5100
 S  V  A  K  I  V  S  D  R  N  L  P  F  V  A  R  Q  M  A  L  1694
                   ▾

CACGCAAATATGGCCTCACAGGTGCATCATAGCCGCTCCAACCCCACCGATATCTACCCC 5160
 H  A  N  M  A  S  Q  V  H  H  S  R  S  N  P  T  D  I  Y  P  1714

TCCAAGTGGATTGCCCGGCTCCGCCACATCAAGCGGCTCCGCCAGCGGATCTGCGAGGAA 5220
 S  K  W  I  A  R  L  R  H  I  K  R  L  R  Q  R  I  C  E  E  1734

GCCGCCTACTCCAACCCCAGCCTACCTCTGGTGCACCCTCCGTCCCATAGCAAAGCCCCT 5280
 A  A  Y  S  N  P  S  L  P  L  V  H  P  P  S  H  S  K  A  P  1754
       ♦
```

```
GCACAGACTCCAGCCGAGCCCACACCTGGCTATGAGGTGGGCCAGCGGAAGCGCCTCATC 5340
 A  Q  T  P  A  E  P  T  P  G  Y  E  V  G  Q  R  K  R  L  I  1774
       ►
TCCTCGGTGGAGGACTTCACCGAGTTTGTGTGAGGCCGGGCCCCTCCCTCCTGCACTGGC 5400
 S  S  V  E  D  F  T  E  F  V  *                                1794
 ►  ►
CTTGGACGGTATTGCCTGTCAGTGAAATAAATAAAGTCCTGACCCCAGTGCACAGACATA 5460
GAGGCACAGATTGC                                               5474
```

Figure 5.12: The nucleotide sequence of TSC2. The predicted protein is shown below the DNA sequence, assuming that translation begins at the first in-frame methionine of the long open reading frame. The 'xxxxxxxx' bar denotes the GAP3 related domain (aa. 1593-1631). The double underlining marks the possible membrane spanning regions. The dotted line indicates a a potential leucine zipper starting at aa. 81. 'r_______r' indicates the repeated motif H A V E/L A L W/L K A at aa. 76-84 and 99-107. Possible N-linked glycosylation sites (@) are marked at aa. 1037, 1205, 1499, and 1628. Serine (S) and threonine (T) residues that are potentially phosphorylated by cAMP- and cGMP- dependent protein kinases (▲) (Glass et al., 1986), protein kinase C (►) (Woodgett et al.,1986), or casein kinase 2 (▼) (Pinna, 1990), and possible tyrosine (Y) kinase phosphorylation sites (#) (Patschinsky et al., 1982) are indicated. Two potential polyadenylation signals at bases 5425 and 5429 (underlined) are indicated. base.
Homo Sapiens tuberin mRNA sequence has been deposited in the EMBL database, accession number X75621.



```
              *   * * * * * * * * *     *   *         * * * * * * * * *   * * * * * * *   *
tuberin       Q A V F H I A T L M P T K D V D K H R C D K K R H L G N D F V S I V Y N D S G
GAP3 (human)  E I M F H V S T K L P Y T E G D A Q Q L Q R K R H I G N D I V A V V F Q D E N
GAP3 (mouse)  E I M F H V S T M L P Y T P N N Q Q Q L L R K R H I G N D I V T I V F Q E P G
```

Figure 5.13. The homology between the predicted protein sequence of tuberin (aa. 1593-1631) and N-terminal domains of human GAP3 and murine GAP3. Identical amino acids are boxed. Asterisks indicate identical, or interchangeable amino acids, which are shared between tuberin and at least one of the GAP proteins. Interchangeable amino acids were identified using the criteria of Dayhoff et al. (1978).

Discussion

We have used a positional cloning strategy to identify a gene on chromosome 16 which is mutated in tuberous sclerosis. A number of questions concerning the biology of TSC and its relationship to other disorders can now be addressed.

The TSC2 gene maps within the candidate region for the unidentified PKD1 gene, causing autosomal dominant polycystic kidney disease type 1 (ADPKD1), as defined by Germino et al. (1992). As polycystic kidneys are a feature common to TSC and ADPKD1 (Bernstein and Robbins, 1991) the possibility of an aetiological link, as proposed by Kandt et al. (1992), must be considered. Renal cysts, however, have been reported in a chromosome 9-linked TSC family (Nellist et al., 1993) and their presence is therefore not limited to chromosome 16-linked TSC. Furthermore, while TSC and ADPKD1 cysts are macroscopically similar, the epithelium lining TSC associated cysts is usually considered to be histologically distinct (Bernstein et al., 1974). Despite these observations it may be tempting to hypothesise that chromosome 16-linked forms of TSC and ADPKD1 are allelic variants. However, we have not found any evidence that this is the case.

The search for possible functional motifs in the sequence of the predicted protein, which we have called tuberin, indicates several regions of interest. Four hydrophobic domains were identified which may be involved in membrane anchorage and four potential glycosylation sites were observed downstream of the last putative transmembrane domain. No sequence at the amino- terminus of the predicted protein matched the signal peptide structure as defined by von Heijne (1985). However, the occurrence of several transmembrane domains without an apparent signal peptide was noted in the cystic fibrosis-related protein CFTR (Riordan et al., 1989). We also noted a periodic array of leucine residues (leucine zipper), a structure associated with protein-protein interaction. Experiments are in progress to determine the cellular localisation of tuberin, which will provide insight into the functional significance of the sequence motifs that have been identified.

Because of the highly variable TSC phenotype, the genetic status of a patient's relatives may remain uncertain, even after extensive diagnostic investigation (Al-Gazali et al., 1989; Fryer et al., 1990). In this situation the identification of the causative mutation would be very helpful. Although a relatively small number of mutations are reported in this study, alternative approaches such as SSCP analysis (Orita et al., 1989) can be applied now that the TSC2 gene sequence is available. Identification of the TSC1 gene on chromosome 9 will also have to be achieved before the full mutational spectrum in TSC and the practicalities of DNA based diagnostics can be completely evaluated.

We have identified multiple deletional mutations affecting different parts of the TSC2 gene in unrelated TSC patients. This pattern, and the reduced expression of TSC2 seen in affected individuals, suggest that constitutional mutations in TSC are likely to be inactivating. The patchy focal nature of TSC associated lesions and the loss of heterozygosity which they exhibit (Green and Yates, 1993) suggest that reduction to the homozygous null state is required before cellular growth and differentiation become disordered. A similar combination of inactivating constitutional and somatic mutations have been clearly demonstrated in the Rb gene in retinoblastoma (Horowitz et al., 1989) and more recently the NF1 gene in neurofibrosarcoma (Legius et al., 1993); it has also been proposed in NF2 (Rouleau et al., 1993) and VHL (Latif et al., 1993). It would seem likely, therefore, that TSC2 also behaves as a tumour suppressor gene as defined by Knudson's theory of carcinogenesis (Knudson, 1971).

Neurofibromin, the product of the NF1 gene, has been shown to be homologous to rasGAP (Xu et al., 1990) and to regulate ras activity (Martin et al., 1990). Ras is a member of a group of homologous GTPases involved in regulation of cell proliferation and differentiation. Since rap1 is another member of this group, it is interesting to note that tuberin shows a small region of homology with rap1GAP (GAP3). Although tuberin shows no homology with neurofibromin or rasGAP, these proteins may be part of the same cellular process. It is possible that they have analogous roles in distinct signalling pathways. Alternatively, tuberin and neurofibromin may act via a common effector, such as the rap1 protein (p21rap1). A model involving neurofibromin-p21rap1 binding has been proposed by Martin et al. (1990) and tuberin may fit into this. However, it should be stressed that these considerations are purely hypothetical. For instance neurofibromin has never been tested for p21rap1 binding. Moreover, the region of tuberin-GAP3 homology is contained within the catalytic domain of GAP3 (Rubinfeld et al., 1992), but it is not known whether this particular region is required for catalytic activity. Further experimental data are needed to establish the significance of the regional homology between tuberin and GAP3.

As the proteins involved in the various phakomatoses are identified, their functions and possible inter-relationships will be established. In particular, it will be interesting to discover the role which tuberin plays in cellular proliferation and differentiation and how this relates to the product of the unidentified TSC1 gene in determining the varied phenotype of TSC.

Experimental Procedures

Pulsed field electrophoresis

High molecular weight DNA was isolated from peripheral blood in agarose plugs by standard methods (Hermann et al., 1987) and digested according to the manufacturers recommendations. Blocks were loaded into the wells of 1% agarose gels and electrophoresis carried out using a BioRad CHEF DR II or a similar apparatus and programs appropriate to the varying resolutions required.

Southern blot analysis

Genomic DNA was extracted from peripheral blood by standard methods. 5-8μg DNA was digested with restriction enzymes, electrophoresed through agarose gels and blotted to nylon filters as described (Sambrook et al., 1989). Probes were labelled by the random-primer method (Feinberg and Vogelstein, 1984). For probes containing repetitive elements, 10ng of labelled DNA was pre-associated with 0.1-1mg denatured sonicated total human DNA in a total volume of 200μl at 65°C for 1-5 hr. prior to hybridisation. If required filters were additionally prehybridized with 100μg/ml denatured sonicated total human DNA and salmon sperm DNA. Filters were hybridised, washed as described (Sambrook et al., 1989) and exposed to autoradiographic film with an intensifying screen at -70°C.

DNA probes and somatic cell hybrids

Some of the probes used in this study have been described previously: MS205.2 (D16S309; Royle et al., 1992); GGG1 (D16S259; Germino et al., 1990); 16AC2.5 (D16S291; Thompson et al., 1992) and N54 (D16S139; Himmelbauer et al., 1991). A number of new probes were also isolated during the course of this study: SM6, a 2.3kb Sau3A fragment from SMII; BFS2, a 1.8kb BssHII fragment of CC1-2; SM9, a 7kb EcoRI fragment of CBFS1; CW9 a 1kb EcoRI/NotI segment of CBFS1; CW15 a 10kb EcoRI/NotI fragment of CW9D; CW24 and CW26 are 0.9kb and 0.4kb, SacII and SacII/SacI fragments, respectively, of CW9D; CW13 and CW12 are EcoRI/NotI fragments of 2.2kb and 2.0kb, respectively, from CW9D; CW18, CW20 are EcoRI/NotI fragments of 3kb and 16kb respectively from CW12I, JH1, a 4.4kb BamHI fragment of CW12I; and CW23 and CW21 are 14kb and 3.5kb NotI/EcoRI fragments, respectively, of JHIK. All new probes except SM6, BFS2, CW26 and CW21 contain repetitive sequences and were hybridised in the presence of denatured, sonicated human DNA (75μg/ml) and washed in 0.05 x SSC, 0.2% SDS at 65°C.

The somatic cell hybrid N-OH1 and the radiation hybrid, Hy145.19 have been described previously (Germino et al, 1990.; Himmelbauer et al, 1991). The P-MWH2A hybrid contains the derivative chromosome 16qter→16p13.3::7q32→7qter and was isolated from a subject, MW, who has a balanced translocation. P-MWH2A was produced by fusing lymphoblastoid cells from MW with APRT deficient mouse erythroleukemia cells by the method of Deisseroth and Hendrick (1979). The breakpoint in this hybrid has been localised to the region between 16AC2.5 and the adjacent ClaI site (see Figure 5.1).

RNA isolation and Northern blot analysis

RNA was extracted from cell-lines and tissues by the acid phenol method of Chomczynski and Sacchi (1987). mRNA was isolated from total RNA using a biotinylated oligo (dT) primer and streptavidin coupled paramagnetic particles (PolyATtract mRNA Isolation System, Promega). RNA was separated in denaturing formaldehyde gels and Northern blotted by standard procedures. Hybridisation and washing of Northern blots was as described for Southerns.

Cosmid walking

Cosmids were obtained from several different libraries: Los Alamos Chromosome 16 specific library (Stallings, et al. 1990) and total genomic cosmid libraries 412 and IG328 (Integrated Genetics) and 961200 (Stratagene). Successive cosmid walks were made by mapping each cosmid, isolating end clones and rehybridising the libraries using conditions to repress repetitive sequences if necessary. A cosmid/genomic EcoRI map was produced and the location of cosmids was checked by mapping on hybrid panels, PFGE and fluorescence in situ hybridisation.

cDNA isolation and characterisation

Screening for cDNAs was performed using standard phage plating, filter lift and clone purification techniques in commercial libraries derived from human fetal brain (Clonetech, Stratagene) and human adult kidney (Clonetech). Filters were lifted as described by Sambrook (Sambrook, 1989). Repetitive sequences were suppressed as described above. After overnight hybridisation at 65°C, filters were washed as described (Sambrook, 1989). All positive clones were subcloned into one of the pBluescript or pUC vectors and sequenced with a Pharmacia A.L.F. or ABI model 373A automated sequencer according to the manufacturers protocol, or manually.

Acknowledgements

We particularly thank Dr Douglas Higgs and Dr T. Mazurczak. Dr A. Krause and Prof T. Jenkins are thanked for arranging reevaluation of patient BO, Russell Snell, Andrew Wilkie and A.J.J. Reuser for helpful discussions. We are grateful to Peter Thompson, for undertaking cytogenetic analyses. We acknowledge Prof P. Harper, Prof. H. Galjaard, Prof. Sir David Weatherall, and Dr. D. Shaw for their support, J. Myring, J. Sloane-Stanley, I. Ceccherini, P. Vyas, D. Blake, R. Boone, Y-G Wu, S. Ramlakhan, M. van der Kraan, P. Simons and J. Pertijs for technical assistance and advice, and Alan Fryer, Stuart Gould, Alan Harris, Charles Shepherd, Prof. H.F.M. Busch, M Wildervanck de Blécourt-Devilee, P. Willems, L. Sunde, J. Zaremba, P. Fleury, E.C. Merkens, J.A. Maat-Kievit, R. Riikonen, Prof. J.J. Cassiman, G. Holmgren, E. Smeets, Prof. J.P. Fryns, Prof. M.F. Niermeijer, C.T.R.M. Schrander-Stumpel, J. Hulsman, H. Stroink, L.A.E.M. Laan, and A.P. Oranje for the clinical help, which was essential with respect to mutation screening. We thank the many other people who have played a role in sustaining this study. This work was financially supported by the Tuberous Sclerosis Association (GB), Action Research, Dutch Kidney Foundation, NWO, the Prevention Fund (Netherlands), the Medical Research Council, Wellcome Trust and the University of Wales. We wish to thank the patients and their organisation (the STSN) for their kind cooperation.

References

Al-Gazali LI, Arthur RJ, Lamb JT, et al. Diagnostic and counselling difficulties using a fully comprehensive screening protocol for families at risk for tuberous sclerosis. J Med Genet 1989; 26: 694-703.

Bernstein J, Brough AJ, McAdams AJ. The renal lesion syndromes of multiple congenital malformations: cerebrohepatorenal syndrome; Jeune asphyxiating thoracic dystrophy; tuberous sclerosis; Meckel syndrome. Birth Defects: Original Article Series 1974; 10: 35-43.

Bernstein J, Robbins TO. Renal involvement in tuberous sclerosis. Ann N Y Acad Sci 1991; 615: 36-49.

Chomczynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem 1987; 162: 156-159.

Dayhoff MO, Schwatz RM, Orcutt BC. In Atlas of protein sequence and structure. 1978 Vol 5 Suppl. 3. Dayhoff MO ed. (Washington: NBRF) pp.345.

Deisseroth A, Hendrick D. Activation of phenotypic expression of human globin genes from non-erythriod cells by a chromosome-dependent transfer to tetraploid mouse erythroleukaemia cells. Proc Natl Acad Sci USA 1979; 76: 2185- 2189.

Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol 1984; 179: 125-142.

Feinberg AP, Vogelstein B. Addendum: a technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 1984; 137: 266-267.

Fryer AE, Chalmers A, Connor JM, et al. Evidence that the gene for tuberous sclerosis is on chromosome 9. Lancet 1987; i: 659-661.

Fryer AE, Chalmers AH, Osborne JP. The value of investigation for genetic counselling in tuberous sclerosis. J Med Genet 1990; 27: 217-223.

Germino GG, Barton NJ, Lamb J, et al. Identification of a locus which shows no genetic recombination with the autosomal dominant polycystic kidney disease gene on chromosome 16. Am J Hum Genet 1990; 46: 925-933.

Germino GG, Weinstat-Saslow D, Himmelbauer H, et al. The gene for autosomal dominant polycystic kidney disease lies in a 750-kb CpG-rich region. Genomics 1992; 13 144-151.

Germino GG, Somlo S, Weinstat-Saslow D, Reeders ST. Positional cloning approach to the dominant polycystic kidney disease gene PKD1. Kidney International 1993; 43 Suppl. 39: 20-25.

Glass DB, El-Maghrabi MR, Pilkis SJ. Synthetic peptides corresponding to the site phosphorylated in 6-phosphofructo-2-kinase / fructose-2,6-biphosphatase as substrates of cyclic nucleotide-dependent protein kinases. J Biol Chem 1986; 261: 2987-2993.

Gomez MR. Tuberous Sclerosis, 2nd edition (Raven Press New York, 1988).

Green AJ, Yates JRW. Loss of heterozygosity on chromosome 16p in hamartomata from patients with tuberous sclerosis. Am J Hum Genet 1993; 53 Suppl.: 244.

Haines JL, Amos J, Attwood J, et al. Genetic heterogeneity in tuberous sclerosis: study of a large collaborative dataset. Ann N Y Acad Sci 1991a; 615: 256-264.

Haines JL, Short MP, Kwiatkowski DJ, et al. Localization of one gene for tuberous sclerosis within 9q32-9q34 and further evidence for heterogeneity. Am J Hum Genet 1991b; 49: 764-772.

Harris PC, Barton NJ, Higgs DR, et al. A long-range restriction map between the α- globin complex and a marker closely linked to the polycystic kidney disease 1 (PKD1). Genomics 1990; 7: 195-206.

Hermann BG, Barlow DP, Lehrach H. A large inverted duplication allows homologous recombination between chromosomes heterozygous for the proximal t complex inversion. Cell 1987; 48: 813-825.

Himmelbauer H, Germino GG, Ceccherini I, et al. Saturating the region of the polycystic kidney disease gene with NotI linking clones. Am J Hum Genet 1991; 48: 325-334.

Horowitz JM, Yandell DW, Park SH, et al. Point mutational inactivation of the retinoblastoma antioncogene. Science 1989; 243: 937-940.

Kandt RS, Haines JL, Smith M, et al. Linkage of an important gene locus for tuberous sclerosis to a chromosome 16 marker for polycystic kidney disease. Nature Genet 1992; 2: 37-41.

Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci USA 1971; 68: 820-823.

Kozak M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucl Acids Res 1987; 15: 8125-8148.

Kwiatkowski DJ, Armour J, Bale AE, et al. Report on the second international workshop on human chromosome 9. Cytogenet Cell Genet 1993; 64: 94-106.

Landschulz WH, Johnson PF, McKnight SL. The leucine zipper: A hypothetical structure common to a new class of DNA binding proteins. Science 1988; 240: 1759-1764.

Latif F, Tory K, Gnarra J, et al. Identification of the von Hippel-Lindau disease tumor suppressor gene. Science 1993; 260: 1317-1320.

Legius E, Marchuk DA, Collins FS, Glover TW. Somatic deletion of the neurofibromatosis type 1 gene in a neurofibrosarcoma supports a tumour suppressor gene hypothesis. Nature Genet 1993; 3: 122-126.

Martin GA, Viskochil D, Bollag G, et al. The GAP- related domain of the neurofibromatosis type 1 gene product interacts with ras p21. Cell 1990; 63: 843-849.

Nellist M, Brook-Carter PT, Connor JM, et al. Identification of markers flanking the tuberous sclerosis locus on chromosome 9 (TSC1). J Med Genet 1993; 30: 224-227.

Northrup H, Kwiatkowski DJ, Roach ES, et al. Evidence for genetic heterogeneity in tuberous sclerosis: one locus on chromosome 9 and at least one locus elsewhere. Am J Hum Genet 1992; 51: 709-720.

Orita M, Iwahana H, Kanazawa H, et al. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc Natl Acad Sci USA 1989; 86: 2766-2770.

Osborne JP, Fryer A, Webb D. Epidemiology of tuberous sclerosis. Ann NY Acad Sci 1991; 615: 125-127.

Patschinsky T, Hunter T, Esch FS, et al. Analysis of the sequence of amino acids surrounding sites of tyrosine phosphorylation. Proc Natl Acad Sci USA 1982; 79: 973-977.

Pinna LA. Casein kinase 2: an 'eminence grise' in cellular regulation? Biochim Biophys Acta 1990; 1054: 267-284.

Povey S, Attwood J, Janssen LAJ, et al. An attempt to map two genes for tuberous sclerosis using novel two-point methods. Ann N Y Acad Sci 1991; 615: 298-305.

Rack KA, Harris PC, MacCarthy AB, et al. Characterization of three de novo derivative chromosomes 16 by "reverse chromosome painting" and molecular analysis. Am J Hum Genet 1993; 52: 987-997.

Riordan JR, Rommens JM, Kerem B-S, et al. Identification of the cystic fibrosis gene: cloning and characterisation of complimentary DNA. Science 1989; 245: 1066-1072.

Rouleau GA, Merel P, Lutchman M, et al. Alteration in a new gene encoding a putative membrane-organizing protein causes neuro-fibromatosis type 2. Nature 1993; 363: 515-521.

Royle NJ, Armour JA, Webb M, et al. A hypervariable locus D16S309 located at the distal end of 16p. Nucl Acids Res 1992; 20: 1164.

Rubinfeld B, Munemitsu S, Clark R, et al. Molecular cloning of a GTPase activating protein specific for the Krev-1 protein p21rap1. Cell 1991; 65: 1033-1042.

Rubinfeld B, Crosier WJ, Albert I, et al. Localisation of the rap1GAP catalytic domain and sites of phosphorylation by mutational analysis. Mol Cell Biol 1992; 12: 4634-4642.

Sambrook J, Fritsch EF, Maniatis T. Molecular Cloning: A Laboratory Manual Second Edition (Cold Spring Harbor New York: Cold Spring Harbour Laboratory Press, 1989).

Sampson JR, Yates JRW, Pirrit LA, et al. Evidence for genetic heterogeneity in tuberous sclerosis. J Med Genet 1989a; 26: 511-516.

Sampson JR, Scahill SJ, Stephenson JBP, et al. Genetic aspects of tuberous sclerosis in the west of Scotland. J Med Genet 1989b; 26: 28-31.

Sander C, Schneider R. Database of homology- derived protein structures and the structural meaning of sequence alignment. Proteins 1991; 9: 56-68.

Smith M, Handa K, He W, Spear G. Loss of heterozygosity for chromosome 16p13.3 markers in renal hamartomas from tuberous sclerosis patients. Am J Hum Genet 1993; 53 Suppl: 366.

Stallings RL, Torney DC, Hildebrand CE, et al. Physical mapping of human chromosomes by repetitive sequence fingerprinting. Proc Natl Acad Sci USA 1990; 87: 6218-6222.

Thompson AD, Shen Y, Holman K, et al. Isolation and characterisation of $(AC)_n$ microsatellite genetic markers from human chromosome 16. Genomics 1992; 13: 402-408.

Trofatter JA, MacCollin MM, Rutter JL, et al. A novel moesin- ezrin- radixin-like gene is a candidate for the neurofibromatosis 2 tumor suppressor. Cell 1993; 72: 791-800.

Van der Hoeve J. Les phakomatoses de Bourneville de Recklinghausen et de von Hippel-Lindau. J Belge Neurol Psychiat 1933; 33: 752-762.

Viskochil D, Buchberg AM, Xu G, et al. Deletions and a translocation interrupt a cloned gene at the neurofibromatosis type 1 locus. Cell 1990; 62: 187-192.

Von Heijne G. Signal sequences. The limits of variation. J Mol Biol 1985; 184: 99-105.

Wilkie AOM, Buckle VJ, Harris PC, et al. Clinical features and molecular analysis of the α thalassemia/mental retardation syndromes. I. Cases due to deletions involving chromosome band 16p13.3. Am J Hum Genet 1990; 46: 1112-1126.

Woodgett JR, Gould KL, Hunter T. Substrate specificity of protein kinase C. Use of synthetic peptides corresponding to physiological sites as probes for substrate recognition requirements. Eur J Biochem 1986; 161: 177-184.

Xu G, O'Connell, P Viskochil, et al. The neurofibromatosis type 1 gene encodes a protein related to GAP. Cell 1990; 62: 599-608.

CHAPTER 6

DISCUSSION

6.1

LANDMARKS IN TSC MAPPING STUDIES: HOW MANY GENES CAN WE MAP BY LINKAGE ANALYSIS?

Following initial reports on the provisional mapping of TSC to chromosome 9 (Fryer et al., 1987), we have mapped the first gene responsible for TSC (TSC1) to the ABO region on 9q34 and refined the localisation by means of multipoint linkage strategies that allowed for locus heterogeneity. The various steps in this delineation process are summarised in Figure 6.1. The first landmark was the linkage report by Fryer et al (1987) and the last was the assignment of TSC1 to a 5 cM region between ABL and ABO, as described in Chapter 3.3. An even smaller TSC1 interval can be concluded by considering the overlaps between the various reported candidate regions. The genetic size of this consensus interval is surprisingly small: 1 cM. The TSC1 locus is involved in 45-65% of all familial cases (Chapter 3). Although most linkage information on the TSC1 map position seems quite straightforward, two apparent exceptions were presented at the Third Chromosome 9 Workshop. The first family reported showed a recombination with DBH and markers proximal to DBH in a patient whose diagnosis did not quite fulfil the Gomez criteria (Table 1.2) (Sampson and Harris, 1994). A French multigeneration family, however, which disease was highly likely to be linked to chromosome 9, showed a recombination in an affected individual, placing the TSC1 gene distal to D9S10 (Pitiot et al., 1994).

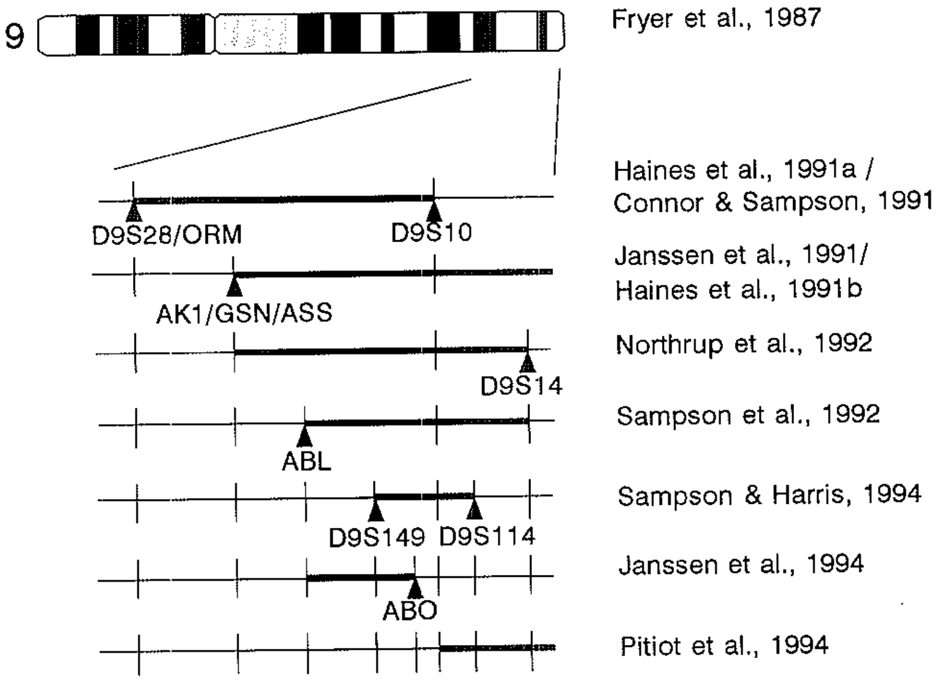


Figure 6.1: The history of TSC1 gene mapping (maps are not drawn to scale). The gene was first assigned to chromosome 9 by Fryer et al. (1987). Thick bars indicate the TSC1 candidate regions as identified or described by Haines et al. (1991a), Connor and Sampson (1991), Janssen et al. (Chapter 3.1), Haines et al. (1991b), Northrup et al. (1992), Sampson et al. (Chapter 3.2), Sampson and Harris (1994), Janssen et al. (Chapter 3.3) and Pitiot et al. (1994). It should be noted that the smallest region of overlap, between D9S149 and ABO, is not supported by the French data, indicated at the bottom.

In addition to the chromosome 9q34 area, regions on chromosomes 11, 12 and 14 have been proposed as sites for other major TSC loci. The search for the TSC2 gene locus is summarised in Figure 6.2. Linkage to chromosome 11 was proposed by the end of 1988, when an affected child was reported with a partial trisomy of chromosome 11q, due to a t(11;22)(q23.3;q11.2) translocation (Clark et al., 1988). The translocation breakpoint was mapped to 11q23.3, near the NCAM locus. A subsequent publication reported on linkage to markers from this region (Smith et al., 1990) and the NCAM gene, encoding a neural cell adhesion molecule, was considered to be a good candidate for TSC2. Initial studies in Rotterdam revealed positive lod scores for the same 11q23.3 markers in a small subset of Dutch families (Janssen et al., 1990). However, significance thresholds could not be reached. A collaborative study, on 128 families contributed by groups from the United States, Great Britain and The Netherlands, could not confirm the chromosome 11 mapping data. Results obtained from a recent re-evaluation of the collaborative linkage data, aimed at excluding linkage to chromosome 11 in certain proportions of the family material, are shown in Figure 2.3. These data indicate that chromosome 11q does not harbour a major gene responsible for tuberous sclerosis (TSC2).

In 1990, suggestive, but insignificant results, focused the attention to chromosome 14. A positive lod score of 1.98, as published by Kandt et al., indicated that 14q might be of interest (Kandt et al., 1991) However, when the family material was analyzed with micro-satellite markers, the positive evidence for linkage disappeared almost completely (M. Pericak-Vance, pers. comm.).

Another possible candidate region has been proposed by Fahsold and his German colleagues in 1991 (Fahsold et al., 1991 a,b). Following the observation of a *de novo* TSC patient with a *de novo* reciprocal translocation t(3;12)(p26.3;q23.3), a linkage study was performed, resulting in a maximum multipoint lod score of 4.56 at the PAH locus on chromosome 12q22-24.1. This finding prompted an extension of the collaborative effort by including chromosome 12q in the linkage studies on the 128 families. The results, schematically represented in Figure 3.3, do not show significant evidence for a TSC locus on chromosome 12. In Chapter 2.2 and 3.2 we concluded that a major TSC locus (TSC2) cannot be assigned to chromosome 12q, although a locus of minor importance between D12S7 and PAH may not be excluded.

In 1992 Kandt and coworkers reported linkage of a second major TSC locus (TSC2) to chromosome 16 (Kandt et al., 1992). This finding was soon followed by the isolation of the responsible gene by the European consortium (Chapter 5). Extremely helpful to the isolation of the gene was the information obtained from family 2338 segregating a t(16;22)(p13.3;q11.21) translocation. This family (Figure 6.3), brought to our attention in 1992, is of Portuguese origin. The karyotype with the translocation has already been mentioned in Chapter 1.3 (Figure 1.5). Intensive clinical and cytogenetic investigations led to a double conclusion with

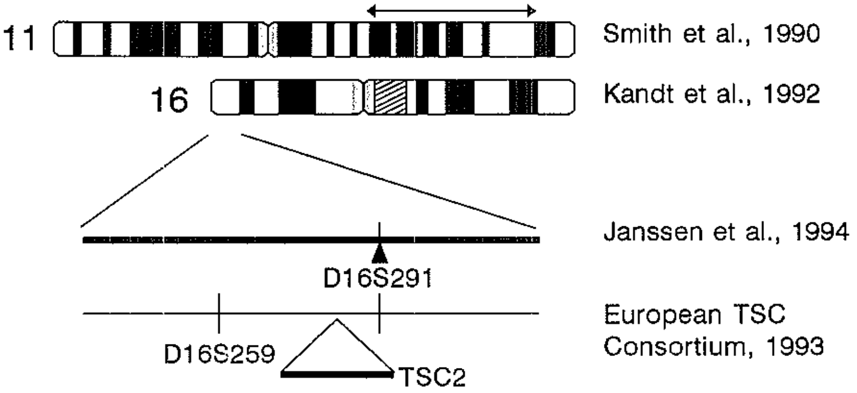


Figure 6.2: The history of TSC2 gene mapping (maps are not drawn to scale). Provisional gene assignments were followed by an assignment to chromosome 16 (Kandt et al. (1991)). This assignment was confirmed by Janssen et al (Chapter 3.3) and the gene was identified by the European TSC Consortium (Chapter 5).

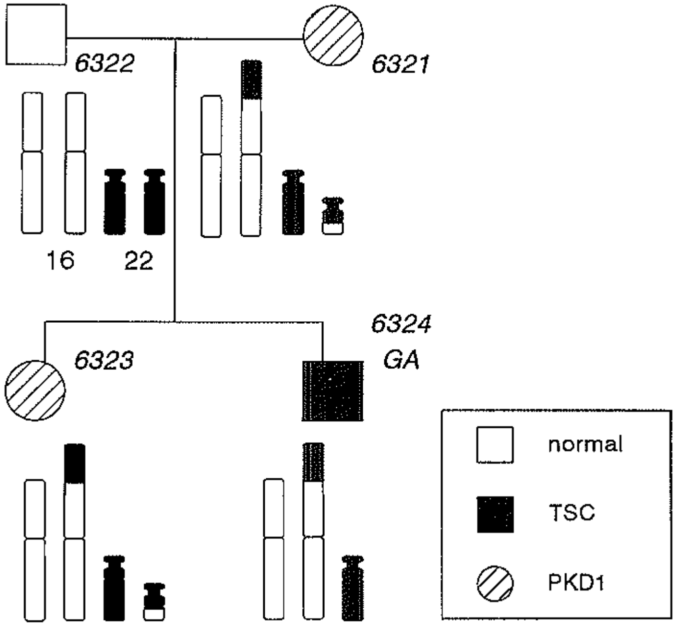


Figure 6.3: Pedigree of family 2338, which segregates a t(16;22)(p13.3;q11.21) translocation, showing the karyotype of each subject. TSC only occurs in the youngest child, carrying the unbalanced form of the translocation and who is therefore monosomic for 16p13.3-16pter and 22q11.21-22pter.

respect to TSC. First it was noted that the TSC phenotype only occurred in the unbalanced form of the translocation, with monosomy for a part of 16p13.3 and 22q. This suggested a pathogenetic mechanism: the presence of one instead of two functional copies of the TSC2 gene leads to the disease. The second conclusion was that the gene maps between the breakpoint and the telomere of chromosome 16. The area was further refined in a comparative study, involving previously reported 16p deletions. All other terminal deletions were smaller and not associated with TSC. This placed the TSC2 gene between the 'NOH' breakpoint, which is the most distal breakpoint reported in patients with the α-thalassaemia/mental retardation syndrome, and the Portuguese breakpoint, which was designated PBP and localised slightly distal to N54 (Chapter 5).

It is doubtful whether the rapid identification of the TSC2 gene, as described in Chapter 5, would have happened without the cooperation of family 2338. On the other hand, one should not overestimate the importance of this piece of evidence. Other lines of evidence, like the linkage data, have also been important. By 1992 five different cytogenetic abnormalities associated with TSC had been brought to our attention. In addition to the t(16;22)(p13.3;q11.21), the partial trisomy 11 and the t(3;12)(p26.3;q23.3) translocation mentioned above, a trisomy 10 (Simon et al., 1979) and a t(16;17)(q24;q24) translocation (Van Hemel, 1993, unpublished results) with a breakpoint in the long arm of chromosome 16 had to be considered. The linkage results were used to select the 16p region out of the 8 chromosomal regions involved in the cytogenetic abnormalities.

In 1993 the requirement of two functional TSC2 alleles (first conclusion) was supported by 'loss of heterozygosity' (LOH) studies in tumours from TSC patients (Chapters 6.2 and 6.6). Allelic loss was found in three angiomyolipomas, a cardiac rhabdomyoma, a cortical tuber and a giant cell astrocytoma (Green et al., 1994). The map position of TSC2 (second conclusion) could soon be narrowed down by a more precise localisation based on PFGE data (Chapter 5). We may therefore conclude that a combination of efforts and techniques led to the localisation and eventually the isolation of the TSC2 gene. This combination involved linkage analysis in TSC families, cytogenetic screening of individual patients, LOH studies on TSC specific tumours and PFGE analysis on *de novo* patients and their parents.

As the TSC1 and TSC2 genes have been mapped to chromosomes 9 and 16 respectively, the question arises whether a third (TSC3) gene might exist. Currently there is no evidence for further heterogeneity in TSC. The work described in

Chapter 3.3 involved a HOMOG3 analysis aimed at determining the proportion of families unlinked to both chromosomes 9 and 16. The peak lod score was obtained at $\alpha_{(TSC3)}$=0, indicating that all 14 families could be assigned to either chromosome 9 or chromosome 16. Analysis of chromosome 16 markers in families from the collaborative set revealed that several families that contributed previously to the support for a chromosome 12 locus, where in fact chromosome 16 linked (R. Fahsold, pers. comm.). Although no evidence for a third locus has been obtained, it is not possible to exclude the existence of a third locus. This is a very general and a purely methodological limitation and therefore not based on the observed segregation patterns (Chapter 2.1). If a third locus exists, it is responsible for the disease in only a very small proportion of TSC families. Even a minor third locus could, however, be relevant for the localisation of the TSC1 gene. In case the map position of a putative TSC3 locus will be found, it will perhaps resolve the discrepancy between the French TSC1 mapping data (Pitiot et al., 1994) and the results obtained by other studies (Nellist et al., 1993, Chapter 3.3). Furthermore this hypothesis might explain the segregation patterns in three recently published British families, which could not be assigned unequivocally to chromosome 9q34 or 16p13.3 (Povey et al., 1994). A putative TSC3 locus, will not only be relevant to the mapping of TSC1, it will also have an impact on genetic counselling, as discussed in Chapter 6.4.

Assuming the existence of a TSC3 locus, its putative location is currently unknown. Unfortunately, the power study on the family material presently available around the world, revealed that the chances of localising TSC3 by means of linkage analysis are extremely small (Chapter 2.2). The best strategy would be to map and clone the breakpoints of the translocations mentioned above and to study genes in the vicinity of the breakpoints with respect to a possible role in TSC. Future reports on chromosomal aberrancies or patients with multiple congenital abnormalities including TSC may be the key to the next breakthrough in TSC mapping studies.

References

Clark RD, Smith M, Pandolfo M, et al. Tuberous sclerosis in a liveborn infant with trisomy due to t(11q23.3:22q-11.2) translocation. Am J Hum Genet 1988;43:44.

Connor M, Sampson J. Recent linkage studies in tuberous sclerosis: Chromosome 9 markers. Ann NY Acad Sci 1991;615:265-273.

Fahsold R, Rott HD, Claussen U, Schmalenberger B. Tuberous sclerosis in a child with de novo translocation t(3;12). Clin Genet 1991a;40:326-328.

Fahsold R, Rott HD, Lorenz P. A third gene locus for tuberous sclerosis is closely linked to the phenylalanine hydroxylase gene locus. Hum Genet 1991b;88:85-90.

Fryer AE, Chalmers A, Connor JM et al. Evidence that the gene for tuberous sclerosis is on chromosome 9. Lancet 1987;I:659-661.

Gomez MR. Tuberous sclerosis, 2nd ed. New York: Raven Press, 1988.

Green AJ, Smith M, Yates JRW. Loss of heterozygosity on chromosome 16p13.3 in hamartomas from tuberous sclerosis patients. Nature Genet 1994;6:193-196.

Haines JL, Amos J, Attwood J, et al. Genetic heterogeneity in tuberous sclerosis: study of a large collaborative dataset. Ann NY Acad Sci 1991;615:256-264.

Haines JL, Short MP, Kwiatkowski DJ, et al. Localization of one gene for tuberous sclerosis within 9q32-9q34, and further evidence for heterogeneity. Am J Hum Genet 1991;49:764-772.

Janssen LAJ, Sandkuyl LA, Merkens EC, et al. Genetic heterogeneity in tuberous sclerosis. Genomics 1990;8:237-242.

Kandt RS, Pericak-Vance MA, Hung W-Y, et al., Linkage studies in tuberous sclerosis: chromosome 9?, 11?, or maybe 14!. Ann N Y Acad Sci 1991;615:284-297.

Kandt RS, Haines JL, Smith M, et al. Linkage of an important gene locus for tuberous sclerosis to a chromosome 16 marker for polycystic kidney disease. Nature Genet 1992;2:37-41.

Nellist M, Brook-Carter PT, Connor JM, et al. Identification of markers flanking the tuberous sclerosis locus on chromosome 9 (TSC1). J Med Genet 1993;30:224-227.

Northrup H, Kwiatkowski DJ, Roach ES, et al. Evidence for genetic heterogeneity in tuberous sclerosis: one locus on chromosome 9 and at least one locus elsewhere. Am J Hum Genet 1992;51:709-720.

Pitiot G, Waksman G, Bragado-Nillson E, et al. Linkage analysis places TSC1 gene distal to D9S10. Ann Hum Genet 1994;58:232-233.

Povey S, Burley MW, Attwood J et al. Tow loci for tuberous sclerosis: one on 9q34 and one on 16p13. Ann Hum Genet 1994;58:107-127.

Sampson JR, Harris P. The molecular genetics of tuberous sclerosis. Hum Mol Genet 1994;3:1477-1480.

Simon M, Kelemen J, Szörényi A, et al. Ein seltener Mosaizismus der Chromosom-10-Trisomie bei einer Patientin met Bourneville-Pringle-Syndrom. Der Hautarzt 1979;30:292-294.

Smith M, Smalley S, Cantor R, et al. Mapping of a gene determining tuberous sclerosis to human chromosome 11q14-11q23. Genomics 1990;6:105-114.

6.2

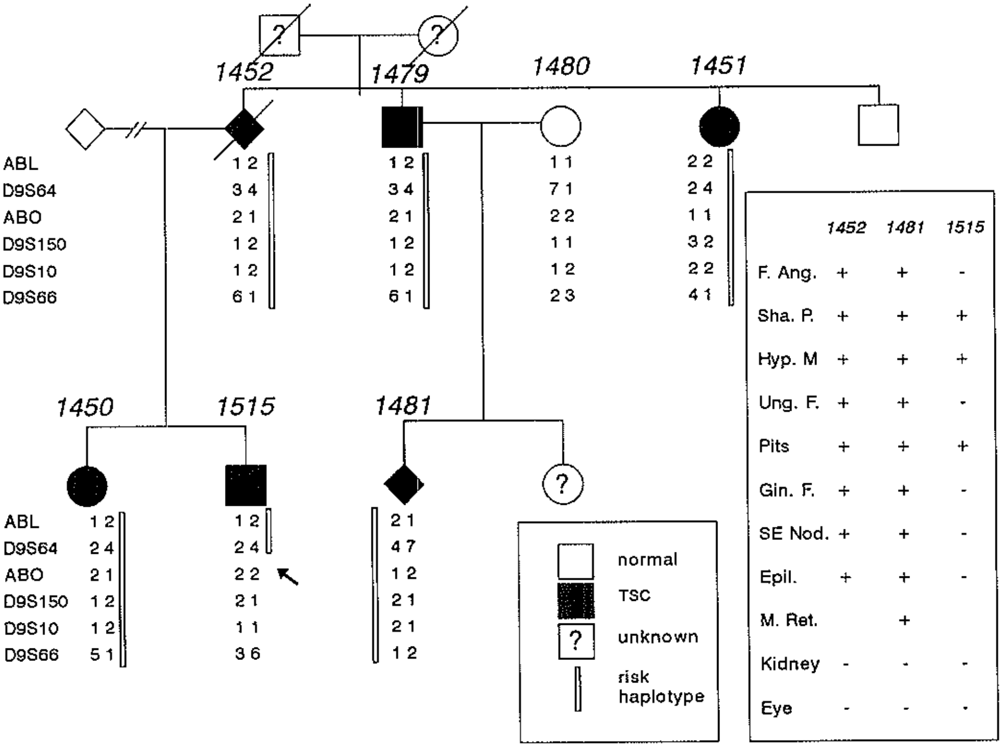
PROSPECTS FOR THE ISOLATION OF THE TSC1 GENE ON 9Q34

Genes in the ABO region

As a logical next step after linkage studies, we initiated genomic cloning efforts. A detailed description of a 600 kb region encompassing ABO was given in Chapter 4. A few years ago, ABO was still believed to be the most distal gene on the long arm of chromosome 9. Its map position, distal to the loci for the ABL (Heisterkamp et al., 1985) and CAN (Von Lindern et al., 1990) oncogenes has been well established. Genes mapping distal to ABO remained unknown, until CEL and TAN1 were mapped telomeric to ABO (Taylor et al., 1991; Ellisen et al., 1991). Our own data do not support the distal localisation of CEL (unpublished results). In Chapter 4 we presented mapping data on 7 genes, placed distal to ABO. The aim of this project was to clarify which genes may serve as a good candidate for TSC1, based on their chromosomal position, and which genes map outside the TSC1 region. Initially we focused our contig assembly efforts on the core of the TSC1 region, around ABO and DBH. At present we are extending this region, mainly in a proximal direction.

The map position of TSC1

Figure 6.1 displays all relevant studies relating to the map position of TSC1. All



	1452	1479	1480	1451
ABL	1 2	1 2	1 1	2 2
D9S64	3 4	3 4	7 1	2 4
ABO	2 1	2 1	2 2	1 1
D9S150	1 2	1 2	1 1	3 2
D9S10	1 2	1 2	1 2	2 2
D9S66	6 1	6 1	2 3	4 1

	1450	1515	1481
ABL	1 2	1 2	2 1
D9S64	2 4	2 4	4 7
ABO	2 1	2 2 ↖	1 2
D9S150	1 2	2 1	2 1
D9S10	1 2	1 1	2 1
D9S66	5 1	3 6	1 2

	1452	1481	1515
F. Ang.	+	+	-
Sha. P.	+	+	+
Hyp. M	+	+	+
Ung. F.	+	+	-
Pits	+	+	+
Gin. F.	+	+	-
SE Nod.	+	+	-
Epil.	+	+	-
M. Ret.		+	
Kidney	-	-	-
Eye	-	-	-

Figure 6.4: The 9q34 haplotypes of family 4219. For reasons of privacy, only a part of this extended pedigree has been depicted. Information on the sex of key persons has been removed. The arrow indicates a key recombination event that occurred between D9S64 and ABO in individual 1515, placing TSC1 proximal of ABO. Markers D9S125 and D9S149 were not informative in the 1452/1515 branch and are therefore not shown. The family is likely to chromosome 9 linked (p>0.999 (Chapter 3.3)). The presence of facial angiofibromas, shagreen patches, typical hypomelanotic macules, ungual fibromas, dental pits, gingival fibromas, subependymal nodules, epilepsy, mental retardation, kidney findings and ophthalmologic findings has been indicated for individuals 1452, 1481 and 1515.

recent linkage reports are consistent with a position in between D9S149 and D9S114 (Sampson and Harris, 1994). Therefore, the cloned area spans not more than half of the consensus region (Chapter 4). However, there is disagreement on the exact position of TSC1, since data are seemingly conflicting. While some groups have found evidence in favour of a position proximal of ABO and DBH, other groups have presented data supporting a location distal to these markers (Sampson and Harris, 1994). The controversy may be due to misclassification of individuals with only minor clinical findings, phenocopy clusters, or to non-linkage of one or more families. The D9S149-D9S150 interval is supported by recombinations in two chromosome 9-linked families. Both families were included in the study reported in Chapter 3.3. Family 1013, previously reported by Nellist et al. (1993), shows a recombination with D9S150 in a completely investigated unaffected individual. The other family, depicted in Figure 6.4, shows a recombination in an apparently affected individual. The latter observation indicates that the TSC1 locus maps proximal to ABO. Although both families are significantly chromosome 9-linked (Chapter 3.3), a more distal position cannot be excluded completely at this stage.

Since linkage data are apparently conflicting and difficult to interpret with certainty, an alternative mapping strategy has been chosen by several research groups. This strategy is based on the tumour suppressor gene model, as provided by Knudson (Knudson, 1971;1973;1985). According to this model, two steps are necessary before a tumour cell can arise. The first step, and perhaps the second as well, is a mutational event. If the mutation occurs in a gamete, a predisposition to the development of cancer may be inherited. Nevertheless, a second event is required before a cell changes into a tumour cell. Patients in whom the second event does not occur, can pass on the gene, but will not develop tumours themselves. In tumours studied by Knudson and his colleagues, the second mutation was often found to be a large deletion. As a consequence of this, informative genetic markers showed loss of one allele in tumour material. Therefore, allelic loss or 'loss of heterozygosity' (LOH) in tumour DNA is a characteristic finding suggesting the presence of a tumour suppressor gene in the deleted area.

Very recently LOH in hamartomas of TSC2 patients has been reported. This indicates that the TSC genes may indeed act as tumour suppressor genes (Green et al., 1994a). Green, Carbonara and coworkers have attempted to narrow down the TSC1 region by mapping allelic loss in TSC related tumours (Green et al., 1994b;

Carbonara et al., 1994). Attempts to find LOH for chromosome 9q34 markers, have indeed been successful. Our own data, obtained by studying tumour material from a chromosome 9 linked family, also show LOH of the TSC1 region and thereby confirm the hypothesis that TSC1 may be a tumour suppressor gene (unpublished results). Unfortunately, LOH studies have not revealed more precise mapping data. On the other hand, even if LOH studies would reveal a consensus region distinct from the region identified by linkage studies, this would not necessarily invalidate the linkage results. Studies on familial breast cancer indicate that relevant tumour suppressor genes - identified by LOH - and mutated causative genes, segregating through a family, do not always coincide. Although the recently cloned BRCA1 gene (Miki et al., 1994) is thought to act as a tumour suppressor gene and LOH for the BRCA1 region on 17q12-q21 has been reported, the LOH overlap region involved in 73% of all breast and ovarian tumours maps to a discrete, more distal part of chromosome 17q (Godwin et al., 1994). We may thus conclude that LOH studies have not contributed so far to a further refinement of the TSC1 region and that the significance of LOH as a gene mapping tool should not be overestimated.

Candidate genes for TSC1

The high number of cloned genes in the ABO/TSC1 area prompts the question whether one of these might be a good candidate gene for TSC1. On the basis of their chromosomal position and possible function, a few genes are indeed interesting candidates.

Epilepsy is one of the most conspicuous symptoms of TSC. The possibility of a relationship between epilepsy in TSC and a DBH gene defect can be raised. Dopamine-beta-hydroxylase converts dopamine to norepinephrine. DBH activity is reduced in brain areas of genetically epilepsy-prone rats (Browning et al., 1989). Furthermore, norepinephrine is known to exert seizure attenuating effects. This makes DBH a realistic candidate. Although it might be possible that parts of DBH function as a tumour suppressor, the tumour suppressor hypothesis slightly weakens the candidacy of DBH. An even stronger argument against DBH as TSC1 gene is provided by comparison of the most crucial key recombinants. While some data favour a TSC1 position proximal to D9S150 and other recombinants position TSC1 distal to DBH and D9S66, none of these studies support the DBH area itself.

Another putative candidate for TSC1 is the human Notch homologue TAN1

(Ellisen et al., 1991). The Drosophila Notch gene encodes a transmembrane protein with homology to EGF. It is involved in cell-cell interactions, especially in the determination of neuroectodermal cell destination during embryonal development (Artavanis-Tsakonas 1988). TSC lesions have been correlated with the unusual presence of neuron-like cells (N-cells) (Davidson et al., 1991), which might perhaps be the result of a disturbed determination of cell destination. A defect TAN1 protein may be responsible for these findings. Like its human homologue, Notch is ubiquitously expressed in the embryo, which suggests a pleiotropic role in development, regulating more than cell fate only. We have performed FISH studies aimed at mapping TAN1 relative to the TSC1 region. TAN1 turned out to map distal to D9S66 and D9S114 (unpublished results) and may therefore be ruled out as TSC1 candidate gene.

A third attractive candidate arose from exon trapping experiments, using cosmids identified in Rotterdam. This candidate, designated VAV2, has been shown to be homologous to the VAV proto-oncogene (Henske et al., 1995). This is of particular interest, because VAV is known to play a role in cell signalling and appears to act as a guanine nucleotide exchange factor for ras. The chromosome 16 linked TSC2 gene shows homology with the rap1 GTPase activating protein, which deactivates rap1, an antagonist of ras. It may thus be that TSC2 and VAV2 play a role in the same pathway, both stimulating ras activity. The hypothesis of VAV2 being the TSC1 gene, has been investigated by Henske and coworkers. They have sequenced several VAV2 transcripts, including 4 transcripts originating from proven TSC1 risk haplotypes. No obvious inactivating mutations were found (Henske et al., 1995). We may conclude that VAV2 is an unlikely candidate for TSC1.

If we wish to consider candidate genes solely on the basis of their position and expression pattern, we must decide whether we accept a single recombination event - like the ABO recombination shown in Figure 6.4 - as conclusive evidence, despite the inconsistent TSC1 mapping data and the possibility that we are dealing with non-linkage. If we decide that multiple recombination events are required, the SURF genes - mapping between DBH and ABO - are true candidates for TSC1, as proposed in Chapter 4. It is also possible that ABO maps in an intron of one of the SURF genes. These grounds clearly indicate that the SURF genes may not be ruled out as TSC1 candidates.

Concluding remarks

Although the evidence presented in Figure 6.4 may be regarded as not completely convincing, because of the rather mild phenotype of individual 1515 and the remote possibility that the family is not segregating a defect TSC1 gene, the data should not be neglected. This means that the D9S149-ABO interval still is the most important region with respect to the search for the TSC1 gene. Unfortunately, very little is known about this region. Interphase FISH experiments aimed at determining the size of the interval have been inconclusive, indicating that the size is considerable, measuring at least several hundreds of kilobases (B. Eussen pers. comm.). The genetic distance between D9S149 and ABO is 1 cM (Povey et al., 1994). Genomic clones available at the moment cover only a small percentage of the area. Currently there are no genes known to map between D9S149 and ABO, but efforts aimed at isolating expressed sequences from the region are in progress.

In theory, all techniques for gene isolation and mutation screening mentioned in Chapter 1.3, such as direct screening of cDNA libraries and exon trapping aimed at gene isolation and southern blotting techniques or SSCP for gene identification, can be applied in the search for TSC1. In practice it is most convenient to use the same techniques that have been applied for TSC2, complemented with more modern or refined techniques when necessary. The availability of large deletions, detectable by pulsed field analysis (PFGE), has been very beneficial to the rapid identification of TSC2. Whether these mutations can be found in TSC1 patients seems unpredictable. The presence or absence of large deletions is determined by genomic sequences within and surrounding the gene. It is not directly related to the gene function. While the gene function of TSC1 and TSC2 may be similar, the genomic organisation will presumably be completely different. Notwithstanding environmental differences, there are important similarities between TSC1 and TSC2. The average family size and the proportion of linked families is exactly identical. Therefore, the mutation frequency of TSC1 and TSC2 will also be about equal, provided that sporadic cases and familial cases have a similar α. The relatively high mutation frequency indicates that we are looking for a gene which is susceptible to mutation, probably because of its size. Apart from its putative tumour suppressor role and its substantial size we can also speculate on the expression pattern of TSC1. The spectrum of tissues and organs involved in both genetic subtypes appears to be identical. If both gene products act together in the same pathway or protein complex, they will indeed show expression in the same cells and tissues. Therefore, we expect the TSC1 gene to be expressed in the same tissues as TSC2, based on the limited clinical differences that have been observed between both genetic subtypes.

References

Artavanis-Tsakonas S. The molecular biology of the Notch locus and the fine tuning of differentiation in Drosphila. Trends Genet 1988;4:95-100.

Browning RA, Wade DR, Marcinczyk M, et al. Regional brain abnormalities in norepinephrine uptake and dopamine beta-hydroxylase activity in the genetically epilepsy-prone rat. J Pharmacol Exp Ther 1989;249:229-235.

Carbonara C, Longa L, Grosso E, et al. 9q34 loss of heterozygosity in a tuberous sclerosis astrocytoma suggests a growth suppressor-like activity also for the TSC1 gene. Hum Mol Genet 1994;3:1829-1832.

Davidson M, Yoshidome H, Stenroos E, Johnson WG. Neuron-like cells in culture of tuberous sclerosis tissue. Ann NY Acad Sci 1991;615:196-210.

Ellisen LW, Bird J, West DC, et al. TAN-1, the human homolog of the Drosophila Notch gene, is broken by chromosomal translocations in T lymphoblastic neoplasms. Cell 1991;66:649-661.

Godwin AK, Vanderveer L, Schultz DC, et al. A common region of deletion on chromosome 17q in both sporadic and familial epithelial ovarian tumors distal to BRCA1. Am J Hum Genet 1994;55:666-677.

Green AJ, Smith M, Yates JRW. Loss of heterozygosity on chromosome 16p13.3 in hamartomas from tuberous sclerosis patients. Nature Genet 1994a;6:193-196.

Green AJ, Johnson PH, Yates JRW. The tuberous sclerosis gene on chromosome 9q34 acts as a growth suppressor. Hum Mol Genet 1994b;3:1833-1834.

Heisterkamp N, Stam K, Groffen J, et al. Structural organization of the bcr gene and its role in the Ph translocation. Nature 1985; 315: 758-761.

Henske EP, Short MP, Jozwiak S, et al. Identification of VAV2 on 9q34 and its exclusion as the tuberous sclerosis gene TSC1. Ann Hum Genet 1995; 59: 25-37.

Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci USA 1971; 68: 820-823.

Knudson AG. Mutation and human cancer. Adv. Cancer Res. 1973; 17: 317-352.

Knudson AG. Hereditary cancer, oncogenes and antioncogenes. Cancer Res 1985; 45: 1437-1443.

Miki Y, Swensen J, Shattuck-Eidens D, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 1994;266:66-71.

Nellist M, Brook-Carter PT, Connor JM, et al. Identification of markers flanking the tuberous sclerosis locus on chromosome 9 (TSC1). J Med Genet 1993;30:224-227.

Povey S, Armour J, Farndon P et al. Report on the third international workshop on chromosome 9. Ann Hum Genet 1994;58:177-250.

Sampson JR, Harris PC. The molecular genetics of tuberous sclerosis. Hum Mol Genet 1994;3:1477-1480.

Taylor AK, Zambaux JL, Klisak I, et al., Carboxyl ester lipase: a highly polymorphic locus on human chromosome 9qter. Genomics 1991;10:425-431.

Von Lindern M, Poustka A, Lerach H, Grosveld G. The (6;9) chromosome translocation, associated with a specific subtype of acute nonlymphocytic leukemia, leads to aberrant transcription of a target gene on 9q34. Moll Cell biol 1990; 10: 4016-4026.

RELEVANCE OF THE WORK TOWARDS THE MAPPING OF OTHER HETEROGENEOUS DISORDERS

In the course of our project we have substantially gained insight into linkage analysis under locus heterogeneity. TSC was not the first disease proven to be heterogeneous. Neither was it the first heterogeneous disorder with identical phenotypes associated with the different loci. For several of these disorders, including TSC, gene locations on more than one chromosome had been proposed. Most of these problems are only accessible by means of linkage analysis. Nevertheless the TSC problem was unique, because it allowed for the application of a novel method that utilised all available data in one single test (ICA).

For some heterogeneous diseases gene locations can be found by performing complementation experiments. In cell fusion experiments the mouse chromosome 4 was found to correct the DNA repair defect (XP-A) in cells from xeroderma pigmentosum patients (Lin and Ruddle, 1981). Xeroderma pigmentosum (XP) is a cancer prone recessive heterogeneous disease caused by a homozygous inactivation of one of at least 7 different XP DNA repair genes (De Weerd-Kastelein et al., 1972;

Vermeulen et al., 1991). Cell fusion and transfection experiments provide a quite direct approach to the cloning of genes, when applicable. These approaches are in fact examples of functional cloning under locus heterogeneity.

Unfortunately, the function of most genes involved in genetically heterogeneous disorders remains unknown. Furthermore, most of these diseases have an autosomal dominant mode of inheritance (Chapter 2.1). Therefore, elucidation of the biochemical defect in these disorders can only be achieved if positional cloning strategies are used. If family material is limited, it is essential to use the available information as efficiently as possible. In cases comparable with the TSC situation, with a few candidate regions and small families showing an indistinguishable phenotype, the imaginary chromosome approach (ICA) presented in Chapter 2 is perfectly suitable. The list of disorders for which the ICA may be helpful contains almost every disease known or suspected for locus heterogeneity. Since the majority of yet insufficiently mapped disease genes falls into this category, an important role for ICA in future mapping studies can be foreseen. The Clinical Genetics Department in Rotterdam is currently involved in studies on 3 disorders for which the approach can be directly applied, namely multiple exostoses, polycystic kidney disease and familial breast cancer. For all these disorders multiple candidate regions are know. With respect to multiple exostoses three candidate regions on chromosomes 8, 11 and 19 have been postulated (Cook et al., 1993; Wu Y-Q et al., 1994; Le Merrer et al., 1994). Polycystic kidney disease is caused by a defect in the chromosome 16 linked gene, the chromosome 4 linked gene, or perhaps a yet unmapped third gene (Chapter 6.4). Genes responsible for familial breast cancer have recently been reported to map to chromosomes 13 and 17, whereas the existence of more genes cannot be excluded (Wooster et al., 1994).

In Chapter 2.3 studies are described that evaluate the robustness of the ICA, compared with the original test (the A-test). The outcome of these analytical evaluations turned out to be quite surprising. We evaluated whether size differences between linked and unlinked families might cause a bias if the ICA was applied. We also evaluated the influence of clinical misclassification. The studies revealed that the ICA was less prone to bias due to misclassification. Moreover, the studies revealed that the conventional A-test is more prone to bias caused by differences in family size than the ICA variant of the test. The latter is surprising, since biases due to differences in family size have never been reported before. We have clearly demonstrated an inconsistency in the A-test, which has an almost negligible effect on the ICA. Under normal circumstances the inconsistency has little effect on the outcome of the conventional A-test. The testAtest program, described in Chapter 2.3, can be used to verify this for each individual analysis.

The ICA method is based on a simplified model of our genomic organisation: The whole genome is regarded as one linear chromosome. Nevertheless, it is not possible to incorporate data from all regions in one single analysis. Therefore, the ICA is not directly applicable to genome wide searches. In theory it is possible to overcome this limitation by dividing the genome in 50 linkage regions and evaluating each combination by ICA, assuming a model with only two trait causing loci. However, the huge number of possible combinations (50 x 50 x 0.5 = 1250) determines the infeasibility of this solution. A good alternative would be the development of a simple and fully automated scoring scheme that cancels the analysis of uninteresting combinations. The development of a program that would perform this procedure is currently under investigation in our group. Furthermore we are currently studying the possibilities of incorporating the ICA method in exclusion mapping studies.

References

Cook A, Raskind W, Blanton SH, et al. Genetic heterogeneity in families with hereditary multiple exostoses. Am J Hum Genet 1993; 53: 71-79.

De Weerd-Kastelein EA, Keijzer W, Bootsma D. Genetic heterogeneity of xeroderma pigmentosum demonstrated by somatic cell hybridisation. Nature (london) New Biology 1972;238:80-83.

Le Merrer M, Legeai-Mallet L, Jeannin PM, et al. A gene for hereditary multiple exostoses maps to chromosome 19p. Hum Mol Genet 1994; 3: 717-722.

Lin PF, Ruddle FH. Murine DNA repair gene located on chromosome 4. Nature 1981;289:191-194.

Vermeulen W, Stefanini M, Giliani S, et al. Xeroderma pigmentosum complementation group H falls into complementation group D. Mutat Res 1991;255:201-208.

Wooster R, Neuhausen SL, Mangion J, et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science 1994; 265: 2088-2090.

Wu Y-Q, Heutink P, de Vries BBA, et al. Assignment of a second locus for multiple exostoses to the pericentromeric region of chromosome 11. Hum Mol Genet 1994; 3: 167-171.

6.4

RELEVANCE OF THE WORK TOWARDS THE ISOLATION OF A GENE RESPONSIBLE FOR POLYCYSTIC KIDNEY DISEASE

Bart Janssen
and the
European Polycystic Kidney Disease Consortium

Parts of this chapter are published in:
Cell 1994;77:881-894.

The studies described in the previous chapters led to the mapping of two TSC genes and the isolation of four candidate genes for TSC2 on chromosome 16. The gene corresponding to cDNA 2A6 was identified as the TSC2 causative gene, while the function of the other three genes was unknown. Might one of these be the gene responsible for autosomal dominant polycystic kidney disease I (PKD1)? Extensive linkage and physical mapping efforts had been performed, indicating that the PKD1 locus maps to a 600 kb interval between GGG1 and SM7 (Harris et al., 1991; Somlo et al., 1992). All four genes described in Chapter 5 map within this region.

Polycystic kidney disease (ADPKD)

Polycystic kidney disease is a common genetic disease, affecting approximately 1 in every 1000 individuals. Its major feature is the development of cystic kidneys that commonly lead to renal failure in adult life. ADPKD accounts for 8-10% of all renal transplantation and dialysis patients in Europe and the United States (Gabow et al., 1992). Other manifestations of the disease include intracranial aneurysms, liver cysts and (although more rarely) cysts in other organs like the pancreas (Gabow 1990). More recently the disease has been associated with an increased prevalence of cardiac valve defects (Hossack et al., 1988), herniae (Gabow, 1990) and colonic diverticulae (Scheff et al., 1980). The progressive renal disease however, is the most important aspect of autosomal polycystic kidney disease (ADPKD). The formation of fluid-filled cysts results in grossly enlarged kidneys. Renal function deteriorates and end stage renal disease occurs in more than 50% of patients by the age of 60 (Gabow et al., 1992).

ADPKD is another example of a genetically heterogeneous disorder. The first and most important locus (PKD1) was mapped to chromosome 16p13.3 (Reeders et al., 1985), long before the TSC2 locus could be assigned to the same region (Kandt et al., 1992). A second locus, designated PKD2, has been assigned to chromosome 4 (Kimberling et al., 1993; Peters et al., 1993). A search for additional loci is still in progress. It is estimated that PKD1 accounts for approximately 85% of all cases (Peters and Sandkuijl, 1992). Like in TSC, there are no conclusive phenotypical differences, that can be used to discriminate between chromosome 16 linked (PKD1) and unlinked families, although the age at onset differs.

Despite the successful delineation of the PKD1 region to an interval of 600 kb by haplotype analysis, it proved difficult to pinpoint the disease gene. The region is very gene dense and alternative mapping techniques, like linkage disequilibrium studies (Pound et al., 1992; Peral et al., 1994), could not contribute to a further delineation of the candidate region. Despite laborious studies on individual transcripts from the region, no information leading to the identification of the trait-causing gene could be obtained.

This situation changed when family 2338, depicted in Figure 6.3, underwent a closer clinical examination. This family, segregating a t(16;22)((p13.3;q11.21) translocation, associated with TSC in the unbalanced form, had already been instrumental in the isolation of the TSC2 gene (Chapter 5). Further analysis revealed that the translocation carriers of the balanced translocation showed the clinical features of ADPKD. Although renal cysts occasionally occur as a symptom of TSC, individuals 6321 and 6323 showed no additional signs of TSC. The cytogenetically normal parents of 6321 had no clinical features of TSC of ADPKD at the age of 67 and 82 years respectively. These findings make the diagnosis TSC in individuals 6321 and 6323 highly unlikely. Therefore, we considered the possibility that the kidney cysts were due to ADPKD and caused by the translocation, which presumably disrupted the PKD1 gene.

One of the genes described in Chapter 5 is a good candidate for the PKD1 gene

FISH studies revealed that the t(16;22) breakpoint mapped less than 40 kb proximal of cosmid ZDs5 and therefore slightly proximal of the TSC2 gene. This was in line with our interpretation of the translocation in family 2338 and the associated phenotypes. The disruption of a gene crossing the breakpoint may be the cause of ADPKD, while monosomy of the region distal of the breakpoint results in TSC. This also allowed us to exclude three of the four isolated genes as candidates for PKD1. Only the gene proximal to TSC2, which corresponds to cDNA clone 3A3, still had to be considered as a candidate for PKD1.

PKD1 lies within a region reiterated elsewhere on chromosome 16

It was previously noted that the region between CW21 and N54 (Figure 5.2) contained a chromosome 16 specific repeat, which also occurred at the centromeric end of the short arm (p13.1) (Germino et al., 1992; Chapter 5). The size of the repeated fragment and the high homology between these repeats seriously hampered cosmid walking in the area seriously. The structure of the repeated area was found to be complex, with each fragment present once in 16p13.3 reiterated two to four times in 16p13.1. Cosmids spanning the area were subcloned and a restriction map was generated. This pattern was compared with a genomic map of the 3A3 region, constructed using a radiation hybrid (Hy145.19) that contains only the distal portion of 16p.

The t(16;22) translocation breakpoint was localised by a series of hybridisations, using a ClaI PFGE blot and conventional Southern blots from family 2338 probed with subclones from the repeated region. The results presented in the Figures 6.5

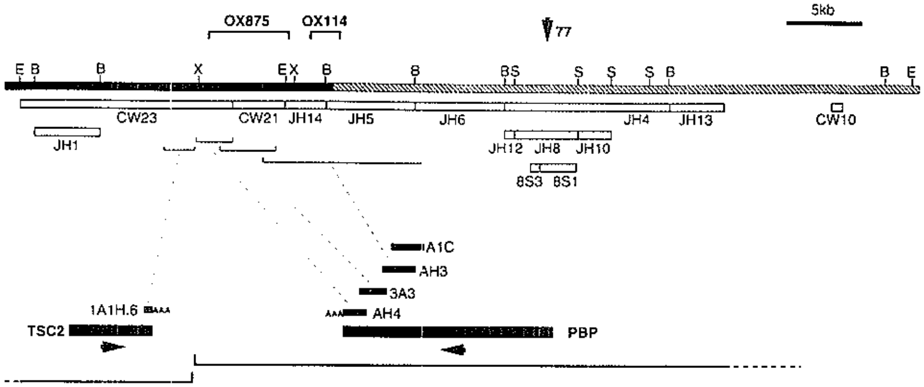


Figure 6.5: A detailed map of the translocation region showing the precise localization of the family 2338 breakpoint (↓77) and the repeated region (hatched). DNA probes (open boxes), the transcripts PBP and TSC2 (closed boxes), and cDNAs are shown below the genomic map. The known genomic extent of each gene is indicated at the bottom of the diagram, and the approximate genomic locations of each cDNA are indicated under the genomic map. The positions of genomic deletions found in PKD1 patients OX875 and OX114 are also indicated. Restriction sites for EcoRI (E) and incomplete maps for BamHI (B), SacI (S), and XbaI (X) are shown.

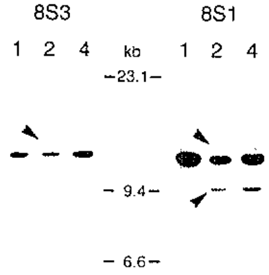


Figure 6.6: Southern blots of BamHI digested DNA from individuals 6322 (1), 6321 (2) and 6324 (4) hybridised with 8S3 and 8S1. 8S3 detects a novel 12kb fragment on the telomeric side of the breakpoint, associated with the der(22) chromosome in 6321, but not 6324; 8S1 identifies a novel 9kb fragment on the centromeric side of the breakpoint, associated with the der(16). The telomeric breakpoint fragment is also seen weakly with 8S1, indicating that the breakpoint lies in the distal part of 8S1.

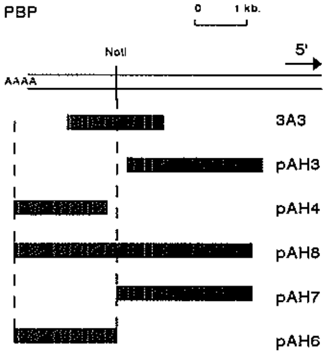


Figure 6.7: cDNA clones obtained from the Clonetech library at the Clinical Genetics Department Rotterdam.

and 6.6 prove that the balanced translocation was not associated with a substantial deletion. Furthermore the breakpoint is shown to map more than 20 kb proximal to the TSC2 locus (Figure 6.5). These results supported the hypothesis that the cysts found in kidneys of members of family 2338 were not due to disruption of the TSC2 gene, but indicated that a separate gene, mapping just proximal to TSC2 was likely to be the PKD1 gene.

Unfortunately, at least a part of this gene maps in the repeated area and the gene may therefore contain sequences that are not unique. Northern blotting experiments, using JH8 as a probe, indeed revealed three bands, indicating that JH8 recognised three partly homologous transcripts (results not shown). A probe from the 16p13.1 region hybridised to the same transcripts plus a fourth somewhat smaller RNA. The 3A3 cDNA probe detected only one of these four transcripts. These results suggest that at least four copies of the repeat exist. Only one of the four transcripts corresponds to the 3A3 cDNA clone, while the other transcripts originate from the 16p13.1 region and have no causative relationship with ADPKD, since no linkage to this region has ever been reported. Genomic probes JH5 and JH13 (Figure 6.5), from the distal and proximal side of the breakpoint respectively, detected the same RNA band as 3A3. The gene encoding this 14 kb messenger was therefore designated polycystic breakpoint (PBP) gene. All cDNAs depicted in Figure 6.7 were shown to contain parts of the 3′ end of the PBP transcript. This 3′ end maps outside the repeated area, immediately proximal to the 3′ end of TSC2.

Mutations in the PBP gene prove that it is responsible for ADPKD

As the PBP gene was transcribed across the repeated region disrupted by the t(16;22) breakpoint in family 2338, in a proximal to distal direction on the chromosome, it was possible that a novel transcript would be found in this family. Probe 8S1 (Figure. 6.5) indeed detected a novel 9 kb messenger. This transcript was only present in the translocation carriers and appeared to be expressed at a higher level than the normal product, transcribed from the other allele. These results confirmed that the translocation disrupts the PBP gene and supports the hypothesis that this is the PKD1 gene.

To prove that the PBP gene is the defective gene at the PKD1 locus, the ADPKD patient material present in Oxford was analyzed for mutations. The 3′ end of the PBP gene was most accessible to study as it maps outside the repeated area.

BamHI digests of DNA from 282 apparently unrelated patients were hybridised with the poly-A containing EcoRI-HindIII fragment of the TSC2 transcript (1A1H.6) (Figure 6.5). This screening procedure was complemented by the analysis of a large EcoRI fragment (41kb), separated by PFGE and hybridised with CW10. Both procedures showed rearrangements in two patients. One patient showed a 5.5 kb genomic deletion at the 3′ end of the PBP gene, resulting in a novel 11 kb transcript. This deletion was transmitted from the affected father (results not shown). The second rearrangement detected by hybridisation was a 2 kb genomic deletion within the PBP gene. Although the transcript was shortened, this could only be demonstrated by RT-PCR, since only 446 bp were missing. Three steps were taken to prove that this frameshift mutation had occurred *de novo*. First, medical records were checked to confirm that both the deceased father (age 78) and the mother (age 73) showed no evidence of ADPKD. Second, haplotypes were constructed, using somatic cell hybrids from the patient. Last, a healthy brother of the patient was demonstrated to have inherited the same paternal haplotype, with one important difference: the PBP gene was intact (results not shown).

Additional RT-PCR studies on 48 apparently unrelated patient, revealed another mutation. The PBP transcript of a member of a large PKD1 family apparently missed 135 bp. Several other family members were tested and shown to have the same mutation, indicating that the abnormal transcript indeed segregated with PKD1. The abnormal transcript was shown to be caused by a G to C transition at +1 of the splice donor site following the 135 bp exon (results not shown). Together, the three intragenic mutations confirm that the PBP gene is the defective gene at the PKD1 locus.

Characterization of the gene

Despite several cDNA library hybridisation experiments, using probes from the unique part of the gene, we did not obtain additional cDNA clones stretching further towards the 5′ end. The full genomic extent of the PKD1 gene is therefore not yet known. Northern blotting experiments show that it extends at least as far as JH13. Several CpG islands have been localised 5′ of the known extent of the PKD1 gene, although there is no direct evidence that any of these are associated with the PBP gene.

The cDNA contig extending 5631 bp to the 3′ end of the transcript was sequenced.

This sequence - depicted in Figure 6.8 - represents about 40% of the transcript. An open reading frame was detected that runs from the 5′ end of the sequence and spans 4842 bp, leaving a 3′ untranslated region (UTR) of 789 bp. The UTR contains the previously described microsatellite marker KG8 (Peral et al., 1994; Snarey et al., 1994), which is of considerable importance to the molecular diagnostics of both TSC and ADPKD. A polyadenylation signal is present at nucleotides 5598-5603, and a poly(A) tail was detected in several independent cDNAs. No significant areas of homology with other available protein sequences were found, neither did we identify important protein motifs.

The PBP gene shows ubiquitous expression and is conserved throughout mammal evolution. The more proximal genes at 16p13.1 occur only in primates (P Harris pers. comm.). It therefore seems that a series of duplication events occurred quite recently in evolution.

Figure 6.8

THE PBP SEQUENCE:

```
  1 ctcaacgaggagcccctgacgctggcgggcgaggagatcgtggcccagggcaagcgctcg
  1  L  N  E  E  P  L  T  L  A  G  E  E  I  V  A  Q  G  K  R  S

 61 gacccgcggagcctgctgtgctatggcggcgcccccgggcccggctgccacttctccatc
 21  D  P  R  S  L  L  C  Y  G  G  A  P  G  P  G  C  H  F  S  I

121 cccgaggctttcagcggggccctggccaacctcagtgacgtggtgcagctcatctttctg
 41  P  E  A  F  S  G  A  L  A  N  L  S  D  V  V  Q  L  I  F  L

181 gtggactccaatccctttccctttggctatatcagcaactacaccgtctccaccaaggtg
 61  V  D  S  N  P  F  P  F  G  Y  I  S  N  Y  T  V  S  T  K  V

241 gcctcgatggcattccagacacaggccggcgcccagatccccatcgagcggctggcctca
 81  A  S  M  A  F  Q  T  Q  A  G  A  Q  I  P  I  E  R  L  A  S

301 gagcgcgccatcaccgtgaaggtgcccaacaactcggactgggctgcccggggccaccgc
101  E  R  A  I  T  V  K  V  P  N  N  S  D  W  A  A  R  G  H  R

361 agctccgccaactccgccaactccgttgtggtccagccccaggcctccgtcggtgctgtg
121  S  S  A  N  S  A  N  S  V  V  V  Q  P  Q  A  S  V  G  A  V

421 gtcaccctggacagcagcaaccctgcggccgggctgcatctgcagctcaactatacgctg
141  V  T  L  D  S  S  N  P  A  A  G  L  H  L  Q  L  N  Y  T  L

481 ctggacggccactacctgtctgaggaacctgagccctacctggcagtctacctgcactcg
161  L  D  G  H  Y  L  S  E  E  P  E  P  Y  L  A  V  Y  L  H  S

541 gagccccggcccaatgagcacaactgctcggctagcaggaggatccgcccagagtcactc
181  E  P  R  P  N  E  H  N  C  S  A  S  R  R  I  R  P  E  S  L

601 cagggtgctgaccaccggccctacaccttcttcatctccccggggagcagagacccagcg
201  Q  G  A  D  H  R  P  Y  T  F  F  I  S  P  G  S  R  D  P  A

661 gggagttaccatctgaacctctccagccacttccgctggtcggcgctgcaggtgtccgtg
221  G  S  Y  H  L  N  L  S  S  H  F  R  W  S  A  L  Q  V  S  V

721 ggcctgtacacgtccctgtgccagtacttcagcgaggaggacatggtgtggcggacagag
241  G  L  Y  T  S  L  C  Q  Y  F  S  E  E  D  M  V  W  R  T  E

781 gggctgctgcccctggaggagacctcgccccgccaggccgtctgcctcacccgccacctc
261  G  L  L  P  L  E  E  T  S  P  R  Q  A  V  C  L  T  R  H  L

 841 accgccttcggcgccagcctcttcgtgccccccagccatgtccgctttgtgtttcctgag
 281  T  A  F  G  A  S  L  F  V  P  P  S  H  V  R  F  V  F  P  E

 901 ccgacagcggatgtaaactacatcgtcatgctgacatgtgctgtgtgcctggtgacctac
 301  P  T  A  D  V  N  Y  I  V  M  L  T  C  A  V  C  L  V  T  Y

 961 atggtcatggccgccatcctgcacaagctggaccagttggatgccagccggggccgcgcc
 321  M  V  M  A  A  I  L  H  K  L  D  Q  L  D  A  S  R  G  R  A

1021 atccccttctgtgggcagcggggccgcttcaagtacgagatcctcgtcaagacaggctgg
 341  I  P  F  C  G  Q  R  G  R  F  K  Y  E  I  L  V  K  T  G  W

1081 ggccggggctcaggtaccacggcccacgtgggcatcatgctgtatggggtggacagccgg
 361  G  R  G  S  G  T  T  A  H  V  G  I  M  L  Y  G  V  D  S  R

1141 agcggccaccggcacctggacggcgacagagccttccaccgcaacagcctggacatcttc
 381  S  G  H  R  H  L  D  G  D  R  A  F  H  R  N  S  L  D  I  F

1201 cggatcgccaccccgcacagcctgggtagcgtgtggaagatccgagtgtggcacgacaac
 401  R  I  A  T  P  H  S  L  G  S  V  W  K  I  R  V  W  H  D  N

1261 aaagggctcagccctgcctggttcctgcagcacgtcatcgtcagggacctgcagacggca
 421  K  G  L  S  P  A  W  F  L  Q  H  V  I  V  R  D  L  Q  T  A

1321 cgcagcgccttcttcctggtcaatgactggctttcggtggagacggaggccaacgggggc
 441  R  S  A  F  F  L  V  N  D  W  L  S  V  E  T  E  A  N  G  G

1381 ctggtggagaaggaggtgctggccgcgagcgacgcagcccttttgcgcttccggcgcctg
 461  L  V  E  K  E  V  L  A  A  S  D  A  A  L  L  R  F  R  R  L

1441 ctggtggctgagctgcagcgtggcttctttgacaagcacatctggctctccatatgggac
 481  L  V  A  E  L  Q  R  G  F  F  D  K  H  I  W  L  S  I  W  D

1501 cggccgcctcgtagccgtttcactcgcatccagagggccacctgctgcgttctcctcatc
 501  R  P  P  R  S  R  F  T  R  I  Q  R  A  T  C  C  V  L  L  I

1561 tgcctcttcctgggcgccaacgccgtgtggtacggggctgttggcgactctgcctacagc
 521  C  L  F  L  G  A  N  A  V  W  Y  G  A  V  G  D  S  A  Y  S
```

(CONTINUED ON NEXT PAGE)

```
1621 acggggcatgtgtcccaggctgagcccgctgagcgtcgacacagtcgctgttggcctggtg
 541  T  G  H  V  S  R  L  S  P  L  S  V  D  T  V  A  V  G  L  V

1681 tccagcgtggttgtctatcccgtctacctggccatcctttttctcttccggatgtcccgg
 561  S  S  V  V  V  Y  F  V  Y  L  A  I  L  F  L  F  R  M  S  R

1741 agcaaggtggctgggagcccgagccccacacctgccgggcagcaggtgctggacatcgac
 581  S  K  V  A  G  S  P  S  P  T  P  A  G  Q  Q  V  L  D  I  D

1801 agctgcctggactcgtccgtgctggacagctccttcctcacgttctcaggcctccacgct
 601  S  C  L  D  S  S  V  L  D  S  S  F  L  T  F  S  G  L  H  A

1861 gaggcctttgttggacagatgaagagtgacttgtttctggatgattctaagagtctggtg
 621  E  A  F  V  G  Q  M  K  S  D  L  F  L  D  D  S  K  S  L  V

1921 tgctggccctccggcgagggaacgctcagttggccggacctgctcagtgacccgtccatt
 641  C  W  P  S  G  E  G  T  L  S  W  P  D  L  L  S  D  P  S  I

1981 gtgggtagcaatctgcggcagctggcacggggccaggcgggccatgggctgggcccagag
 661  V  G  S  N  L  R  Q  L  A  R  G  Q  A  G  H  G  L  G  P  E

2041 gaggacggcttctccctggccagcccctactcgcctgccaaatccttctcagcatcagat
 681  E  D  G  F  S  L  A  S  P  Y  S  P  A  K  S  F  S  A  S  D

2101 gaagacctgatccagcaggtccttgccgagggggtcagcagcccagcccctacccaagac
 701  E  D  L  I  Q  Q  V  L  A  E  G  V  S  S  P  A  P  T  Q  D

2161 acccacatggaaacggacctgctcagcagcctgtccagcactcctggggagaagacagag
 721  T  H  M  E  T  D  L  L  S  S  L  S  S  T  P  G  E  K  T  E

2221 acgctggcgctgcagaggctgggggagctggggccacccagcccaggcctgaactgggaa
 741  T  L  A  L  Q  R  L  G  E  L  G  P  P  S  P  G  L  N  W  E

2281 cagccccaggcagcgaggctgtccaggacaggactggtggagggtctgcggaagcgcctg
 761  Q  P  Q  A  A  R  L  S  R  T  G  L  V  E  G  L  R  K  R  L

2341 ctgccggcctggtgtgcctccctggcccacgggctcagcctgctcctggtggctgtggct
 781  L  P  A  W  C  A  S  L  A  H  G  L  S  L  L  L  V  A  V  A

2401 gtggctgtctcagggtgggtgggtgcgagcttccccccgggcgtgagtgttgcgtggctc
 801  V  A  V  S  G  W  V  G  A  S  F  P  P  G  V  S  V  A  W  L

2461 ctgtccagcagcgccagcttcctggcctcattcctcggctgggagccactgaaggtcttg
 821  L  S  S  S  A  S  F  L  A  S  F  L  G  W  E  P  L  K  V  L

2521 ctggaagccctgtacttctcactggtggccaagcggctgcacccggatgaagatgacacc
 841  L  E  A  L  Y  F  S  L  V  A  K  R  L  H  P  D  E  D  D  T

2581 ctggtagagagcccggctgtgacgcctgtgagcgcacgtgtgccccgcgtacggccaccc
 861  L  V  E  S  P  A  V  T  P  V  S  A  R  V  P  R  V  R  P  P

2641 cacggctttgcactcttcctggccaaggaagaagcccgcaaggtcaagaggctacatggc
 881  H  G  F  A  L  F  L  A  K  E  E  A  R  K  V  K  R  L  H  G

2701 atgctgcggagcctcctggtgtacatgctttttctgctggtgaccctgctggccagctat
 901  M  L  R  S  L  L  V  Y  M  L  F  L  L  V  T  L  L  A  S  Y

2761 ggggatgcctcatgccatgggcacgcctaccgtctgcaaagcgccatcaagcaggagctg
 921  G  D  A  S  C  H  G  H  A  Y  R  L  Q  S  A  I  K  Q  E  L

2821 cacagccgggccttcctggccatcacgcggtctgaggagctctggccatggatggcccac
 941  H  S  R  A  F  L  A  I  T  R  S  E  E  L  W  P  W  M  A  H

2881 gtgctgctgccctacgtccacgggaaccagtccagcccagagctggggcccccacggctg
 961  V  L  L  P  Y  V  H  G  N  Q  S  S  P  E  L  G  P  P  R  L

2941 cggcaggtgcggctgcaggaagcactctacccagaccctcccggccccagggtccacacg
 981  R  Q  V  R  L  Q  E  A  L  Y  P  D  P  P  G  P  R  V  H  T

3001 tgctcggccgcaggaggcttcagcaccagcgattacgacgttggctgggagagtcctcac
1001  C  S  A  A  G  G  F  S  T  S  D  Y  D  V  G  W  E  S  P  H

3061 aatggctcggggacgtgggcctattcagcgccggatctgctgggggcatggtcctggggc
1021  N  G  S  G  T  W  A  Y  S  A  P  D  L  L  G  A  W  S  W  G

3121 tcctgtgccgtgtatgacagcgggggctacgtgcaggagctgggcctgagcctggaggag
1041  S  C  A  V  Y  D  S  G  G  Y  V  Q  E  L  G  L  S  L  E  E

3181 agccgcgaccggctgcgcttcctgcagctgcacaactggctggacaacaggagccgcgct
1061  S  R  D  R  L  R  F  L  Q  L  H  N  W  L  D  N  R  S  R  A

3241 gtgttcctggagctcacgcgctacagcccggccgtggggctgcacgccgccgtcacgctg
1081  V  F  L  E  L  T  R  Y  S  P  A  V  G  L  H  A  A  V  T  L

3301 cgcctcgagttcccggcggccggccgcgccctggccgccctcagcgtccgcccctttgcg
1101  R  L  E  F  P  A  A  G  R  A  L  A  A  L  S  V  R  P  F  A

3361 ctgcgccgcctcagcgcggggctctcgctgcctctgctcacctcggtgtgcctgctgctg
1121  L  R  R  L  S  A  G  L  S  L  P  L  L  T  S  V  C  L  L  L

3421 ttcgccgtgcacttcgccgtggccgaggcccgtacttggcacagggaagggcgctggcgc
1141  F  A  V  H  F  A  V  A  E  A  R  T  W  H  R  E  G  R  W  R

3481 gtgctgcggctcggagcctgggcgcggtggctgctggtggcgctgacggcggccacggca
1161  V  L  R  L  G  A  W  A  R  W  L  L  V  A  L  T  A  A  T  A

3541 ctggtacgcctcgcccagctgggtgccgctgaccgccagtggacccgtttcgtgcgcggc
1181  L  V  R  L  A  Q  L  G  A  A  D  R  Q  W  T  R  F  V  R  G

3601 cgcccgcgccgcttcactagcttcgaccaggtggcgcacgtgagctccgcagcccgtggc
1201  R  P  R  R  F  T  S  F  D  Q  V  A  H  V  S  S  A  A  R  G

3661 ctggcggcctcgctgctcttcctgcttttggtcaaggctgcccagcacgtacgcttcgtg
1221  L  A  A  S  L  L  F  L  L  L  V  K  A  A  Q  H  V  R  F  V

3721 cgccagtggtccgtctttggcaagacattatgccgagctctgccagagctcctgggggtc
1241  R  Q  W  S  V  F  G  K  T  L  C  R  A  L  P  E  L  L  G  V

3781 accttgggcctggtggtgctcggggtagcctacgcccagctggccatcctgctcgtgtct
1261  T  L  G  L  V  V  L  G  V  A  Y  A  Q  L  A  I  L  L  V  S

3841 tcctgtgtggactccctctggagcgtggcccaggccctgttggtgctgtgccctgggact
1281  S  C  V  D  S  L  W  S  V  A  Q  A  L  L  V  L  C  P  G  T

3901 gggctctctaccctgtgtcctgccgagtcctggcacctgtcacccctgctgtgtgtgggg
1301  G  L  S  T  L  C  P  A  E  S  W  H  L  S  P  L  L  C  V  G

3961 ctctgggcactgcggctgtggggcgccctacggctgggggctgttattctccgctggcgc
1321  L  W  A  L  R  L  W  G  A  L  R  L  G  A  V  I  L  R  W  R

4021 taccacgccttgcgtggagagctgtaccggccggcctgggagccccaggactacgagatg
1341  Y  H  A  L  R  G  E  L  Y  R  P  A  W  E  P  Q  D  Y  E  M

4081 gtggagttgttcctgcgcaggctgcgcctctggatgggcctcagcaaggtcaaggagttc
1361  V  E  L  F  L  R  R  L  R  L  W  M  G  L  S  K  V  K  E  F

4141 cgccacaaagtccgctttgaagggatggagccgctgccctctcgctcctccaggggctcc
1381  R  H  K  V  R  F  E  G  M  E  P  L  P  S  R  S  S  R  G  S

4201 aaggtatccccggatgtgcccccacccagcgctggctccgatgcctcgcacccctccacc
1401  K  V  S  P  D  V  P  P  P  S  A  G  S  D  A  S  H  P  S  T

4261 tcctccagccagctggatgggctgagcgtgagcctgggccggctggggacaaggtgtgag
1421  S  S  S  Q  L  D  G  L  S  V  S  L  G  R  L  G  T  R  C  E

4321 cctgagccctcccgcctccaagccgtgttcgaggccctgctcacccagtttgaccgactc
1441  P  E  P  S  R  L  Q  A  V  F  E  A  L  L  T  Q  F  D  R  L

4381 aaccaggccacagaggacgtctaccagctggagcagcagctgcacagcctgcaaggccgc
1461  N  Q  A  T  E  D  V  Y  Q  L  E  Q  Q  L  H  S  L  Q  G  R

4441 aggagcagccgggcgcccgccggatcttcccgtggcccatccccgggcctgcggccagca
1481  R  S  S  R  A  P  A  G  S  S  R  G  P  S  P  G  L  R  P  A

4501 ctgcccagccgccttgcccgggccagtcggggtgtggacctggccactggccccagcagg
1501  L  P  S  R  L  A  R  A  S  R  G  V  D  L  A  T  G  P  S  R

4561 acaccctcgggccaagaacaaggtccaccccagcagcacttagtcctccttcctggcggg
1521  T  P  S  G  Q  E  Q  G  P  P  Q  Q  H  L  V  L  L  P  G  G

4621 ggtgggccgtggagtcggagtggacaccgctcagtattactttctgccgctgtcaaggcc
1541  G  G  P  W  S  R  S  G  H  R  S  V  L  L  S  A  A  V  K  A

4681 gagggccaggcagaatggctgcacgtaggttccccagagagcaggcaggggcatctgtct
1561  E  G  Q  A  E  W  L  H  V  G  S  P  E  S  R  Q  G  H  L  S

4741 gtctgtgggcttcagcactttaaagaggctgtgtggccaaccaggacccagggtcccctc
1581  V  C  G  L  Q  H  F  K  E  A  V  W  P  T  R  T  Q  G  P  L

4801 cccagctcccttgggaaggacacagcagtattggacggtttctagcctctgagatgctaa
1601  P  S  S  L  G  K  D  T  A  V  L  D  G  F

4861 tttatttccccgagtcctcaggtacagcgggctgtgcccggccccaccccctgggcagat
4921 gtcccccactgctaaggctgctggcttcagggagggttagcctgcaccgccgccaccctg
4981 cccctaagttattacctctccagttcctaccgtactccctgcaccgtctcactgtgtgtc
5041 tcgtgtcagtaatttatatggtgttaaaatgtgtatatttttgtatgtcactattttcac
5101 tagggctgaggggcctgcgcccagagctggcctcccccaacacctgctgcgcttggtagg
5161 tgtggtggcgttatggcagcccggctgctgcttggatgcgagcttggccttgggccggtg
5221 ctgggggcacagctgtctgccaggcactctcatcaccccagaggccttgtcatcctccct
5281 tgccccaggccaggtagcaagagagcagcgcccaggcctgctggcatcaggtctgggcaa
5341 gtagcaggactaggcatgtcagaggaccccagggtggttagaggaaaagactcctcctgg
5401 gggctggctcccagggtggaggaaggtgactgtgtgtgtgtgtgtgtgcgcgcgcgacgc
5461 gcgagtgtgctgtatggcccaggcagcctcaaggccctcggagctggctgtgcctgcttc
5521 tgtgtaccacttctgtgggcatggccgcttctagagcctcgacaccccccccaaccccgc
5581 accaagcagacaaagtcaataaagagctgtctgactgcAAAAAAAAAAAAAAAA
```

Figure 6.8: The partial nucleotide sequence of the PKD1 transcript extending 5631 bp to the 3′ end of the gene. The corresponding predicted protein is shown below the sequence and extends from the start of the nucleotide sequence. The microsatellite repeat KG8 is in the 3′ untranslated region between nucleotide 5430 and 5448.

Concluding remarks

We have characterised a gene, corresponding to cDNA 3A3, and presented compelling evidence that mutations in this so-called PBP gene give rise to the typical phenotype of ADPKD. Whereas the 3′ end of the gene contains unique coding sequence, the 5′ end is shared with at least three other genes. This part of the gene is repeated at the genomic level. This causes considerable practical implications for isolating and characterizing a full-length transcript and for the detection of PKD1 mutations. Genomic sequencing and exon linking strategies, using RT-PCR are currently being applied aimed at the upstream extension of the sequence. Other research groups are currently focusing their efforts on non-primate species, since these lack the proximal repeat block, providing a good chance of rapidly obtaining a full-length sequence. Hopefully the characterisation of the full-length sequence and its product will shed more light on its function. For the time being the function and the etiologic mechanism remains unknown. The identification of the PKD1 gene certainly is a first step toward understanding of the molecular pathology of this complex disorder.

References

Gabow PA. Autosomal dominant polycystic kidney disease: more than a renal disease. Am J Kidney Dis 1990;16:403-413.

Gabow PA, Johnson AM, Kaehny WD, et al. Factors affecting the progression of renal disease in autosomal-dominant polycystic kidney disease. Kidney Int 1992;41:1311-1319.

Germino GG, Weinstat-Saslow D, Himmelbauer H, et al. The gene for autosomal dominant polycystic kidney disease lies in a 750-kb CpG-rich region. Genomics 1992; 13: 144-151

Harris PC, Thomas S, Ratcliffe PJ, et al. Rapid genetic analysis of families with polycystic kidney disease 1 by means of a microsatellite marker. Lancet 1991;1484-1487.

Hossack KF, Leddy CL, Johnson AM, et al. Echocardiographic findings in autosomal dominant polycystic kidney disease. N Engl J Med 1988319:907-912.

Kandt RS, Haines JL, Smith M, et al. Linkage of an important gene locus for tuberous sclerosis to a chromosome 16 marker for polycystic kidney disease. Nature Genet 1992;2:37-41.

Kimberling WJ, Kumar S, Gabow PA, et al. Autosomal dominant polycystic kidney disease: localization of the second gene to chromosome 4q13-q23. Genomics 1993; 18: 467-472.

Peral B, Ward CJ, San Millán JL, et al. Evidence of linkage disequilibrium in the Spanish polycystic kidney disease 1 population. Am J Hum Genet 1994; 54: 899-908.

Peters DJM, Spruit L, Saris JJ, et al. Chromosome 4 localization of a second gene for autosomal dominant polycystic kidney disease. Nature Genet 1993; 5: 359-362.

Peters DJM, Sandkuijl LA. Genetic heterogeneity of polycystic kidney disease in Europe. Contributions Nephrol 1992; 97: 128-139.

Pound SE, Carothers AD, Pignatelli PM, et al. Evidence for linkage disequilibrium between D16S94 and the adult onset polycystic kidney disease (PKD1) gene. J Med Genet 1992; 29: 247-248.

Reeders ST, Breuning MH, Davies KE, et al. A highly polymorphic DNA marker linked to adult polycystic kidney disease on chromosome 16. Nature 1985; 317: 542-544.

Scheff RT, Zuckerman G, Harter H, et al. Diverticular disease in patients with chronic renal failure due to polycystic kidney disease. Ann Int Med 1980; 92: 202-204.

Snarey A, Thomas S, Schneider MC, et al. Linkage disequilibrium in the region of the autosomal dominant polycystic kidney disease gene (PKD1) Am J Hum Genet 1994; 55: 365-371.

Somlo S, Wirth B, Germino GG, et al. Fine genetic localization of the gene for autosomal dominant polycystic kidney disease (PKD1) with respect to physically mapped markers. Genomics 1992; 13:152-158.

6.5

IMPLICATIONS FOR TSC PATIENTS AND THEIR RELATIVES

Until 1993, molecular TSC research offered merely moral support to the patient community. The patients responded by cooperating enthusiastically in anticipation of diagnostic possibilities. Since the identification of the TSC2 gene, it is feasible to offer a more concrete service to a growing part of the TSC population.

Diagnostic possibilities

If the gene defect in a family can be assigned significantly to either chromosome 9 or chromosome 16, molecular diagnostics may be offered. This novel achievement has been made available thanks to the work of two international consortia, which resulted in a precise localisation of both genes between flanking markers (Chapters 3 and 5). On chromosome 16 the marker KG8 maps immediately adjacent to the 3′ end of the recently identified TSC2 gene (Chapter 5). On chromosome 9 the TSC1 gene maps in a relatively large interval, but a large number of informative markers flanking this interval are available (Chapter 4 and 6.1). Due to the significance of the linkage data, it is no longer necessary to demonstrate linkage by a lod score larger than 3.0 prior to using linkage data in a diagnostic test. In a genome wide search, the *a priori* chance of finding linkage is estimated to be 1/50. In TSC the *a priori* chance is increased to about 1/2, since only two locations are known. Therefore an odds ratio threshold of 1/2 x 0.05= 1/40 corresponding to a lod score of at least 1.6 can be regarded significant at a $p<0.05$ level. This lod score can be reached when a family contains 6 or more fully informative meioses. In case markers from both TSC regions show positive lod scores the equation

$$w_i = \frac{\alpha \cdot 10^{(Z_{i(x_1)})}}{\alpha \cdot 10^{(Z_{i(x_1)})} + (1-\alpha) \cdot 10^{(Z_{i(x_2)})}}$$

can be applied to calculate the significance of the assignment (w represents the posterior probability of being linked; α is the *a priori* chance of being linked; $Z_{i(X1)}$ denotes the lod score at the first (TSC1) locus). If $Z>1.6$ and $w>0.95$ molecular diagnostics by means of linkage analysis may be offered. If the proband inherited the low-risk haplotype with no recombination between the flanking markers, a remaining risk R of

$$R = (w \cdot (0.5 \cdot \theta)^2\ ) + ((1-w) \cdot 0.5)$$

has to be used in counselling (w=probability of being linked; Θ=recombination frequency between the flanking markers).

Only a minority of patients have 6 or more relatives at 50% risk for TSC. In fact, most patients are sporadic cases, carrying a *de novo* mutation. Since the identification of the TSC2 gene a proportion of familial and sporadic patients can be served by mutation analysis. At present, we are applying three techniques, aimed at finding mutations in TSC2 at three different levels. Pulsed field gel electrophoresis (PFGE) techniques are applied in order to detect deletions larger than 20 kb and large rearrangements. Since PFGE turns out to be occasionally unreliable, we are currently investigating whether FISH might replace PFGE at this level. Deletions between 20 kb and 1 kb are effectively detected by conservative southern blotting techniques. SSCP is being performed to screen for the smallest mutations. By December 1994, 168 unrelated TSC patients had been examined by our group. In 8 cases a mutation had been detected. Five of these patients showed a deletion at southern blot level, whereas 3 showed a point mutation detected by SSCP and direct sequencing. The low yield of 4.8% is due to the fact that only few exons have been investigated by SSCP yet. In theory the maximum yield is about 50%, since another 50% is assumed to be due to a mutation in TSC1. In theory, all screening methods mentioned in Chapter 1.3 can be applied. Their cost, implementation difficulties and yield will determine whether or not they will be applied. One very promising test is the protein truncation test (PTT), which combines RT-PCR with an in vitro translation. This technique can be used to detect truncated transcripts and premature stopcodons, caused by frameshifts and nonsense mutations. If SSCP studies reveal mutational hot spots within the gene, sequencing these specific parts of the gene will become an attractive option.

If one fails to find the causative mutation in a small family there still is a possibility to classify the family as TSC1 or TSC2 type and to identify the risk haplotype, although this possibility is still mainly theoretical. One would have to take and study biopsies from hamartomas of the index patient. If LOH can be found we assume that allelic loss occurred in the normal haplotype. As discussed in the next chapter this mechanism has been confirmed in two proven TSC1 families showing LOH of the normal allele. However, it is yet unknown whether exceptions to the rule exist. Therefore it seems appropriate to consider this potential diagnostic method with caution.

A mild phenotype: reduced penetrance or mosaicism?

TSC is a disorder with a variable phenotype. Most patients show a number of the pathognomonic signs listed in Table 1.2. On the other hand there are mildly affected patients, who show only atypical signs. At the end of the spectrum we find a small number of apparently unaffected gene carriers, usually described as cases of non-penetrance. Our current opinion is that non-penetrance in TSC occurs rarely and that alternative explanations have to be examined. Our mutation screening program has recently identified one individual, with aspecific CT findings, mild papules in the nasolabial fold and dental pits, who contained a TSC2 deletion in only a proportion of his blood cells (Verhoef et al., 1995). His affected child shows the same mutation in all cells, indicating that the germline of the father contains the mutation as well. We concluded from this that somatic mosaicism occurs in TSC and that some apparently unaffected parents of *de novo* patients may be mosaic for the (*de novo*) mutation. There is no information on the frequency of this phenomenon. Mosaicism is unlikely to serve as an acceptable explanation for all cases in which non-penetrance has previously been assumed

(Webb et al., 1991).

If an affected child is born to unaffected parents, it is possible that one of the parents is a non-penetrant gene carrier. Therefore a recurrence risk for the next child of about 1% seems correct. However, the occurrence of mosaicism demonstrated above, indicate the hazardous nature of such a low recurrence risk. Since no paternal age effect has ever been reported for TSC, we have to take the occurrence of mitotic rather than meiotic mutations into serious account. If mutations often occur in mitosis, germ-line mosaicism and somatic mosaicism may be a rule, rather than an exception.

A severe phenotype: TSC with infantile polycystic kidney disease

Whether there is a pattern connecting genotype to clinical phenotype is subject of much debate. The large intrafamilial variation among individuals carrying the same mutation, implies that correlation will be difficult to find (Verhoef et al., 1995). In the studies described in Chapter 5 we came across only one patient whose phenotype might be explained by the extent of the deletion. This patient, WS-53, was noted to have grossly enlarged polycystic kidneys within the first few months of life. The large deletion, detected by PFGE, involved both TSC2 and the PBP gene. This finding led the Welsh TSC research group to embark on a next study, aimed at the ascertainment of other TSC patients who had presented during early infancy with severe polycystic kidneys (Brook-Carter et al, 1994). Five more patients were identified who showed a similar phenotype. All six showed enlarged polycystic kidneys before six months of age, suggestive of a prenatal onset. All patients were shown to have a deletion involving both TSC2 and the PBP gene, deleting at least half the PBP coding sequence.

The breakpoint in family 2338 extends less far 5′ compared with the six patients described by Sampson. Perhaps this explains the relatively mild PKD1 phenotype in this family. This suggests that the phenotype with severe infantile polycystic kidney disease is mainly due to complete inactivation of PBP, rather than to a synergistic effect of the presence of mutations in both genes (Brook-Carter et al., 1994).

Renal cysts are often found in TSC patients. However, involvement of PKD1 in the cyst formation could not be demonstrated for the majority of these patients

(Chapter 5). Some of these families do not even show linkage to chromosome 16 (Nellist et al., 1993). Therefore, the patients with severe infantile polycystic kidneys have to be classified as a special subgroup of TSC2 patients.

Concluding remarks

The identification of TSC2 and the refined localisation of TSC1 have provided new possibilities for molecular diagnostics in TSC. Furthermore our understanding of the etiology of the disease is increasing. It is hoped that our findings will eventually lead to the development of new therapies aimed at the prevention and treatment of the most important symptoms of the disease.

References

Brook-Carter PT, Peral B, Ward CJ, et al. Deletion of the TSC2 and PKD1 genes associated with severe infantile polycystic kidney disease - a contiguous gene syndrome. Nat Genet 1994;8:328-332.

Nellist M, Brook-Carter PT, Connor JM, et al. Identification of markers flanking the tuberous sclerosis locus on chromosome 9 (TSC1). J Med Genet 1993;30:224-227.

Verhoef S, Vrtel R, Van Essen T, et al. Somatic mosaicism explains clinical variation in a family with tuberous sclerosis complex. Lancet 1995;345:202.

Webb DW, Osborne JP. Non-penetrance in tuberous sclerosis. J Med Genet 1991; 28: 417-419.

6.6

THE TSC2 GENE PRODUCT TUBERIN, ITS POSSIBLE FUNCTION AND SUGGESTIONS FOR FURTHER RESEARCH

TSC2 has been shown to encode a 1784 aa protein, designated tuberin. The 5.5 kb transcript contains very small 5′ and 3′ untranslated regions. Furthermore it contains a double polyadenylation signal that may explain the observed differential polyadenylation (Figure 6.9). The transcript contains 40 exons, ranging in size from less than 50 to almost 300 bp and covering a genomic region of about 45 kb (M. Nellist, M. Maheshwar, pers. comm.). The smallest intron was found to be 83 bp long, while the largest intron spans 4 kb. The promotor has not been identified with certainty yet. Reduced expression and altered transcripts have been observed in TSC patients.

The calculated molecular mass of the protein is 198 kD (Chapter 5). For various reasons, it is attractive to hypothesize that tuberin is a multidomain protein that interacts with many other proteins, for instance in a large complex. If an interaction with the TSC1 gene product can be demonstrated, this would explain the locus heterogeneity of the disorder. Another protein that might interact with

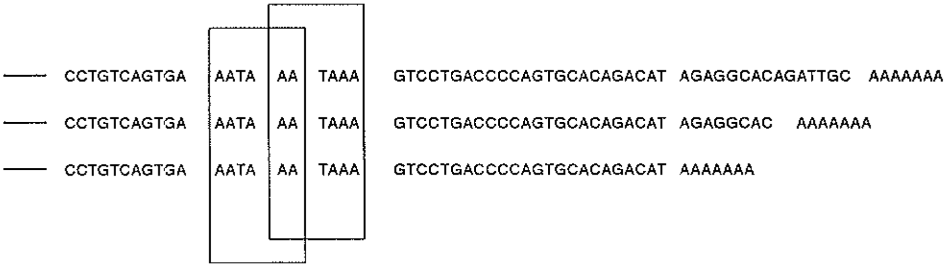

CCTGTCAGTGA AATA AA TAAA GTCCTGACCCCAGTGCACAGACAT AGAGGCACAGATTGC AAAAAAA
CCTGTCAGTGA AATA AA TAAA GTCCTGACCCCAGTGCACAGACAT AGAGGCAC AAAAAAA
CCTGTCAGTGA AATA AA TAAA GTCCTGACCCCAGTGCACAGACAT AAAAAAA

Figure 6.9: Differential polyadenylation of the TSC2 transcript. In four independent cDNA clones we observed three different polyadenylation sites. The upper line represents the sequence obtained from the 3' end of cDNA clone 1A1. Subsequent lines represent the sequences observed in clones 4.9 and 2A6 (2nd line), and AH5 (3rd line). The cleavage sites are proceeded by two potential polyadenylation signals (boxed). Both signals match 100% with the AATAAA consensus sequence. The longest distance between the last polyadenylation signal and the polyadenylation site is 39 bases (clone 1A1)

tuberin is Rap1. Of considerable interest is a small region of homology to the GTPase activating protein rap1GAP (GAP3), which suggests that tuberin itself may have a Rap1GAP activity, or plays another role in the Rap1 signalling pathway. Eventually the identification of the various domains of tuberin and the identity of the interacting proteins, will shed light on the cellular function of tuberin. So far, no conclusive evidence relating to the cellular function of tuberin has been obtained.

The TSC genes function as tumour suppressors

Loss of heterozygosity (LOH) has now been demonstrated at marker loci in both the TSC1 and TSC2 regions (Green et al., 1994a,b; Carbonara et al., 1994). The pattern of LOH at 16p13.3 is in keeping with the position of the cloned TSC2 gene. On 9q LOH has been demonstrated to affect the normal allele in chromosome 9-linked families (Chapter 6.2). So far LOH has been shown in renal angiomyolipomas, giant cell astrocytomas, a cortical tuber and a cardiac rhabdomyoma. These findings support the hypothesis that both TSC genes act as

tumour suppressor genes. For TSC2 the hypothesis is strengthened by the observation that two active alleles of the TSC2 gene are required in order to obtain a normal non-TSC phenotype. Unfortunately, origin and presumed clonality of LOH-showing cells has not yet been demonstrated. Moreover, no TSC2 expression studies in tumours have been performed. An LOH study on an angiomyolipoma of a proven TSC2 patient (5771) revealed no evidence of LOH (results not shown). Patient 5771 and his affected child (5773) have a small deletion in the TSC2 gene (Chapter 5). The mutation itself was used as marker in the LOH test. Surprisingly, both alleles were present in the angiomyolipoma at an equal ratio. Presumably a small, still undetected, mutation has occurred in the normal allele. With respect to the presumed tumour suppressor role of TSC2 it will be important to see whether the gene is also involved in other (non-TSC) tumour types and whether LOH can be found in TSC-like tumours sporadically occurring in patients not affected with TSC. Sporadic tumours are supposed to occur if somatic mutations inactivate both (normal) alleles in a cell of an individual who is not genetically predisposed. Such findings would complete the evidence in favour of the tumour suppressor hypothesis. Although not yet completely documented, LOH has recently been reported for sporadic hamartomas (Carbonara et al., 1994)
LOH affecting the normal TSC1 allele has been demonstrated in two familial cases (Carbonara et al., 1994; Chapter 6.2). We may thus provisionally conclude that both TSC1 and TSC2 act as a tumour suppressor gene.

The possible involvement of the neural crest in the pathogenesis of TSC

The neural crest is a transient embryonic structure, which forms from the neural ectoderm at the time of neural tube closure. Shortly after this formation, crest cells begin a migratory process along characteristic pathways giving rise to a variety of differentiated cell types. Both intrinsic cell lineage information and environmental cues are thought to play a role in determining the fate of these cells. Early in development, these cells can be divided into distinct populations based on their axial level of origin. Cranial neural crest cells differentiate into facial cartilage and cranial ganglia. Trunk neural crest cells normally give rise to melanocytes, neurons, and schwann of the peripheral nervous system and chromaffin cells of the adrenal medulla. Abnormal neural crest migration and development has been postulated for a number of congenital disorders, such as neurofibromatosis 1 and 2, neuroblastoma and phaeochromocytoma. TSC specific lesions mainly occur in tissues that are normally populated by neural crest derived cells. Although not all tissues affected by TSC are of neural crest origin, many interact with the neural crest during their development. Furthermore, these lesions are populated by characteristic neuron-like cells, often referred to as N-cells. Hence, many research groups, investigating the cell biology of TSC, have considered TSC as a neural crest disorder. This simplified model has been instrumental to studies on the cellular mechanisms underlying the disease (Lalier, 1991; Johnson et al., 1991).

To what extent do our recent findings modify, falsify, or support these models? It should be noted that the tumour suppressor model and the neural crest related model are not incompatible. The hypothesis that a tumour suppressor gene serves a key role in the correct routing of a migrating cell, as proposed for neurofibromatosis 1 (Schafer et al., 1993), is quite plausible. For TSC, a two-hit mechanism causing abnormal neural crest migration has been proposed by Johnson, long before LOH was demonstrated at the DNA level (Johnson et al., 1991; Green et al., 1994). The work described in Chapter 5 contributes only little data pertaining to this issue. In the absence of a signal peptide and obvious transmembrane domains it is difficult to develop a model relating tuberin to cellular migration and cell-cell interactions. However, it is very well possible that other components in the same pathway perform an extracellular signalling function. The TSC1 gene product might be a good candidate for this, but other (more complicated) models involving additional components may also be plausible.

The TSC2 gene has been shown to be conserved throughout the evolution of the higher vertebrates (Chapter 5). Southern blots containing DNA from various species, including primates, rodents, a marsupial, a reptile, a fish, a fly, a worm and many other species, were hybridised to TSC2 cDNAs and washed at various stringencies (Figure. 6.10). No signal was obtained from fish, worm or Drosophila DNA. With respect to the evolutionary conservation of TSC, it is relevant to question whether the occurrence of TSC2 parallels the evolution of the neural crest.

Although the neural crest development is very similar among species of higher vertebrates, primitive vertebrates, such as fish, show important differences. The craniofacial development is exceedingly different. Fish have a small skull and the first five somites are not incorporated in the head, as in birds and mammals. In fish the trunk neural crest is still able to give rise to cartilage. Another difference is the neural crest involvement in the development of the dorsal fin, which is only

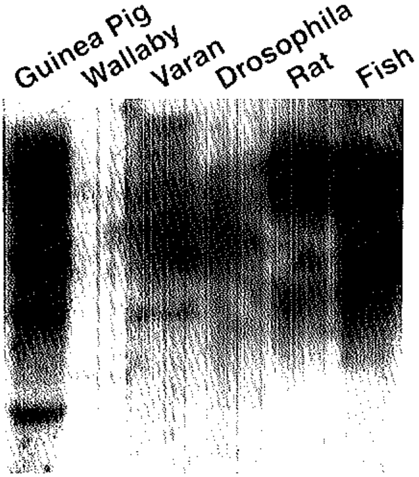


Figure. 6.10: Zoo-blot containing DNA from guinea pig (rodent), wallaby (marsupial), varan (reptile), Drosophila (insect), rat (rodent) and fish, probed with the 3' TSC2 cDNA clone 2A6 and washed until 1 x SSC at 65 °C.

present among amphibians and more primitive vertebrates (reviewed by Peters-van der Sanden, 1994). The neural crest cells of for instance zebrafish embryos are significantly larger and fewer in number than those in other vertebrate embryos. Furthermore these cells migrate along a different pathway.

Our findings are compatible with published data suggesting that tuberin may play a role in neural crest cell migration and differentiation. The tumour suppressor theory and the analogous two-hit model for neural crest derived N-cells, as proposed by Johnson (1991) imply that TSC mutations may act recessively at the cellular level. The theory predicts that nothing abnormal happens unless a second hit occurs. Unfortunately there is no actual proof that such a strict interpretation of Knudson's rule may be applied to the TSC situation. It is possible that tuberin has multiple functions. As a consequence of this, TSC2 mutations may act recessive in some cells, but have a dominant effect in others. A realistic possibility is that TSC mutations do not only affect the migration of neural crest cells, but also the migration of neural cells in the brain. Figure 6.11 shows a simplified scheme of the development of malformations in TSC. It has been shown that hamartomas result from a second hit. It remains unclear however, whether this occurs before or after the development of the neural crest and whether N-cells also show LOH. Some TSC manifestations in the central nervous system, like epilepsy and mental retardation, cannot be directly correlated with pathological findings (Jambaqué et al., 1991). Other mechanisms, like haploinsufficiency may provide a better explanation for these manifestations than a two-hit model. The involvement of the neural crest in this mechanism is unlikely.

In conclusion, it is possible that tuberin serves a supporting role in neural crest cell migration. The lesions seen in TSC can be explained by hypothesizing the presence of incorrectly migrated and abnormally differentiated neural crest cells. However, the amount of data in favour of this hypothesis is still limited. Furthermore, the precise effect of the inherited mutation in the absence of a second hit remains unknown.

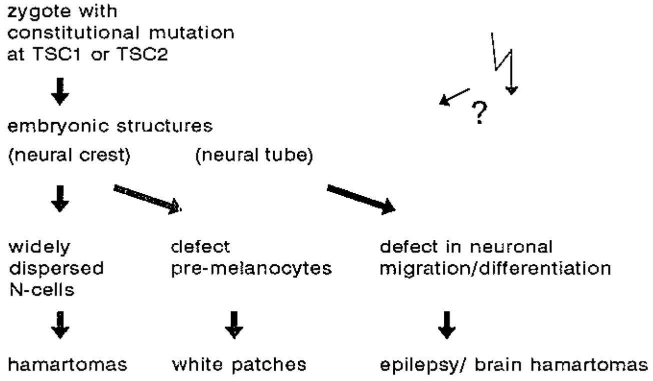

zygote with constitutional mutation at TSC1 or TSC2
?
embryonic structures
(neural crest)
(neural tube)
widely dispersed N-cells
defect pre-melanocytes
defect in neuronal migration/differentiation
hamartomas
white patches
epilepsy/ brain hamartomas

Figure 6.11: The embryonic development of malformations in TSC (simplified model). Each phase in the development of a lesion is indicated by an arrow. While some steps proceed without additional mutations, other (yet unknown) steps require a second hit. It should be noted that the relationship between the occurrence of N-cells and the formation of hamartomas is not completely certain yet.

Animal models for TSC2

The availability of animal models is of extreme importance to cell biological, pharmacological, and clinical studies on hereditary disorders. If such a model is not available, transgenic animals are usually made in order to provide such a model. Until the identification of the TSC2 gene, TSC research lacked a useful animal model.

The Eker rat was the first known animal model of inherited renal adenomas and adenocarcinomas (Eker et al., 1981). Most tumours develop from the proximal convoluted tubules. Although kidney manifestations are most prominent, other organs like the spleen and uterus are also predisposed to tumour formation. The Eker rat was recently shown to have a mutation in rTsc2, the homologue of TSC2 on rat chromosome 10 (Yeung et al., 1994; Kobayashi et al., 1995). The finding was guided by the synteny that extends between human chromosome 16p13.13 to 16p13.3 and the proximal part of rat chromosome 10, containing the Eker locus. rTsc2 is largely homologous to human TSC2. Its transcript has an identical size (5.5 - 5.8 kb) and is expressed in a wide spectrum of tissues including kidney, brain , heart, lung, muscle, liver and spleen. The GAP related domain showed only one conservative valine to isoleucine substitution at aa 1626. The Eker mutation involves a replacement of the last 1.5 kb at the 3′ end of the transcript by 3.5 - 5 kb of unknown sequence, resulting in a longer messenger of about 7.5 kb (Yeung et al., 1994). At the DNA level this might be explained as an insertion in the intron positioned between human exon 30 and 31 at codon 1272. The exact nature of the Eker gene product has not yet been completely verified. The GAP3 related domain is missing from the aberrant transcript. Rats homozygous for the Eker mutation are not viable (Hino et al., 1993). Apparently the Eker mutation is an inactivating mutation. Yeung examined 29 tumours and tumour cell lines for LOH. Allelic loss was found in 21 cases. Northern analysis on one of the lines that did not show LOH, revealed only expression of the abnormal 7.5 kb messenger, indicating that an undetected inactivating mutation had occurred. These findings confirm the tumour suppressor hypothesis with respect to TSC2 (Yeung et al., 1994).

Undoubtedly, the Eker rat will be an important animal model for future studies on tuberous sclerosis. The phenotypical difference between the Eker model and TSC in humans is striking. There are no indications of clinical epilepsy or other CNS involvement in the Eker mutation. Another striking aspect is the malignant nature of the kidney tumours. The development of renal cell carcinomas instead of angiomyolipomas indicates that tuberin may also play a more general role in renal oncogenesis. Embryological studies on the Eker rat, using TSC2 specific antibodies or RNA FISH techniques, will be very important. Hopefully these studies will teach us more about the developmental artefact that leads to TSC, the possible involvement of the neural crest in this process, and perhaps the role of TSC2 in renal carcinogenesis.

Despite the importance of the Eker mutant, the development of additional transgenic animal models with domain-specific mutations may still be necessary. Transgenic animals may also help to solve the uncertainties about the developmental stage at which effective second hits occur. The replacement of the endogenous TSC2 genes by a transgene cloned behind an inducible promoter, like the tetracycline-responsive promoter described by Furth and colleagues (1994), will enable us to observe the effect of a second hit at any developmental stage. The Cre-loxP system may also be useful in this context. loxP site specific recombination will inactivate the TSC2 transgene in any tissue where the Cre enzyme - encoded by a second transgene - is expressed (Orban et al., 1992; Gu et al., 1994).

Concluding remarks

Apart from the search for the TSC1 gene on chromosome 9, future studies will focus on the intra cellular position of tuberin and the tissue specific expression of tuberin in normal and tumour cells. Antibodies instrumental to these studies are already available, although their specificity remains to be demonstrated. Additional Polyclonal antibodies are currently being developed. Together with the ongoing search for TSC2 mutations, these studies will give us insight into the possible functions of tuberin and the domains required to perform these functions.

However, additional studies are required for building a comprehensive model for the normal function of tuberin in the cell and the etiology of TSC. It might be advantageous to study the interaction of tuberin with other proteins. A useful molecular technique aimed at cloning cDNA fragments, encoding interacting proteins, is the yeast two-hybrid system (Chien et al., 1991). This method is based on the presence of two discrete functional domains in the yeast protein GAL4, a protein that acts an activator of the yeast GAL1 gene. Recombinant strains expressing the GAL4 'DNA binding domain' and the GAL4 'activating domain' as separate proteins do not show GAL1 expression and similar strains with a GAL1-lacZ fusion gene instead of a GAL1 gene, do not stain blue. The latter can be used for making a library of cDNAs expressing proteins interacting with tuberin. While DNA encoding tuberin is cloned in an expression vector behind the GAL4 DNA binding domain, a library of unknown cDNAs is cloned behind the activating domain of GAL4. Co-transformation of both plasmids into a yeast expression strain will lead to activation of a GAL1-lacZ fusion gene if an interacting protein is encoded by the unknown cDNA. The GAL4 DNA binding domain will always bind to the promoter in front of the GAL1-lacZ gene, but blue colonies will only appear if the DNA binding domain and the activating domain are brought together by a protein-protein interaction of the proteins under investigation. Therefore, only a few colonies in the yeast library will stain blue.

Another possibility to obtain interacting proteins is to perform a series of immuno co-precipitation experiments. These experiments can be used to obtain more definite data on the function of e.g. the GAP3 related domain in tuberin. At present the relationship is purely based on sequence homology. A Rap1 binding assay may confirm that tuberin indeed binds to Rap1. As a next step, a putative GTPase activating (GAP) activity can be measured, as described by Maruta and coworkers (Maruta et al., 1991).

The experiments proposed here are not more than a few examples of studies that can be initiated immediately. An enormous number of additional experiments will have to be performed before the final goal can be reached: a sufficient understanding of the etiology and pathogenesis of TSC. At each stage we should consider whether the obtained information can be used to enhance the pace of TSC1 gene isolation studies and how the results can be used to improve the diagnostic protocols and the treatment of the patients.

References

Carbonara C, Longa L, Grosso E, et al. 9q34 loss of heterozygosity in a tuberous sclerosis astrocytoma suggests a growth suppressor-like activity also for the TSC1 gene. Hum Mol Genet 1994;3:1829-1832.

Chien C-T, Bartel PL, Sternglanz R, Fields S. The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. Proc Natl Acad Sci USA 1991; 88: 9578-9582.

Eker, R. Mossige J, Johannessen JV, Aars H. Hereditary renal adenomas and adenocarcinomas in rats. Diag. Histopath. 1981;4:99-110.

Furth PA, Onge LS, Böger H, et al., Temporal control of gene expression in transgenic mice by a tetracycline-responsive promoter. Proc Natl Acad Sci USA 1994; 91: 9302-9306.

Green AJ, Smith M, Yates JRW. Loss of heterozygosity on chromosome 16p13.3 in hamartomas from tuberous sclerosis patients. Nature Genet 1994;6:193-196.

Green AJ, Johnson PH, Yates JRW. The tuberous sclerosis gene on chromosome 9q34 acts as a growth suppressor. Hum Mol Genet 1994;3:1833-1834.

Gu H, Marth JD, Orban PC, et al. Deletion of a DNA polymerase ß gene segment in T cells using cell type-specific gene targeting.

Hino O, Klein-Szanto AJP, Freed JJ, et al. Spontaneous and radiation-induced renal tumours in the Eker rat model of dominantly inherited cancer. Proc Natl Acad Sci USA 1993;90:327-331.

Jambaqué I, Cusmai R, Curatolo P, et al. Neuropsychological aspects of tuberous sclerosis in relation to epilepsy and MRI findings. Dev Med Child Neurol 1991;33:698-705.

Johnson WG, Yoshidome H, Stenroos ES, Davidson MM. Origin of the neuron-like cells in tuberous sclerosis tissues. Ann N Y Acad Sci 1991; 615: 211-219.

Kobayashi T, Hirayama Y, Kobayashi E, et al. A germline insertion in the tuberous sclerosis (Tsc2) gene gives rise to the Eker rat model of dominantly inherited cancer. Nat Genet 1995;9:70-74

Lalier TE, Cell lineage and cell migration in the neural crest. Ann N Y Acad Sci 1991; 615: 158-171.

Maruta H, Holden J, Sizeland A, D'Abaco G. The residues of Ras and Rap that determine their GAP specificities. J Biol Chem 1991;266: 1161-11668.

Orban PC, Chui D, Marth JD. Tissue- and site-specific DNA recombination in transgenic mice. Proc Natl Acad Sci USA 1992; 89: 6861-6865.

Peters - van der Sanden M. The Hindbrain neural crest and the development of the enteric nervous system 1994. (PhD thesis, Erasmus University Rotterdam) pp. 15-32.

Raible DW, Wood, A, Hodson W, et al. Segregation and early dispersal of neural crest cells in the embryonic zebrafish. Developmental dynamics 1992;195:29-42.

Schafer GL, Ciment G, Stocker KM, Baizer L. Analysis of the sequence and embryonic expression of chicken neurofibromin mRNA. Mol Chem Neuropathol 1993; 18: 267-278.

Yeung RS, Xiao G-H, Jin F, et al. Predisposition to renal carcinoma in the Eker rat is determined by germline mutation of the tuberous sclerosis 2 (TSC2) gene. Proc Natl Acad Sci USA 1994;91:11413-11416.

SUMMARY

Linkage analysis is the most frequently used method for determination of the chromosomal localisation of disease causing genes. It involves the search for genetic markers that cosegregate with the disease trait. Occasionally, the analysis of cytogenetic abnormalities has provided complementary mapping data. If a disease shows locus heterogeneity, linkage studies are considerably complicated, as is the case in the autosomal dominant disorder tuberous sclerosis (TSC). Locus heterogeneity is defined as a situation in which only a proportion of families shows linkage to a marker in the vicinity of a disease locus, while other families do not. If locus heterogeneity is the case, a mutation in one out of multiple disease related genes causes the disease. In such a situation a modest collection of families is unlikely to be sufficient for successful localisation of one or more genes. Given the complexity of biochemical pathways, it may be expected that locus heterogeneity plays a role in a large proportion of genetic diseases.

In the case of locus heterogeneity, the major difficulty is to distinguish cases of non-linkage from families in which linkage is difficult to detect, due to one or more recombinations between the disease and a genetic marker. In the latter group information on the position of the disease gene can be obtained by mapping the recombination spots. Several statistical methods are available for the analysis of these data. The most widely used method is the admixture test (A-test). If multiple candidate regions are known, we prefer a variant of the A-test that we call the 'imaginary chromosome approach' (ICA). This powerful method combines the information from all available regions into one single analysis. We have demonstrated that the A-test and the ICA are robust, provided that linked and unlinked families show a similar size distribution. In a series of simulation studies we determined the power of the methods to detect linkage and heterogeneity in tuberous sclerosis.

Tuberous sclerosis complex (TSC) is a condition recognised particularly for its combination of neurological and cutaneous manifestations, like epilepsy, mental retardation and facial angiofibromas. However, the hamartomatous lesions that typify TSC can affect virtually all organs. Specific (pathognomonic) signs of TSC are: facial angiofibromas, multiple ungual fibromas, cortical tubers, subependymal nodules, giant cell astrocytoma and multiple retinal astrocytoma. The disorder has an estimated prevalence of 1/6000.

Preliminary evidence of linkage between TSC and markers on chromosome 9q34 was already found in 1987, but subsequently disputed. Alternative localisations on chromosome 11q or 12q have been suggested. We performed a collaborative linkage study on 128 families from Irvine, Boston, Houston, Durham (USA), Cardiff, London (UK), Erlangen (Germany) and Rotterdam (The Netherlands). The simulation study indicated that this set of families was sufficient to obtain significant evidence for linkage and heterogeneity, especially when the ICA was used. On average TSC families are very small, in most cases there are less than 2 informative meioses, which reflects the negative effect of TSC on reproductive fitness. The linkage analysis revealed that one TSC determining locus (TSC1), accounting for approximately 50% of the families studied, maps to the region of ABO and D9S10 on 9q34, but no evidence was found to support the existence of major loci on 11q or 12q. A locus elsewhere in the genome, responsible for the disease in non-9 linked families, still had to be found. The size distribution of chromosome 9 linked families was similar to that of non-linked families. This indicates that a dramatic difference in the severity of the major genetic types of TSC is unlikely.

Our next aim was to assemble YAC and cosmid contigs (groups of overlapping clones), covering the TSC1 region. We have mapped about 200 cosmids in 6 large contigs (40 to 265 kilobases (kb)). All contigs map near ABO and D9S10 and may therefore be regarded as relevant to the cloning of TSC1. Recently consensus has been achieved over D9S149 and D9S114 as markers flanking the TSC1 interval. Therefore, our YAC-cosmid contig of about 300 kb over the ABO locus and an adjacent cosmid contig of similar size over DBH, cover the core of the TSC1 region. Ongoing cosmid walking efforts aimed at extending the contigs supplemented by exon trapping and cDNA screening efforts are expected to result in the isolation of the TSC1 gene.

In 1992 linkage to markers on chromosome 16p13.3 in five non-9 linked families was reported by Kandt in the USA. An ICA linkage analysis, using 14 families from Cardiff and Rotterdam, confirmed the hypothesis that either TSC1 on 9q34 or TSC2 on 16p13.3 was responsible for the disease in each of these families. The combined analysis of both chromosomal regions revealed an overall lod score (under heterogeneity) of 8.92. We assigned the TSC1 locus to a region flanked by the Abelson oncogene and the ABO locus. This delineates the TSC1 region further to an interval of about 1-2 cM between D9S149 and ABO. On chromosome 16 the peak lod score was achieved at marker D16S291, which is also tightly linked to the locus PKD1 for autosomal dominant polycystic kidney disease (ADPKD). Later, DNA studies on TSC tumours revealed that small parts of chromosome 9 or 16, carrying the normal TSC gene, were often missing from these tumours. This indicated that both the TSC1 and TSC2 genes might be 'tumour suppressor genes'.

In the summer of 1992 two TSC research groups from Cardiff and Rotterdam and two ADPKD research groups from Oxford and Leiden formed a consortium aimed at cloning the TSC2 and PKD1 gene(s). Crucial to the consortium's efforts was the discovery of a Portuguese family segregating a translocation between chromosomes 16 and 22. In the family both TSC and ADPKD were manifest. TSC (and ADPKD) was present in one family member in whom the tip of chromosome 16 was missing (due to an unbalanced chromosome rearrangement), while family members carrying the balanced form of the translocation had only ADPKD. This suggested that either one locus was responsible for both TSC2 and ADPKD or, more likely, that TSC2 lay distal to the PKD1 locus, which was itself disrupted by the translocation. Equally important in pinning down the precise position of TSC2 was confirmation that ATR-16 (α-thalassaemia and mental retardation) patients with a slightly smaller part missing from the end of chromosome 16 did not have TSC. Through these observations it was deduced that TSC2 was probably located in a region of only 200 kb distal of the putative locus for PKD1.

By studying the 200kb area in 255 unrelated TSC patients we identified 5 large deletions, detected by pulsed-field gel electrophoresis. These mapped within a 120 kb region from which 4 candidate genes were isolated. One gene (named TSC2) was interrupted by all 5 deletions and closer examination of other patients revealed several intragenic mutations. In normal individuals, a 5.5 kb TSC2 transcript is expressed in a wide range of tissues. Its protein product (designated tuberin) contains 1784 amino acids and contains a region of homology to the Rap1 activator Rap1-GAP. A presumed function of tuberin in the Rap1 signalling pathway or an analogous pathway is in agreement with its proposed function as a tumour suppressor gene. The work on the identification of TSC2 has prepared the way for molecular diagnostics. By December 1994, 8 mutations had been detected in 168 unrelated patients analyzed in Rotterdam. Most TSC2 mutations are small and a screening program, applying sensitive techniques, started only recently. This, together with the enormous variety of different mutations, explains the low proportion of detected mutations.

Together with TSC2, we also isolated a candidate gene for PKD1. The gene maps immediately adjacent to TSC2 in a tail-to-tail orientation. It contains the translocation breakpoint - mentioned above - and is therefore called the Polycystic BreakPoint (PBP) gene. A number of intragenic mutations, found in ADPKD patients, provided further evidence that PBP is the PKD1 gene. Due to the presence of PBP adjacent to TSC2 it is possible that a single mutation affects both genes. This suggests a possible role for PBP in the development of (infantile) renal cystic disease in some TSC patients.

It is foreseen that our findings will be relevant to genetic counselling in TSC. We also expect that our work will provide insight into the pathogenesis of TSC, like the etiology of epilepsy and the development of cysts and angiomyolipomas in the kidney. The recent finding that the rat homologue of TSC2 is mutated in a classical rat model of renal carcinogenesis may imply that tuberine plays a more general role in carcinogenesis in the kidney.

In brief, statistical methods and molecular biological techniques were used in order to get insight into the pathogenesis of TSC. Furthermore, we also partly unravelled the molecular basis of another genetically heterogeneous disorder: ADPKD. It is hoped that the statistical methods described here will be beneficial to the (improved) localisation of the remaining genes for TSC, ADPKD and other heterogeneous disorders and that fine mapping of these genes will be rapidly followed by gene identification.

SAMENVATTING

In de genetica is koppelingsonderzoek de meest gebruikte methode voor het bepalen van de chromosomale lokalisatie van ziektegenen. Er wordt gezocht naar genetische markers die met de ziekte overerven en dus 'koppeling' met het ziektegen vertonen. In sommige gevallen zal daarnaast bestudering van chromosomale afwijkingen extra informatie kunnen opleveren. Indien een erfelijke aandoening locus heterogeniteit vertoont, zal dit het koppelingsonderzoek aanzienlijk compliceren, zoals bij de autosomaal dominant overervende aandoening tubereuze sclerosis (TSC) het geval is. Men spreekt van locus heterogeniteit indien slechts een deel van de families koppeling vertoont en een ander deel duidelijk niet. Er bestaan dan meerdere genen en genproducten (eiwitten), waarvan de correcte werking essentieel is voor het *niet* krijgen van de ziekte. In geval van locus heterogeniteit, zal een bescheiden collectie families doorgaans ontoereikend zijn voor de lokalisatie van één of meer ziektegenen. Met het oog op de complexiteit van biochemische pathways, mag men verwachten dat locus heterogeniteit een belangrijke rol speelt bij het merendeel van de genetische aandoeningen.

In geval van locus heterogeniteit is het moeilijk om goed onderscheid te maken tussen families die absoluut geen koppeling vertonen en families waarbij de koppeling wordt gemaskeerd door één of meer recombinaties tussen de genetische marker en het ziekte locus, veroorzaakt door uitwisseling van chromosomale segmenten ('crossing-over'). Door telkens de recombinatieplaats nauwkeurig in kaart te brengen kan informatie worden verkregen over de plaats van het gen. Verschillende statistische methodes zijn beschikbaar voor de analyse van deze data, verkregen van een heterogene aandoening. De meest populaire methode is de admixture test (A-test). Als meerdere kandidaatgebieden bekend zijn, prefereren we de 'ICA' (Imaginary Chromosome Approach), een door ons geïntroduceerde variant van de A-test. Deze methode stelt onderzoekers in veel gevallen in staat statistisch significante uitspraken te doen, aangezien alle beschikbare informatie afkomstig van de verschillende kandidaat gebieden kan worden verenigd in één

enkele analyse. We hebben aangetoond dat de A-test en de ICA betrouwbaar zijn, mits gekoppelde en ongekoppelde families qua grootte gelijk verdeeld zijn.

Kenmerkend voor tubereuze sclerosis complex (TSC) is de combinatie van neurologische en huidaandoeningen, zoals epilepsie, geestelijke achterstand en angiofibromen in het gelaat. Daarnaast kunnen goedaardige tumoren - de voor TSC typerende hamartomas - in vrijwel alle organen voorkomen. Typische (pathognomonische) kenmerken van TSC zijn: faciale angiofibromen, meerdere unguale fibromen, corticale tubers, subependymale noduli, reuscel astrocytomen en meerdere astrocytomen in de retina. De prevalentie van de ziekte wordt geschat op 1/6000.

De koppeling tussen de ziekte TSC en genetische markers op chromosoom 9 (9q34) werd voor het eerst beschreven in 1987, maar pas later algemeen aanvaard. Destijds werden alternatieve lokalisaties op de chromosomen 11 en 12 voorgesteld. In samenwerking met onderzoeksgroepen uit Boston, Houston, Durham (Verenigde Staten), Cardiff, Londen (Groot Brittannië) en Erlangen (Duitsland) hebben we koppelingsonderzoek verricht op 128 families. Computersimulaties toonden aan dat deze set families voldoende informatie bevatte voor het verkrijgen van statistische significantie, met name bij gebruik van de ICA. De plaats van het ziektelocus op chromosoom 9 kon met koppelingsonderzoek nauwkeurig worden bepaald. Het locus (TSC1) ligt nabij het ABO bloedgroeplocus en de marker D9S10 en bleek betrokken bij de ziekte in ongeveer 50% van de families. Er werd geen bewijs gevonden voor een locus op chromosoom 11 of 12, terwijl locus heterogeniteit wel afdoende bewezen werd. Hieruit werd geconcludeerd, dat een tweede TSC gen elders in het genoom gezocht moest worden. De families bleken gemiddeld vrij klein te zijn, hetgeen samenhangt met de invloed van TSC op de 'reproductive fitness'. De families vertoonden een grootteverdeling die voor de aan chromosoom 9 gekoppelde families gelijk was aan die voor de ongekoppelde families. Wij nemen daarom aan dat er geen belangrijk verschil in ernst tussen de 9q34 gekoppelde en ongekoppelde vorm van TSC bestaat.

Vervolgens werd gewerkt aan het in handen krijgen van fragmenten van chromosoom 9, gekloneerd in zogenaamde YACS en cosmides. Uit de TSC1 regio rond ABO en D9S10 hebben we ongeveer 200 cosmides verkregen, die verdeeld konden worden in 6 groepen van onderling overlappende cosmides. Deze groepen bestreken 40 tot 265 duizend DNA basen (40-265 kb). Onlangs werd overeenstemming bereikt over de markers D9S149 en D9S114 als TSC1 flankerende markers. Onze twee grootste groepen kloons van ieder ongeveer 300 kb liggen in het hart van de aldus afgebakende TSC1 regio. Meer recente inspanningen zijn gericht op het uitbreiden van het gekloneerde gebied door vergroting van het aantal verzamelde cosmides. Bovendien wordt gezocht naar genen door middel van beproefde methoden, die bekend staan als 'cDNA screening' en 'exon trapping'.

In 1992 werd door Kandt in de Verenigde Staten gerapporteerd dat vijf families, ongekoppeld aan chromosoom 9, koppeling vertoonden met chromosoom 16 markers (16p13.3). Een ICA analyse, gebruik makend van 14 families uit Cardiff en Rotterdam, bevestigde de hypothese dat hetzij TSC1, hetzij TSC2 op chromosoom 16 verantwoordelijk is voor de ziekte in de geteste families (Z_{max}=8.92). Het TSC1 gen werd toegewezen aan een chromosomaal gebied tussen het Abelson oncogen en ABO, waardoor een verdere inperking van de TSC1 regio tot het interval tussen D9S149 en ABO mogelijk werd. Op chromosoom 16 werd het markerlocus D9S291 geïdentificeerd als de meest waarschijnlijke positie van het gen. Reeds bekend was dat deze marker nauwe koppeling vertoont met polycysteuze nierziekte (ADPKD). Later werd aangetoond dat in TSC tumoren vaak een stuk van chromosoom 9 of 16 met daarop het normale TSC gen ontbreekt. Hieruit werd afgeleid dat het hier gaat om 'tumor suppressor genen'.

In de zomer van 1992 werd een consortium gevormd met een tweede TSC groep uit Cardiff en twee ADPKD groepen uit Oxford en Leiden, met als doel het kloneren van de genen verantwoordelijk voor beide ziekten (respectievelijk TSC2 en PKD1). Van cruciaal belang was een chromosomale translocatie tussen de chromosomen 16 en 22, die voorkwam in een Portugese familie met zowel TSC als ADPKD. TSC (en ADPKD) kwam voor bij een familielid bij wie het uiteinde van chromosoom 16 ontbrak. Dit terwijl dragers van de translocatie bij wie geen chromosomaal materiaal verloren was gegaan, alleen ADPKD vertoonden. We concludeerden hieruit dat het TSC2 gen in het ontbrekende stuk chromosoom gezocht moest worden; waarbij het PKD1 gen (verantwoordelijk voor ADPKD) het breekpunt markeerde. Door het consortium werden ook ATR-16 patiënten (α-thalassemie gecombineerd met mentale retardatie) bestudeerd die meer terminaal gelegen fragmenten van chromosoom 16p misten en géén TSC vertoonden. Hieruit kon worden afgeleid dat het TSC2 gen zich binnen 200 kb van PKD1 bevindt.

255 TSC patiënten werden onderzocht op grote ontbrekende stukken DNA (deleties) binnen 200 kb vanaf PKD1, hetgeen 5 deleties van enkele tientallen kilobasen aan het licht bracht. Samen beslaan deze deleties een gebied van 120 kb, waarin 4 genen aangetroffen werden. Één gen, TSC2 genaamd, werd door alle vijf

de deleties getroffen. Nader onderzoek bij andere patiënten liet ook kleinere intragene deleties in TSC2 zien. Normaliter wordt van het TSC2 DNA een RNA messenger van 5.5 kb afgeschreven, die wordt vertaald in een eiwit van 1784 aminozuren. Dit eiwit, tuberine genaamd, komt in een groot aantal weefsels tot expressie en vertoont een gedeeltelijke homologie met de Rap1 activator Rap1-GAP. Het lijkt aannemelijk dat tuberine een rol kan vervullen in de Rap1 signaaltransductieweg, gezien de tumor-suppressor functie van tuberine. Dankzij de identificatie van het TSC2 gen is de weg naar moleculaire diagnostiek geopend. Tot nu toe werd voornamelijk naar grote mutaties gezocht; tot december 1994 zijn er 8 mutaties gevonden bij 168 in Rotterdam bestudeerde patiënten. Deze resultaten suggereren dat het merendeel van de mutaties subtiele veranderingen in het DNA zijn. Mutatie screening met gevoelige methoden, gericht op het opsporen van kleine mutaties, is onlangs gestart.

Naast TSC2, werd ook een kandidaatgen voor PKD1 gevonden. De eindpunten van beide genen liggen zeer dicht bij elkaar. Het gen bevat het hierboven genoemde translocatiebreekpunt en is daarom genoemd naar dit Polycystic BreekPunt (PBP). Intragene mutaties, gevonden in ADPKD patiënten, bevestigen dat het hier gaat om het PKD1 gen. De positie van PBP in de onmiddellijke nabijheid van TSC2 suggereert dat PBP mogelijk een rol kan spelen in het (op de kinderleeftijd) ontstaan van niercysten, indien beide genen door dezelfde mutatie worden getroffen.

Voorzien kan worden dat onze bevindingen relevant zullen zijn voor de erfelijkheids advisering in TSC. Bovendien verwachten we dat ons werk inzicht zal geven in de pathogenese van TSC, zoals de etiologie van epilepsie en het ontstaan van angiomyolipomen en cysten in de nier. De recente bevinding dat het ratte-equivalent van TSC2 gemuteerd is in een klassiek rattemodel van niercarcinogenese kan impliceren dat tuberine een meer algemene rol speelt in carcinogenese in de nier.

Resumerend kunnen we stellen dat met behulp van statistische methoden en moleculair biologische technieken een beter inzicht is verkregen in het ontstaan van TSC. Daarnaast werd een bijdrage geleverd aan de ontrafeling van de moleculaire bassis van een andere genetisch heterogene aandoening: ADPKD. We hopen dat de hier beschreven statistische methoden bij kunnen dragen aan een (verfijnde) lokalisatie van de resterende genen verantwoordelijk TSC, ADPKD en andere heterogene aandoeningen en dat verfijnde lokalisatie spoedig mag worden gevolgd door gen identificatie.

AUTHORS, INSTITUTES AND CONSORTIA

MGC Department of Clinical Genetics,
Academic Hospital Rotterdam Dijkzigt and Erasmus University Rotterdam,
Dr Molewaterplein 50, 3015 GE Rotterdam, The Netherlands.
Bart Janssen (L.A.J.),
Lodewijk A. Sandkuijl,
Dick Lindhout,
Dicky J.J. Halley,
E.C. Merkens,
Mieke van der Est,
Wout Deelen,
Senno Verhoef,
Arjenne Hesseling
Marjon van Slegtenhorst
Mark Nellist
Sarvan Ramlakhan
Caroline Hermans
Ans Van den Ouweland
Bert Eussen

Institute of Medical Genetics,
University Hospital of Wales,
Cardiff CF4 4XN, Wales.
Julian R. Sampson,
Lodewijk A. Sandkuijl,
Ian Daniels,
Phillip Brook-Carter,
Mark Nellist

MRC Human Biochemical Genetics Unit,
University College London,
London NW1 2HE, England.
S. Povey,
J. Attwood

Department of Pediatrics,
University of California Irvine,
Irvine, California 92717.
P. Flodman,
M. Smith

Neurogenetics Division,
Massachusetts General Hospital,
Boston, Massachusetts.
J.L. Haines,
P. Short,
J. Amos

Department of Neurology,
Academic Medical Centre,
Amsterdam, The Netherlands.
P. Fleury

Experimental Medicine Division,
Brigham and Women's Hospital
Boston, Massachusetts.
David Kwiatkowski

Roswell Park Cancer Institute,
Human Genetic Department
Buffalo, New York.
Pieter de Jong

The Tuberous Sclerosis Collaborative Group.

Boston:
J. Amos, H. Bovey, J. Haines,
D. Kwiatkowski, P. Short
(Brigham and Women's Hospital, and Division of Experimental Medicine, Molecular Neurogenetics Laboratory, Massachusetts General Hospital, and Center for Human Genetics, Boston University, Boston, Massachusetts)

Cardiff:
P. Brook-Carter, J.M. Connor,
I. Fenton, A. Hockey, M. Nellist,
J.R. Sampson, L. Sheffield, G. Trench
(address above)

Durham:
R.J.M. Gardner, R. Kandt,
M. Pericak-Vance, A.D. Roses
(Duke University Medical Center, Division of neurology, Durham, North Carolina)

Erlangen:
R. Fahsold, P. Lorenz, H. Rott
(Institut für Humangenetik, Erlangen, Germany)

Houston:
S.H. Blanton, H. Northrup
(University of Texas Medical School, Houston, Texas)

Irvine:
K. Handar, P. Flodman, M. Smith
(address above)

London:
F. Benham, M.W. Burley, A.E. Fryer,
G. Gillet, D. Hunt, R. Mueller,
J. Osborne, S. Povey, M. Super,
D. Webb
(address above)

Rotterdam:
P. Fleury, D.J.J. Halley,
A. Hesseling-Janssen, L.A.J. Janssen,
D. Lindhout, C. Merkens, S. Verhoef
(address above)

The European Chromosome 16 Tuberous Sclerosis Consortium.

Group 1:
Mark Nellist, Bart Janssen,
Phillip T. Brook-Carter,
Arjenne L.W. Hesseling-Janssen,
Magitha M. Maheshwar,
Senno Verhoef,
Ans M.W. Van den ouweland,
Dick Lindhout, Bert Eussen,
Isabel Cordeiro, Heloisa Santos,
Dicky J.J. Halley, Julian R. Sampson
(Institute of Medical Genetics, Cardiff (address above);
MGC Department of Clinical Genetics, Rotterdam (address above);
Genetics Unit, Hospital Santa Maria, 1699 Lisbon, Portugal (IC, HS))

European chromosome 16 TSC consortium (continued):

Group 2:
Christopher J. Ward, Belén Peral,
Sandra Thomas, Jim Hughes,
Peter C. Harris
(MRC Molecular Haematology Unit, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DU, England)

Group 3:
Jeroen H. Roelfsema, Jasper J. Saris,
Lia Spruit, Dorien J.M. Peters,
Johannes G. Dauwerse,
Martijn H. Breuning
(MGC Institute of Human Genetics,
Leiden University, 2333 AL leiden,
The Netherlands)

(MN, BJ and CJW contributed equally to the work)

The European Polycystic Kidney Disease Consortium.

Group 1:
Christopher J. Ward, Belén Peral,
Jim Hughes, Sandra Thomas,
Vicki Gamble, Angela B. MacCarthy,
Jackie Sloane-Stanley,
Veronica J. Buckle, Lyndal Kearney,
Douglas R. Higgs, Peter J. Ratcliffe,
Peter C. Harris
(MRC Molecular Haematology Unit, Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, England)

Group 2:
Jeroen H. Roelfsema, Lia Spruit,
Jasper J. Saris, Hans G. Dauwerse,
Dorien J.M. Peters,
Martijn H. Breuning
(Department of Human Genetics,
Leiden University, 2333 AL leiden,
The Netherlands)

Group 3:
Mark Nellist, Phillip T. Brook-Carter,
Magitha M. Maheshwar,
Isabel Cordeiro, Heloisa Santos,
Pedro Cabral, Julian R. Sampson
(Institute of Medical Genetics,
Cardiff (address above);
Genetics Unit, Hospital Santa Maria, and
Department of Neurology, Hospital
D. Estefânia, 1699 Lisbon, Portugal (IC, HS, PC))

Group 4:
Bart Janssen,
Arjenne L.W. Hesseling-Janssen,
Ans M.W. van den Ouweland,
Bert Eussen, Senno Verhoef,
Dick Lindhout, Dicky J.J. Halley
(Department of Clinical Genetics, Rotterdam (address above))

CURRICULUM VITAE

The author of this thesis (Lambertus Antonius Jacobus Janssen) was born on the 19th of April 1962 in Haarlem. After completing his secondary education at the St. Maartenscollege in Maastricht in 1981, he moved to Leiden to study Biology at the Rijksuniversiteit.

June 25, 1985	Bachelor's degree in Biochemistry, Rijksuniversiteit Leiden.
June 1985-October 1986	Laboratory of Biochemistry, Leiden. (under supervision of Dr. B.J.C. Cornelissen)
November 1986-August 1987	Yeast Genetics Laboratory, Leiden. (under supervision of Dr. B.J.M. Zonneveld)
September 1987-March 1988	Gist-brocades NV, Delft. (under supervision of Dr. G.C.M. Selten)
April 26, 1988	Doctoral degree in Biochemistry, Rijkuniversiteit Leiden.

The work described in this thesis was performed between 1988 and 1994 in the Clinical Genetics Department Rotterdam and was the subject of two different projects.

October 1988-October 1992	Research project: 'Development of molecular genetic diagnostics of TSC and investigation of clinical and genetic heterogeneity'. (under supervision of Dr. D.J.J. Halley)
November 1992-	Research project: 'Genetic mapping of disease genes under locus heterogeneity'. (under supervision of Dr. D.J.J. Halley and L.A. Sandkuijl)

PUBLICATIONS

(All publications marked with an asterisk are included in this thesis)

Janssen LAJ, Sandkuijl LA, Merkens EC, Maat-Kievit JA, Sampson JR, Fleury P, Hennekam RC, Grosveld GC, Lindhout D, Halley DJJ. Genetic heterogeneity in tuberous sclerosis. Genomics 1990; 8: 237-242.

Haines JL, Amos J, Attwood J, Bech-Hansen NT, Burley M, Conneally PM, Connor JM, Fahsold R, Flodman P, Fryer A, Halley DJJ, Jewell A, **Janssen LAJ**, Kandt R, Northrup H, Osborne J, Pericak-Vance M, Povey S, Sampson J, Short MP, Smith M, Speer M, Trofatter JA, Yates JRW. Genetic heterogeneity in tuberous sclerosis. Study of a large collaborative dataset. Ann N Y Acad Sci 1991; 615: 256-264.

Povey S, Attwood J, **Janssen LAJ**, Burley M, Smith M, Flodman P, Morton NE, Edwards JH, Sampson JR, Yates JRW, Haines JL, Amos J, Short MP, Sandkuijl LA, Halley DJJ, Fryer AE, Bech-Hansen T, Mueller R, Al-Ghazali L, Super M, Osborne J. An attempt to map two genes for tuberous sclerosis using novel two-point methods. Ann N Y Acad Sci; 1991; 615. 298-305

* **Janssen LAJ**, Povey S, Attwood J, Sandkuijl LA, Lindhout D, Flodman P, Smith M, Sampson JR, Haines JL, Merkens EC, Fleury P, Short P, Amos J, Halley DJJ. A comparative study on genetic heterogeneity in tuberous sclerosis: evidence for one gene on 9q34 and a second gene on 11q22-23. Ann N Y Acad Sci; 1991; 615: 306-315

Fleury P, **Janssen B**, Merkens C, Sandkuijl L, Lindhout D, Halley D, Sampson J, Connor M, Smith M, Haines J, Amos J, Kwiatkowski D, Short P, Northrup H, Blanton S. Linkage heterogeneity in tuberous scerosis: a collaborative study. In: Fetal and Perinatal Neurology; 1992: pp 197-202. ed. Fakayama Y. et al. Karger, Basel.

* **Janssen LAJ**, Sandkuijl LA, Sampson JR, Halley DJJ. Computer simulation of linkage and heterogeneity in tuberous sclerosis: a critical evaluation of the collaborative family data. J Med Genet; 1992; 29: 867-874.

* Sampson JR, **Janssen LAJ**, Sandkuijl LA, and the Tuberous Sclerosis Collaborative Group. Linkage investigation of three putative tuberous sclerosis determining loci on chromosomes 9q, 11q and 12q. J Med Genet 1992; 29: 861-866.

Hagemeijer A, Buijs A, Smit E, **Janssen B**, Creemers G-J, Van der Plas D, Grosveld G. Translocation of BCR on chromosome 9: a new cytogenetic variant detected by FISH in two Philadelphia Ph-negative, BCR-positive CML patients. Genes Chromosomes and Cancer 1993; 8: 237-245.

* **The European Chromosome 16 Tuberous Sclerosis Consortium.** Identification and characterization of the tuberous sclerosis gene on chromosome 16. Cell 1993; 75: 1305-1315.

The European Polycystic Kidney Disease Consortium. The polycystic kidney disease 1 gene encodes a 14 kb transcript and lies within a duplicated region on chromosome 16. Cell 1994; 77: 881-894.

* **Janssen B**, Sampson J, Van der Est M, Deelen W, Verhoef S, Daniels I, Hesseling A, Brook-Carter P, Nellist M, Lindhout D, Sandkuijl L, Halley D. Refined localization of TSC1 by combined analysis of 9q34 and 16p13 data in 14 tuberous sclerosis families. Hum Genet 1994; 94: 437-440.

Heutink P, Haitjema T, Breedveld GJ, **Janssen B**, Sandkuijl LA, Bontekoe CJM, Westerman CJJ, Oostra BA. Linkage of hereditary haemorrhagic telangiectasia to chromosome 9q34 and evidence for locus heterogeneity. J Med Genet 1994; 31: 933-936.

- * **Janssen B**, Halley D, Sandkuijl L. Linkage analysis under locus heterogeneity: biased parameter estimates using the admixture approach. (Submitted).

- * Van Slegtenhorst M, **Janssen B**, Nellist M, Ramlakhan S, Hermans C, Hesseling A, Van den Ouweland A, Kwiatkowski D, Eussen B, Sampson J, De Jong P, Halley D. Cosmid contigs from the tuberous sclerosis candidate region on chromosome 9q34. Eur J Hum Genet (in press).

- * **Janssen B**, Halley D, Lindhout D, Sandkuijl L. Analysis of locus heterogeneity in practice. (to be submitted).

EPILOGUE

Each year an impressive number of disease causing genes is being isolated from cDNA libraries after determination of the chromosomal position by positional cloning techniques. However, the isolation of two disease genes in one single experiment is a very rare event. In March 1993, after a long series of non-productive experiments, both the TSC2 and PKD1 (PBP) genes were isolated from a fetal brain cDNA library, using a ±17 kb genomic probe. Usually, an experiment like this can be done by a single investigator. Considering the relevance of this particular experiment, it was gratifying that all members of the TSC team in Rotterdam were involved in different parts of the experiment. This was necessary because of time constraints and the large number of positive clones. As a result of this, the credits have to be shared between Mark Nellist, Arjenne Hesseling, myself and our supervisors Dicky Halley and Ans Van den Ouweland. Furthermore, it should be realised that we would not have found any mutation if Anneke Maat, Corinne Merkens, Senno Verhoef, Dick Lindhout and a large number of other clinicians did not collect the patient material. The clinical support provided by Senno and his predecessors was excellent.

Although most cDNA clones were isolated by our team, we may not neglect the contributions made by other members of the consortium. From start to finish the groups from Cardiff, Lisbon, Oxford, Leiden and Rotterdam worked together closely and each group contributed significantly to the success. We are grateful to Isabel Cordeiro, Heloisa Santos and Julian Sampson for bringing the Portuguese family with the crucial translocation to our attention. Furthermore we thank the Welch group, headed by Julian Sampson, for the contribution of PFGE and intragenic deletions and the 5′ sequence of the transcript. We acknowledge the Leiden group, led by Martijn Breuning, for the cosmids that played a key role in the isolation of the gene. The Oxford group, supervised by Peter Harris, should be recognised for providing a very useful set of genomic clones, cDNA clones from adjacent genes and for the RNA work they kindly performed on our patients. These Northern blots results provided one of the first hints that this gene might be the one we searched for so long. The TSC2 and PKD1 (PBP) genes were identified and described by the same cohesive consortium.

The identification of TSC2 was the final step in a process that started in 1988. It is essential to mention the people who initiated the project in 1988: Dicky Halley, Gerard Grosveld, Paul Fleury, Dick Lindhout and the Dutch parents and patients organisation STSN. Together with the British TSA and the American NTSA, the STSN played a special role throughout the whole project. Their members were not only the subject of our studies, but also prominent supporters of our efforts. Els den Hollander - secretary of the STSN - frequently volunteered when special tasks had to be fulfilled. When we desperately needed a more powerful computer, the NTSA provided the required funds. We always tried to interact with the patients organisations as good as possible and found their responses very stimulating.

A series of linkage studies led to the mapping of two disease causing genes. Initial results demonstrating the existence of a TSC2 gene on chromosome 16p were kindly provided by Ray Kandt, prior to publication. His findings were confirmed by loss of heterozygosity in the region, as reported by Andrew Green, and by our own linkage studies. In fact, the TSC project would have ended in 1990, if the linkage studies would not have been successful. Therefore we owe our gratitude to Lodewijk Sandkuijl, who helped us with the 'ICA', when locus heterogeneity was hampering progress. During this period the Prevention Fund of the Netherlands funded the project. When prolongation of this grant became uncertain, Professor Galjaard guaranteed the required funding. Fortunately, the Preavention Fund approved the requested extension of our grant in 1990. We thank Professor Galjaard for saving the project from a premature end. In 1992, we were able to avoid a similar situation by diversifying the project: a NWO grant enabled us to investigate the theoretical aspects of locus heterogeneity in TSC and other heterogeneous disorders, while the remainder of our funds could be used for the molecular cloning.

Between the covers of this thesis many experiments are described. Many individuals have been typed with a substantial number of markers. A major proportion of this work has been performed by our technical staff (in alphabetical order): Wout Deelen, Mieke van der Est, Bert Eussen, Caroline Hermans and especially Arjenne Hesseling. I certainly owe them a lot. Furthermore I would like to thank Jan Pertijs, Leon Testers, Magna Van der Kraan and Patrick Simons for

their tissue culture work, Guido Breedveld for providing blots with useful reference (GTS) families, Willem van Loon for assisting with the library work, Radek Vrtel for sharing his results from the TSC2 mutation screening work and the various clinicians (listed in Chapter 5), who sent us blood samples from the TSC families they examined. Since 1994 Marjon Van Slegtenhorst has joined our team. I would like to thank her, together with Sarvan Ramlakhan and Mark Nellist for their contribution to the TSC1 work, as described in Chapter 4.

The success of experiments is somehow determined by the availability of clean glassware and good coffee. Elly Hofman, Jopie Bolman and Joke Bolman should be acknowledged for this. The analysis of resulting data requires computer hardware and properly installed software, which seems impossible without the help of Peter van Vuuren, Nick Pearson, Bob Koudenburg and Lodewijk Sandkuijl. The quality of the results of the experiments is in part determined by the skills of the photographers, in our case Mirko Kuit, Ruud Koppenol, Tom de Vries Lentsch and Tar van Os. Furthermore I like to thank our secretaries Jeannette Lokker, Jacqueline du Parant and Jolanda van Deursen. I have always appreciated the company of my room mates, Annemieke Verkerk and Peter Heutink, with whom I had a lot of fruitful discussions. I also like to thank Arnold Reuser, Carel Meijers and Ans Van den Ouweland for numerous helpful discussions.

The TSC linkage studies involved the analysis of data from many institutes all over the world. I am glad that we could collaborate in this way. Therefore, I like to thank our colleagues from Boston, especially David Kwiatkowski and Jonathan Haines, our friends from Cardiff, especially Julian Sampson, the TSC group from Durham, especially Ray Kandt and Margaret Pericak-Vance, the collaborators in Erlangen, especially Raimund Fahsold, our colleagues from Houston, especially Hope Northrup, our colleagues from Irvine, especially Moyra Smith and our partners in London, especially Sue Povey and Jonathan Wolfe.

A last and therefore special place in this list of acknowledgements should be reserved for my supervisors and my family (especially my wife Helma). The continuous support of these people was crucial to my work. Dicky, Lodewijk and Dick, you taught me the essential molecular biological, statistical and clinical aspects of human genetic research. I hope you like the results. Helma, remember that my work is not the most important thing in my life. By the way, this is the last sentence; I am coming home now.