MOLECULAR CHARACTERISATION OF THE *EXTRACELLULAR MATRIX PROTEIN 1* GENE IN LIPOID PROTEINOSIS IN SOUTH AFRICA

Wesley George van Hougenhouck-Tulleken

A dissertation submitted to the Faculty of Health Sciences, University of the Witwatersrand, in fulfilment of the requirements for the degree of Master of Science in Human Genetics

Johannesburg, 2004

DECLARATION

I, Wesley van Hougenhouck-Tulleken, declare that this dissertation is my own work. It is being submitted for the degree of Master of Science in Human Genetics in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other University.

---------- ----------

Wesley van Hougenhouck-Tulleken Date

DEDICATION

I would like to dedicate this dissertation to all the people who have helped me over the last two and a bit years. There are too many of you to name individually, but please know that I am eternally grateful for all your help!

Thank you!

PRESENTATIONS AND PUBLICATIONS ARISING FROM THIS STUDY

1) van Hougenhouck-Tulleken, W. G. and Ramsay, M. *Testing for a ECM1 founder mutation in Lipoid Proteinosis patients from Namaqualand: A rapid and accurate dHPLC detection method*. Research Day, University of the Witwatersrand Medical School, 7th August 2002. (Poster Presentation)
2) Ramsay, M., van Hougenhouck-Tulleken, W. G., Jenkins, T. and McGrath, J. *ECM1 founder mutation in South African patients with lipoid proteinosis.* XIX International Congress of Genetics, Victoria, Australia, 6-11 July 2003. (Poster Presentation)
3) van Hougenhouck-Tulleken, W. G., Jenkins, T., and Ramsay, M. *Molecular Aspects of Lipoid Proteinosis in the South African Population.* Tenth Biennial Southern African Society of Human Genetics Congress, Durban, South Africa, 11-14 May 2003. (Oral Presentation)
4) van Hougenhouck-Tulleken W, Chan I, Hamada T, Thornton H, Jenkins T, McLean WH, McGrath JA, Ramsay M. *Clinical and molecular characterization of lipoid proteinosis in Namaqualand, South Africa.* British Journal of Dermatology. 2004 August; 151(2):413-23.

ABSTRACT

Lipoid proteinosis (LP) (OMIM 247100) is a rare autosomal recessive disorder that is caused by mutations within the *extracellular matrix protein 1* gene (*ECM1*). The *ECM1* gene has been shown to play a role in angiogenesis and connective tissue matrix generation, especially in skin and bone. The role of ECM1 in normal skin development and maintenance is further highlighted by its role in LP and in lichen sclerosis where autoantibodies are raised against ECM1.

LP usually presents in the first year of life with a faint or hoarse cry and is due to a hyaline-like material deposited in the mucous membranes of the vocal cords. Gradually (over years) there is diffuse skin infiltration and general skin thickening with a yellow, waxy appearance. There is excessive scarring with scars often appearing at sites of minor injury or stress. In many cases, the eyelids show typical beaded papules. In some cases, calcification of certain aspects of the temporal lobes have been observed, and may or may not be associated with variable neurological, psychiatric and neuropsychological sequelae. Although the prevalence of LP in South Africa is unknown, the disproportionately high number of case reports originating from South Africa indicates that LP is unusually common in certain South African populations, most notably the Coloured population of Namaqualand and the Afrikaans-speaking White population. This may be due to a possible LP founder effect that occurred early during the European colonisation of South Africa.

The founder effect was investigated in the South African LP patients by conducting *ECM1* mutation and linked marker analysis. The data supported a LP founder effect as the Q276X mutation in exon seven of *ECM1* was present in the homozygous state in all LP patients investigated. In addition, the Q276X mutation was associated with a single founder haplotype of 19-12-23-22 (ND1-D1S2343-D1S305-D1S2624). These markers were in significant linkage disequilibrium with each other and with the Q276X mutation.

As variation within ECM1 may alter properties of skin such as healing and scar formation, *ECM1* exons two through five and the first part of exon six were investigated for nucleotide variation using denaturing high performance liquid chromatography (dHPLC) and direct DNA sequencing in three different South African populations. Eight nucleotide variants were identified, of which six were cytosine to thymine transitions. Seven of the eight variants identified were either intronic or synonymous, with one variant being a missense variant, changing a methionine residue to a threonine residue (T130M).

ACKNOWLEDGEMENTS

1) First and foremost, I would like to thank all the LP patients, their families and all other volunteers who took part in this study and made it possible. Thank you!
2) To my supervisor, Professor Michele Ramsay, thank you for giving me so many truly wonderful opportunities, unending support and sagely advice with this project over the last two years!
3) To Dr. Tony Lane, my most heartfelt thanks for all the assistance with the statistics of this project. Thank you for always taking the time to correct my many mistakes!
4) To my collaborators, Professor John McGrath, Dr. Takahiro Hamada and Dr. Ien Chan, thank you for all the help, discussion and assistance with this project. It has helped me immensely!
5) To my girlfriend Wanjiru, thank you for your superb support, your willingness to listen and for all your help with editing this thesis!
6) To my family and friends, thank you for all your unending help, support and good will!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

1q21	chromosome 1, long arm, cytogenetic band 12
A	adenine
ABI	Applied Biosystems
ACD	acid citrose dextrose
bp	base pairs
BSA	bovine serum albumin
C	cytosine
cm	centimetres
ddH_2O	double distilled water
ddNTP	dideoxribose nucleic acid base
dHPLC	Denaturing High Performance Liquid Chromatography
DNA	deoxyribose nucleic acid
dNTP	deoxyribose nucleic acid base
dsDNA	double stranded DNA
ECM1	extracellular matrix protein 1 gene
ECM1	extracellular matrix protein 1 protein
F	phenylalanine
G (in G415S)	glycine
G	guanine
GDB	The Genome Database
H	histidine
I	isoleucine
Kb	kilobases
L	leucine
LP	lipoid proteinosis
M	methionine
Mb	megabases
mRNA	messenger ribose nucleic acid
NCBI	The National Center for Biotechnology Information
ND1	Novel Dinucleotide 1
nm	nanometer
nt	nucleotide
^oC	degrees centigrade
OMIM	Online Mendelian Inheritance in Man
P	proline
PAS	periodic acid–Schiff
PCR	polymerase chain reaction
ppm	parts per million
Q	glutamine
R	arginine
RFLP	restriction fragment length polymorphism
RT-PCR	reverse transcription polymerase chain reaction

S	serine
ssDNA	single stranded DNA
T (in T130M)	threonine
T	thymine
Taq	*Thermus aquaticus*
TBE	Tris-EDTA-Boric acid buffer
TRF	Tandem Repeat Finder
UCSC	The UCSC genome bioinformatics browser
V	volts
W	tryptophan
WT	wild type
X (in ^{o}C)	variable temperature in degrees centigrade
X	stop codon
μm	micrometer

CHAPTER 1 INTRODUCTION

Lipoid proteinosis (LP) (OMIM 247100), is a rare autosomal recessive disorder caused by mutations within the extracellular matrix protein 1 gene (*ECM1*).[1] LP is also commonly known as hyalinosis cutis et mucosae, lipoproteinosis or lipoid proteinosis of Urbach and Wiethe. LP was first described by Urbach and Wiethe in 1929.[2]

The gene mutated in LP has been shown by Hamada *et al.*[1] to be the *ECM1* gene. Currently, 24 LP mutations are known,[3] (I. Chan, personal communication) most of which have been found in the homozygous state in LP patients. Very few compound heterozygotes have been detected. The mutation responsible for LP in South Africa has been identified as the Q276X mutation in exon seven of the *ECM1* gene. This mutation has not been described in any other population.

While the genetic cause of LP is now known, the exact pathogenesis of LP is not well understood. The clinical manifestations have however been well documented, but show a great deal of variation.[3,4] The most common clinical features of LP are a hoarse voice, excessive scarring of the skin, a widespread warty hyperkeratosis, beaded eyelid papules (moniliform blepharosis), a thickened sub-lingual frenulum and extensive scarring of the skin[3,5-8]. These symptoms may be partially explained by the deposition of a hyaline-like material under the skin or within the relevant tissue. An additional observation is that some LP patients exhibit psychological or neurological problems. This may be correlated with “bean-shaped” calcifications in the region of the anteromesial temporal lobes.[9]

Although rare worldwide, the apparently high prevalence of LP in South Africa is mirrored by the largest number of LP case reports originating from South Africa. The majority of these LP case reports are from the Namaqualand region of South Africa. All confirmed South African LP patients have White ancestors, making it highly probable that European colonists introduced LP into South Africa. This founder effect has been substantiated by extensive genealogical studies on the South African LP patients in the late 1960s and early 1970s by authors such as Heyl and Botha.[10-13]

1.1 The Clinical Diagnosis of LP

The clinical symptoms of LP vary substantially, not just between population groups, but also within population groups. This makes a definitive diagnosis very difficult to perform. To complicate matters, LP is rare, with multiple LP cases in a single population or between populations occurring rarely. This makes comparative clinical studies very difficult to conduct. When a large group of patients is identified, it is a reasonable expectation that there will be significant age differences between the patients. This complicates comparative clinical LP studies, as the clinical symptoms of LP are presumed to progress with age,[4] with the symptoms becoming more detectable and pronounced with time. However there is a plethora of case reports on LP patients in the literature, and from these sources, it is possible to extract the most universal clinical symptoms of LP, some of which are presented in Figure 1-1.

A clinical diagnosis should be suspected when a patient presents with a hoarse voice that has been present since birth or early childhood, generally thickened and scarred skin, beaded eyelid papules and a thickened sub-lingual frenulum. Other confirmatory symptoms are infiltrated warty skin papules especially on the elbows and extensor aspects of the forearms and oval bilateral, symmetrical calcifications of the anteriomedial aspect of the temporal lobes (Figure 1-1).

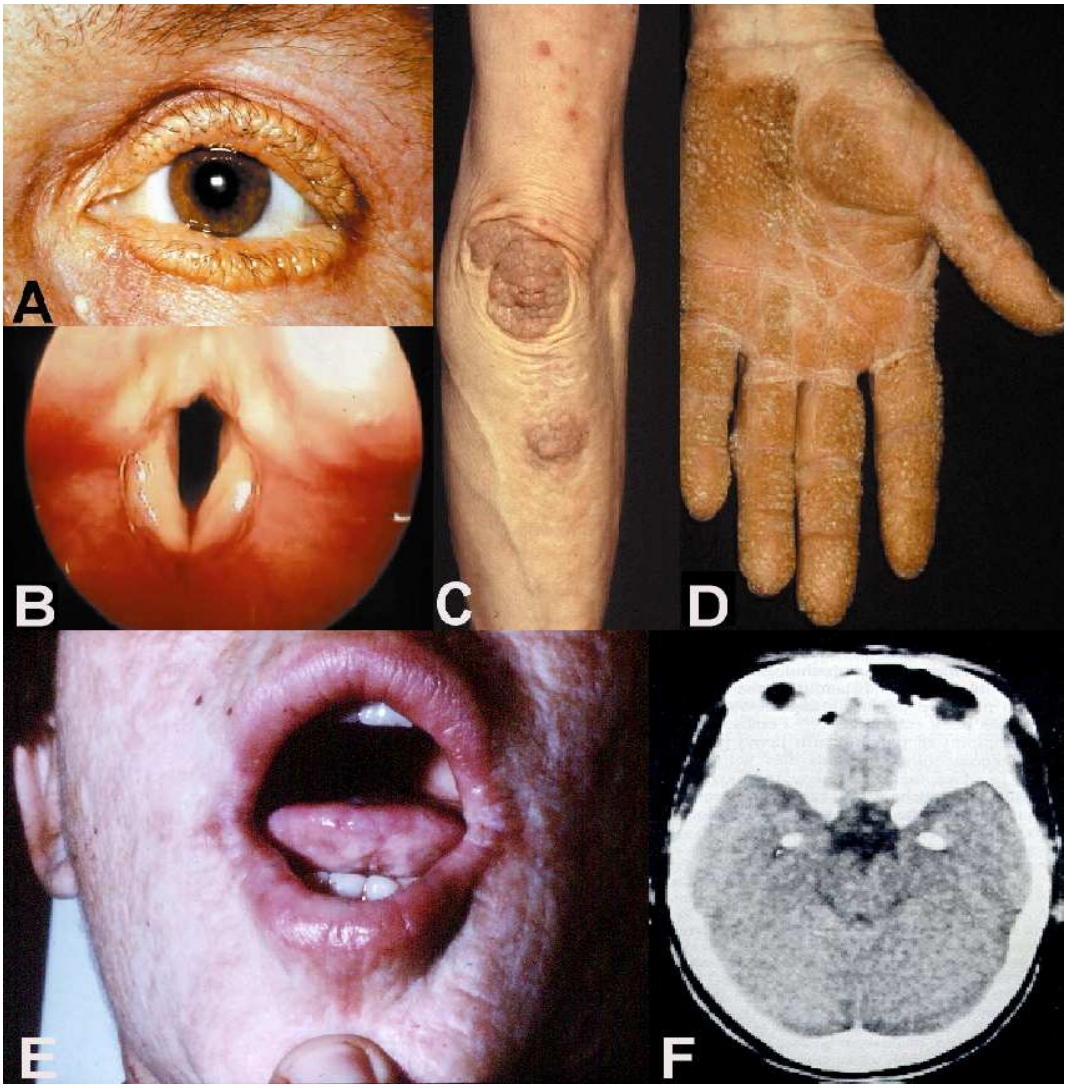


Figure 1-1: Common LP manifestations

A: Beaded eyelid papules. B: Thickened vocal cords causing the typical hoarse voice. C: Warty, infiltrated skin papules on the elbow and scarring of the arm. D: Warty hyperkeratosis of the palm. Images A to D are taken from Hamada *et al.*, 2002.[1] E: Thickened sub-lingual frenulum leading to reduced tongue movement and inability to protrude the tongue normally from the mouth. Note the excessive scarring of the face and thickened lips. Courtesy of Prof. T. Jenkins. F: Oval bilateral, symmetrical calcifications of the anteriomedial aspect of the temporal lobes. Taken from Emsley and Paster, 1985.[14]

1.1.1 Skin Manifestations of LP

Next to the hoarse voice, the most noticeable clinical manifestation of LP is excessive scarring and infiltration of the skin, which usually develops within the first years of life.[8] There are recurrent pustules or bullae on the face and on the exposed surfaces of the arms and legs, which can result in white variloform or acneiform scars. As the patient ages,

skin lesions can include yellow, waxy papules, nodules and verrucoid plaques that are commonly on the elbows and knees.[15,16] Eventually the skin takes on a characteristic yellowish colour, has a waxy texture and is thickened (hyperkeratosis), especially at areas where the skin is stretched constantly, for example the elbows, hands and knees.[8,17] The skin of the hands and elbows may show infiltrated, warty papules (Figure 1-1 D, Figure 1-2 D), which may be exacerbated by excessive manual labour or prolonged exposure to detergents. Scarring of the skin occurs typically after trauma, but there have been cases reported where the scarring has occurred spontaneously. Scarring usually begins in childhood and is particularly noticeable on the face.[17] Interestingly, it has been reported that scarring may not occur after surgery, such as tracheostomies, or after vaccination. However, scarring will often occur at sites of minor stresses such as grazes or the stretching of the skin.[4]

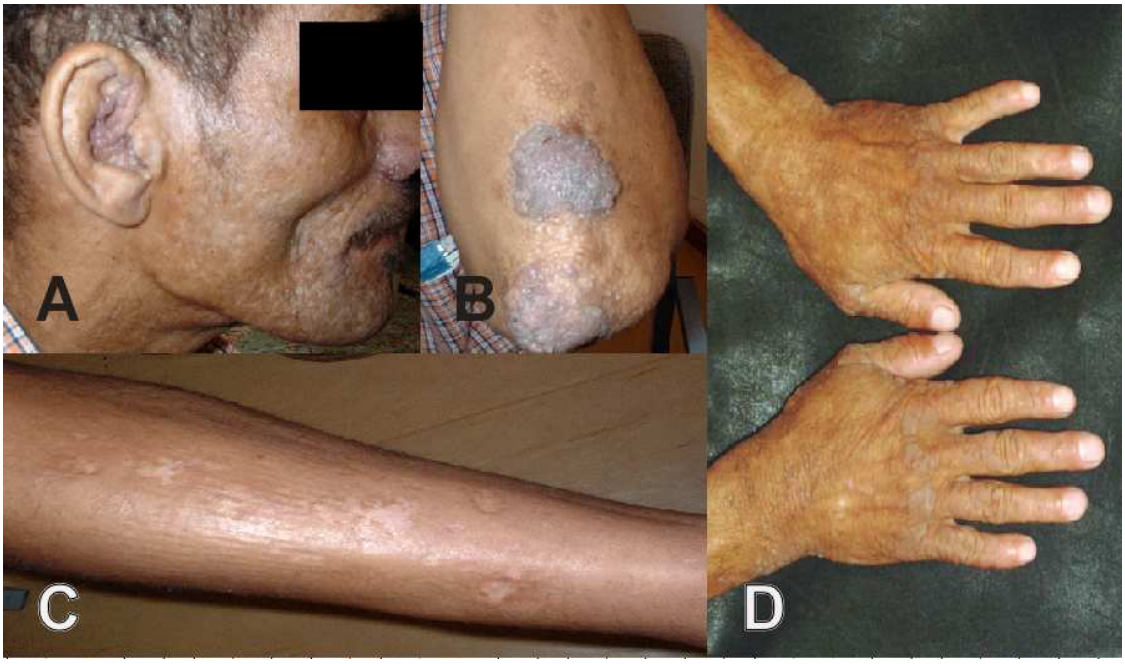


Figure 1-2: Common LP skin manifestations

A: Excessive scarring of the face with warty, infiltrated papules in the patients ears. B: Warty, infiltrated papules of the elbows. C: Typical pockmark scarring of the calf. D: Hyperkeratosis of the hands. Images are of South African LP patients only.

1.1.2 Psychological and Neurological Manifestations of LP

It has been reported that certain LP patients may show some form of psychological or neurological problems, with the most common symptom being minor nervous disturbances and/or epilepsy. It has also been noted that some LP patients have oval or bean-shaped calcifications in the anteriomedial aspect of their temporal lobes (Figure 1-1 F).[14] While this may be the reason behind the psychological problems seen in LP patients, the correlation between the calcifications and the psychological problems has not been thoroughly investigated.[3]

1.1.3 Biochemical and Histological Manifestations of LP

The main manifestations of LP are related to the deposition of an amorphous hyaline-like material in connective tissues. This hyaline-like material displays a periodic acid–Schiff (PAS) positive reaction, is resistant to enzyme digestion and is eosinophillic. The hyaline material is often deposited around the walls of small and medium sized blood vessels in the papillary dermis, which then shows a typical "onion-skin" proliferation of the adventitia. Under electron microscopy, the deposit is comprised of small granules and short filaments that are electron lucent.[6]

Harper *et al.*[18] have shown that LP skin has significantly less collagen per unit dry weight than normal skin. The total collagen content of dried LP skin was 36%, compared to normal skin having 52-53%. However, a generalized increase in basement membrane collagens was found in LP patients, with elevated levels of collagen type III with respect to collagen type I. Immunohistochemistry has shown that there is a patchy and diffuse distribution of collagen type III and a relative increase in collagen types IV and V. This is associated with an "onion skin" appearance around endothelial basement thickening. Estimation of collagen cross-links showed an abnormal pattern with a preponderance of the keto-imine form not normally associated with skin. Harper *et al.*[18] puts forward that these results suggest that LP involves a primary perturbation of collagen metabolism. Also, no abnormal collagens were purified from LP skin after pepsin degradation.

Bauer *et al.*[19] reported numerous vacuoles in LP fibroblasts that are not present in normal fibroblasts. When LP fibroblasts were examined under electron microscopy, the fibroblasts displayed striking similarities to known lysosomal storage disorders and this lead Bauer *et al.*[19] to postulate that the material in LP patients represents a complex glycoprotein present in increased amounts and is likely to be as a result of a single or multiple enzyme defects.

It is important to note here that a limited number of cases (usually not more than three patients) were investigated in each histological and/or biochemical study that was carried out on LP skin and hence it is not advisable to extrapolate these results to all cases and to generalise about a situation in LP.

1.2 LP in South Africa

The relatively common occurrence of LP in South Africa has led many researchers[10-13] to conclude that there has been a founder effect that occurred early on in South Africa's history.[20] The populations affected by LP in South Africa are the Afrikaans-speaking Coloured population of Namaqualand and the Afrikaans-speaking White population.

As LP has not yet been documented in the Black or Khoisan populations of South Africa, it is reasonable to conclude that LP is not common in the South African Black population or the Khoisan population of Namaqualand. Therefore the White population is the most likely population to have introduced the LP mutation into the South African population. This is supported by genealogical studies[10-13,20] that have traced the pedigrees of many of the White and Coloured families with LP back to the following ancestors of European descent: Gerrit or Elsje Cloete, Jan van den Heever and Piet Engelbrecht. The predominance of the surname Cloete and the fact that many LP family pedigrees investigated by Heyl[11,12] and Gordon *et al.*[13,20] have shown at least one common ancestor in either Gerrit or Elsje Cloete, allows one to conclude that this Cloete sib pair is likely to be responsible for introducing the LP causing mutation into the South African population. Gerrit and Elsje Cloete's ancestry traces back to their great-grandfather, Jacob Cloete

from Cologne, Germany. Jacob Cloete's son, who was also named Jacob Cloete, was one of the initial settlers of the Cape colony with Jan van Riebeck in 1652, and is thought to be the progenitor of the large number of Cloete families in South Africa today.

In a separate study on the Rhenish Mission in South Africa, Strassberger[21] mentions a Jasper Cloete who arrived in the settlement at Komaggas in Namaqualand in 1790 when he was rejected and expelled from his family farm in the Kammies mountains. His three White half-brothers rejected him, as he was a "half-caste" or Coloured. It is not clear as to how or if Jasper is related to Jacob Cloete. However Jasper may be responsible for introducing the LP mutation into the Coloured population of Namaqualand, while his half-brothers, together with the other descendants of Elsje and Gerrit Cloete, are likely to be the progenitors of the individuals with LP in the White population of South Africa.

This presents strong evidence that the mutation responsible for causing LP was introduced into South Africa with the very first White settlers and that the original founder may have entered South Africa as early as 1652.

1.3 Founder Disorders in South Africa

The first White settlers arrived in South Africa in 1652 and at this time the White population was relatively small. This population was likely to have been a representative sample that did not differ significantly from the parent European population. However, due to the small population size of the original White settlers, it is probable that the frequency of many genotypes would have been distorted, raising (or lowering) the frequency of some rare genotypes and lowering the frequency of other common genotypes. This raising of rare genotype frequencies due to a small subsection of a population migrating from the parent population and establishing a new population with the same genotypes but at differing frequencies, is termed a founder effect.

The founder effect can be more or less 'fixed' if the new population expands rapidly. If the population size remains moderate, consanguinity is likely to be relatively common

and the frequency of recessively inherited conditions will be increased. The above scenario is likely to have occurred in South Africa, as there are many South African founder effect examples. Genetic disorders such as variegate porphyria,[22,23] familial hypercholesterolemia,[24] pseudoxanthoma elasticum[25,26] and Fanconi anemia[27] are remarkably common in South Africa, and the relatively high frequency of each disorder is thought to be primarily due to a founder effect. The possible addition of LP to this list is therefore not surprising, as Gorden[13,20] and Heyl[10,11] have shown that all LP patients they studied could be traced to a sib pair who arrived in South Africa shortly after 1652. However, further molecular genetic work will be needed to substantiate Gordon and Heyl's findings.

1.4 Identification of *ECM1* as the "LP" gene

Early in 2002 Hamada *et al.*[1] announced that they had identified the gene that was mutated in LP patients. The gene mutated was the *ECM1* gene, and it was noted that all the populations with LP that were studied had a mutation unique to each population. Compound heterozygotes were very rare.[1,3]

In order to identify the LP gene, Hamada *et al.*[1] used a standard genome wide linkage analysis to identify a 2.3 centiMorgan interval on chromosome 1q21 that showed strong linkage at marker D1S498, with a maximum two point LOD score of 21.85 ($\theta = 0$). The critical interval was then focused on a ~6.5Mb section of 1q21 between markers D1S2344 and D1S2343 using observed recombinations within the haplotypes. This interval contained 68 known genes or *in silico* predicted genes. A positional candidate gene approach was then used to identify which of these 68 genes was the LP gene. To narrow down the number of positional candidates, the mRNA from LP fibroblasts was examined for mRNA levels of each candidate gene. If there is a homozygous nonsense mutation within one of the candidate genes, the resultant mRNA from that gene is thought to have a shortened half-life due to a process known as nonsense mediated mRNA decay. This would result in low levels of the candidate gene's mRNA present in an affected tissue.[28] This is contrasted with mRNA levels of the same gene in an

unaffected individual's tissue.[29] Therefore, as LP is primarily a skin disorder and as there have been reports that fibroblasts are abnormal in LP patients,[19] mRNA derived from LP fibroblasts were investigated by Hamada *et al.* for nonsense mediated mRNA decay of the 68 positional candidate genes. This was accomplished using reverse transcription PCR (RT-PCR). The results were compared to results similarly obtained from normal fibroblast mRNA. *ECM1* mRNA was found to be decreased in LP fibroblast mRNA relative to *ECM1* mRNA from normal fibroblasts. *ECM1* was therefore a good candidate gene and was investigated further.

Direct sequencing of *ECM1* from genomic DNA from LP individuals allowed Hamada *et al.*[1] to identify three nonsense mutations: a single nucleotide insertion, a single nucleotide deletion and a large (1.1kb) deletion. Each mutation was found in a separate population and each was present in the homozygous state in affected individuals, in the heterozygous state in obligate carriers and none of these mutations were found in 80 ethnically diverse normal controls. Further studies have now identified that there are at least 24 mutations that are pathogenic for LP.[3] (I. Chan personal communication)

1.5 The *ECM1* gene

The *ECM1* gene is located on chromosome 1q21.2, is comprised of 11 exons and spans 5454 nucleotides. It is located close to, but not within, the epidermal differentiation complex on chromosome 1.[30] *ECM1* has two alternatively spliced exons, exon 5a and exon 7 (Figure 1-3). This results in three ECM1 isoforms, namely ECM1a, which contains all exons except for exon 5a, ECM1b, which contains all exons except for exons 5a and 7, and ECM1c, which contains all 11 exons.

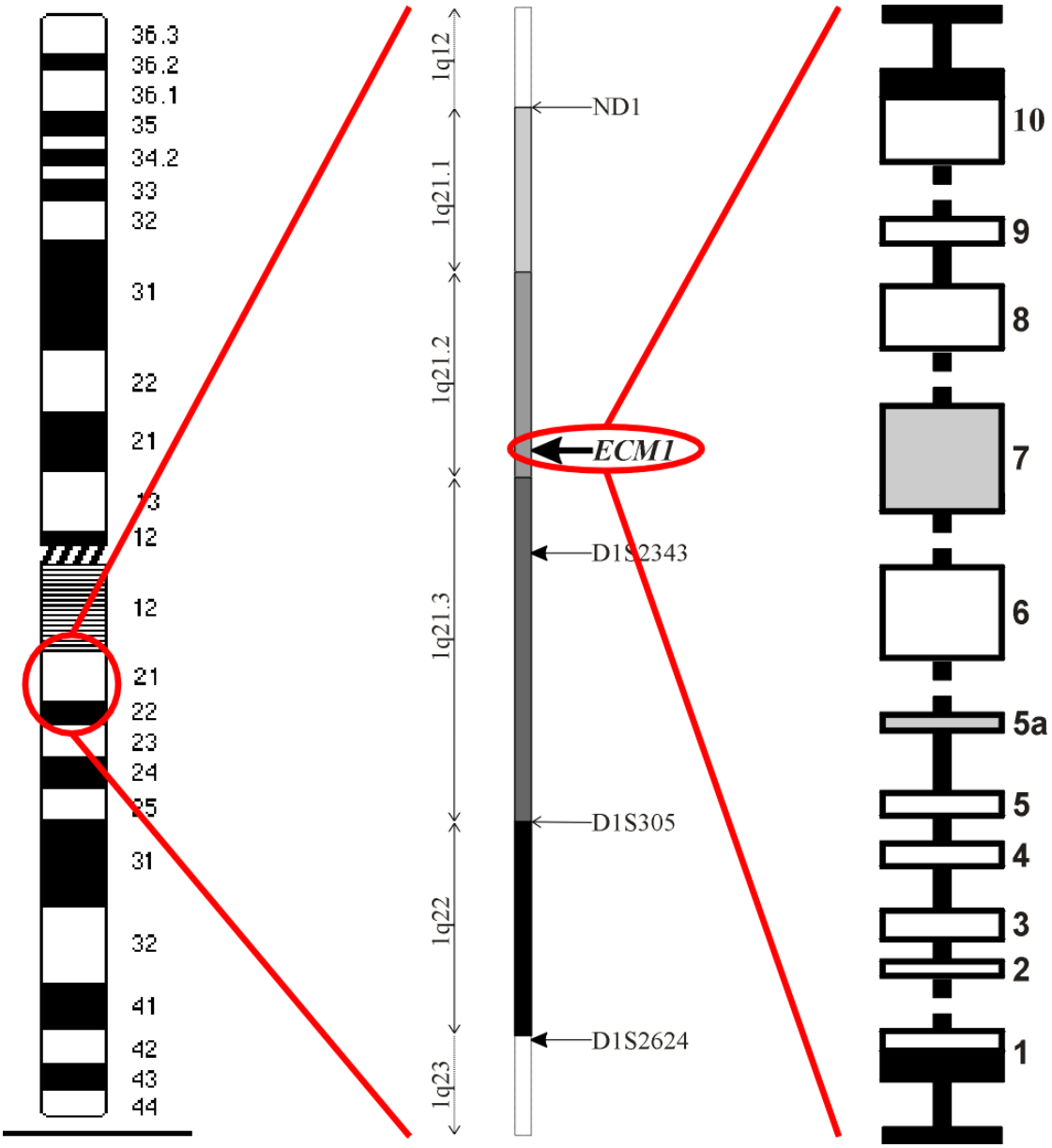


Figure 1-3: Genomic positioning of *ECM1*

The image on the left shows chromosome 1 with the banding patterns obtained with standard giemsa staining. The image in the middle shows an enlarged picture of the 1q12 to 1q23 area of chromosome 1 and indicates where *ECM1* is found in band 1q21.2. The middle image shows the position of the four microsatellite markers used in this study. All data for the middle image were compiled from the UCSC genome browser July 2003 freeze. The rightmost image is an enlarged image of the *ECM1* gene. The numbers refer to the different exons of *ECM1*; the broken lines indicate areas not drawn to scale. The shaded exons indicate exons that are alternately spliced (exons 5a and 7) while the blacked out parts of exons 1 and 10 indicate the 5' and 3' untranslated regions, respectively.

ECM1a is expressed predominantly in the heart and placenta, but is also found in the skin, liver, ovary, kidney, lung, pancreas, testis, muscle and colon,[31,32] while expression

of ECM1b has been described only in skin and tonsils.[30,33] ECM1c expression has only been detected in two cancer cell lines[31] and is thought to constitute ~15% of total *ECM1* mRNA expression in those cells. In other studies, *ECM1* has been implicated in promoting angiogenesis in the chorioallantoic membrane of the chicken egg and to promote endothelial proliferation in the mouse,[34] while *ECM1a* has been shown to regulate bone alkaline phosphatase activity.[35] Very little is known about the expression pattern of ECM1b, except that it is expressed after cultured keratinocytes have almost completely differentiated (ECM1a is expressed throughout all three stages of differentiation).[35] The expression profile of ECM1c is unknown. Additionally, increased *ECM1* expression elucidated by microarray experiments has been reported in cartilage formation,[36] dendritic cell differentiation and maturation,[37] as well as in grade I, II and IV glioblastoma multiformes.[38] Apart from dendritic cell differentiation and maturation, these processes involve extracellular matrix generation and vascularisation of the respective matrix/tissue, which further accentuates *ECM1*'s role in angiogenesis and metabolism of extracellular matrices in general.

The ECM1 protein is made up of five domains, of which the first domain is a cysteine free putative signal peptide at the N terminus consisting of 19 amino acids. The second domain is also a cysteine free domain (from amino acid position 20 to 150), while the third and fourth domains are two tandemly repeated domains (positions 151 to 279 and 280 to 405, respectively) each with a cysteine repeat pattern of CC-(X_{7-10})-C.[30] This pattern is strikingly conserved between the mouse and human, with the number of cysteines being almost identical (28 residues in the human protein compared to 29 residues in the mouse protein. The extra cysteine residue in the mouse is found in the signal peptide). The last domain is a C-terminal region that also contains cysteines arranged in the CC-(X_{7-10})-C pattern. The CC-(X_{7-10})-C pattern is predicted to form double loop domains, which are involved in ligand binding in the serum albumins.[30](*Op. Cit*)

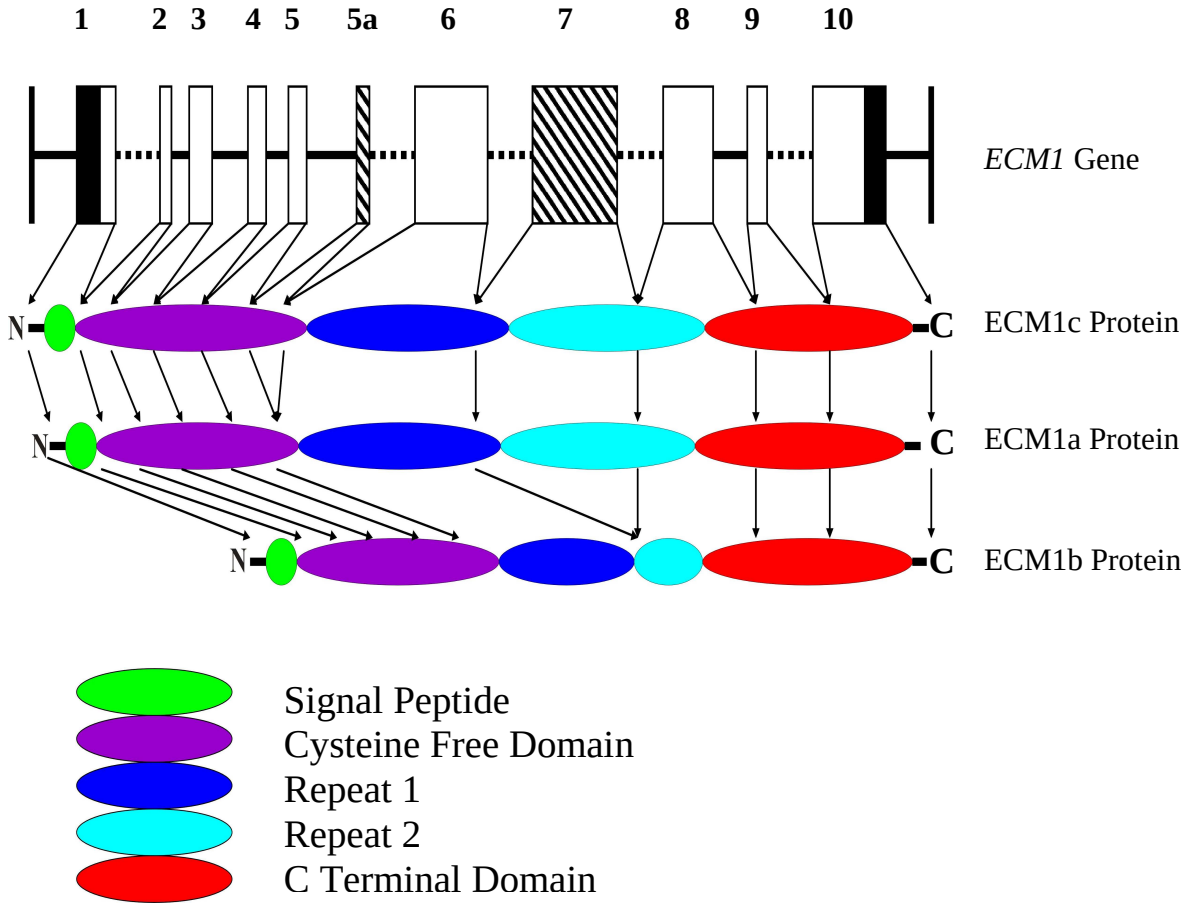


Figure 1-4: Schematic representation of the three ECM1 protein isoforms showing their respective domains in relation to *ECM1* exons one to ten

The conservation of the cysteine pattern between mouse and humans implies that the ECM1 protein is under severe structural constraints for these cysteine double-loop domains and hence these structures are likely to be important for ECM1 function. ECM1b has exon seven alternatively spliced out and this removes the second tandemly repeated domain, leaving only two CC-(X_{7-10})-C patterns. This may alter the ligand binding properties of ECM1b. It is interesting to note that the majority of LP causing mutations are within the seventh exon or occur before it. This implies that the ECM1a isoform is the isoform that plays the major biological role in humans. With regard to the South African LP mutation (Q276X), the ECM1b isoform would have the mutation spliced out and could theoretically function normally. However this does not "rescue" the LP patient from the pathological affect of the Q276X mutation, which therefore highlights the biological importance of ECM1a. The addition of exon 5a into the ECM1c

protein is predicted to increase the length of the N terminal domain by 12 amino acids, but not to affect the cysteine pattern. The function of ECM1c is therefore arguably identical to that of ECM1a.

With regards to known mutations that are pathogenic for LP, all 24 mutations result in the perturbation of ECM1a and ECM1c expression, while ECM1b expression is only affected by the mutations outside exon seven (15/24 of LP mutations lie outside *ECM1* exon seven). This again highlights the biological importance of ECM1a and possibly ECM1c, as ECM1b cannot rescue the LP phenotype when a mutation occurs within *ECM1* exon seven, which could be alternatively spliced out before translation.

Recently, Mongiat *et al.*[31] published a paper in which he reports that ECM1 interacts with domain V of perlecan in a specific manner. Perlecan itself is a modular heparan sulfate proteoglycan that is an intrinsic constituent of basement membranes. It has been implicated in roles of vasculogenesis, thrombosis, tumorigenesis, and angiogenesis and acting in concert with other extracellular matrix molecules, it maintains structural integrity of vessel walls.[31] Interestingly, ECM1 and perlecan are co-expressed in similar regions of the dermis, and notably within the collagen-rich regions of the upper dermis. This co-expression pattern is interesting as the basement membranes immediately around vessels in the dermis in LP patients show a thickening or re-duplication. Therefore the lack of ECM1 interacting with perlecan in LP individuals may be a direct or indirect cause.

1.6 Implications of *ECM1* as the LP gene

Once the *ECM1* gene had been identified as the "LP" gene, various interesting points became evident. Hamada *et al.*[1] noted that there was strong linkage to marker D1S498, with a two point LOD score of 21.85 ($\theta = 0$) with six LP families. The LOD score dropped to 3.45 (which is a borderline value for LOD scores, a LOD score of 3.00 and above is considered significant) when five of the families were removed from the analysis. These five families totalled 28 individuals, of which 22 were actually a cohort

of South African LP individuals who were assumed to be related because of the proposed founder effect put forward by Heyl and Gordon.[10-13,20]

The study by Hamada *et al.*[1] also noted that mutations within the alternatively spliced exon seven of *ECM1* seemed to result in a milder phenotype than phenotypes that arose from mutations outside exon seven. Subsequent studies by Hamada *et al.*[3] have noted that there is considerable phenotypic variation of LP between and within families. Also, in a recent genotype-phenotype correlation study on Namaqualand LP individuals, it was noted that while the LP individuals studied were all homozygous for the Q276X mutation (i.e. they were genotypically identical), there was considerable phenotypic variation (J. McGrath, personal communication). The lack of genotype-phenotype correlation may be partially explained by the observation made by Hamada *et al.*[3] who noted that partial reinstatement of *ECM1* expression was made possible by restoration of the open reading frame of *ECM1* mRNA by the splicing out of exons three and four. The patient was homozygous for the mutation 243 del G (nt 1732 del G) in exon four and the clinical manifestation of LP in this individual was noted to be mild. This restoration of the reading frame was termed aberrant by Hamada *et al.* as many other alternatively spliced *ECM1* transcripts were seen that did not reinstate the open reading frame. Therefore, the lack of genotype-phenotype correlation seen in the Namaqualand LP patients and LP patients worldwide may in part be explained by partial rescue of *ECM1* expression by the in frame splicing out of the mutated exon. Therefore variation that alters *ECM1* splicing may be able to reduce the symptoms of LP and hence explain the lack of genotype-phenotype correlation. Additionally, variation in other genetic factors whose protein products play a role in the same physiological pathways as ECM1 cannot be excluded as a source of the discordance between genotype and phenotype in LP individuals.

ECM1 is involved in many important physiological processes such as angiogenesis, which is an important factor in tumorigenesis and connective tissue matrix formation such as skin and bone formation. Variation that alters the activity or amounts of ECM1 produced will be of great scientific and possibly medical interest. Association studies could be done to examine the role of ECM1 in scar formation, in solid tumour

vascularisation and in other common skin and bone abnormalities, such as acne and osteoporosis.

AIMS

The objective of this study was to investigate and provide scientific support for the LP founder effect within South Africa and to identify polymorphic variants within the *ECM1* gene in representative South African populations. To accomplish this, the following aims were pursued:

1) To identify the gene mutations responsible for LP in South African LP patients from different populations.
2) To estimate the carrier frequency of the Q276X mutation in the Namaqualand Coloured population, the Gauteng White population and the South African White population.
3) To identify a founder haplotype(s).
4) To determine the extent of linkage disequilibrium around the Q276X mutation in LP patients.
5) To screen *ECM1* exons two through five and the first part of exon six for variants.

CHAPTER 2 SUBJECTS AND METHODS

2.1 SUBJECTS

All samples from Namaqualand are from individuals referred to as Coloureds, who are all Afrikaans-speakers. The Coloured individuals are descendants of the admixture between the early Caucasoid immigrants of South African and the indigenous Khoisan peoples of Namaqualand.[39] The term Coloured is loosely defined and the Coloured population in Namaqualand is genetically distinct from the Coloured populations found in Gauteng and other parts of South Africa.

Blood samples were collected by registered doctors and nurses, after informed consent. Samples were collected from 29 Coloured LP patients, from 28 of their immediate family members and 100 randomly selected control individuals from the Coloured population in Namaqualand. A group of seven White LP patients and five of their immediate family members were also investigated, and blood was drawn from these individuals after informed consent. The White LP families reside in Gauteng, Kwa-Zulu Natal and Mpumalanga. The majority are Afrikaans-speakers.

In addition to this, 50 random White DNA samples and 20 random Black DNA samples were obtained from the Human Molecular Genetics Laboratory random sample collection.

Ethics approval for this study was obtained from the Human Research Ethics Committee of the University of the Witwatersrand, Johannesburg, South Africa (Protocols M01-01-19 and M02-04-31, Appendix 5-10).

2.2 METHODS

2.2.1 DNA Extraction

Where DNA was not readily available, whole blood was collected in ACD tubes and genomic DNA was extracted using a protocol adapted from Miller *et al.*[40] (Appendix 5.1).

2.2.2 Polymerase chain reaction (PCR), gel electrophoresis and *Bfa*I restriction endonuclease digestion for Q276X screening

2.2.2.1 PCR

To screen for the mutation Q276X, PCR was employed, as it is a simple and reliable technique that is capable of exponentially amplifying small sections of genomic and other DNA. PCR is a simple 5-step technique that is accomplished by denaturing the template DNA, annealing the primers to the target sequences and then the elongation of the primers using a thermal stable DNA polymerase from the bacterium *Thermus aquaticus*. It is the repetition of these three steps that results in the exponential amplification of the target sequence.

A typical PCR reaction consists of the following 5 steps:

Step 1: Denaturing step at 94^{o}C for ~5 minutes.

This step denatures all double stranded (ds)DNA in the reaction to single stranded (ss)DNA, thereby allowing the primers access to the target sequence.

The next three steps (steps 2-4) are usually repeated 25 to 30 times.

Step 2: Denaturing step at 94^{o}C for ~30 seconds.

To denature dsDNA to ssDNA.

Step 3: Annealing step at ~55^{o}C (usually between 45^{o}C and 65^{o}C) for ~30 seconds.

The annealing temperature is primer specific, and each primer set will have a unique annealing temperature.

Step 4: Elongation step at 72^{o}C for ~30 seconds.

This step is carried out at the *Thermus aquaticus* DNA polymerases (*Taq*) optimum temperature. This enzyme polymerises complimentary dNTPs to the 3’

hydroxyl group of the primer, and thereby extends the primer in a 5' to 3' direction.

Step 5: Elongation step at 72^{o}C for ~10 minutes.

This is a final primer extension step.

The primers are central to the PCR reaction. They must be designed to flank the region of interest. Therefore two primers are needed for a complete PCR reaction: a forward and a reverse primer. In each cycle (steps 2-4), the two primers anneal to their complimentary ssDNA sequences, after which the *Taq* polymerase covalently bonds complimentary dNTPs (to the target strand) to the 3' hydroxyl group of the primers.[41-43] A generic PCR protocol is shown in Table 2-1.

Table 2-1: Generic PCR protocol

Thermal Cycling Conditions	**Description**	**Stock Concentration**	**Volume**
94^{o}C → 5 minutes 94^{o}C → 30 seconds X^{o}C → 30 seconds 72^{o}C → 30 seconds (30x) 72^{o}C → 10 minutes 24^{o}C → hold	Forward primer	10 pM	1µl
	Reverse primer	10 pM	1µl
	PCR Buffer	10x	2.5µl
	dNTPs	1.25 mM	2.5µl
	ddH_2O	N/A	16.8µl
	Taq DNA Polymerase	5 Units/µl	0.2µl
	DNA	≥ 100ng/µl	1µl
	TOTAL	N/A	25µl

X^{o}C is the optimal annealing temperature for the chosen primer pair. Increasing or decreasing the number of cycles can increase or decrease the yield of amplified DNA, respectively.

Exon seven of the *ECM1* gene was PCR amplified with exon seven specific primers situated in the introns flanking the exon (*ECM1* Ex 7F 5' TTA TCT GCC TGC CCA GTG TC 3', *ECM1* Ex 7R 5' ACA TGG ATG GAT GGA CTG GC 3') according to the protocol in Table 2-1 with an annealing temperature of 58.5 ^{o}C. This resulted in a PCR product of 548bp. Primers were designed using the computer program DNASTAR PrimerSelect™.

2.2.2.2 Gel electrophoresis

To determine if the PCR reaction was successful, a small volume of PCR product (usually 5µl) was electrophoresed on an agarose gel. The PCR product was mixed with Ficoll dye (this helps the sample to sink into the well, and was used as an indicator of how far the smaller DNA fragments have run in the gel), was loaded into a well, and a current was applied to the gel (the gel was placed in a tank containing 1x TBE buffer, which conducts the current). The current varies with the conductivity of the 1xTBE buffer, but the voltage was set at 120V. The electrodes were spaced 30cm apart, with the gel placed between them. As DNA is negatively charged, the samples were placed in the area of the gel closest to the negative electrode, so that the sample will pass through the gel as it moves towards the positive electrode.[43]

Ethidium bromide was added to the gel as it was prepared. When the PCR product was in the gel, the ethidium bromide would intercalate into the dsDNA of the PCR product, and when exposed to ultra-violet light (260nm), it would fluoresce. This was used to detect the presence of DNA in the gel and the presence was recorded by photographing the image of the gel illuminated by ultra-violet light.

In order to confirm that the correct PCR product had been amplified, a size standard was loaded onto each gel. The size standard contains DNA fragments of known sizes, and the size of the PCR fragment was estimated by comparing it to the size standard. The size standards used in this study were the 1kb and the 1kb plus ladders (Figure 2-1 lane 1 and 6, respectively).

To confirm that the PCR of *ECM1* exon 7 was successful, 5µl of the PCR product was electrophoresed on a 2% agarose gel for approximately 30 minutes, and sized using the 1kb or 1kb plus ladders. If a single band was present at the expected size range of 548bp (Figure 2-1, lane2), the remaining PCR product (20µl) was digested with the restriction endonuclease *BfaI*.

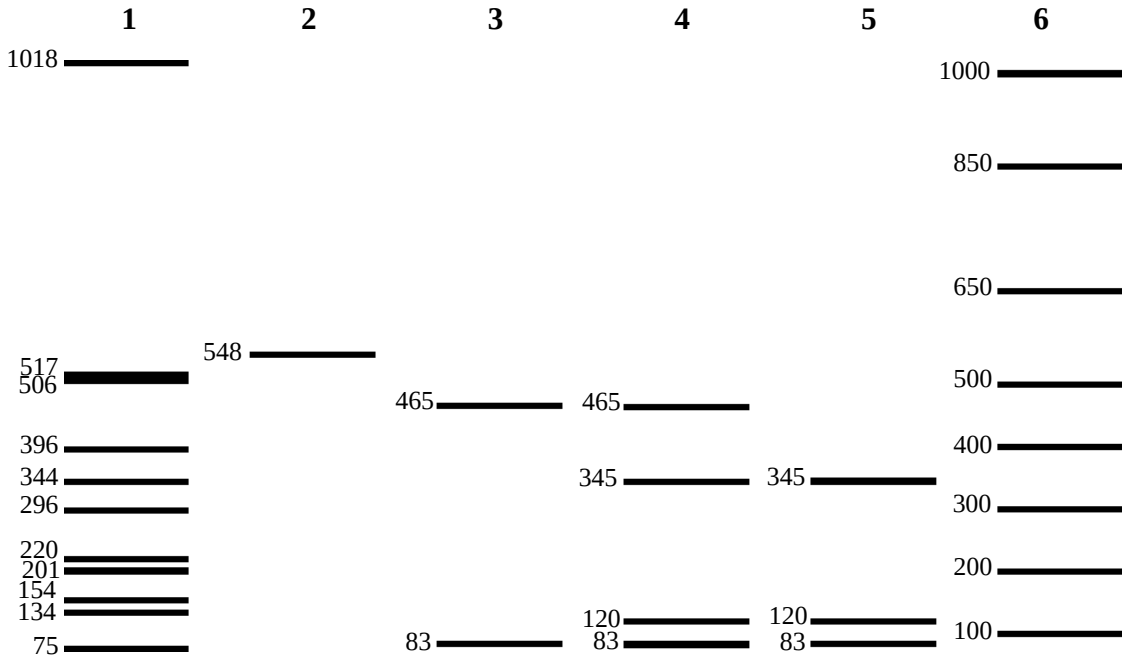


Figure 2-1: Schematic representation of the expected *ECM1* Ex 7 restriction digestion pattern run on a 2% agarose gel. All sizes are in base pairs.

Lane 1: A section of the 1kb ladder. Lane 2: *ECM1* Ex7 uncut PCR product. Lane 3: Homozygous WT *BfaI* digestion pattern. Lane 4: Heterozygous WT/Q276X *Bfa*I digestion pattern. Lane 5: Homozygous Q276X *Bfa*I digestion pattern. Lane 6: A section of 1kb plus ladder. All numbers are in base pairs. WT: wild type (non-Q276X) allele.

2.2.2.3 *Bfa*I restriction endonuclease digestion

Type II restriction endonucleases are enzymes produced by micro-organisms that recognize specific DNA sequences, bind to these sequences and cut the DNA strands at that site. Restriction endonucleases are useful tools in molecular biology, as they will only cut at their specific recognition site. If the recognition site is altered in any way, such that just one base of the recognition site is changed, the restriction endonuclease will not recognize or cut that altered recognition site.

The restriction endonuclease used to detect the Q276X mutation was *Bfa*I, which recognizes and cuts the sequence 5'CTAG3'. The mutation Q276X is a C → T transition, changing a 5'CCAG3' sequence to a 5'CTAG3' sequence, thereby creating a new *Bfa*I recognition sequence. In PCR products amplified from normal individuals without the

mutation Q276X, there is one *Bfa*I recognition site and when digested with *Bfa*I, results in two fragments sized 83bp and 465bp (Figure 2-2A). In PCR products amplified from LP individuals (who are homozygous for the mutation Q276X), the Q276X mutation introduces a second *Bfa*I restriction site (Figure 2-2B) and therefore digestion with *Bfa*I will result in three fragments sized 83bp, 365bp and 120bp. The initial *Bfa*I restriction site is present in PCR products amplified from all individuals, and therefore acts as a positive internal control for digestion with *Bfa*I.

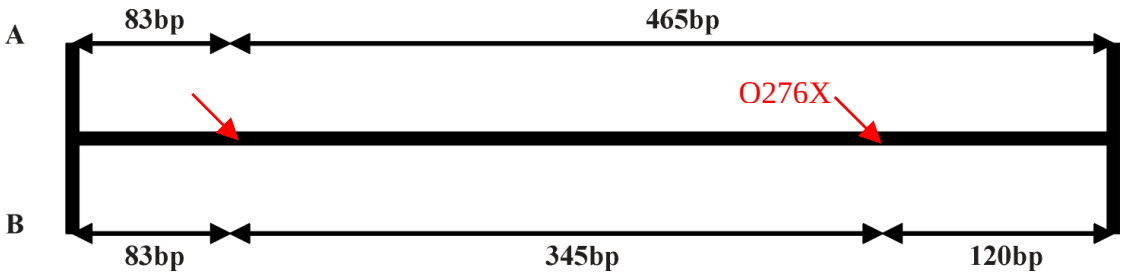


Figure 2-2: *Bfa*I restriction map of the 548bp PCR fragment

*Bfa*I recognition sites are shown with red arrows. The mutation Q276X is shown. Part A indicates how *Bfa*I will cut the 548bp fragment when the mutation Q276X is absent, while part B indicates how *Bfa*I will cut the fragment if the mutation Q276X is present.

If PCR of a sample was successful, the sample would then be dialyzed on a 0.025µm filter paper suspended in a petridish of ddH_2O for two hours to remove any salts that could inhibit the digestion of the PCR product. Once purified, the PCR product would then be digested with *Bfa*I according to the protocol in Table 2-2.

Table 2-2: Conditions of ECM1 Ex 7 Restriction Digestion with *Bfa*I

Thermal Conditions	Description	Stock Concentration	Volume
37°C for four days	NEBuffer 4	10x	3µl
	ddH_20	N/A	3µl
	BSA	10%	3µl
	*Bfa*I	5 Units/µl	1µl
	Purified *ECM1* Ex 7 PCR product	N/A	20µl
	TOTAL	N/A	30µl

No STAR activity (where a restriction enzyme cuts a DNA strand at a site that does not correspond to its recognition sequence[44,45]) has been noted for the *Bfa*I enzyme, but it was noted that *Bfa*I commonly only partially digests PCR products (P. Walsh, personal communication;http://www.neb.com/neb/tech/tech_resource/restriction/properties/primer_extension.html).[46] It was therefore necessary to incubate the samples for four days instead of the usual three hours or overnight. The resultant products were electrophoresed on a 2% agarose gel submerged in 1xTBE buffer at 120V with the electrodes spaced 30cm apart. The gel was electrophoresed for at least 30 minutes and then viewed by trans-illumination of ultraviolet light (260nm) on the Gel Documentation system. The image was then captured electronically.

If the sample digested with *Bfa*I was a non-Q276X homozygote, the expected fragment pattern after gel electrophoresis is shown in Figure 2-1 lane 3, while if the sample digested was a Q276X homozygote, the expected fragment pattern is shown in Figure 2-1 lane 5. If the sample digested was a heterozygote for the mutation Q276X, the expected fragment pattern is shown in Figure 2-1 lane 4.

2.2.3 Dinucleotide microsatellite genotyping

To determine the extent of linkage disequilibrium (LD) around the *ECM1* gene in the South African LP population, six dinucleotide microsatellite markers (D1S2863, D1S2696, D1S2636, D1S2343, D1S305 and D1S2624) around the *ECM1* gene were initially typed in all affected individuals, their immediate family-members when available and in 80 random ethnically and geographically matched controls (50 Coloured and 30 White individuals). The markers were chosen to span an estimated area of 10.5Mb of chromosome one, with *ECM1* near the centre of this area. However, three of these markers had to be discarded at a later stage (D1S2863, D1S2696 and D1S2636) as they were found to be too far away (> 50 megabases (Mb), or across the centromere) from *ECM1* and hence would not be informative. This was due to the UCSC genome bioinformatics browser (http://genome.ucsc.edu) reassembling the available sequence of chromosome one, as fresh sequence was made available, with the newer version or

"freeze" moving the markers further away than was originally anticipated. To facilitate effective coverage of the genomic region surrounding *ECM1*, three of the initial markers: D1S2343, D1S305 and D1S2624 that were telomeric to *ECM1* were used while a single new dinucleotide microsatellite marker was identified centromeric to *ECM1*. This new marker was named Novel Dinucleotide 1 (ND1). Figure 2-3 illustrates the position of the dinucleotide markers relative to *ECM1*.

ND1 was identified by downloading approximately 5Mb of sequence centromeric to *ECM1* from the UCSC website and inputting the sequence into the computer program "Tandem Repeat Finder" (TRF).[47] TRF can identify all repeating sequences within a specific input sequence, and the user can then pick any of these to suit their needs. For this study, a dinucleotide microsatellite 3.82Mb from *ECM1* was chosen. This was done as preliminary results from the markers telomeric to *ECM1* showed strong LD at a distance of 3.80Mb from the Q276X mutation. This marker became known as ND1 and primers were then designed to flank the repeat sequence of ND1 using the software program DNASTAR PrimerSelect™.

All of the markers were confirmed by BLAT searches against the UCSC genome bioinformatics browser and information pertaining to them, with the exception of marker ND1, can be found at the Genome Database (GDB) (http://gdb.org), the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/) and the UCSC genome bioinformatics browser (http://genome.ucsc.edu).

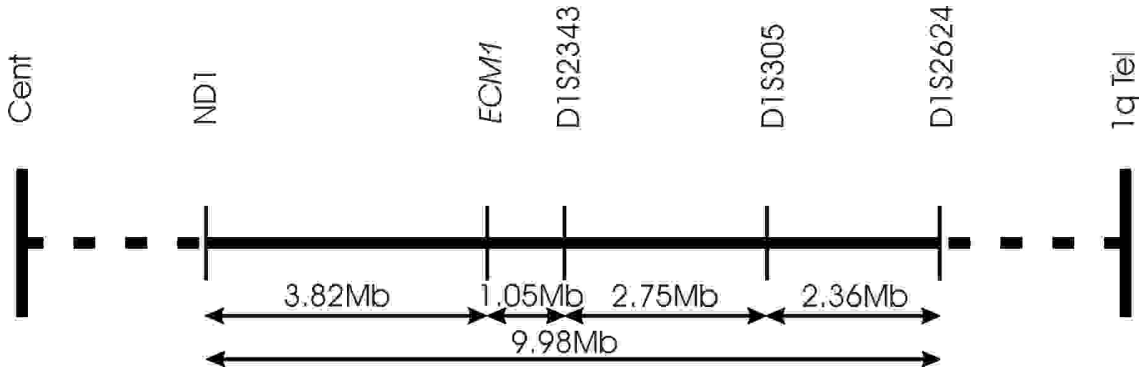


Figure 2-3: Schematic representation of the microsatellite markers surrounding *ECM1*.

Distances estimated from the UCSC website. Mb denotes megabases. Broken lines indicate areas not drawn to scale.

Table 2-3: Primer sequences of the four microsatellite markers used in this study

Marker Name	Primer Name	Sequence
ND1	ND1 F	5' HEX TGT GGG GAG AAA GTG GAT AGT T 3'
	ND1 R	5' CTC ATT TGC TTG GGG TGT CT 3'
D1S2343	D1S2343 F	5' HEX GTT TCT TGG GTG GAT CAC TTA AGC CT 3'
	D1S2343 R	5' CTA GCA TAT TCG TCC TGA ACT AA 3'
D1S305	D1S305 F	5' FAM CCA GNC TCG GTA TGT TTT TAC TA 3'
	D1S305 R	5' CTG AAA CCT CTG TCC AAG CC 3'
D1S2624	D1S2624 F	5' HEX CTG CGT CTC TTC CCT CCA TAC ACA 3'
	D1S2624 R	5' AGC GTC CTG CAC AGA GTC CAA CC 3'

F indicates forward primers, while R indicates reverse primers.

Each marker was PCR amplified with fluorescently labelled primers (Table 2-3), for each individual under the conditions set out in Tables 2-1 and 2-4.

Table 2-4: Microsatellite marker PCR conditions

Marker Name	Annealing Temperature	Expected Product Size
ND1	58.5 ^{o}C	211bp – 229bp (15 repeats to 24 repeats)
D1S2343	55^{o}C	246bp – 282bp (10 repeats to 28 repeats)
D1S305	56^{o}C	167bp – 195bp (18 repeats to 32 repeats)
D1S2624	66^{o}C	327bp – 351bp (12 repeats to 24 repeats)

One microlitre of each resultant product was mixed with 2.2µl dextran-formamide dye and 0.3µl of GeneScan™ 500 Rox™ Size Standard. The resultant mix was then denatured at 98 ^{o}C for two minutes, and then immediately put on ice. Either 1.5µl or 1.7µl of

denatured sample mix was loaded onto a 36 or 24 lane (respectively) denaturing 4.3% polyacrylamide 36cm gel and electrophoresed using an ABI 377 Sequencer. The raw fluorescent data was captured by the ABI™ 377 collection software v.2.1 with filter set A and then analysed with the Genescan v.3.1.2 and Genotyper v.2.5 software programs. To avoid errors introduced by intergel variability, a homozygote and heterozygote of known size was added to each gel as a size control.

Ideally, for each marker an individual who is homozygous at that marker should be sequenced, thereby allowing the number of repeats to be directly counted and associated with a specific allele size estimated by the Genescan v.3.1.2. This can then be used as a benchmark whereby all other allele sizes for that marker can directly be correlated to a repeat number. However, due to time constraints, this was not done, and instead a number of repeats was assigned to a particular allele size by estimating this from the published sequence at the UCSC website (http://genome.ucsc.edu).

Family pedigrees were constructed for the LP families and phase of the alleles for each marker was determined using the pedigrees, thus producing haplotypes. Phase and hence haplotypes could not be deduced from the random Coloured or White individuals as there was no pedigree information available. Haplotypes were then estimated using the Arlequin v1.1 software program, which is available at http://lgb.unige.ch/arlequin/.

2.2.4 Variation screening in the *ECM1* gene

2.2.4.1 PCR

To screen the coding region of *ECM1* for polymorphisms, each of *ECM1*'s exons were PCR amplified from 60 random individuals. The 5' and 3' untranslated regions were excluded, as they do not code for amino acids. Exon 5a was not included in this study as it was discovered to be a coding exon in humans only in the first half of 2003.[30,31]

The primers were designed using the software program DNASTAR PrimerSelect™, and since the amplimers were intended for denaturing high performance liquid

chromatography (dHPLC) analysis, the primers were designed such that they produced amplimers no greater than 400bp. The reason for this is discussed on pages 29 and 30.

Each *ECM1* exon was PCR amplified in 60 random individuals, 20 individuals from three distinct South African population groups, namely the Black population, the Coloured population, and the White population. The PCR conditions are described in Tables 2-1 and 2-6 and use the primers described in Table 2-5.

Table 2-5: Primer sequences used to identify variants within the *ECM1* gene

***ECM1* Exon**	**Primer Name**	**Sequence**
***ECM1* Ex 1**	***ECM1* Ex 1 F**	5' GAGGGCGGGAGATCACACCAG 3'
	***ECM1* Ex 1 R**	5' CAAGCCCAGACCCCCTCCTAAGT 3'
***ECM1* Ex 2**	***ECM1* Ex 2 F**	5' ACGGGGCAGGTGAATACAGAG 3'
	***ECM1* Ex 2 R**	5' TAGAAGCAAGAGGGGAGACATCAG 3'
***ECM1* Ex 3**	***ECM1* Ex 3 F**	5' GGCACTGTGGGCTCTGATGTCTCC 3'
	***ECM1* Ex 3 R**	5' GGCCTCTTCCCTCCTTTCCACCTG 3'
***ECM1* Ex 4**	***ECM1* Ex 4 F**	5' TCAGGTGGAAAGGAGGGAAGA 3'
	***ECM1* Ex 4 R**	5' TGAGGGCACCAAGGAGCAG 3'
***ECM1* Ex 5**	***ECM1* Ex 5 F**	5' CTCCCTCACCTCTATCCCACTATG 3'
	***ECM1* Ex 5 R**	5' GAGCCCACCGTCTTGTCTGC 3'
***ECM1* Ex 6pI**	***ECM1* Ex 6pI F**	5' TCGCTTCTCTTTTCCTTTCAGTTA 3'
	***ECM1* Ex 6pI R**	5' AAGGCAGATTTGGTTCAGATTGT 3'
***ECM1* Ex 6pII**	***ECM1* Ex 6pII F**	5' GACCAGAGCCATCCAGAACC 3'
	***ECM1* Ex 6pII R**	5' CCCCCGGCATCAAGAAC 3'
***ECM1* Ex 7pI**	***ECM1* Ex 7pI F**	5' GCCAGGGGAGCAGAGGACAACC 3'
	***ECM1* Ex 7pI R**	5' GCGTGGCACAGAGCGGAAGC 3'
***ECM1* Ex 7pII**	***ECM1* Ex 7pII F**	5' CTCCCCAGCCACACTACCAG 3'
	***ECM1* Ex 7pII R**	5' GGACAGCCACAAGCAGATGA 3'
***ECM1* Ex 8pI**	***ECM1* Ex 8pI F**	5' TCTAGTTGCCAGGGACGATAAGG 3'
	***ECM1* Ex 8pI R**	5' CCGACGGGCAAAGCACTCAT 3'
***ECM1* Ex 8pII**	***ECM1* Ex 8pII F**	5' CTGTGACCGGGAGTATGCTGTG 3'
	***ECM1* Ex 8pII R**	5' TTCCCAAAGGTGTCCCAAAAG 3'
***ECM1* Ex 9**	***ECM1* Ex 9 F**	5' ATATCCCAACCCCATCTGA 3'
	***ECM1* Ex 9 R**	5' AAAAACACCTCCCCCACTA 3'
***ECM1* Ex 10**	***ECM1* Ex 10 F**	5' CTCCCCACCCCATCATCTGTTTTA 3'
	***ECM1* Ex 10 R**	5' GGCTCTGGGGTGACTCATTCTTCC 3'

Table 2-6: dHPLC PCR conditions

Marker Name	Annealing Temperature	Expected Product Size
***ECM1* Ex 1**	66 ^{o}C	268bp
***ECM1* Ex 2**	62^{o}C	269bp
***ECM1* Ex 3**	69^{o}C	233bp
***ECM1* Ex 4**	62^{o}C	257bp
***ECM1* Ex 5**	62.5 ^{o}C	251bp
***ECM1* Ex 6pI**	59.6 ^{o}C	341bp
***ECM1* Ex 6pII**	60.6 ^{o}C	330bp
***ECM1* Ex 7pI**	69.3 ^{o}C	381bp
***ECM1* Ex 7pII**	61.1 ^{o}C	400bp
***ECM1* Ex 8pI**	64.4 ^{o}C	204bp
***ECM1* Ex 8pII**	62.8 ^{o}C	266bp
***ECM1* Ex 9**	55.7 ^{o}C	292bp
***ECM1* Ex 10**	66.5 ^{o}C	286bp

The resultant PCR product was then electrophoresed on a 2% agarose gel at 120V, with electrodes spaced 30cm apart for approximately 30 minutes and photographed under UV (260nm) light to confirm that the correct amplimer had been PCR amplified.

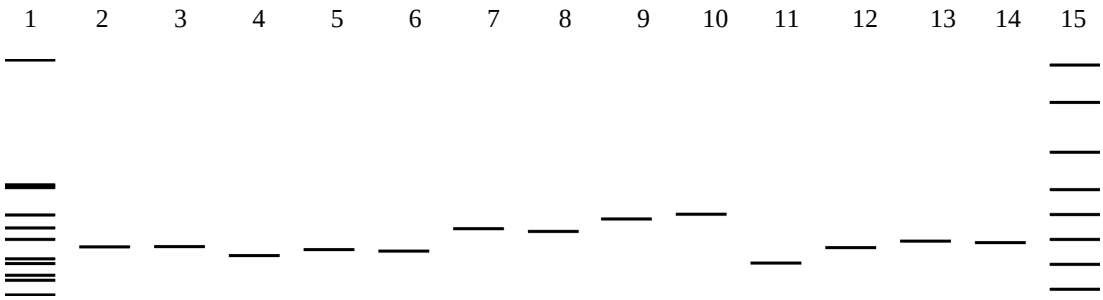


Figure 2-4: Schematic representation of the expected *ECM1* PCR products for exons one to ten run on a 2% agarose gel

Lane 1: 1kb ladder, fragment sizes stated in Figure 2-1; Lanes 2 – 14: PCR products of *ECM1* exons one to ten, with product sizes stated in Table 2-6, 1kb plus in lane 15.

2.2.4.2 Denaturing high performance liquid chromatography (dHPLC)

dHPLC is a simple and reliable technique that is based on the difference in retention time of the heteroduplexes and homoduplexes on a stationary phase or column. A homoduplex is defined as a stretch of dsDNA with both strands totally complimentary to each other,

i.e. there is no miss-pairing. A heteroduplex is defined as a stretch of dsDNA with both strands being almost totally complimentary to each other, but with one or more miss-pairing base pair. Homo- and heteroduplexes are formed by denaturing the dsDNA at 95^{o}C for five minutes to produce ssDNA. The ssDNA is then slowly cooled to 23^{o}C over 45 minutes to facilitate reannealing of the ssDNA. As all strands are either totally or almost totally complimentary to each other, they will re-anneal randomly, forming homoduplexes and heteroduplexes (Figure 2-5). Therefore, homoduplexes are formed when two identical amplimers are mixed and denatured. Heteroduplexes are formed when two similar but not identical amplimers are mixed and denatured. Similarly, homoduplexes can be formed when a single homozygous amplimer is denatured and heteroduplexes can be formed when a single heterozygous amplimer is denatured. [48]

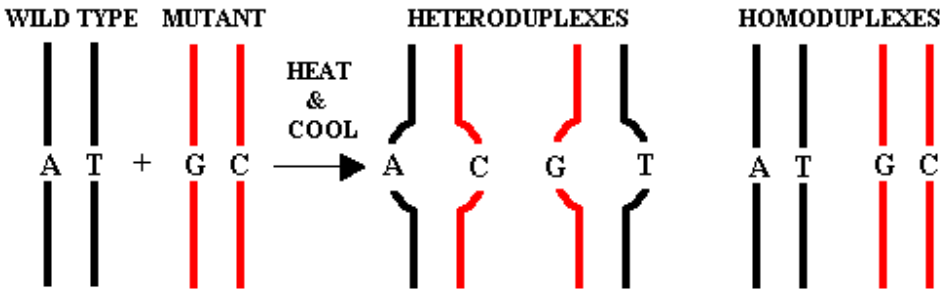


Figure 2-5: Homoduplex and heteroduplex formation

"Wild type" and "mutant" can be read as variant A and variant B

The published DNA sequence of each amplimer from the UCSC genome browser was entered into the WAVEMAKER™ v.4.1 System software, which then calculated and set up the optimal analysis conditions (temperature and acetonitrile concentration) for each amplimer (Table 2-7). The Transgenomic WAVE® DNA fragment analysis system then automatically loaded the sample into the stationary phase which electrostatically bound the amplimer. This was done in a low-stringency mobile phase (low acetonitrile concentration). Within heteroduplexes, "bubbles" would be formed where there was miss-pairing. These "bubbles" decreased the binding affinity of the heteroduplexes (relative to the homoduplexes) to the stationary phase. The Transgenomic WAVE® DNA fragment analysis system would then initiate a stringency gradient (based on acetonitrile

concentration) which is ramped to a high concentration over time (Table 2-7). The increasing levels of acetonitrile would increasingly interfere with the bonds between the stationary phase and the amplimer, eventually dislodging it from the stationary phase. The amplimer would then pass through a UV detector set at 260nm, which detected the presence and quantity of DNA. The data from the UV detector was translated into a graph of retention time vs. absorbance level.

As the heteroduplexes would bind less strongly to the stationary phase, they would be eluted off the stationary phase before the homoduplexes. This theoretically would result in four peaks being detected by the UV detector, with the two heteroduplexes forming the first two peaks and the homoduplexes forming the latter two peaks. However, this does not always occur as the difference between the retention time between the two heteroduplexes and similarly between the two homoduplexes is so slight that the two heteroduplexes elute at the same time and similarly with the homoduplexes. This gave rise to two peaks instead of the expected four. If there were two different variants within a specific exon, a sample with variant A would have a different retention time of the heteroduplexes and a different peak pattern when compared to a sample with variant B.

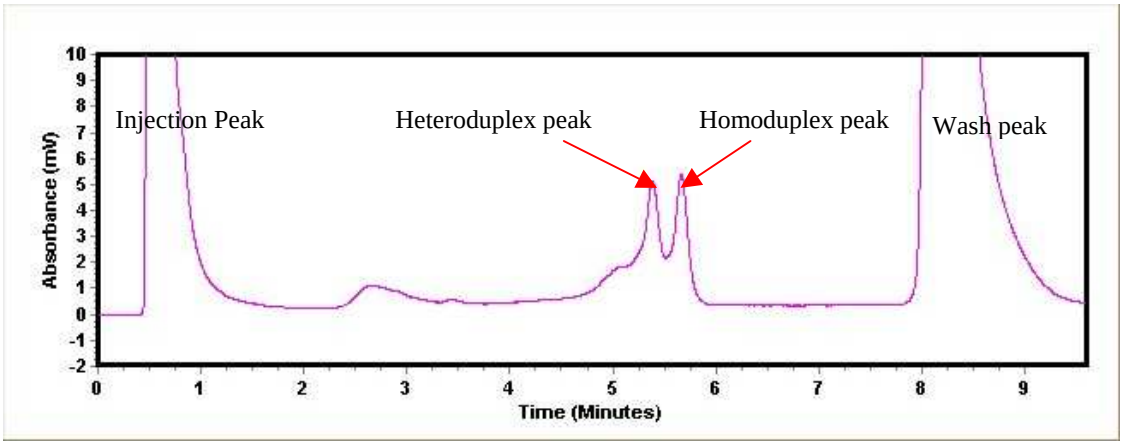


Figure 2-6: dHPLC chromatogram showing a typical double peak pattern generated by heteroduplex and homoduplex elution off the stationary phase

Each PCR primer pair was designed such that the expected amplimer size was 400bp or less. This was done because as one analyses larger amplimers, these amplimers tend to

contain clusters or “domains” that have a high GC content (guanine and cytosine residues). This forms areas within the amplimer that have a higher Tm than other areas with lower GC content. These GC rich domains increase the Tm needed to analyse the entire amplimer with dHPLC and this means that the domains with lower Tm's cannot be analysed as accurately as the domains with higher Tm's. Obviously, as the length of the amplimer increases, the larger the chance of having two or more domains with different Tm's. To avoid this, amplimers should be less than 400bp in length or alternatively the amplimer should be analysed a number of times, at each domain’s optimum Tm.[48] This meant that exons six, seven and eight each were amplified using two overlapping PCR primer sets, named *ECM1* Ex 6pI, *ECM1* Ex 6pII, *ECM1* Ex 7pI, *ECM1* Ex 7pII, *ECM1* Ex 8pI and *ECM1* Ex 8pII.

Once the PCR product for dHPLC analysis was available, either one of two approaches was taken. Either the sample was mixed with an equal amount of a control homozygous sample (determined by dHPLC) or the sample was analysed alone. The former was done to identify homozygotes that were different from the known homozygote, while the latter was done to identify samples that were heterozygous only. Initially the mixing of the samples was done to identify all variants, that is, variants in both the heterozygous and homozygous states. However, the homozygous control sample was finished before the analysis could be completed, and the standard heterozygous detection method was employed on the remainder of the samples. This resulted in exons one through eight part one being analysed with the mixing method, while exons eight part two through ten were analysed with the non-mixing method.

If the mixing method was to be used, each PCR product was electrophoresed on a 2% agarose gel to confirm that product was present. If the sample and the control PCR product were present, they were mixed in a one to one ratio and denatured. If the mixing method was not used, the samples were denatured without mixing. Once denatured and re-annealed, the samples were then loaded and analysed on the Transgenomic WAVE® DNA fragment analysis system with conditions stipulated in Table 2-7.

Table 2-7: dHPLC conditions for each ECM1 exon analysed

Exon	Oven Temperature (°C)	Start Buffer B%	End Buffer B%	Time Shift (min)
Exon 1	61.9	50	62	0
Exon 2	62.0	50	62	0
Exon 3	60.9	49	61	0
Exon 4	62.4	50	62	0
Exon 5	62.1	50	62	0
Exon 6pI	58.4	52	64	0
Exon 6pII	61.8	54	66	-1
Exon 7pI	61.9	53	65	0
Exon 7pII	62.3	56	68	-1
Exon 8pI	62.2	47	59	0
Exon 8pII	61.7	50	62	0
Exon 9	59.6	51	63	0
Exon 10	60.0	51	63	0

Once each sample had been analysed on the Transgenomic WAVE® DNA fragment analysis system, the peak patterns were analysed for any double peaks or time shifts. Within each exon, the samples that showed variation were then analysed to see if they had the same peak retention time and peak pattern. It was assumed that if a certain sample showed a similar retention time and pattern to a second sample's retention time and pattern, that these two samples contained the same variant. Therefore, one sample of each dHPLC pattern was sequenced to identify the variant.

2.2.4.3 Sanger dideoxy DNA sequencing

The Sanger dideoxy DNA sequencing method or cycle sequencing is a modified PCR technique that utilizes both fluorescently labelled dideoxynucleotides (ddNTPs) and unlabeled deoxynucleotides (dNTPs).[43,49] The four different ddNTPs are each fluorescently labelled with a different fluorescent molecule, which will fluoresce a different colour when in the excited state. The ddNTPs each lack a free 3' hydroxyl group, and hence when they are incorporated into a synthesizing strand, they terminate the synthesis of the strand because they lack the 3' hydroxyl group that is needed to form a phosphodiester bond with the next dNTP (or ddNTP) 5' phosphate group. Since both ddNTPs and dNTPs are present in the reaction mix, they compete for position in the

synthesizing strand, and are incorporated into it randomly, thereby generating strands of random length, each with a fluorescently labelled ddNTP attached to the 3' end.

In each cycle sequencing reaction, only one primer of a primer set is used and therefore two reactions are set up for each primer set, one containing only the forward primer, and the other containing only the reverse primer. This would generate a "forward" and a "reverse" sequence for the area that the primers flanked. In this way the forward and reverse sequences of each sample are comparable, and thereby confirm the obtained sequences.

The PCR product to be sequenced was first purified using Nucleospin® Extract columns according to the protocol in Appendix 5.2. The purified PCR product was then cycle sequenced twice, using the forward primer only and then the reverse primer only to generate forward and reverse sequence for each sample. The cycle sequencing was done using the Sanger dideoxy sequencing method, according to the protocol in Table 2-8.

Table 2-8: Cycle sequencing reaction conditions

Thermal Cycling Conditions	**Volume**	**Description**	**Stock Concentration**
96°C → 30 seconds 50°C → 15 seconds 60°C → 4 minutes } 25x 24°C → hold	5µl	Purified PCR product	N/A
	4µl	Big Dye v.3.1	N/A
	1µl	Primer	3.3pM
	4µl	Big Dye v3.1 Buffer	2.5x
	6µl	ddH_2O	N/A
	20µl	**Total**	N/A

Once the sample had been cycle sequenced, the resultant product was purified using SigmaSpin™ Post-Reaction Purification columns according to the protocol in Appendix 5-3. These columns selectively remove small molecules such as unincorporated dyes, nucleotides and salts from the reaction, while maximizing the recovery of single and double stranded DNA molecules larger than 20bp. The purified cycle sequenced product was then dried in a rotary evaporator in a vacuum. Once the sample has been dried, it was

resuspended in 2.5µl dextran-formamide dye and the sample was then loaded onto a denaturing 4.3% polyacrylamide gel and electrophoresed using the ABI 377 Sequencer. The raw fluorescent data was captured with the ABI™ 377 collection software v.2.1 and then analysed with the Sequencing Analysis v.3.4.1 software program. This generated an electropherogram, which represents the sequence of the sample that was cycle sequenced. The electropherogram was analysed with the sequence analysis software DNASTAR SeqMan™II. A variant was confirmed if it was seen in both forward and reverse sequence.

2.2.5 Statistical Analyses

2.2.5.1 Hardy-Weinberg equilibrium and carrier frequency calculations

Hardy-Weinburg equilibrium assumes that the variant or mutation being investigated does not incur a selective advantage or disadvantage to an individual, and assumes that there is no migration, mutation or random genetic drift occurring in the population studied.[50,51] This was assumed to be true for the Q276X mutation and the Namaqualand Coloured population of South Africa, as LP individuals generally have a normal life expectancy, and it is likely that negligible migration into or out of Namaqualand occurs. This may be because of the poor socio-economic status of the general Namaqualand population that was observed by ourselves on visiting the area.

To assess whether the ECM1 genotype frequencies deviated from Hardy-Weinberg expectation in the Coloured population, the observed genotype frequencies were compared to expected genotype frequencies (predicted by the Hardy-Weinberg law from the sample allele frequencies). To ascertain significance, a Chi-squared test was then completed using the observed and expected values, as shown in Table 2-9. The expected values were calculated using the following equations:

$p + q = 1$

$p^2 + 2pq + q^2 = 1$

where

p is the frequency of the non-Q276X allele

q is the frequency of the Q276X allele

p^2 is the frequency of homozygous non-Q276X genotypes

2pq is the frequency of the Q276X/non-Q276X heterozygous genotypes

q^2 is the frequency of homozygous Q276X genotypes

If either p or q is known, then q or p can be estimated by $p + q = 1$. Hence p^2, 2pq and q^2 can be estimated.

Table 2-9: Chi-Squared calculation

Genotype	Observed (O)	Expected (E)	$(O - E)^2/E$
Non-Q276X homozygote	x_1	p^2	$(O - E)^2/E$
Q276X heterozygote	x_2	2pq	$(O - E)^2/E$
Q276X homozygote	x_3	q^2	$(O - E)^2/E$
Total	N	N	χ^2_1
			P

N is the total number of individuals investigated, x_1 to x_3 are the observed numbers of non-Q276X homozygotes, Q276X heterozygotes and Q276X homozygotes respectively.

If the P value is greater than 0.05, then the Q276X mutation in the Coloured population did not differ significantly from the expectations of Hardy-Weinberg equilibrium and the population was considered to be in Hardy-Weinberg equilibrium. The converse would be considered true if the P value was less than 0.05.

2.2.5.2 Haplotype Estimation

Where haplotypes could not be inferred from pedigrees, haplotypes and haplotype frequencies were estimated from genotypic data using the software program Arlequin v1.1 (Appendix 6.1 and 6.2)[52]. Arlequin uses an Expectation-Maximum (EM) algorithm to generate maximum-likelihood estimates of haplotype frequencies from the genotypic, phase unknown data.

The EM algorithm has the following steps:

1) Start with random estimates of haplotype frequencies.
2) Assuming Hardy-Weinberg equilibrium, use the above estimates to calculate expected genotype frequencies. This is the E step.
3) The relative genotype frequencies are used as weights for their constituent haplotypes in a gene counting procedure. This leads to new haplotype frequency estimates. This is the M step.
4) Steps two to three are repeated until the estimated haplotype frequencies reach equilibrium.

This is a reasonable method for estimating haplotypes and their frequencies when phase cannot be determined from genotypic data. However, the EM method can erroneously estimate haplotypes and their frequencies. This can occur because there are theoretically many different haplotypes and haplotype frequencies at which the EM algorithm can reach equilibrium. To avoid this, Arlequin will try to identify the most likely haplotypes and frequencies (i.e. those that reach an equilibrium with the lowest variance). Alternatively, one could run the program many times and use the most common haplotypes that are estimated. This was not done here because of the computing time necessary to estimate the haplotypes in this study (~24 hours).

2.2.5.3 Linkage disequilibrium

Linkage disequilibrium (LD) is a term used to describe a scenario where two alleles at different loci are associated with each other more often than the population frequencies of those alleles would predict.[49] For example, if in a population (in Hardy-Weinberg equilibrium) there were two loci, A and B, each with alleles 1 and 2 with the frequencies:

A1: 0.4 B1: 0.7

A2: 0.6 B2: 0.3

The predicted frequency of haplotype A1 B2 in that population would be 0.4 x 0.3 = 0.12. If the observed frequency of A1 B2 was 0.8, that would show that there was a discrepancy between the observed and expected frequencies. If this was a

significant discrepancy and hence chance could be ruled out, then the alleles A1 and B2 could be said to be in LD. If not, then they could be said to be in linkage equilibrium, where the observed frequencies do not significantly differ from the expected frequencies.

LD within the random Coloured, random White, Coloured LP and White LP groups was investigated with the D' statistic calculated using Arlequin v1.1.[52] Additionally, LD was calculated in the Coloured LP and White LP groups using the δ statistic, as described by Devlin and Risch.[53] This was determined to be either significant or not using the Fisher's Exact test and the χ^2 test, respectively.

2.2.5.3.1 D' Analysis

D' refers to the decimal fraction of the LD (D) relative to the maximum possible LD ($D_{max)}$ for the observed frequencies present. Traditionally D' is calculated using the equation D/D_{max}. As this is tedious for large data sets, the computer program Arlequin v1.1 was employed to calculate the D' statistic. Observed haplotypic data was used where possible, however, in the random samples pedigree information was not available and estimated haplotypes were then used.

In order to use the D' statistic, a contingency table is needed to estimate haplotype frequencies, as shown in table 2-11 for a 2x2 contingency table. However, in order to do this, an initial contingency table must be constructed with the observed numbers of alleles (Table 2-10). In this study, the disease allele was the Q276X mutation and the normal allele was any allele without the Q276X mutation. Two groups were used to construct this table: the LP group (Coloured or White) and a random control group (Coloured or White, respectively). The LP group information was entered only into the disease allele column, while the control group information was entered only into the normal allele column.

Table 2-10: Layout and notation for a 2x2 contingency table for use in haplotype frequency estimation

Marker	Disease Allele	Normal Allele	Total
A1	n_{11}	n_{12}	n_{1+}
A2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

A1 is the allele suspected to be associated with the disease allele while A2 is any other allele other than A1. n_{11} is the number of observed A1-Disease allele haplotypes, n_{12} is the number of observed A1-Normal haplotypes, n_{21} is the number of observed A2-Disease haplotypes and n_{22} is the number of observed A2-Normal haplotypes. n_{+1} and n_{+2} are the sums of their respective columns while n_{1+} and n_{2+} are the sums of their respective rows. n is the sum of n_{+1}, n_{+2}, n_{1+} and n_{2+}.

Table 2-11: Layout and notation for a 2x2 contingency table for use in δ calculation

	Disease Allele	Normal Allele	Total
A1	π_{11}	π_{12}	π_{1+}
A2	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	$1(\pi)$

π_{11}, π_{21}, π_{12} and π_{22} are estimated frequencies estimated from Table 2-10.

$\pi_{11} = n_{11}/n$

$\pi_{12} = n_{12}/n$

$\pi_{21} = n_{21}/n$

$\pi_{22} = n_{22}/n$

π_{+1} and π_{+2} are the sums of their respective columns while π_{1+} and π_{2+} are the sums of their respective rows. π is the sum of π_{+1}, π_{+2}, π_{1+} and π_{2+}, and by definition, π should always add up to 1.

D' was calculated using the equation $D' = (\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/(\pi_{+1}\pi_{2+})$ for each allele of each marker and LD is scored between markers and not alleles. As this is a tedious calculation, the software program Arlequin was employed for these calculations.

2.2.5.3.2 δ Analysis

δ is calculated according to the equation $\delta = (\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/(\pi_{+1}\pi_{22})$ and was calculated in this fashion for the most common allele of each marker in both LP groups. The δ was used as it is more robust in case-control scenarios than D' is,[53] and the calculation here was employed to compare Q276X homozygotes with non-Q276X homozygotes.

2.2.5.3.3 dHPLC Analysis

Raw data from dHPLC analysis was scored as to the presence of heteroduplexes or homoduplexes. This was then summarised into a table as shown in table 2-12 and significant deviation from the null hypothesis (all three samples are from the same population) was tested for using the Chi squared test using two degrees of freedom.[54]

Significance was taken at the 5% significance level ($\alpha = 0.05$), which was then adjusted to a table wide significance level[55] (table wide significance level = α/k, where k = the number of comparisons being made). In this case k = 6 (three population groups and two data groups, namely heteroduplexes and homoduplexes) and therefore P must be less than or equal to α/k (0.05/6 = 0.0083) to be significant at the 5% significance level.

Additionally, heteroduplex dHPLC data for each sample population was compared to the other sample populations. Heteroduplex dHPLC data was summarised for the Random Black, Random Coloured and Random White sample populations as shown in Table 2-13.

Table 2-12: Example χ^2 calculation for dHPLC results for an individual exon

Observed			
	Heteroduplexes	**Homoduplexes**	**Total**
Random Black group	Observed number	Observed number	Row Total (R_1)
Random Coloured group	Observed number	Observed number	Row Total (R_2)
Random White group	Observed number	Observed number	Row Total (R_3)
Total	Column Total (C_1)	Column Total (C_2)	Grand Total (N)
Frequency	Column frequency (C_1/N)	Column frequency (C_2/N)	Total Frequency (1)
Expected			
Random Black group	(C_1/N)xC_1	(C_2/N)xC_2	R_1
Random Coloured group	(C_1/N)xC_1	(C_2/N)xC_2	R_2
Random White group	(C_1/N)xC_1	(C_2/N)xC_2	R_3
Total	Column Total (C_1)	Column Total (C_2)	Grand Total (N)
Chi squared (χ^2) Calculation			
Random Black group	(Observed number – Expected number)^2/(Expected number)	(Observed number – Expected number)^2/(Expected number)	
Random Coloured group	(Observed number – Expected number)^2/(Expected number)	(Observed number – Expected number)^2/(Expected number)	
Random White group	(Observed number – Expected number)^2/(Expected number)	(Observed number – Expected number)^2/(Expected number)	
χ^2 = The sum of all six blocks		**P = χ^2 value with two degrees of freedom**	

Table 2-13: Example table for calculating population pairwise χ^2 results for each of ECM1 exons

Exon	Random Population Sample 1	Random Population Sample 2
Exon 1	OH	OH
Exon 2	OH	OH
Exon 3	OH	OH
Exon 4	OH	OH
Exon 5	OH	OH
Exon 6	OH	OH
Exon 7	OH	OH
Exon 8	OH	OH
Exon 9	OH	OH
Exon 10	OH	OH

OH indicates observed heteroduplexes.

Population Sample 1's observed heteroduplexes for each exon were used as the observed numbers and Population Sample 2's observed heteroduplexes for each exon were used as the expected numbers in a Chi squared test with one degree of freedom. A significance level of 5% ($\alpha = 0.05$) was used. This was adjusted to table wide significance levels by using $k = 2$, i.e. a sample was only significant at the 5% level if the P value was less than or equal to 0.025 ($\alpha / k = 0.05/2$).

2.3 Data integrity

In an attempt to minimize errors, raw *Bfa*I, microsatellite, dHPLC and sequence data was analysed and corroborated by at least one colleague from the Human Molecular Genetics laboratory. Where discrepancies were found the sample was either re-analysed or changed to the new value after discussion.

CHAPTER 3 RESULTS

3.1 Q276X mutation screening

In total, 170 individuals were tested for the Q276X mutation by PCR amplifying exon seven of the *ECM1* gene and digesting the resultant amplimer with the *Bfa*I restriction endonuclease. The 170 individuals were comprised of 36 South African LP individuals (29 Coloured and seven White) and 33 members of their immediate family (28 Coloured and five White), 100 random Coloured control individuals from Namaqualand and one Black non-LP individual. The Black individual was referred to us because of a diagnosis of epilepsy and chronic hoarseness. This individual was included in this study to exclude a diagnosis of LP at the molecular level for this individual. Figure 3-1 shows a typical result following *Bfa*I digestion.

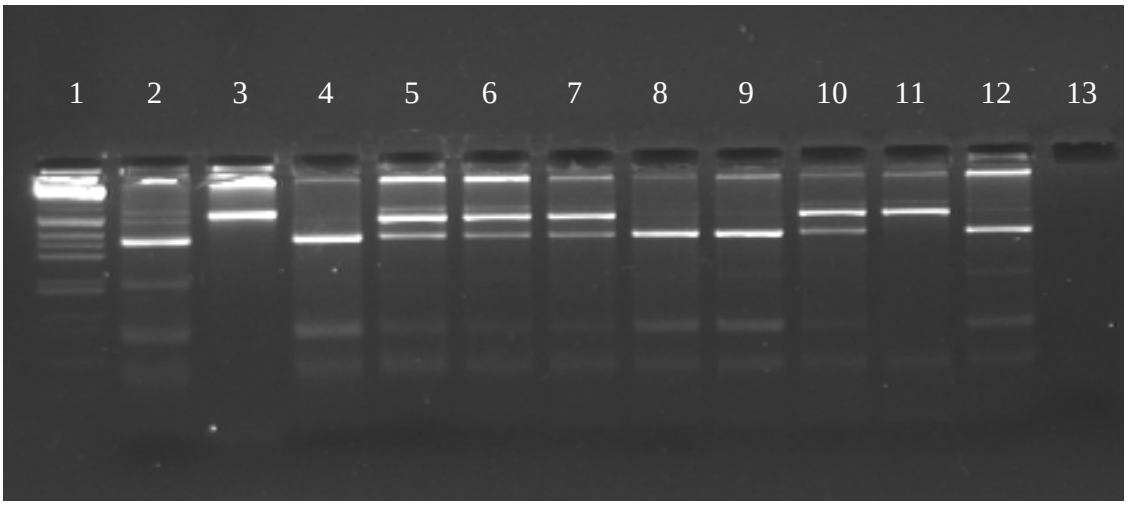


Figure 3-1: Agarose gel showing ECM1 exon seven PCR products digested with the *BfaI* restriction endonuclease.

Lane 1 contains the 1kb size standard. Three controls were used on this gel, a known Q276X homozygote (Lane 2), an undigested control (Lane 3) and a digested sample with no DNA or 'Blank" (Lane 13). Lanes 4, 8, 9 and 12 show the typical Q276X/Q276X homozygote digestion pattern. Lanes 5 – 7 and 10 show the typical Q276X/WT heterozygote digestion pattern, while lane 11 shows the WT/WT homozygote digestion pattern. Fragment sizes are shown in Figure 2-1.

All 36 LP individuals were found to be homozygous for the Q276X mutation and all parents of LP children (obligate heterozygotes) were found to be heterozygous for the Q276X mutation. This provides strong evidence for the suspected founder effect for LP

in South Africa, however further investigation is needed. The Q276X mutation may have arisen multiple times due to recurrent mutations. To provide further support that the Q276X mutation arose once within the South African population either by a single founder, a founder family or mutation, and hence that the high frequency achieved by the mutation Q276X is due to a founder effect, haplotype analysis of LP individuals is needed.

3.1.1 The Q276X mutation in the Namaqualand population

Within the 100 random Coloured controls tested for the Q276X mutation, 11 were found to be carriers of the Q276X mutation, making the carrier frequency one in nine individuals in the Namaqualand region of South Africa (f = 11/100 = 0.11, 11:100 ≈ 1:9). Assuming Hardy-Weinberg equilibrium, the Q276X allele frequency (q) is therefore 0.055 (Table 3-1) and gives an estimated incidence of LP homozygotes of one in every 370 births (q = 0.055, incidence $= (2pq)^2 \times 0.25$).

Table 3-1: Observed allele frequencies (p and q) and estimated genotype frequencies (p^2, q^2 and 2pq) for the Q276X and wild type (WT) alleles in the random Namaqualand population (N = 100)

Allele/Genotype	Variable	Value
WT allele frequency	p	89/200 (0.945)
Q276X allele frequency	q	11/200 (0.055)
WT homozygote frequency	p^2	0.945^2 (0.893)
WT/Q276X heterozygote frequency	2pq	$2 \times 0.945 \times 0.055$ (0.104)
Q276X homozygote frequency	q^2	0.055^2 (0.003)

The Q276X mutation was not significantly different from the expectations of Hardy-Weinberg equilibrium (χ^2_1= 0.339, P = 0.591) in the Coloured population of Namaqualand (Table 3.2).

Table 3-2: χ^2_1 investigation of the Q276X mutation within the Coloured population

Genotype	Observed (O)	Expected (E)	$(O-E)^2/E$
WT/WT	89 (0.890)	89.303	0.001
Q276X/WT	11 (0.110)	10.395	0.035
Q276X/Q276X	0 (0.000)	0.302	0.302
Total	100.000	100.00	$\chi^2_1 = 0.339$
			P = 0.591

3.2 Linked markers analysis

Four linked microsatellite dinucleotide markers: ND1, D1S2343, D1S305 and D1S2624, which together span 9.98Mb around the *ECM1* gene, were investigated. These markers were typed in all LP patients (29 Coloured and seven White), their first-degree family members where available (28 Coloured and five White), 50 random Coloured controls and 30 random White controls.

For marker ND1, a repeat number of 19 was correlated with a fragment size of 219bp, a fragment size of 250bp for marker D1S2343 was correlated with a repeat number of 12, a fragment size of 177bp was correlated with a repeat number of 23 for D1S305 and a fragment size of 347bp was correlated with a repeat size of 22 for marker D1S2624 (Figures 3-2 to 3-5). Sizes were calculated relative to the 500 Rox™ Size Standard that was loaded into each lane with sample.

3.2.1 Microsatellite Analysis

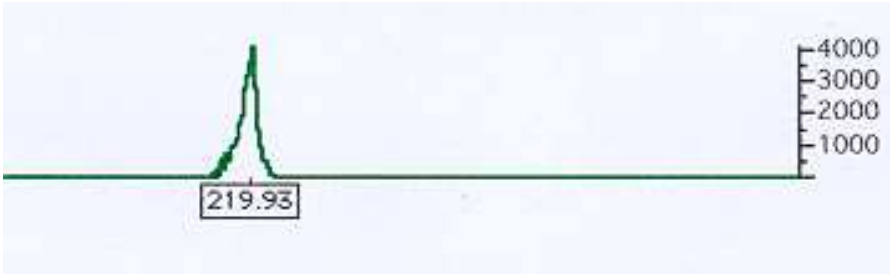


Figure 3-2: Electropherogram of the ND1 dinucleotide microsatellite marker showing an individual homozygous for the 19 repeat allele.

X axis: Fragment size in bp, estimated from the internal 500 Rox™ Size Standard. Y axis: Fluorescence intensity.

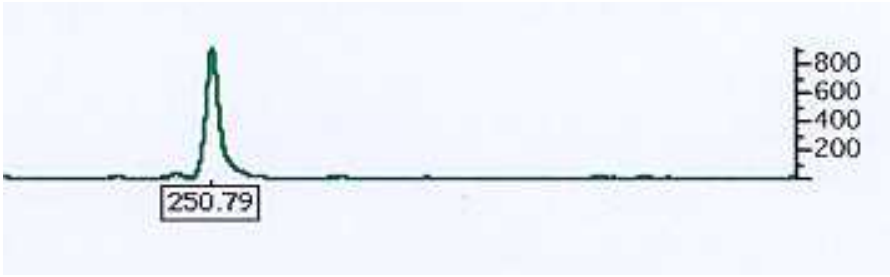


Figure 3-3: Electropherogram of the D1S2343 dinucleotide microsatellite marker showing an individual homozygous for the 12 repeat allele

X axis: Fragment size in bp, estimated from the internal 500 Rox™ Size Standard. Y axis: Fluorescence intensity.

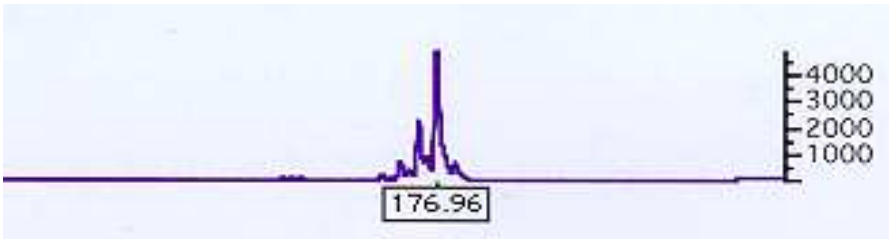


Figure 3-4: Electropherogram of the D1S305 dinucleotide microsatellite marker showing an individual homozygous for the 23 repeat allele

X axis: Fragment size in bp, estimated from the internal 500 Rox™ Size Standard. Y axis: Fluorescence intensity.

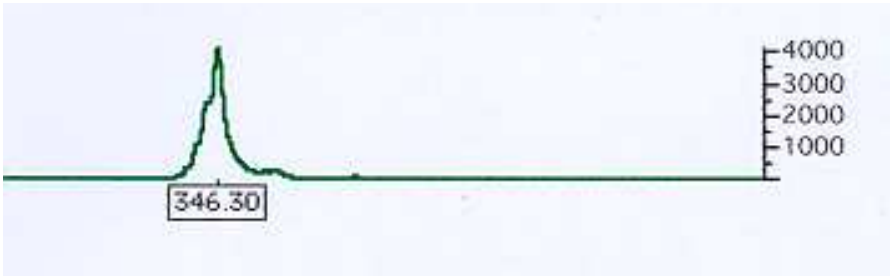


Figure 3-5: Electropherogram of the D1S2624 dinucleotide microsatellite marker showing an individual homozygous for the 22 repeat allele

X axis: Fragment size in bp, estimated from the internal 500 Rox™ Size Standard. Y axis: Fluorescence intensity.

A summary of all the microsatellite marker results is in Appendix 5.4. Each microsatellite marker was highly polymorphic in both the random Coloured and random White populations (Table 3-3), and the observed heterozygosity was very similar between the two population groups analysed.

Table 3-3: Observed heterozygosity in the random Namaqualand and random White populations

Marker	Random Coloured population observed heterozygosity	Random White population observed heterozygosity
ND1	0.74	0.67
D1S2343	0.86	0.87
D1S305	0.90	0.93
D1S2624	0.75	0.77

When the random population allele frequencies are compared with the allele frequencies observed in LP patients, a striking difference is seen (Figures 3-6 to 3-9). This is attributed to the high number of LP patients who are homozygous for the particular allele. The common LP alleles are allele 19 in the ND1 marker system, allele 12 in the D1S2343 marker system, allele 23 in the D1S305 marker system and allele 22 in the D1S2624 marker system.

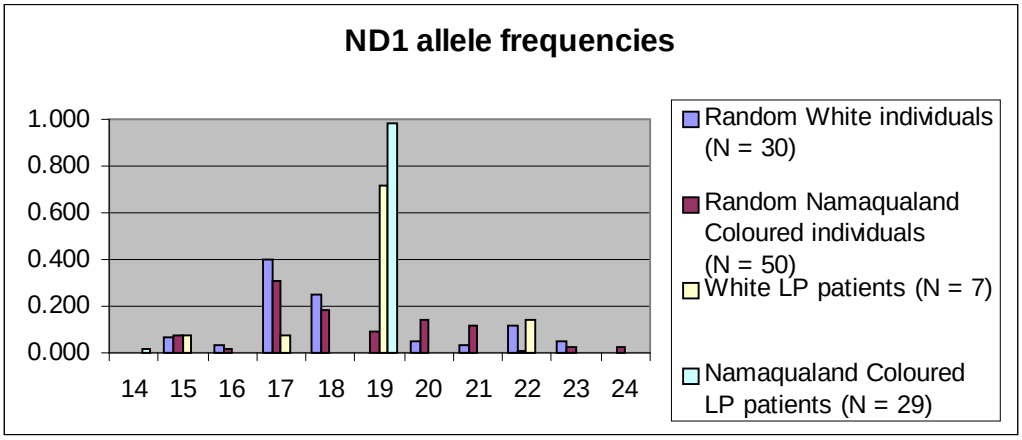


Figure 3-6: Histogram showing the allele frequencies for marker ND1 in the four groups analysed

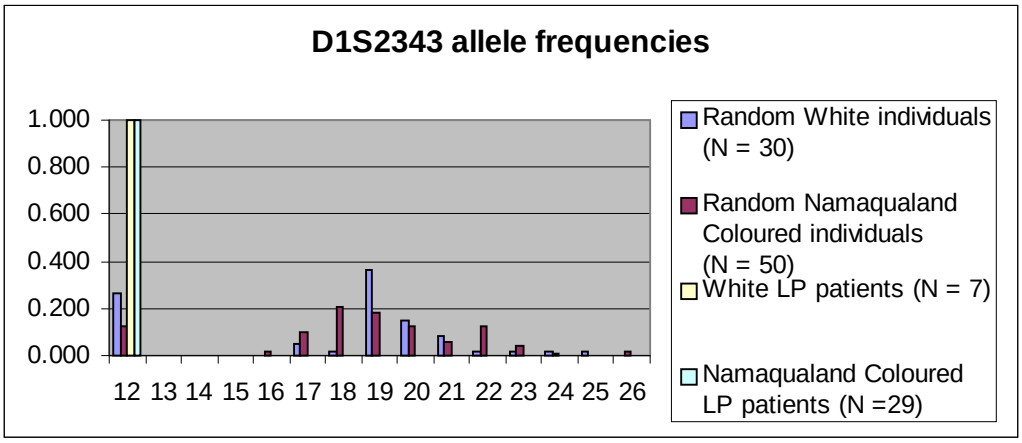


Figure 3-7: Histogram showing the allele frequencies for marker D1S2343 in the four groups analysed

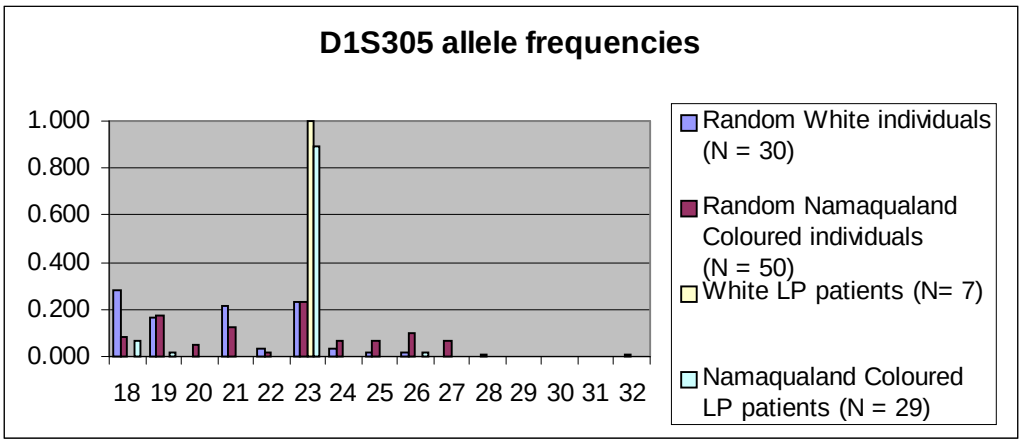


Figure 3-8: Histogram showing the allele frequencies for marker D1S305 in the four groups analysed

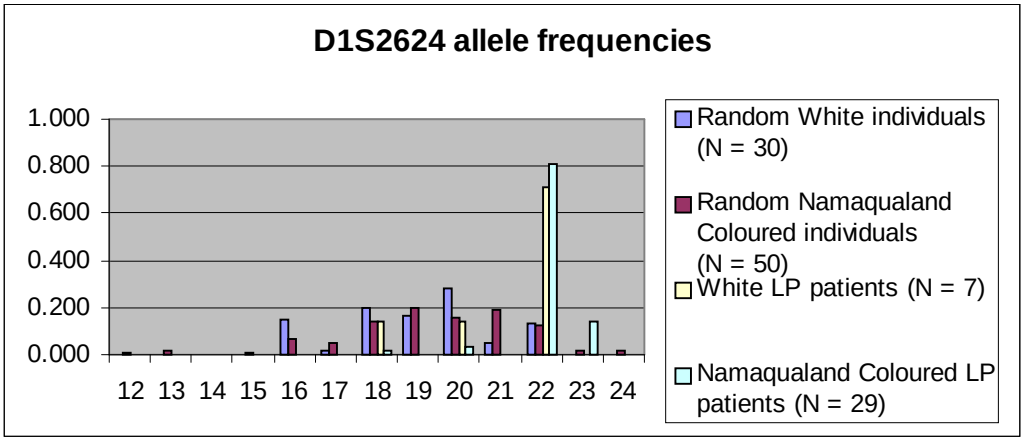


Figure 3-9: Histogram showing the allele frequencies for marker D1S2624 in the four groups analysed

3.2.2 Haplotype analysis

Pedigrees were compiled for the LP patients and their family members (Appendix 5.5). Haplotypes were deduced since phase could be determined in most of these pedigrees, (Table 3-6). As family data was not available for the random Coloured and random White groups, haplotypes were estimated using the software program Arlequin (Appendix 5.6 and Appendix 5.7). Due to restrictions of analysis with Arlequin, *ECM1* mutation status could not be included in the haplotype estimation, as the Q276X mutation is a single nucleotide variant, and cannot be included in microsatellite analysis.

When the alleles are compiled into haplotypes, a striking founder haplotype was seen in both the Coloured and White LP patients (Table 3-6), while in first-degree family members there were very few common haplotypes seen (Table 3-7). Where there were affected siblings, only one affected individual per family was included in the analysis. The most common haplotype seen, and the haplotype most likely to be the founder haplotype, is 19-12-23-22 (ND1-D1S2343-D1S305-D1S2624), or haplotype LPh1.

Table 3-4: ECM1-associated haplotypes in 24 (n=48) unrelated Coloured LP patients from Namaqualand and six (n = 12) unrelated Caucasoid LP patients. Conserved haplotypes are in shaded areas.

Haplotype Number	Number of Haplotypes	ND1 Alleles	*ECM1 Q276X*	D1S2343 Alleles	D1S305 Alleles	D1S2624 Alleles
NAMAQUALAND COLOURED LP HAPLOTYPES						
LPh1	34	19	Y	12	23	22
LPh2	6	19	Y	12	23	23
LPh3	2	19	Y	12	23	20
LPh4	2	19	Y	12	18	22
LPh5	1	19	Y	12	23	18
LPh6	1	19	Y	12	19	22
LPh7	1	19	Y	12	26	23
LPh8	1	14	Y	12	23	22
SOUTH AFRICAN WHITE LP HAPLOTYPES						
LPh1	8	19	Y	12	23	22
LPh9	2	17	Y	12	23	18
LPh10	1	15	Y	12	23	20
LPh11	1	22	Y	12	23	22

A core haplotype of 19-12 (ND1-D1S2343), spanning 4.87Mb around *ECM1*, was present in 92% of the LP haplotypes (98% of the Coloured LP haplotypes and 67% of the White LP haplotypes).

Table 3-5: ECM1-associated haplotypes from non-Q276X alleles deduced from unaffected Coloured (n = 26) and White first degree family members (n = 5). Haplotype similar to the founder haplotype is shaded.

Haplotype Number	Number of Haplotypes	ND1 Alleles	*ECM1 Q276X*	D1S2343 Alleles	D1S305 Alleles	D1S2624 Alleles
NAMAQUALAND COLOURED HAPLOTYPES						
Ph1	2	17	N	26	26	23
Ph2	2	21	N	18	27	20
Ph3	1	15	N	21	19	19
Ph4	1	16	N	17	19	19
Ph5	1	17	N	12	19	20
Ph6	1	17	N	12	21	20
Ph7	1	17	N	12	22	18
Ph8	1	17	N	12	24	18
Ph9	1	17	N	17	23	18
Ph10	1	17	N	18	23	22
Ph11	1	17	N	18	23	23
Ph12	1	17	N	19	18	19
Ph13	1	17	N	20	19	22
Ph14	1	17	N	20	23	20
Ph15	1	17	N	20	24	22
Ph16	1	17	N	21	18	21
Ph17	1	18	N	18	26	20
Ph18	1	18	N	20	23	20
Ph19	1	18	N	22	19	23
Ph20	1	18	N	22	19	20
Ph21	1	19	N	20	23	20
Ph22	1	20	N	18	20	19
Ph23	1	20	N	19	18	21
Ph24	1	23	N	19	23	18
SOUTH AFRICAN WHITE HAPLOTYPES						
Ph25	1	17	N	12	22	22
Ph26	1	17	N	19	18	18
Ph27	1	19	N	19	21	22
Ph28	1	19	N	12	23	19
Ph29	1	21	N	19	21	22

Within the haplotypes estimated by Arlequin (Appendix 5.6) for both the random Coloured and White populations, only one of the 124 (81 Coloured and 43 White)

estimated haplotypes showed similarity to the founder haplotype. This was found within the random Coloured population and was 19-12-23-23 and is only one mutational step away from the founder haplotype of 19-12-23-22.

3.2.3 Linkage Disequilibrium (LD) Analysis

3.2.3.1 D' Analysis

In order to compare LD between the four different groups (namely the Coloured LP patients, White LP patients, Random Coloured population and random White population), D' was calculated with Arlequin v1.1 using haplotypic data only. Table 3-8 summarizes the D' results. Where two or more siblings were affected with LP, data from only one was taken so as not to artificially inflate LD associated with that particular haplotype common to both siblings.

Table 3-6: Pairwise locus D' estimation of LD in the Coloured and White populations using haplotypic data

Locus	ND1	D1S2343		D1S305		D1S2624	
ND1	*	A	B	A	B	A	B
		C	D	C	D	C	D
D1S2343		*		A	B	A	B
				C	D	C	D
D1S305				*		A	B
						C	D
D1S2624						*	

A: Coloured LP patients, B: White LP patients, C: Random Coloured population, D: Random White population. Red squares indicate significant D' LD ($P < 0.05$), Green squares indicate borderline significance D' LD ($0.10 > P > 0.05$) and white squares indicate no significant D' LD ($P > 0.10$). P values were calculated using the Fisher's exact test.

As can be seen in Table 3-8, each locus in both LP groups is in significant LD with each other, while only D1S2343 and D1S2624 are in significant LD with each other in the random Coloured population. The rest of the markers in the random Coloured population and all of the markers in the random White population are not in significant LD with each other ($P > 0.05$). As would be expected, the smaller, more consanguineous Coloured population showed more LD than the larger, less consanguineous White population.

3.2.3.2 δ analysis

In order to estimate LD between the Q276X mutation and each allele of the founder haplotype (LPh1) in the Coloured and White LP patient groups, the δ statistic was used. This was then investigated for significance with the χ^2 test with one degree of freedom. If the χ^2_1 test produced a P value less than 0.05, that was then taken as a significant result. Table 3-9 summarizes the δ results obtained.

Table 3-7: Summary of the LD results between each marker and the Q276X mutation from the Coloured and White LP patients

	Coloured LP patients		White LP patients	
Locus	**δ**	**P**	**δ**	**P**
ND1 Allele 19	0.98	1.84×10^{-28}	0.71	8.41×10^{-12}
D1S2343 Allele 12	1.00	7.10×10^{-27}	1.00	8.11×10^{-7}
D1S305 Allele 23	0.87	6.10×10^{-16}	1.00	1.73×10^{-7}
D1S2624 Allele 22	0.78	2.30×10^{-18}	0.67	7.82×10^{-6}

Using the δ statistic, it was seen that there is strong LD between each allele of the founder haplotype and the Q276X mutation in both the Coloured and White LP groups.

The microsatellite marker results, haplotype results and LD results indicate that the Q276X mutation is likely to have arisen only once on the LPh1 haplotype background. This is strong molecular evidence that the high incidence of LP in South Africa is due to a founder effect and not due to recurrent mutation.

3.3 Screening of the *ECM1* gene for polymorphisms

3.3.1 dHPLC analysis

In order to identify the most common variants within *ECM1*, 60 random individuals, 20 individuals each from three distinct populations (the South African Black population, the Namaqualand Coloured population and the South African White population) were analysed with dHPLC. Each exon was optimised for PCR and the expected fragment sizes, as illustrated in Figure 2-4, were obtained. Once PCR optimisation was completed, then dHPLC conditions were optimised for each PCR product, a summary of which can be seen in Table 2-7. Once the dHPLC system was fully optimised, either a mixing or a non-mixing method was followed. The mixing method was used for exons one through to eight pI, and exons eight pII to ten were analysed with the non-mixing method.

The screening of the *ECM1* gene with dHPLC resulted in the identification of many potential variants (Appendix 5.8). The frequencies of all individuals with variants seen per *ECM1* exon were calculated, and investigated for significance using the χ^2 calculation, with the null hypothesis stating that there was no difference between the three populations. As an example, the results for exon three are shown on the following page (Table 3-10, Figure 3-10).

Table 3-8: Example χ^2 calculation for dHPLC results from *ECM1* exon three

Observed			
	Heteroduplexes	**Homoduplexes**	**Total**
Observed in the random Black group	0 (0.00)	20 (1.00)	20
Observed in the random Coloured group	1 (0.05)	19 (0.95)	20
Observed in the random White group	11 (0.55)	9 (0.45)	20
Total	12	48	60
Frequency	0.20	0.80	1
Expected			
Expected in the random Black group	4 (0.2)	16 (0.80)	20
Expected in the random Coloured group	4 (0.20)	16 (0.80)	20
Expected in the random White group	4 (0.20)	16 (0.80)	20
Total	12	48	60
Chi squared (χ^2) Calculation			
$(O-E)^2/E$ (Random Black group)	4.000	1.000	
$(O-E)^2/E$ (Random Coloured group)	2.250	0.563	
$(O-E)^2/E$ (Random White group)	12.250	3.063	
$\chi^2_2 = 23.125$		**$P = 9.516 \times 10^{-6}$**	

When exon ten was analysed in this manner, the observed and hence expected values for heteroduplexes was zero. This became a problem when calculating the χ^2 values, as the equation was divided by the number of expected heteroduplexes. As the expected number of heteroduplexes was zero, this gave an undefined result. Therefore, to avoid this, all observed values for exon ten were adjusted to 1.

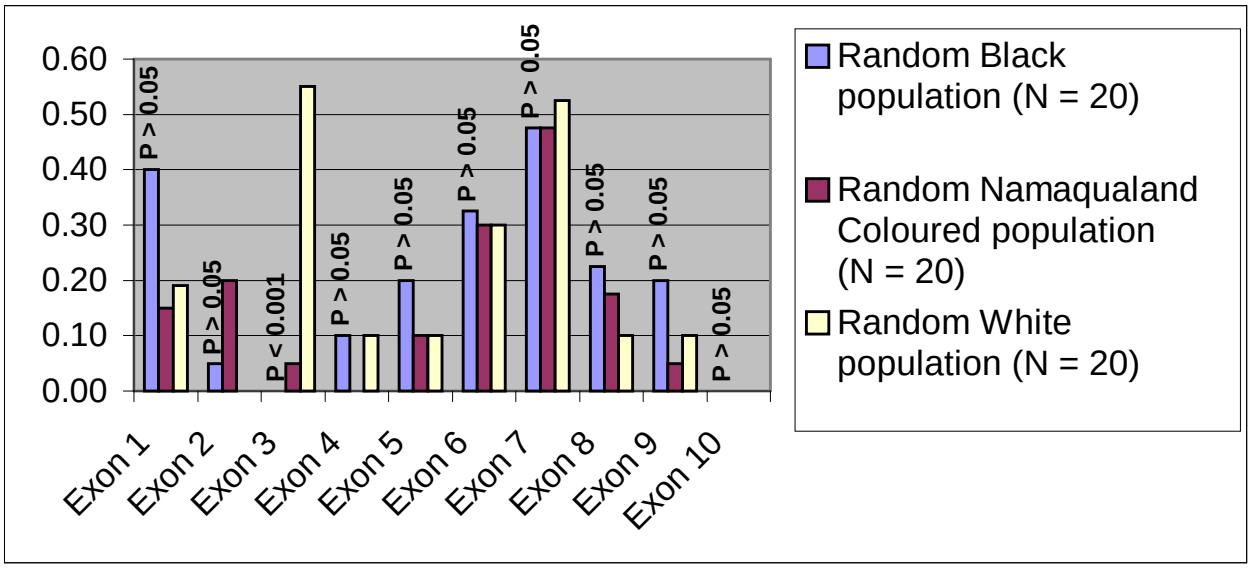


Figure 3-10: dHPLC heteroduplex frequencies identified per ECM1 exon.

χ^2 P values with two degrees of freedom are shown above each exon

Table 3-9: Summary of population pairwise χ^2_1 results for each of *ECM1* exons

	Random Black group vs. random White group	Random Black group vs. random Coloured group	Random White group vs. random Coloured group
Exon 1	-	++	-
Exon 2	-	-	-
Exon 3	+++	-	+++
Exon 4	-	-	-
Exon 5	-	-	-
Exon 6	-	-	-
Exon 7	-	-	-
Exon 8	+	-	-
Exon 9	-	++	-
Exon 10	-	-	-

- indicates no significant difference, + indicates significant difference at $P < 0.05$, ++ indicates significant difference at $P < 0.01$ and +++ indicates significant difference at $P < 0.001$. Significance levels were adjusted to table wide significance levels.[55] Where undefined results were obtained, the expected values were adjusted to from zero to one.

As can be seen in Figure 3-10, only exon three shows significant differences between all three populations, while pairwise analysis between each population (Table 3-9) of exons one, three, eight and nine showed significant difference. However, these results must be

interpreted with caution, as the sample size used for each population was very small (N = 20).

Interestingly, exons three and seven in the White population show heteroduplex frequencies greater than 0.5. The theoretical maximum heteroduplex (heterozygote) frequency obtainable should be 0.5, where q = p = 0.5 and hence 2pq = 0.5 (Figure 3-11).

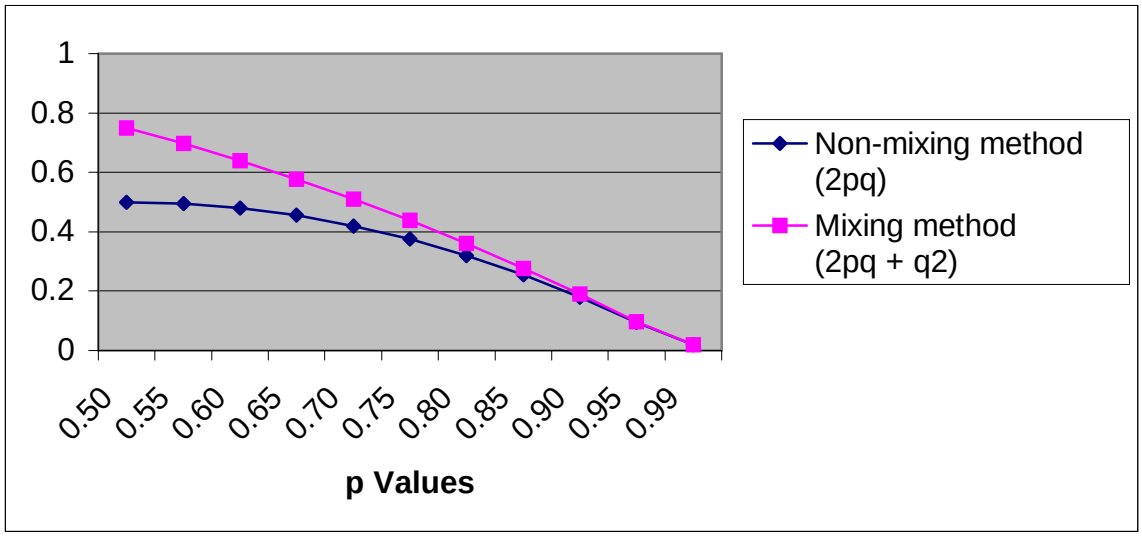


Figure 3-11: Estimated heteroduplex frequencies detected by the mixing method and non-mixing method graphed against varying frequencies of p (common allele frequency)

The mixing method was used to identify variants in exons three to eight pI, and this must be taken into account. Assuming two alleles (p and q) at one locus and Hardy-Weinberg equilibrium, the non-mixing would detect only individuals in the heterozygous form (2pq), while the mixing method would detect individuals in the heterozygous form and in the rarer homozygous form ($2pq + q^2$), assuming that q is the rarer allele and the control homozygote is homozygous for the common allele (p). Therefore the maximum detectable heteroduplex frequency using the mixing method is 0.75 (p = 0.5 and q = 0.5, Figure 3-11). This could explain the observed frequencies of greater than 0.5 in exons three and seven. Additionally, there could be more than one variant locus within the amplimer analysed, leading to an observed heteroduplex frequency of greater than 0.5.

Due to random genetic drift and different population histories, one would expect to see differences in heteroduplex frequencies (not necessarily significant) between different populations. However in exons six and seven, very similar heteroduplex frequencies were seen. This is likely to be due to chance.

Interestingly, no variation was observed in exon ten, suggesting that there are no common variants in this exon. This was unexpected and may point to some form of selection acting on exon ten. This however did not explain the absence of synonymous variants or intronic variation. Again, random genetic drift, leading to fixation of alleles is the most likely explanations for this observation.

3.3.2 Direct dideoxy sequencing

Exons two to six were investigated with sequencing, and only one representative sample was sequenced of each heteroduplex pattern. Therefore, all individuals showing variation were not confirmed by DNA sequencing, and hence the accuracy of the detection of variants by dHPLC cannot be assessed here.

A summary of all variants examined by DNA sequencing is given in Table 3-10 and the actual sequence electropherograms together with dHPLC chromatograms of the sequenced variants are shown in Figures 3-12 to 3-19.

Table 3-10: Summary of variation found within the ECM1 gene

Exon/Intron	Variation seen	Amino acid change
Intron 1	nt 1314 C→T	N/A
Intron 1	nt 1400 C→T	N/A
Intron 2	nt 1486 G→A	N/A
Intron 3	nt 1673 A→G	N/A
Exon 4	nt 1823 C→T	P77P
Exon 5	nt 2057 C→T	H122H
Intron 5	nt 2655 C→A	N/A
Exon 6	nt 2773 C→T	T130M

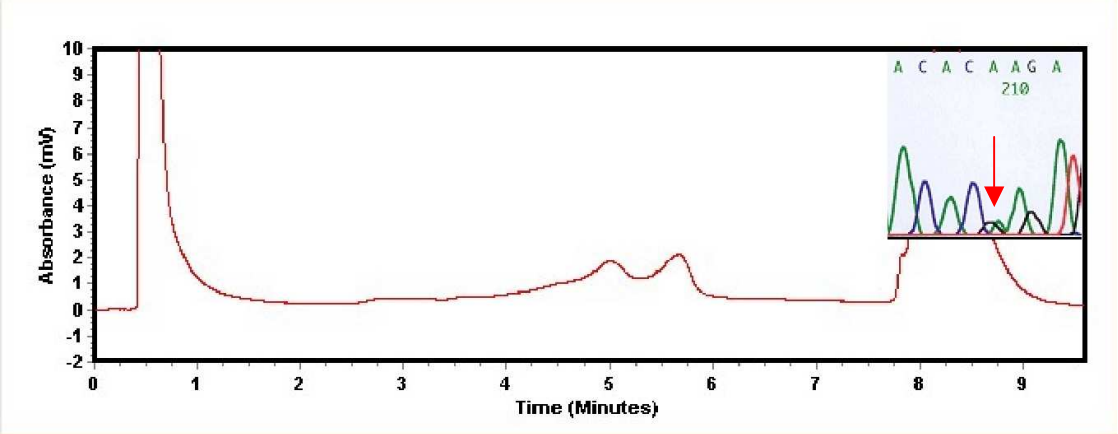


Figure 3-12: nt 1314 C→T dHPLC heteroduplex chromatogram. Inset is nt 1314 C→T heterozygote reverse sequence electropherogram. The red arrow indicates the position of the variant

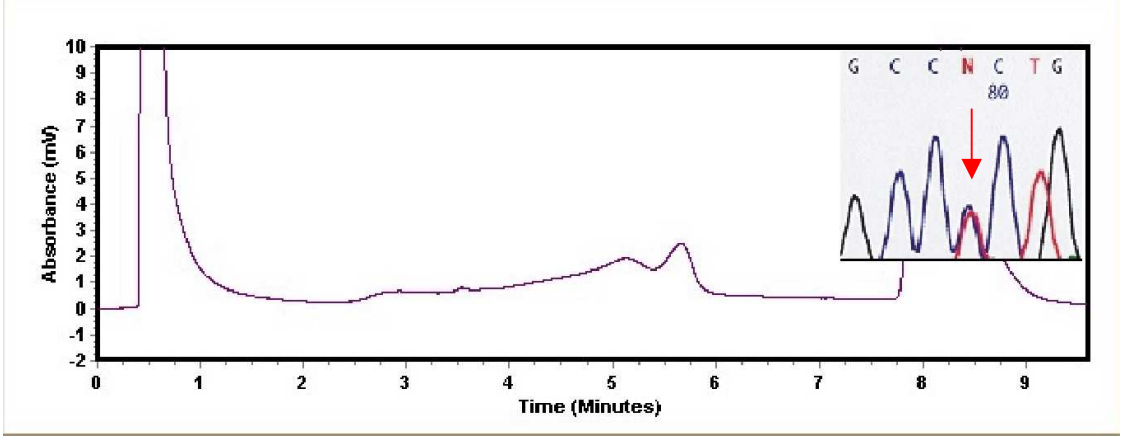


Figure 3-13: nt 1400 C→T dHPLC heteroduplex chromatogram. Inset is nt 1400 C→T heterozygote forward sequence electropherogram. The red arrow indicates the position of the variant

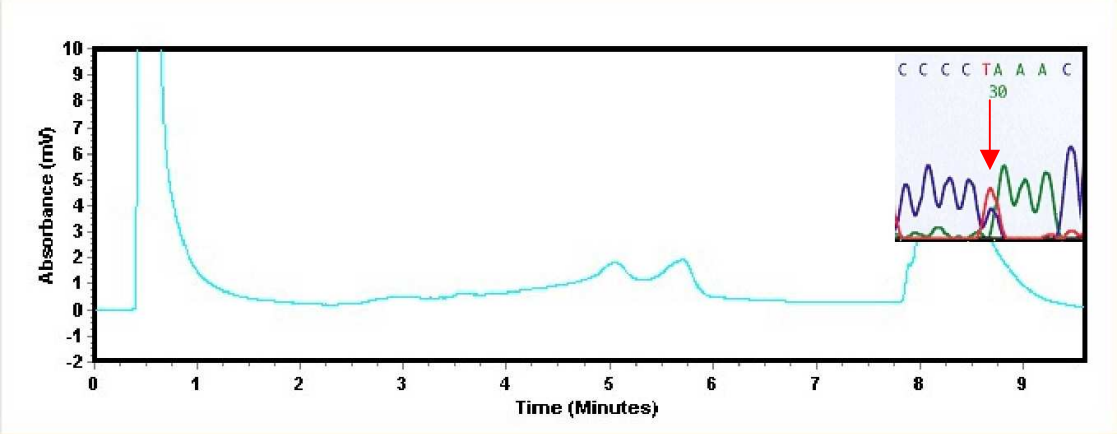


Figure 3-14: nt 1486 G→A dHPLC heteroduplex chromatogram. Inset is nt 1486 G→A heterozygote reverse sequence electropherogram. The red arrow indicates the position of the variant

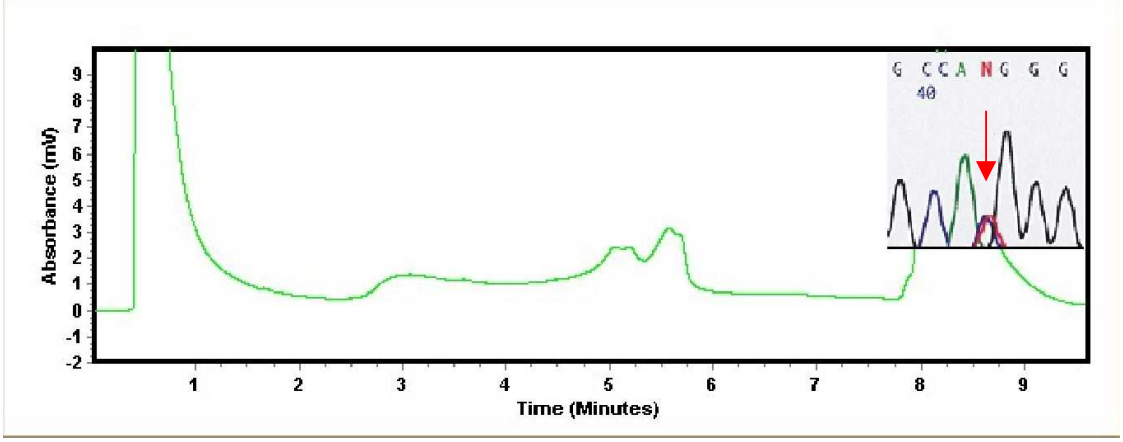


Figure 3-15: nt 1673 A→G dHPLC heteroduplex chromatogram. Inset is nt 1673 A→G heterozygote reverse sequence electropherogram. The red arrow indicates the position of the variant

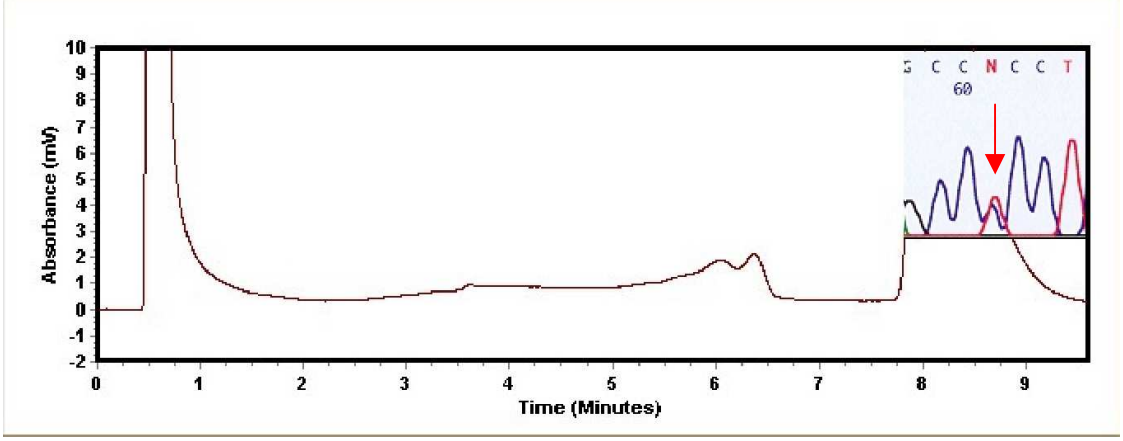


Figure 3-16: nt 1823 C→T dHPLC heteroduplex chromatogram. Inset is nt 1823 C→T heterozygote forward sequence electropherogram. The red arrow indicates the position of the variant

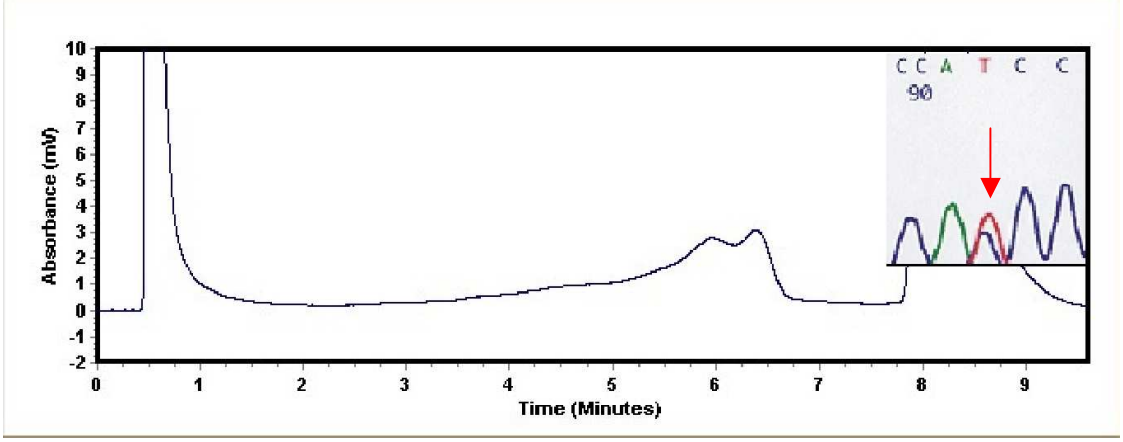


Figure 3-17: nt 2057 C→T dHPLC heteroduplex chromatogram. Inset is nt 2057 C→T heterozygote forward sequence electropherogram. The red arrow indicates the position of the variant

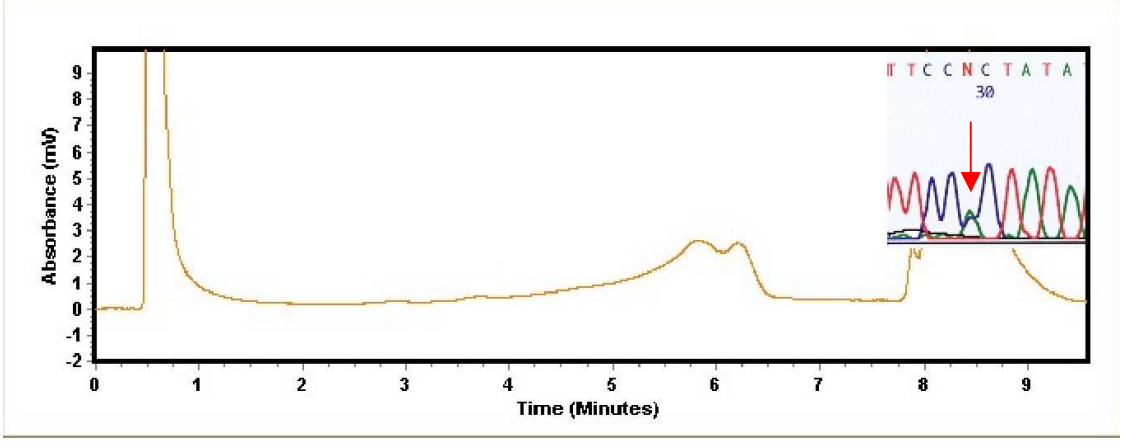


Figure 3-18: nt 2655 C→A dHPLC heteroduplex chromatogram. Inset is nt 2655 C→A heterozygote forward sequence electropherogram. The red arrow indicates the position of the variant

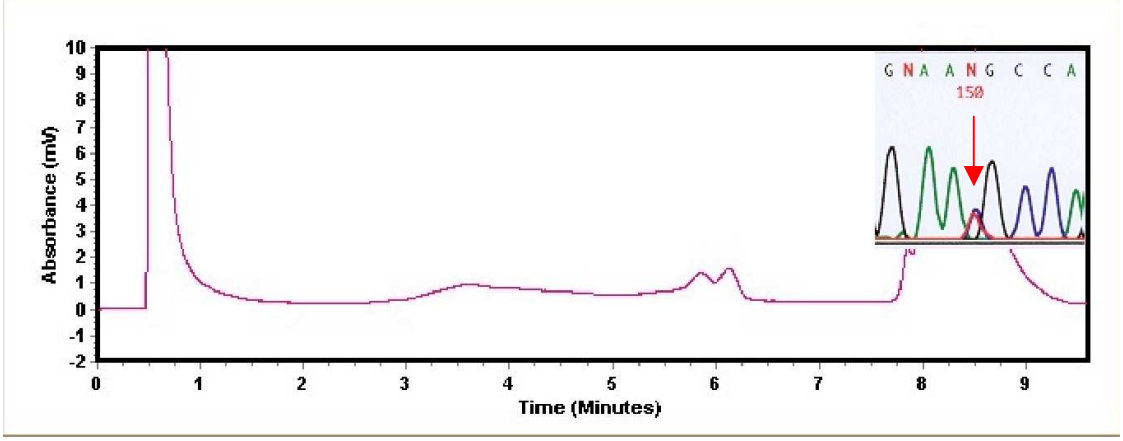


Figure 3-19: nt 2773 C→T dHPLC heteroduplex chromatogram. Inset is nt 2773 C→T heterozygote forward sequence electropherogram. The red arrow indicates the position of the variant

As can be seen in figures 3-12 to 3-19, each heteroduplex chromatogram shows a different retention time and peak pattern, albeit the differences are small. This showed that dHPLC was a useful method for detecting variation in PCR amplimers, but that it was not specific enough to qualify what the variation was. To identify the variant, a method such as endonuclease digestion or direct DNA sequencing would be needed.

CHAPTER 4 DISCUSSION

The relatively common occurrence of LP in South Africa can now be reasonably explained as the result of a founder effect within the Namaqualand Coloured population and White population of South Africa. The presence of the Q276X mutation in the homozygous state in all affected individuals and in the heterozygous state in all obligate heterozygotes, the presence of a single founder haplotype, the presence of four microsatellite markers (spanning 9.98Mb around the *ECM1* gene) in highly significant LD with the Q276X mutation in LP patients and the genealogical studies conducted by Gordon and Heyl,[10-13,20] together leaves little doubt of the LP founder effect in South Africa.

4.1 The Q276X mutation in South Africa

It is unknown when LP was first diagnosed in South Africa, but the Q276X mutation is hypothesised to have been introduced into the South African population with the first Caucasian settlers in 1652 or shortly thereafter.[10] Interestingly, the Q276X mutation has not yet been found in any other population than the South African population (I. Chan, personal communication), but it has been traced to the early German settlers of South Africa. The Q276X mutation has however not been found in Germany, or any other European country, although a cell line developed from a German LP patient has been identified (Coriell Cell Repositories, http://locus.umdnj.edu/nigms/nigms_cgi/display.cgi?GM13241). This cell line has not yet been investigated for the Q276X mutation. German LP patients have not been documented in the literature, and it may be that the Q276X mutation or any other LP mutations are rare within that population. Therefore, it would be interesting to search the German population for LP individuals and to test them for the Q276X mutation.

When exactly the Q276X mutation occurred is unclear, but from the South African founder effect it is likely that the mutational event occurred outside South Africa, in Gerrit and Elsje Cloete's paternal or maternal line and was passed on through them to

their descendants. As the number of descendants from Gerrit and Elsje Cloete increased, so the frequency of the Q276X mutation increased. This increased the likelihood of any random couple within South Africa conceiving a child affected with LP. The result of which is that today, the estimated incidence of LP in the Coloured Namaqualand population the estimated incidence is 1 in 370 live births.

Within the Coloured Namaqualand population, the estimated carrier rate of the Q276X mutation is remarkably high (1 in 9). A founder effect can be considered to be the main reason for the high prevalence of the Q276X mutation within this population. This is compounded in the Namaqualand Coloured population because it is very small (~100 000 individuals, www.statssa.gov.za). In light of these factors, genetic drift would have a powerful affect on the Coloured population of Namaqualand. Also, this population is divided into many small towns that are separated, in many cases, by large distances. This would limit the amount of migration between towns and this is likely to increase the number of intermarriages within families. This would increase the number of loci that are identical by descent, and elevate the chance of obtaining a double dose of the Q276X mutation. To further compound this issue, the general population of Namaqualand is very underprivileged. Therefore it is very hard for these people to migrate out of the Namaqualand area as their situation makes it very difficult to obtain good employment or to attend tertiary education institutions. Also, Namaqualand is not a highly commercialised or industrialized region of South Africa, with two thirds of the labour market either unemployed or not economically active (www.statssa.gov.sa, 2001 census). As a result of this, very little immigration into Namaqualand is likely to occur, and consequently there will be very little gene flow entering the Namaqualand population, allowing the Q276X mutation frequency to remain high in this population.

The Q276X mutation results from a C to T transition at position 3468 within exon seven of *ECM1*. This changes a CAG codon (Glutamine) to a TAG codon (Stop) and hence results in the premature termination of the ECM1a and c protein isoforms, but not the b isoform. Treatment aimed at inducing the removal of or splicing out of exon seven is not likely to succeed as this happens ubiquitously in the skin of LP individuals due to

alternative splicing (ECM1b). As ECM1b does not rescue the LP phenotype, ECM1a is most likely the more biologically active and important isoform of *ECM1*. LP must therefore be caused by perturbation of expression of ECM1a. Symptomatically, LP presents with excessive scarring of the skin and hence ECM1a must play an important role in minimising scar formation or in proper scar healing. Therefore ECM1a presents itself here for potential use as an anti-scarring medication and could be used perhaps as a topical cream, especially on burn victims and on patients to reduce post-operative scars.

4.2 The LP founder haplotype

The LP founder haplotype (LPh1) 19-12-23-22 (ND1-D1S2343-D1S305-D1S2624) was very common in the LP groups studied, was rare in the normal Coloured population group studied (0/24, table 3-5) and was common in the White population studied (1/5, table 3-5). This is generally expected with a founder effect. The fact that a haplotype similar to the founder haplotype is seen in the White unaffected group analysed is an indication that the Q276X mutation is likely to have arisen within Europe before colonisation of South Africa in 1652.

It is also interesting to note that there were very few haplotypes similar to LPh1 within the random Coloured individuals. This is surprising as the carrier rate for the Q276X mutation within this population is estimated at just over five percent (1/9 individuals are carriers), and hence the LPh1 haplotype that is associated with the Q276X mutation would be expected to be seen at a frequency of roughly five percent. Within the random Coloured population, only one haplotype similar to LPh1 was observed among the 81 haplotypes estimated by Arlequin. Therefore, within this population, a haplotype similar to LPh1 was observed in just over one percent of the population. However, this discrepancy may be accounted for by the fact that a small sample size was used (n=48 diploid individuals, of which only two were observed to be Q276X carriers, equalling just over four percent) and that the haplotypes were estimated by Arlequin and not observed haplotypes. In addition to this, while errors due to improper data capturing, incorrect

assigning of data and other forms of data biasing were avoided where possible, they cannot totally be excluded as a source of error.

4.3 Linkage Disequilibrium in LP individuals

As a founder effect was suspected with regards to LP in South Africa and this was supported by the presence of a single LP mutation associated with one founder haplotype, high levels of LD were expected between the markers used and the Q276X mutation itself. This was proven to be true, as significant LD was detected between all four microsatellite markers in both LP groups using the D' statistic and highly significant LD was also detected between all four microsatellite markers and the Q276X mutation using the δ statistic in both LP groups.

The high level of LD detected in this study was further accentuated by the fact that dinucleotide microsatellite markers were used to detect LD. Dinucleotide microsatellite markers are probably the fastest mutating class of markers within the genome, with mutation rates of 4.5×10^{-4} per allele transfer[56] not uncommon. When using these markers to detect LD, it would be expected that detectable LD would decrease relatively rapidly as each marker mutated away from the original allele that was in LD. In contrast, when using single nucleotide polymorphisms (SNPs) much larger areas in LD could be detected, as SNPs generally have a very low mutation rate and LD detected with SNPs is likely to decay only due to recombination. LD detected by microsatellites is likely to decay due to recombination and mutation at each locus. A corollary of this is that SNPs are more sensitive at detecting LD, and that if LD is detected with microsatellites, the LD most probably will have originated relatively recently. Additionally, if LD is detected with microsatellite markers, if the same genomic segment is investigated with SNP markers, it is likely that LD will be detected over a larger genomic segment.

It is not known when the mutational event that created the Q276X mutation occurred, but it has been shown that the Q276X mutation was introduced into South Africa around or very soon after 1652. Therefore the Q276X mutation has been in South Africa for 352

years or approximately 17 generations. This is a short genetic time, and therefore the original haplotype associated with the Q276X mutation has not had much time to decay into new haplotypes. This will lead to high amounts of LD seen within the LP patients of South Africa with respect to the markers used and the Q276X mutation.

As expected, in the larger, less consanguineous White population, no significant LD could be detected in this study, while the smaller Coloured population of Namaqualand showed more significant LD. However, taking into account the population history of the Coloured population, the amount of LD seen was lower than expected. Perhaps this is an indication that the Namaqualand population is more genetically heterogeneous than was previously thought, that there is more immigration into this population than was previously thought, or that perhaps the general population of Namaqualand is becoming aware of the problem of LP in their community and are actively avoiding consanguineous relationships or seeking out partners whose families have no history of LP. Perhaps all the above suggestions are acting synergistically to increase the level of genetic heterogeneity seen in the Namaqualand population.

4.4 Variation within the *ECM1* gene

In order to identify common *ECM1* variants, a screening approach was taken and therefore *ECM1* was screened using dHPLC for common variants by using a small sample size (20 individuals from three populations). Variants that were detected by dHPLC were then sequenced to determine the actual base change.

dHPLC analysis in some cases is unable to differentiate between different variants within an amplimer.[48] It is therefore essential to identify the underlying DNA mutation and to screen all individuals with variant dHPLC patterns with a more specific detection method. One method that is highly specific is restriction fragment length polymorphism (RFLP). RFLP uses a restriction endonuclease that cuts the amplimer at a specific DNA sequence. Mutations can abolish or create new restriction endonuclease recognition sites, hence a specific restriction endonuclease either will not cut or will cut that sequence,

respectively. Restriction endonucleases are specific enough that if a sequence differs by just one base from its recognition sequence, then it will not cut the DNA at that site. Therefore, if time were not a limiting factor, the variants identified in this study would have been specifically examined for variant frequencies. This would have been done in a larger sample set for each of the three populations investigated.

Two variations of the dHPLC screening method were used in this study, namely the mixing and the non-mixing methods. The choice of moving away from the mixing method to the non-mixing method for identifying variants with dHPLC will be criticized by some, as this disallows the direct comparison between exons one to eight pI (analysed with the mixing method) and exons eight pII to ten (analysed with the non-mixing method). This is because the mixing method identifies more heteroduplex patterns than the non-mixing method will. However, the differences in frequencies identified will be small and are likely not to be significant. If one assumes that there is only one variant within a genomic segment being analysed by dHPLC and assuming the variation is in Hardy-Weinberg equilibrium, the mixing method will identify $(2pq + q^2)N$ variants (assuming q to be the frequency of the rarer allele and the reference homozygote is homozygous for the common allele or p^2). Similarly, the non mixing method will identify only $(2pq)N$ variants within a group of N individuals. Therefore, the difference in the number of heteroduplexes identified by the two methods will be q^2N, which would be small as it is derived from the rarer allele.

If we take the Q276X mutation within the random Namaqualand Coloured population sample as an example, which has been shown to be in Hardy-Weinburg equilibrium, the mixing method would theoretically identify $(0.104 + 0.003) \times 100 = 10.7$ or 11 variants while the non-mixing method would theoretically identify $(0.104) \times 100 = 10.4$ or 10 variants. Practically, if the 100 random Namaqualand Coloured samples were used for dHPLC analysis to identify the Q276X mutation, 11 dHPLC heteroduplex patterns would have been identified using the non-mixing method while 11 dHPLC heteroduplex patterns would have been identified using the mixing method. This is because no Q276X

homozygotes were identified within the 100 random samples. Therefore, the frequencies obtained from the two methods are likely to be comparable.

Should there be a second variant within the genomic segment analysed that creates a similar dHPLC heteroduplex pattern to the Q276X pattern, this will create a false frequency for the Q276X mutation. This however is a problem inherent to both the mixing and non-mixing methods, and therefore, based on the information given above, the two methods may tentatively be used interchangeably. However, as the rarer allele increases in frequency, the difference between the two methods becomes larger and hence the larger the error becomes. Table 4-1 estimates the differences that would be observed between the two methods for a range of q values.

Table 4-1: Error calculation for the percentage of heteroduplexes detected using dHPLC analysis for a range of q values

p	0.99	0.90	0.80	0.70	0.60	0.51
q	0.01	0.10	0.20	0.30	0.40	0.49
2pq	0.0198	0.18	0.32	0.42	0.48	0.4998
q^2	0.0001	0.01	0.04	0.09	0.16	0.2401
Mixing Method (%)	1.99	19.00	36.00	51.00	64.00	73.99
Non-Mixing Method (%)	1.98	18.00	32.00	42.00	48.00	49.98
Difference (%)	0.01	1.00	4.00	9.00	16.00	24.01

p and q are the frequencies of the common and rare allele respectively, while 2pq is the expected frequency of the heterozygote and q^2 is the expected frequency of the rare homozygote. Mixing method is the percentage of heteroduplexes identified using the mixing method and similarly non-mixing is the percentage of heteroduplexes identified using the non-mixing method in a group of 100 individuals. Difference is the expected percentage of heteroduplexes missed by the non-mixing method in relation to the mixing method.

If the frequency data obtained from dHPLC in this study are investigated for an underestimation in heteroduplex frequency, exons one through seven will have negligible underestimation of heteroduplex frequencies as they were analysed with the mixing method, while exons eight to ten should be investigated for underestimation due to the non-detection of homozygous rare alleles (q^2). The average variation frequencies (averaged from all three populations, and rounded up to the nearest 0.1) for exons eight,

nine and ten are 0.20, 0.20 and 0.00, respectively. Therefore 2pq for exons eight, nine and ten are 0.20, 0.20 and 0.00, respectively. If data from Table 4-1 are used, we can expect that between one and four percent of heteroduplexes were missed for exons eight and nine, while no variants from exon ten were missed.

All the exons of *ECM1* were successfully screened by dHPLC in the 60 individuals. Sequencing results were obtained for only exons two, three, four five and the first part of exon six. Sequencing for variants in the other exons was not completed due to time constraints. All variants identified were either intronic or reflected synonymous variation, with the exception of the T130M missense mutation. Variation within *ECM1* exon seven would have been especially interesting as it is the ECM1 isoform that contains exon seven (ECM1a) that plays the more important biological role. Variant frequencies were not investigated at the DNA level in all samples, again due to time constraints.

The T130M mutation is a missense mutation that changes a threonine amino acid to a methionine. This variation is seen in the fourth exon and is therefore in the cysteine free domain of ECM1. It is therefore theoretically unlikely to alter the ligand binding properties of ECM1. While the T130M substitution changes a polar hydrophillic amino acid to a non-polar hydrophobic amino acid, this is not likely to alter the properties of ECM1 significantly. This is thought because the T130M substitution has been identified in at least three different studies, namely this study, Hamada *et al.*'s study,[3] and the T130M substitution is a known variant in the SNPper database (http://snpper.chip.org). Using the *Bsm*I restriction endonuclease, Hamada *et al.* found that the T130M variation was polymorphic, with the T allele having a frequency of 0.61 and the M allele having a frequency of 0.39 in a group of 36 individuals from mixed ethnic backgrounds.

Notably, the majority of the variants identified in this study (6/8) are cytosine (C) to thiamine (T) transitions or the complimentary version of C→T transitions, namely guanine (G) to adenine (A) transitions. Incidentally, the Q276X mutation results from a C→T transition. Additionally, of the base changes causing pathogenic mutations for LP, nine of the eleven are due to C→T or G→A transitions. (I. Chan, personal

communication, Table 4-2). Table 4-2 summarises all known variation within the *ECM1* gene. Of the 30 base substitutions in Table 4-2, 21 (70%) are due to C→T transitions (this includes G to A transitions). The most common mechanism of inducing a C to T transition occurs when the C (especially the C of CpG dinucleotides) is methylated and subsequently deaminated to a T.[49] Methylation is a common form of gene regulation and is associated specifically with the down regulation of transcription.[57] This may suggest that *ECM1* is under tight transcriptional control via methylation.

Table 4-2: Summary of all known variation within the *ECM1* gene (as of 12/2004)

SNPper SNP rs number	Relative position and nucleotide change	Amino acid change	Description
rs1050867	nt 39 C→G	n/a	SNPper
rs4995236	nt 1122 G→T	n/a	SNPper
n/a	nt 1314 C→T	n/a	TS
n/a	nt 1400 C→T	n/a	TS
n/a	nt 1448 C→T	Q32X	IC
n/a	nt 1486 G→A	n/a	TS
n/a	nt 1590 C→T	R53X	IC
n/a	nt 1673 A→G	n/a	TS
n/a	nt 1732 del G	n/a	IC
n/a	nt 1823 C→T	P77P	TS
n/a	nt 1875 C→T	Q95X	IC
n/a	nt 2031 C→T	Q114X	IC
n/a	nt 2057 C→T	H122H	TS
n/a	nt 2655 C→A	n/a	TS
rs3737240	nt 2773 C→T	T130M	SNPper, WvH-T, H
n/a	nt 2782 ins C	n/a	IC
n/a	nt 2788 del T	n/a	IC
n/a	nt 2822 del 3 ins 16	n/a	IC
n/a	nt 2823 ins AA	n/a	IC
n/a	nt 2864 G→A	W160X	IC
rs1050874	nt 2876 C→G	L164L	SNPper
n/a	nt 2883 T→A	F167L	IC
n/a	nt 2883 T→C	F167I	IC
rs4970979	nt 3258 C→T	n/a	SNPper
n/a	nt 3274 del TG	n/a	IC
n/a	nt 3324 del A	n/a	IC
n/a	nt 3369 C→T	R243X	IC

n/a	nt 3431 del C	n/a	IC
n/a	nt 3468 C→T	Q276X	IC
n/a	nt 3558 del A	n/a	IC
n/a	nt 3678 C→T	Q346X	IC
n/a	nt 3719 G→A	W359X	IC
rs875514	nt 3973 C→G	n/a	SNPper
n/a	nt 4249 ins C	n/a	IC
n/a	nt 4312 del C	n/a	IC
rs13294	nt 4405 A→G	G415S	SNPper, H
rs945715	nt 4949 C→T	n/a	SNPper
n/a	nt 5067 del A	n/a	IC
rs1050901	nt 5320 A→G	G528R	SNPper
rs1050904	nt 5342 C→T	S535F	SNPper
rs3209760	nt 5361 A→G	X541W	SNPper
rs1050911	nt 5418 C→T	n/a	SNPper
n/a	Exon 9 – 10 del	n/a	IC

TS indicates variation identified during this study, IC indicates LP causing mutations (I. Chan personal communication), H indicates SNPs identified by Hamada *et al.*[3] and SNPper indicates known variation taken from the CHIP Bioinformatics Tools website using the SNPper database (http://snpper.chip.org). The relative position is measured from the first nucleotide of the 5' untranslated region (UTR) to the last nucleotide of the 3' UTR.

4.5 Future studies on LP and *ECM1*

The gene that is mutated in LP, *ECM1*, has now not only been identified, but investigation into the properties of ECM1 has now begun. Investigation of the *ECM1* gene in the South African population has progressed with regards to mutation types, mutation frequency and normal variation found within it. A large genomic segment around *ECM1* (9.98Mb) has been investigated and found to be in significant LD with the Q276X mutation and with each marker in South African LP individuals. Therefore future studies with regard to LP could now be aimed at association studies with common skin scarring disorders and cancer, with treatment development and defining the physiological dynamics of ECM1 in normal and abnormal systems. An example of where this has already occurred is with lichen sclerosis, which exhibits some clinical similarity to LP.[58] Oyama *et al.*[58] have reported that lichen sclerosis patients have autoantibodies against ECM1. Oyama *et al.* suggest that this perturbation of ECM1 function is likely to explain some of the clinical symptoms that are similar in LP and lichen sclerosis.

4.5.1 Genetic studies

As there is now evidence that partial rescue of the LP phenotype is possible due to aberrant splicing out of the exon with the mutation,[3] *ECM1* splicing should be investigated in more depth and variation that is near or within *ECM1* should be investigated for effects on splicing. The fact that the human body has a mechanism whereby it recognises the lack of a certain protein (ECM1 in this case) and attempts to bypass the mutated exon with blind splicing out of exons, has interesting implications for the evolution of the human genome and genomes in general. If a mutation that induces skipping of certain exons is more beneficial to the individual and induces a positive selective advantage in the individual affected, then that variant and hence the skipping of the exon would be favoured and over time the variant and exon skipping may become fixed in the population. This may already have happened to *ECM1* exon 5a, as expression of this exon has only been detected in cancer cells. However, it may be that exon 5a is only needed on rare occasions in humans, and for the majority of the cases it is spliced out during mRNA processing.

As there is a remarkable amount of variability in the severity of LP symptoms seen in the Namaqualand LP population (J. McGrath, personal communication) and taking into account that all Namaqualand LP individuals are genotypically identical for the Q276X mutation, the genotype-phenotype discrepancy suggests that there are likely to be other modifiers, both genetic and environmental, that can reduce or enhance the pathogenic effect of the Q276X mutation in the homozygous state. Therefore, on the genetic side, the genes whose proteins interact with ECM1 or which play a role in the same physiological processes as ECM1 should be investigated for variation. Once identified, association studies may then be conducted to determine if the variants are responsible for the lack of genotype-phenotype correlation in LP patients.

Additionally, as lack of ECM1 is now implicated in at least two skin disorders (LP and lichen sclerosis), it is likely that functional variants within *ECM1* will have an impact on normal skin. For example, variant A in *ECM1* may predispose a person to excessive scar formation while variant B may delay proper skin healing.

The strong LD seen within the Coloured and White LP groups observed with dinucleotide microsatellite markers suggest that there may be a much larger genomic segment that would show LD if SNP markers were used to detect LD. Many other studies have shown that there are other founder effects within the South African White population, a few examples include porphyria varigata (PV)[23], familial hypercholesterolaemia (FH)[24] and pseudoxanthoma elasticum (PXE).[26] As these disorders have reached relatively high frequencies because of founder effects in South Africa, the genomic areas around the genes responsible for PV, FH and PXE are likely to show high LD with markers in those regions. If this is extrapolated, it is likely that many other locus pairs, which may or may not be associated with genetic disorders, will show significant LD in the White population of South Africa. In contrast, the Coloured populations of South Africa have not been as extensively investigated as the White population have been. The Coloured populations would be of great interest, as they are hybrid populations derived from the admixture of the White population and the Black, Khoisan or other populations found in South Africa.[39] Therefore the Coloured populations of South Africa would each be expected to show high amounts of LD at many locus pairs, some of which would be similar to the White population. Therefore it would be interesting to investigate the Coloured populations of South Africa for LD and then to compare and contrast the findings from the Coloured populations to the findings of the South African White population.

The White population of South Africa is larger than the many small Coloured populations and therefore would be expected to show less LD, even though the majority of the White population of South Africa arose from relatively few individuals. This is borne out by the observation that this study has found the Namaqualand Coloured group to have more markers in significant LD in than the White group. Extrapolating from this, the Coloured populations of South Africa may be more likely to demonstrate LD and may make excellent target populations on which to conduct association studies for polygenic and multifactorial disorders.

4.5.2 Protein studies

Future studies on LP and hence *ECM1* clearly need to be aimed at elucidating the exact mode of action of the mutated ECM1 protein or how exactly normal biology is perturbed by the absence of ECM1. In order to do this, proteins that the normal ECM1 protein interacts with *in vivo* need to be identified and then these proteins need to be fully characterized with regards to their physiological functions. Further to this, the calcium binding domains of ECM1 need to be properly investigated and their effect on or interactions with other proteins elucidated. A hypothetical physiological pathway could then be compiled, which could be used to identify drug targets. This could be used not only for treatment of LP, but in all dermatological functions that ECM1 is involved in.

4.5.3 New treatments for LP

A novel idea for treatment of LP patients is to substitute the ECM1a protein in the affected organs. A method of accomplishing this would be to genetically modify bacteria to produce ECM1a. The resultant ECM1a protein could possibly be used as a treatment for LP, and may be administered to a patient as a topical cream, as the majority of the clinical symptoms are dermatological in nature. This would be viable only if the skin is permeable to a protein of ECM1's size and provided it is capable of localising correctly within the extracellular matrix. An aerosolised form of ECM1a may be of benefit in the reduction of the infiltration of the vocal cords, leading to a reduction in hoarseness. As the hyaline deposits do not grow significantly over time, there may be a form of "hyaline homeostasis" whereby excess hyaline material is removed from the affected organs while more is deposited. Therefore, if the cause of the hyaline deposition is removed, the "hyaline homeostasis" mechanism may remove the residual hyaline material, perhaps returning the affected area back to normal. Other symptoms, such as the calcification of the temporal lobes could not be treated in this manner. Systemic gene therapy would be more applicable in this situation, provided the addition of ECM1 protein was able to reverse the calcification process. Naturally, these therapies would have to undergo extensive testing in order to investigate the proof of concept and to evaluate the effectiveness of such a treatment before implementation.

4.6 Implications of this study

The test for the Q276X mutation is a simple restriction digestion of a PCR product that now can be used as part of a routine diagnostic service in any molecular genetics laboratory. The RFLP test using the *Bfa*I restriction endonuclease can simply and easily differentiate between a Q276X homozygote, heterozygote and a non-Q276X homozygote. Therefore it is now possible to rapidly confirm a clinical diagnosis of LP at the molecular level, thereby allowing the patient to understand the genetic basis of their disorder. He/she can now make more informed choices with regards to their health and importantly, he/she can make informed reproductive choices.

Additionally, the *Bfa*I RFLP test now allows for carrier screening, especially in at-risk populations. This study has shown that of the 36 South African LP patients, all were homozygous for the Q276X mutation. Therefore, if a South African individual is negative for the Q276X mutation after a *Bfa*I RFLP test, then we can be relatively confident that this person is not likely to suffer from LP nor be a carrier of an LP mutation. This study cannot exclude the possibility that there may be another, rare, LP mutation at a very low frequency in the South African population nor can it exclude the possibility of a *de novo* LP mutation occurring within an individual. In these scenarios, a proper clinical diagnosis is essential, and then molecular screening of the individual's *ECM1* gene to find the mutation could commence, if there is a strong suspicion of an LP diagnosis.

If the *Bfa*I RFLP test was not available for whatever reason, linked marker analyses using the ND1, D1S2343, D1S305 and D1S2624 markers could be performed. However, these markers span 9.98Mb, and hence the chance of recombination is roughly 10% per meiosis in the genomic region covered by these markers.[49] Therefore, the marker D1S2343 should be given preference when diagnosing LP in an individual, as it is the marker that is situated the closest to the *ECM1* gene, followed by D1S305, ND1 and lastly D1S2624. Therefore, preference should be given in that order when testing for LP in a family where the Q276X mutation status is unknown. The alleles that are associated with the Q276X mutation are 12 (D1S2343), 23 (D1S305), 19 (ND1) and 22 (D1S2624).

If there is a combination of two or more of these alleles, then further investigation is warranted, and the *Bfa*I RFLP test is highly recommended.

With regards to a couple who are both known to be Q276X carriers and are therefore at one in four risk of conceiving a child affected with LP, pre-natal screening can now be offered as an option. If the foetus were diagnosed with LP at the molecular level, the couple would have the option of aborting the affected foetus, should they so wish. This option has not yet been explored in the Namaqualand population, and since LP is rarely life threatening, it is uncertain whether a prenatal test would ever be requested. This study has established carrier rates of the Q276X mutation for the populations affected in South Africa. This gives a genetic counsellor and hence the patient more accurate information regarding the chances of being an LP carrier, or the chances of any couple taken at random from the South African White or Namaqualand Coloured population of conceiving a child affected with LP.

In summary, the LP population of South Africa, especially the Coloured population of Namaqualand, has made a large contribution to the understanding of LP. The South African LP community is likely to carry on contributing to the understanding of LP by making themselves available and amenable to novel treatments for LP. The medical and scientific community of South Africa should seize this opportunity to understand the molecular pathogenic aetiology of LP in greater detail. If they do, and are able to conceive of and create a treatment for LP, this will result in the lessening of the burden of disease in people who are in need.

CHAPTER 5 Appendices

5.1 Salting-out protocol for extracting DNA from whole blood

Collect blood in ACD or EDTA tubes.
Centrifuge sample and remove plasma, left with red blood cells and white blood cells.
Transfer to Nunc tube.
Fill each tube to the 45/50ml mark with cold Sucrose-Triton X Lysing buffer.
Invert the tube to mix.
Centrifuge for 10 minutes at 4^{o}C (2300rpm).
Pour off supernatant fluid.
Wash in 20-25ml Sucrose-Triton X Lysing Buffer.
Put on ice for 5 minutes.
Centrifuge for 5 minutes at 4^{o}C (2300rpm).
Pour off the supernatant fluid.
Add 3ml T20E5.
Add 200µl 10% SDS.
Add 500µl Proteinase K Solution.
Mix well by inversion.
Put at 42^{o}C to 50^{o}C overnight (no need to agitate).
Add 1ml saturated NaCl.
Mix well by inversion.
Put on ice for 5 min.
Centrifuge for 30 minutes at 2500rpm.
A white pellet should be visible which contains proteins that have been precipitated by the salt. If no pellet is visible, centrifuge again. If still no pellet is visible, add another 1ml saturated NaCl.
Transfer the supernatant containing the DNA to a new tube.
Add two volumes of absolute ethanol and keep at room temperature.
Agitate gently and spool, fish or precipitate DNA. If precipitating, do not spin too long to avoid salts and proteins precipitating as well.
Wash DNA in 70% ice cold ethanol.
Air-dry DNA.
Dissolve DNA in appropriate volume (±300µl) of TE buffer or ddH_2O.

5.2 NucleoSpin® Extract protocol for direct purification of PCR products

Adjust PCR sample volume to 50µl with TE buffer (pH 7.5) if the PCR sample volume is less than 50µL.
Mix four volumes of buffer NT2 with one volume of PCR sample.
Insert the NucleoSpin® Extract column into a 2ml collecting tube.
Load sample onto the column and centrifuge for 1 minute at 11000 x g.
Discard the flow-through and place the NucleoSpin® Extract column back into the 2ml collecting tube.
Add 600µl buffer NT3.
Centrifuge for 1 minute at 11000 x g.

Discard the flow-through and place the NucleoSpin® Extract column back into the 2ml collecting tube.
Add 200µl buffer NT3.
Centrifuge for 2 minutes at 11000 x g.
Incubate the NucleoSpin® Extract columns for 2 – 5 minutes at 70^{o}C.
Place the NucleoSpin® Extract column into a clean 1.5ml microcentrifuge tube.
Add 25 – 50µl elution buffer NE.
Incubate the NucleoSpin® Extract columns for 1 minute at room temperature.
Centrifuge for 1 minute at 11000 x g.

The elute contains the purified PCR product and is ready to be used as template for subsequent sequencing reactions.

5.3 SigmaSpin™ Post-Reaction Purification Protocol for purification of cycle sequencing products

Loosen the column cap by half a turn and snap off the bottom closure.
Place the column into a collection tube and centrifuge at 750 x g for 2 minutes.
Discard the elute and the collection tube.
Place the column into a new microcentrifuge tube.
Pipette the cycle sequencing reaction directly into the center of the column.
Centrifuge the above assembly for 4 minutes at 750 x g.
Discard the column but retain the elute.

The eluted contains the purified cycle sequencing product in water containing approximately 25ppm preservative.

5.4 Microsatellite marker results

Table 5-1: Microsatellite results for the random Coloured individuals (N = 60, LP 1 - LP 50), Coloured LP patients and their family members (N = 157, LP 101 - LP 157), White LP patients and their family members (N = 12, LP 502 - LP 513) and the random White individuals (N = 30, RW 251 – 280).

Marker	Nov Di 1	D1S2343	D1S305	D1S2624
Code	Alleles	Alleles	Alleles	Alleles
LP 1	17	12	19	24
	20	16	25	24
LP 2	20	NR	NR	NR
	24	NR	NR	NR
LP 3	15	19	19	21
	17	26	26	21
LP 4	17	21	19	20
	17	21	21	22
LP 5	NR	23	NR	NR
	NR	23	NR	NR
LP 6	19	NR	NR	NR
	21	NR	NR	NR
LP 7	18	NR	20	NR
	18	NR	26	NR
LP 8	21	22	20	20
	21	22	26	22
LP 9	19	20	24	20
	22	20	27	21
LP 10	16	12	21	20
	20	20	24	22
LP 11	17	12	19	18
	17	19	23	18
LP 12	17	12	23	20
	19	22	25	22
LP 13	17	12	19	18
	23	19	23	19
LP 14	17	19	20	19
	17	19	32	21
LP 15	20	12	19	20
	21	19	21	22
LP 16	17	18	23	NR
	17	19	26	NR
LP 17	17	12	23	16
	19	17	23	16
LP 18	15	16	21	19
	15	23	24	22

LP 19	18	17	19	15
	20	18	27	18
LP 20	18	17	22	19
	20	20	23	21
LP 21	15	19	19	18
	18	21	21	20
LP 22	16	18	23	19
	19	19	27	21
LP 23	18	18	19	19
	19	19	25	20
LP 24	17	18	19	20
	18	20	24	21
LP 25	15	19	19	20
	17	26	26	22
LP 26	18	17	24	13
	18	22	25	22
LP 27	21	18	19	16
	21	20	21	19
LP 28	18	18	21	20
	23	23	23	20
LP 29	17	18	19	18
	20	22	20	21
LP 30	17	19	18	18
	19	22	21	17
LP 31	20	17	27	17
	20	18	27	19
LP 32	19	18	19	16
	20	22	27	19
LP 33	17	19	20	19
	21	22	23	21
LP 34	17	22	23	13
	18	17	23	20
LP 35	17	18	24	19
	18	21	26	22
LP 36	17	18	19	19
	18	20	25	21
LP 37	15	21	19	16
	17	24	27	18
LP 38	18	18	23	18
	18	18	23	21
LP 39	18	18	18	19
	21	20	23	19
LP 40	17	18	18	20
	17	19	23	22

LP 41	17	18	18	19
	21	19	25	22
LP 42	17	12	21	18
	20	12	23	22
LP 43	18	18	21	20
	23	19	23	21
LP 44	18	12	21	19
	20	22	26	19
LP 45	17	17	18	16
	18	18	21	18
LP 46	17	12	23	22
	19	19	23	23
LP 47	15	20	23	12
	15	22	26	16
LP 48	19	17	23	18
	21	19	28	20
LP 49	17	18	18	19
	21	22	25	21
LP 50	17	12	18	17
	17	20	19	20
LP 51	17	17	23	19
	20	20	26	23
LP 52	20	18	24	19
	21	21	26	19
LP 53	17	20	18	21
	24	17	22	21
LP 54	20	NR	NR	NR
	24	NR	NR	NR
LP 55	NR	NR	NR	21
	NR	NR	NR	21
LP 56	NR	NR	NR	NR
	NR	NR	NR	NR
LP 57	NR	NR	NR	18
	NR	NR	NR	18
LP 58	NR	NR	NR	17
	NR	NR	NR	17
LP 59	NR	NR	NR	18
	NR	NR	NR	21
LP 60	NR	NR	NR	21
	NR	NR	NR	21
LP 101	19	12	23	20
	19	12	23	22
LP 102	19	12	23	20
	17	20	24	22

LP 103	19	12	23	23
	19	12	26	23
LP 104	18	12	19	20
	19	22	26	23
LP 105	16	12	19	19
	19	17	23	22
LP 106	19	12	23	22
	19	12	23	22
LP 107	19	12	18	22
	19	12	23	22
LP 108	14	12	23	22
	19	12	23	22
LP 109	14	12	18	19
	17	19	23	22
LP 110	19	12	23	22
	19	12	23	22
LP 111	19	12	23	20
	19	12	23	22
LP 112	18	12	23	20
	19	20	23	22
LP 113	19	12	23	22
	19	12	23	23
LP 114	17	12	23	18
	19	12	24	22
LP 115	19	12	23	22
	19	12	23	22
LP 116	17	12	19	22
	19	20	23	22
LP 117	17	12	22	18
	19	12	23	22
LP 118	19	12	23	18
	19	12	23	22
LP 119	17	12	23	22
	19	18	23	23
LP 120	17	12	23	18
	19	17	23	22
LP 121	19	12	23	22
	19	12	23	22
LP 122	19	12	23	20
	19	20	23	22
LP 123	19	12	23	22
	19	12	23	22
LP 124	19	12	23	22
	19	12	23	23

LP 125	17	12	18	22
	19	21	23	21
LP 126	19	12	18	21
	21	19	23	23
LP 127	17	19	18	21
	21	21	18	21
LP 128	19	12	19	22
	19	12	23	23
LP 129	18	12	19	22
	19	22	23	23
LP 130	19	12	23	18
	23	19	23	22
LP 131	18	12	19	22
	19	22	23	23
LP 132	17	12	23	22
	19	18	23	22
LP 133	19	12	23	22
	19	12	23	22
LP 134	19	12	20	19
	20	18	23	22
LP 135	18	12	23	20
	19	18	26	22
LP 136	19	12	23	22
	19	12	23	22
LP 137	19	12	18	22
	19	12	23	22
LP 138	19	12	18	22
	21	18	27	20
LP 139	19	12	18	22
	19	12	23	22
LP 140	19	12	18	22
	19	12	23	22
LP 141	19	12	23	22
	19	12	23	22
LP 142	19	12	23	22
	19	12	23	22
LP 143	17	12	23	20
	19	20	23	22
LP 144	19	12	23	22
	19	12	23	22
LP 145	19	12	23	22
	19	12	23	22
LP 146	19	12	23	22
	19	12	23	22

LP 147	19	12	23	22
	19	12	23	22
LP 148	18	12	23	22
	19	26	26	23
LP 149	19	12	23	22
	19	12	23	22
LP 150	17	12	19	18
	18	18	23	20
LP 151	17	12	19	20
	19	12	23	22
LP 152	17	12	19	20
	19	12	23	22
LP 153	19	12	23	22
	19	12	23	23
LP 154	19	12	23	22
	19	12	23	23
LP 155	17	12	21	20
	19	12	23	22
LP 156	15	21	19	19
	18	26	26	23
LP 157	19	12	23	22
	19	12	23	23
LP 502	15	12	23	20
	19	12	23	22
LP 503	19	12	23	22
	19	12	23	22
LP 504	19	12	23	22
	19	12	23	22
LP 505	19	12	23	22
	19	12	23	22
LP 506	17	12	23	22
	19	12	22	22
LP 507	15	12	23	20
	19	12	23	19
LP 508	19	12	21	18
	21	19	23	22
LP 509	17	12	18	18
	19	19	23	22
LP 510	17	12	21	18
	19	19	23	22
LP 511	17	12	23	18
	19	12	23	22
LP 512	19	12	23	18
	22	12	23	22

LP 513	19	12	23	18
	22	12	23	18
RW 251	18	12	18	18
	18	12	19	20
RW 252	17	12	18	19
	22	20	23	19
RW 253	16	12	18	20
	17	19	21	22
RW 254	17	19	19	18
	20	20	23	20
RW 255	17	12	21	20
	22	17	23	22
RW 256	18	19	18	16
	23	24	23	16
RW 257	17	12	18	19
	22	21	19	20
RW 258	17	19	23	18
	22	19	23	20
RW 259	18	20	18	20
	18	21	19	22
RW 260	23	20	21	20
	23	25	26	20
RW 261	16	19	18	18
	17	20	24	21
RW 262	17	12	19	18
	17	19	21	22
RW 263	17	12	18	18
	21	19	19	21
RW 264	17	20	18	19
	18	22	23	19
RW 265	17	12	18	16
	20	19	23	18
RW 266	15	17	18	18
	17	19	22	20
RW 267	18	18	23	17
	18	19	23	19
RW 268	17	12	21	16
	17	19	24	16
RW 269	18	19	18	16
	22	20	21	18
RW 270	15	19	21	19
	18	19	23	19
RW 271	20	12	18	19
	22	19	23	21

RW 272	17	12	21	20
	22	20	23	22
RW 273	17	12	18	20
	17	21	19	22
RW 274	15	21	19	16
	18	23	21	16
RW 275	18	12	18	20
	18	20	21	22
RW 276	17	19	18	18
	18	21	21	20
RW 277	18	19	22	16
	21	19	25	18
RW 278	15	12	19	18
	17	19	21	20
RW 279	17	17	18	19
	17	19	23	20
RW 280	17	12	19	20
	17	19	21	22

NR indicates no result

5.5 LP pedigrees

LP 149		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22

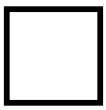
LP 157		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	23

LP 504		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22

LP 144		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22

LP 110		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22

LP 107		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	18	23
D1S2624	22	22



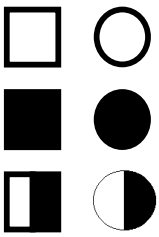
LP 501		
Marker	**Allele 1**	**Allele 2**
Nov Di1	17	20
Q276X	N	N
D1S2343	17	21
D1S305	19	21
D1S2624	19	19

LP 503		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22

The above eight individuals were not part of a pedigree as they are isolated cases from independent families.

The alleles for all individuals, except for LP 501 (where gametic phase cannot be determined), are shown as haplotypes.

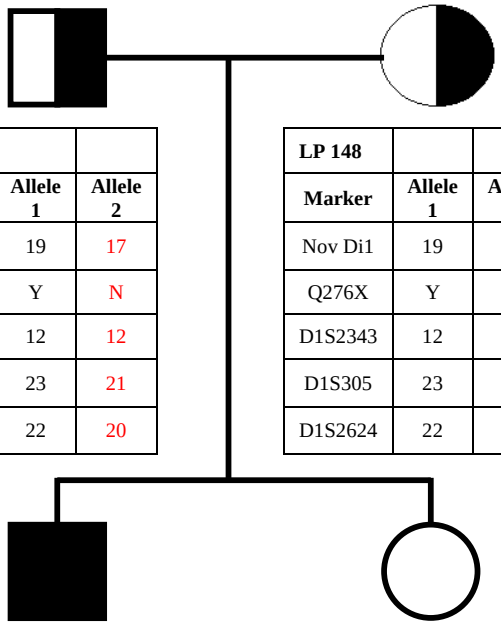
Key:



Unaffected male and female, respectively

LP male and female, respectively

LP carrier male and female, respectively



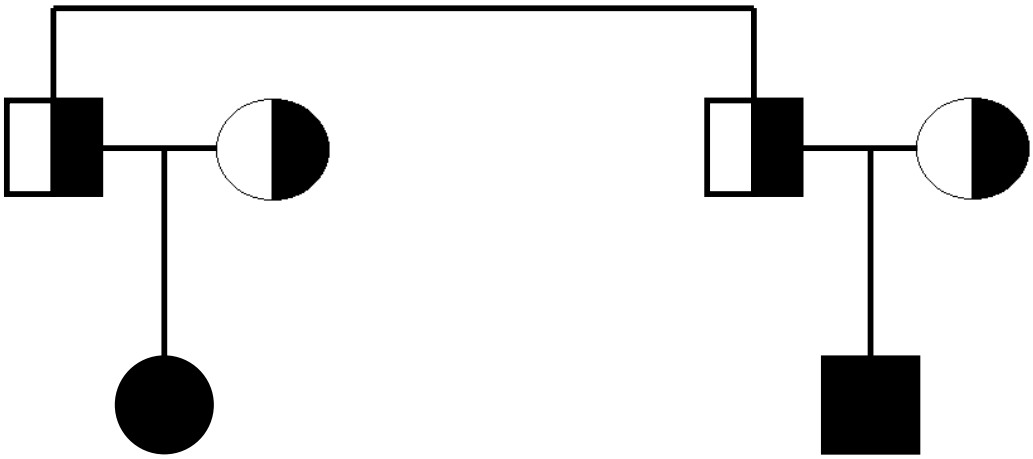
LP 155		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	12
D1S305	23	21
D1S2624	22	20

LP 148		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	18
Q276X	Y	N
D1S2343	12	26
D1S305	23	26
D1S2624	22	23

LP 147		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22

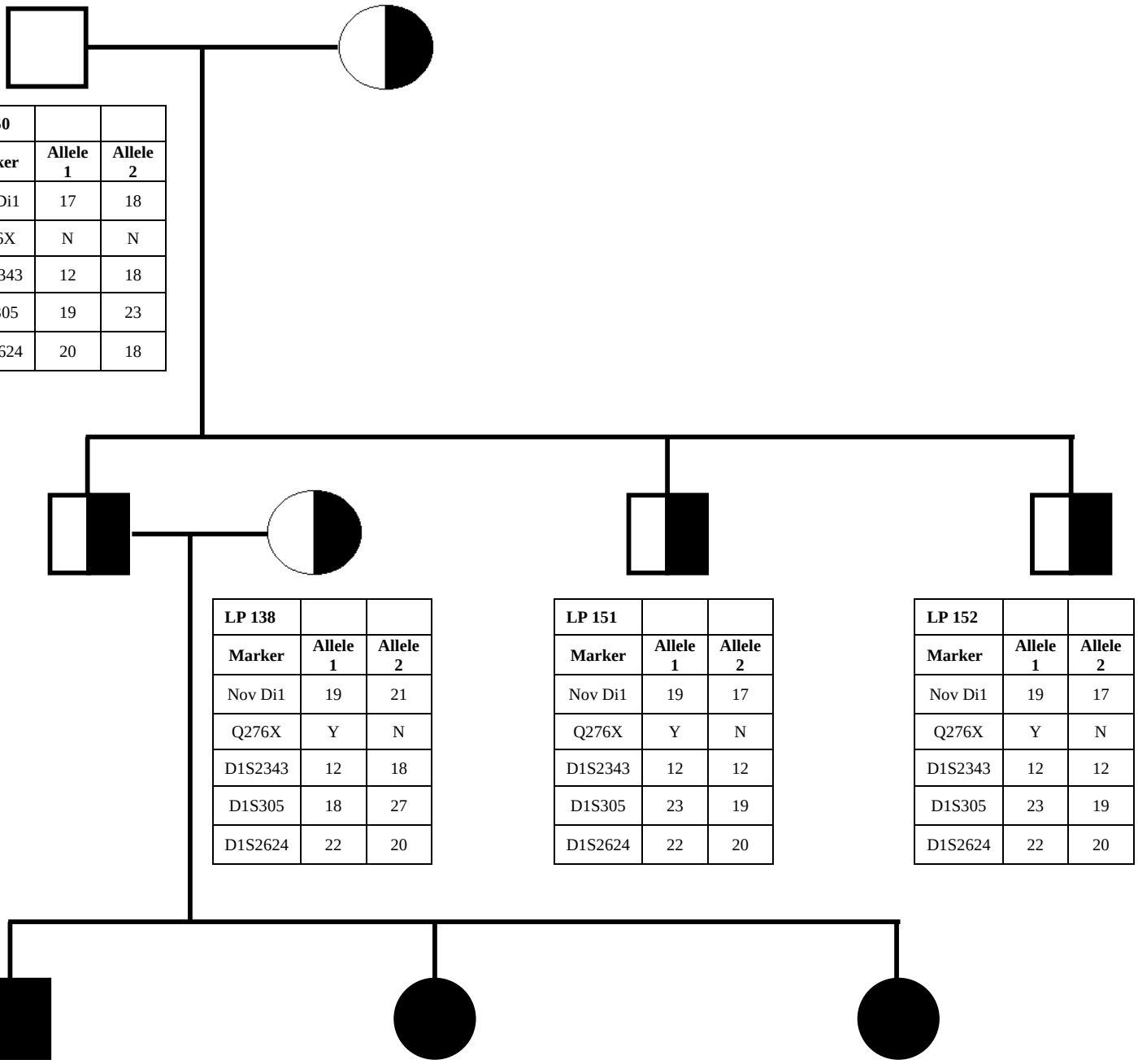
LP 156		
Marker	**Allele 1**	**Allele 2**
Nov Di1	15	18
Q276X	N	N
D1S2343	21	26
D1S305	19	26
D1S2624	19	23

Red in this pedigree indicates non-mendelian inheritance, and is likely to be caused by non-paternity.



LP 153		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	23

LP 154		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	23



LP 150		
Marker	**Allele 1**	**Allele 2**
Nov Di1	17	18
Q276X	N	N
D1S2343	12	18
D1S305	19	23
D1S2624	20	18

LP 138		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	21
Q276X	Y	N
D1S2343	12	18
D1S305	18	27
D1S2624	22	20

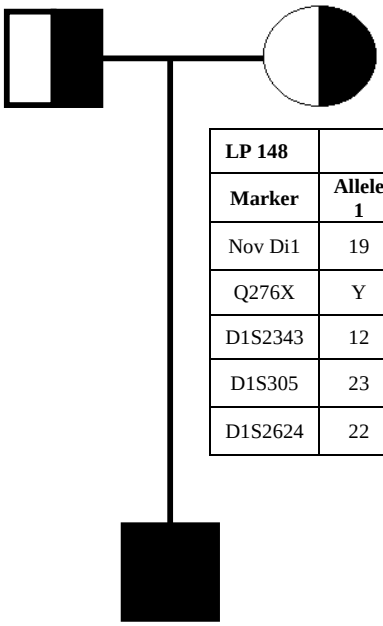
LP 151		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	12
D1S305	23	19
D1S2624	22	20

LP 152		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	12
D1S305	23	19
D1S2624	22	20

LP 137		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	18	23
D1S2624	22	22

LP 139		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	18	23
D1S2624	22	22

LP 140		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	18	23
D1S2624	22	22

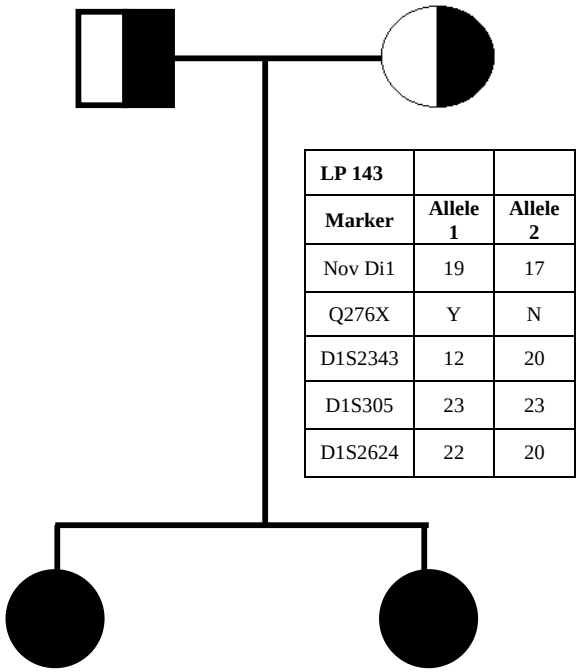


LP 148		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	18
Q276X	Y	N
D1S2343	12	26
D1S305	23	26
D1S2624	22	23

LP 147		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22

LP 145		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22

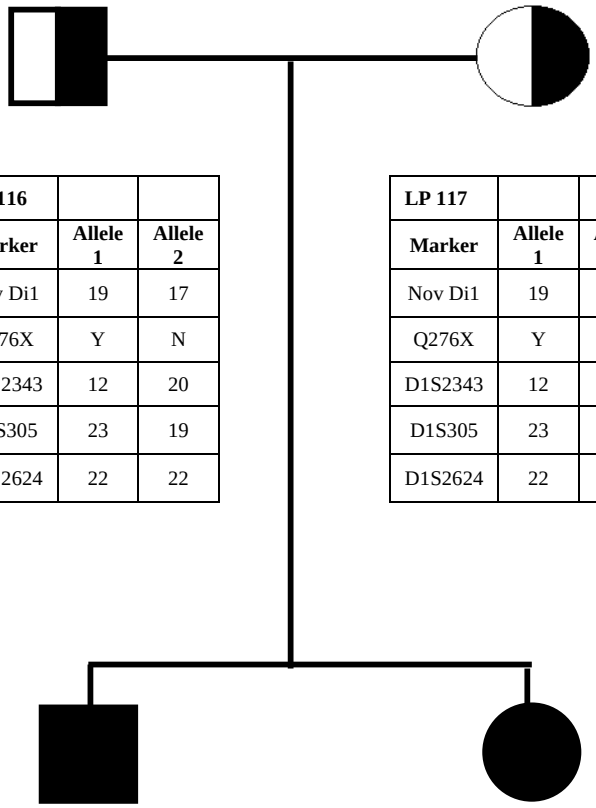
LP 146		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22



LP 143		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	20
D1S305	23	23
D1S2624	22	20

LP 141		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22

LP 142		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22

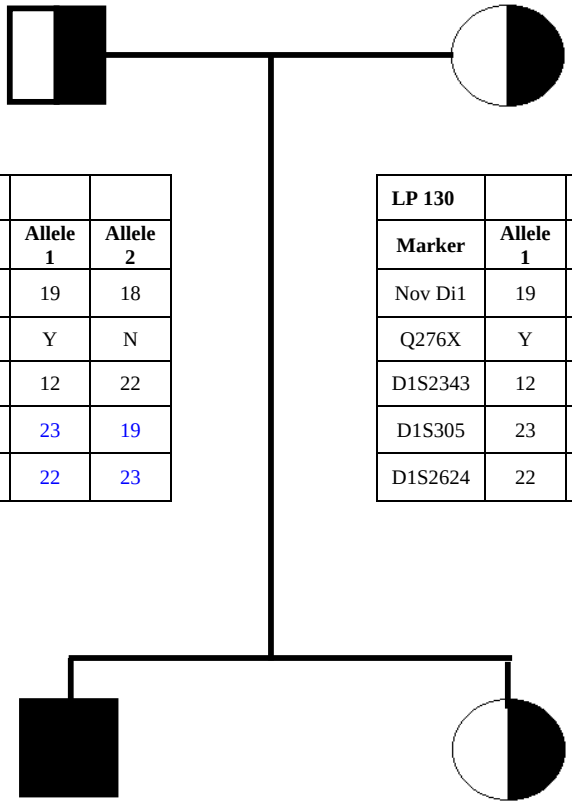


LP 116		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	20
D1S305	23	19
D1S2624	22	22

LP 117		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	12
D1S305	23	175
D1S2624	22	18

LP 115		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22

LP 136		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22



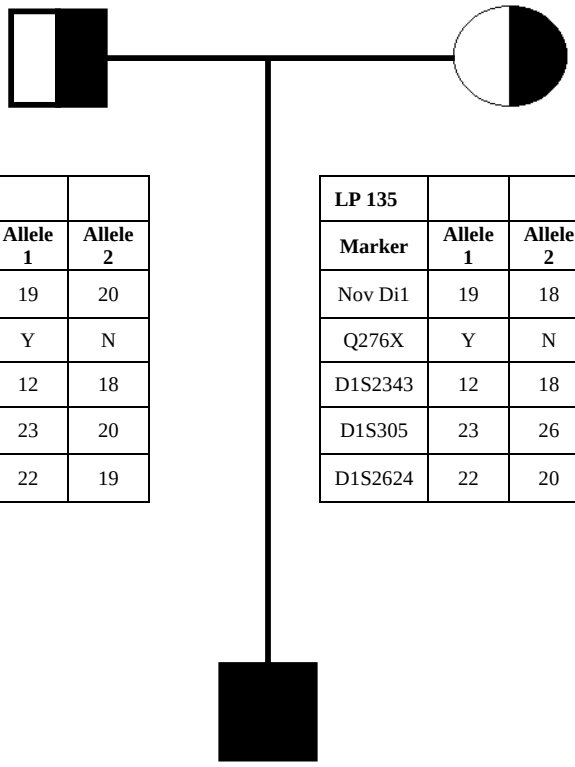
LP 129		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	18
Q276X	Y	N
D1S2343	12	22
D1S305	23	19
D1S2624	22	23

LP 130		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	23
Q276X	Y	N
D1S2343	12	19
D1S305	23	23
D1S2624	22	18

LP 128		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	19
D1S2624	22	23

LP 131		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	18
Q276X	Y	N
D1S2343	12	22
D1S305	23	19
D1S2624	22	23

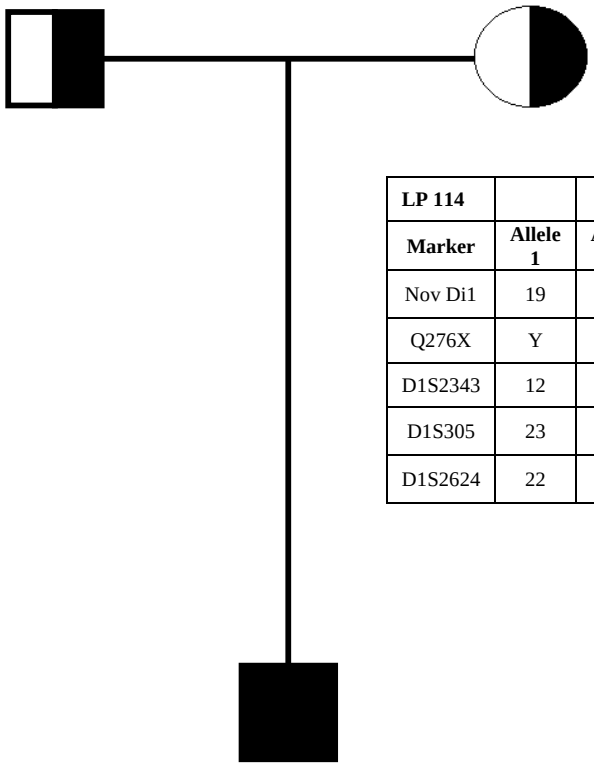
Blue in this pedigree indicates ambiguity of haplotypes with regards to D1S305 and D1S2624 in LP 128 and LP 131. This can be accounted for if there was a recombinational event during gametogenesis in LP 129.



LP 134		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	20
Q276X	Y	N
D1S2343	12	18
D1S305	23	20
D1S2624	22	19

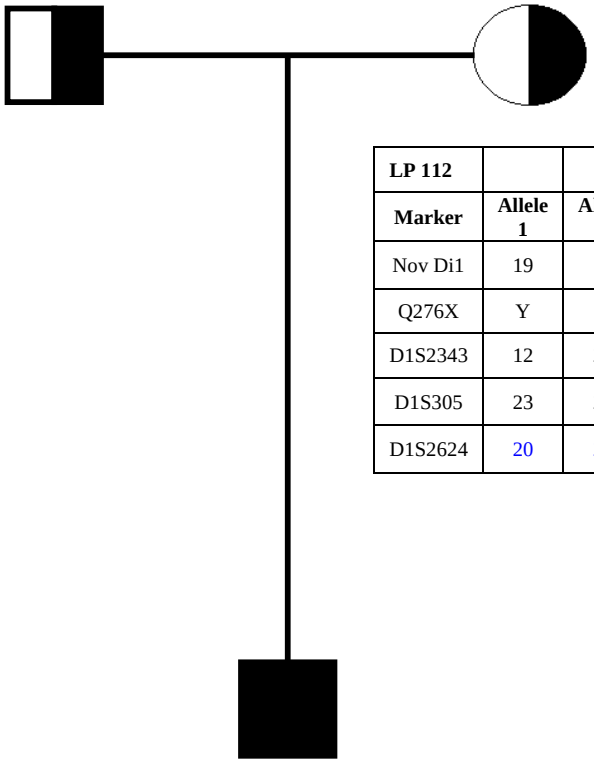
LP 135		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	18
Q276X	Y	N
D1S2343	12	18
D1S305	23	26
D1S2624	22	20

LP 133		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22



LP 114		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	12
D1S305	23	24
D1S2624	22	18

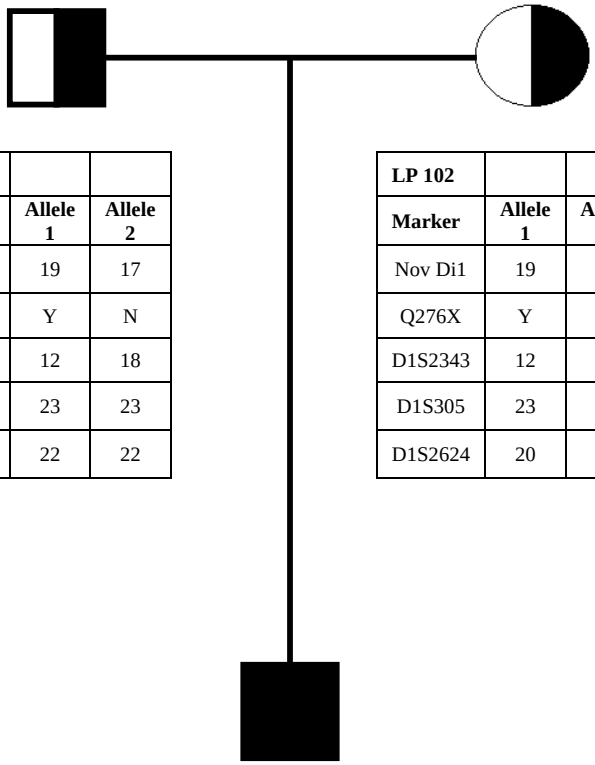
LP 113		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	23



LP 112		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	18
Q276X	Y	N
D1S2343	12	20
D1S305	23	23
D1S2624	20	22

LP 111		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	20	22

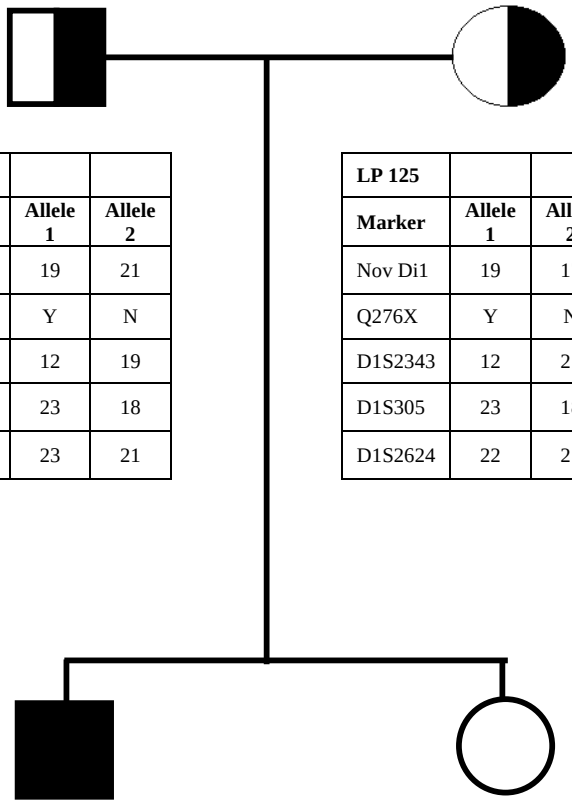
Blue in this pedigree indicates that phase cannot be determined for marker D1S2624 in LP 111, as both his alleles at D1S2624 are identical to his mothers and hence it is not clear which allele he has inherited from his mother.



LP 132		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	18
D1S305	23	23
D1S2624	22	22

LP 102		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	20
D1S305	23	24
D1S2624	20	22

LP 101		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	20	22

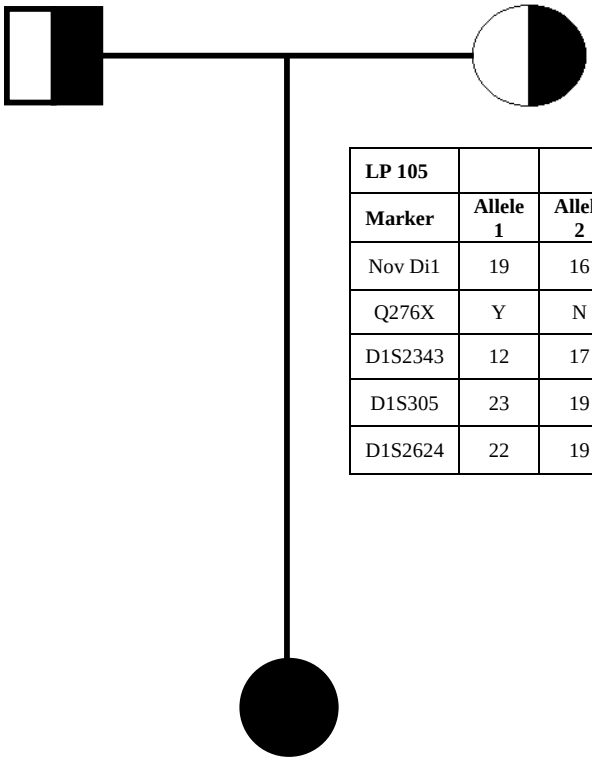


LP 126		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	21
Q276X	Y	N
D1S2343	12	19
D1S305	23	18
D1S2624	23	21

LP 125		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	21
D1S305	23	18
D1S2624	22	21

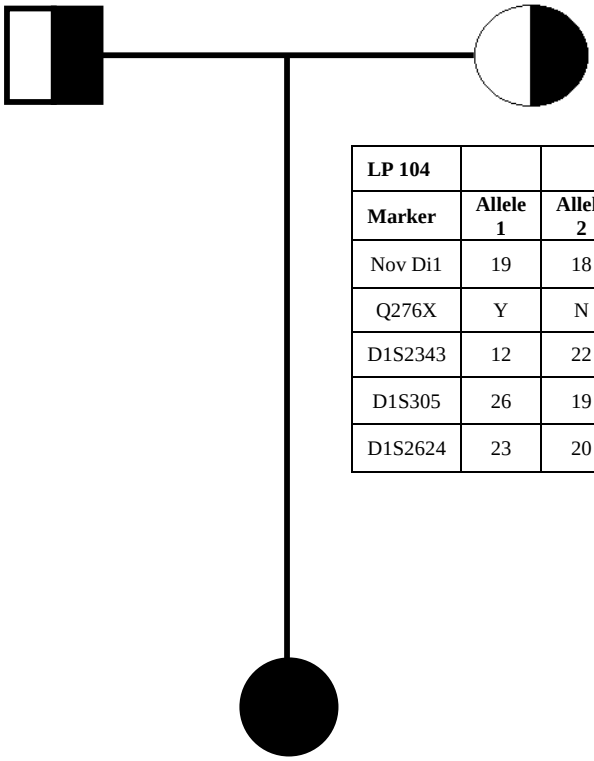
LP 124		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	23	22

LP 127		
Marker	**Allele 1**	**Allele 2**
Nov Di1	21	17
Q276X	N	N
D1S2343	19	21
D1S305	18	18
D1S2624	21	21



LP 105		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	16
Q276X	Y	N
D1S2343	12	17
D1S305	23	19
D1S2624	22	19

LP 106		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22

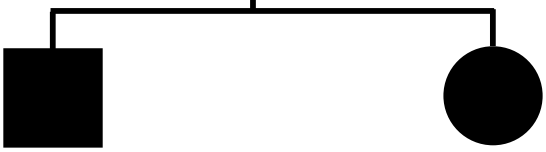


LP 104		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	18
Q276X	Y	N
D1S2343	12	22
D1S305	26	19
D1S2624	23	20

LP 103		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	26
D1S2624	23	23

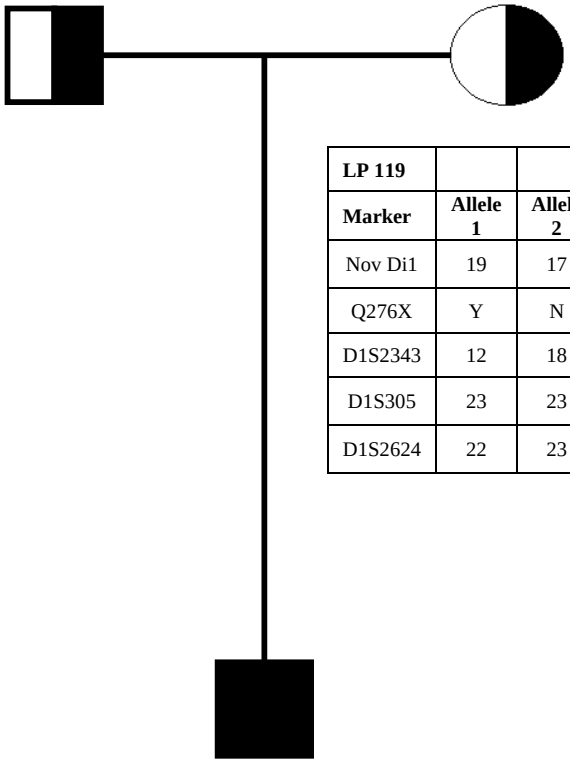
LP 122		
Marker	Allele 1	Allele 2
Nov Di1	19	19
Q276X	Y	N
D1S2343	12	20
D1S305	23	23
D1S2624	22	20

LP 120		
Marker	Allele 1	Allele 2
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	17
D1S305	23	23
D1S2624	22	18



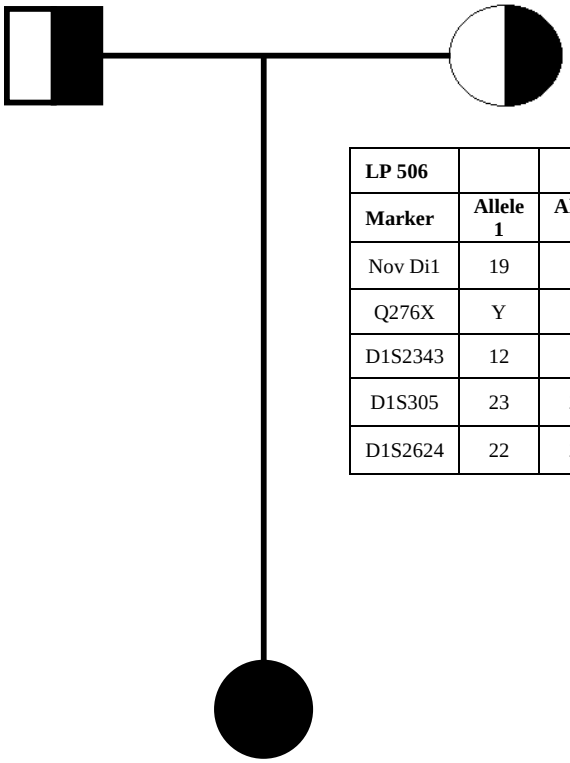
LP 123		
Marker	Allele 1	Allele 2
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22

LP 121		
Marker	Allele 1	Allele 2
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22



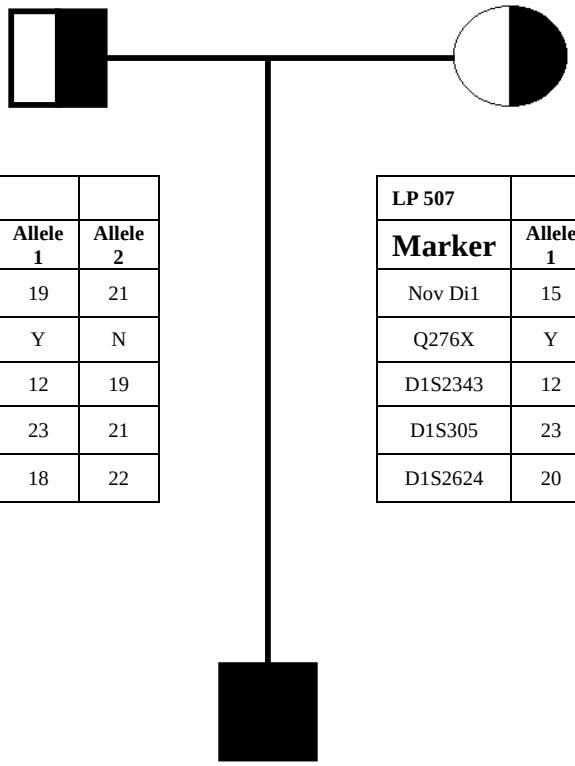
LP 119		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	18
D1S305	23	23
D1S2624	22	23

LP 118		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	18



LP 506		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	12
D1S305	23	22
D1S2624	22	22

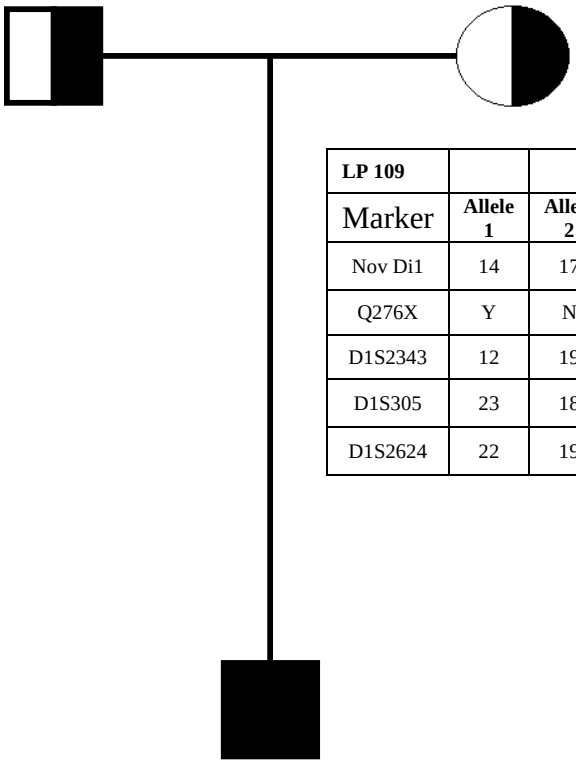
LP 505		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22



LP 508		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	21
Q276X	Y	N
D1S2343	12	19
D1S305	23	21
D1S2624	18	22

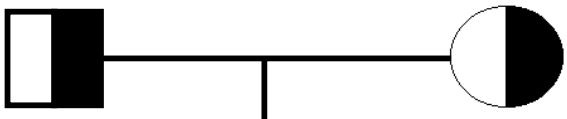
LP 507		
Marker	**Allele 1**	**Allele 2**
Nov Di1	15	19
Q276X	Y	N
D1S2343	12	12
D1S305	23	23
D1S2624	20	19

LP 502		
Marker	**Allele 1**	**Allele 2**
Nov Di1	19	15
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	18	20



LP 109		
Marker	Allele 1	Allele 2
Nov Di1	14	17
Q276X	Y	N
D1S2343	12	19
D1S305	23	18
D1S2624	22	19

LP 108		
Marker	Allele 1	Allele 2
Nov Di1	14	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	22	22



LP 509		
Marker	Allele 1	Allele 2
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	19
D1S305	23	18
D1S2624	22	18

LP 510		
Marker	Allele 1	Allele 2
Nov Di1	19	17
Q276X	Y	N
D1S2343	12	19
D1S305	23	21
D1S2624	22	18

LP 511		
Marker	Allele 1	Allele 2
Nov Di1	17	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	18	22

Blue in this pedigree indicates that phase could not be determined for markers Nov Di1 and D1S2624, as LP 509, LP 510 and LP 511 all share identical alleles at these two loci.

LP 513		
Marker	Allele 1	Allele 2
Nov Di1	22	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	18	18

LP 512		
Marker	Allele 1	Allele 2
Nov Di1	22	19
Q276X	Y	Y
D1S2343	12	12
D1S305	23	23
D1S2624	20	22

Red in this pedigree indicates a discrepancy at marker D1S2624 in a sib pair that is homozygous for the Q276X mutation. The alleles at the D1S2624 locus would be expected to be similar. Possible reasons for the discrepancy are mutation, recombination or lab error.

5.6 Arlequin sample files

5.6.1 Random Namaqualand population genotypic LD estimation and haplotype estimation project file

```
[Profile]

Title="Random Namaqualand Population Haplotype Estimation"
    NbSamples= 1        #Number of sample in the Project.
    DataType= MICROSAT
    GenotypicData= 1
    LocusSeparator= WHITESPACE
    GameticPhase= 0
    RecessiveData= 0
    RecessiveAllele= null
    MissingData= '?'
# Some advenced settings the experienced user can unkomment
#     Frequency= ABS            # - {ABS, REL}
#     CompDistMatrix= 0         # - {0, 1}
#     FrequencyThreshold= 1.0e-5 # - (Any real number between 1.0e-7 and 1.e-2)
#     EpsilonValue= 1.0e-7      # - (Any real number between 1.0e-12 and 1.0e-7)

[Data]

[[Samples]]

        SampleName="Random Namaqualand Population"
        SampleSize= 48
        SampleData= {
 #Sample of genotypic data (4 loci):
        LP1     1       17 12 19 24
                        20 16 25 24
        LP3     1       15 19 19 21
                        17 26 26 21
        LP4     1       17 21 19 20
                        17 21 21 22
        LP8     1       21 22 20 20
                        21 22 26 22
        LP9     1       19 20 24 20
                        22 20 27 21
        LP10    1       16 12 21 20
                        20 20 24 22
        LP11    1       17 12 19 18
                        17 19 23 18
        LP12    1       17 12 23 20
                        19 22 25 22
        LP13    1       17 12 19 18
```

		23 19 23 19
LP14	1	17 19 20 19
		17 19 32 21
LP15	1	20 12 19 20
		21 19 21 22
LP17	1	17 12 23 16
		19 17 23 16
LP18	1	15 16 21 19
		15 23 24 22
LP19	1	18 17 19 15
		20 18 27 18
LP20	1	18 17 22 19
		20 20 23 21
LP21	1	15 19 19 18
		18 21 21 20
LP22	1	16 18 23 19
		19 19 27 21
LP23	1	18 18 19 19
		19 19 25 20
LP24	1	17 18 19 20
		18 20 24 21
LP25	1	15 19 19 20
		17 26 26 22
LP26	1	18 17 24 13
		18 22 25 22
LP27	1	21 18 19 16
		21 20 21 19
LP28	1	18 18 21 20
		23 23 23 20
LP29	1	17 18 19 18
		20 22 20 21
LP30	1	17 19 18 18
		19 22 21 17
LP31	1	20 17 27 17
		20 18 27 19
LP32	1	19 18 19 16
		20 22 27 19
LP33	1	17 19 20 19
		21 22 23 21
LP34	1	17 22 23 13
		18 17 23 20
LP35	1	17 18 24 19
		18 21 26 22
LP36	1	17 18 19 19
		18 20 25 21
LP37	1	15 21 19 16

```
                    17 24 27 18
        LP38   1    18 18 23 18
                    18 18 23 21
        LP39   1    18 18 18 19
                    21 20 23 19
        LP40   1    17 18 18 20
                    17 19 23 22
        LP41   1    17 18 18 19
                    21 19 25 22
        LP42   1    17 12 21 18
                    20 12 23 22
        LP43   1    18 18 21 20
                    23 19 23 21
        LP44   1    18 12 21 19
                    20 22 26 19
        LP45   1    17 17 18 16
                    18 18 21 18
        LP46   1    17 12 23 22
                    19 19 23 23
        LP47   1    15 20 23 12
                    15 22 26 16
        LP48   1    19 17 23 18
                    21 19 28 20
        LP49   1    17 18 18 19
                    21 22 25 21
        LP50   1    17 12 18 17
                    17 20 19 20
        LP51   1    17 17 23 19
                    20 20 26 23
        LP52   1    20 18 24 19
                    21 21 26 19
        LP53   1    17 20 18 21
                    24 17 22 21
}
```

5.6.2 Random White population genotypic LD estimation and haplotype estimation project file

```
 [Profile]

Title="Random White Population Haplotype Estimation"
    NbSamples= 1        #Number of sample in the Project.
    DataType= MICROSAT
    GenotypicData= 1
    LocusSeparator= WHITESPACE
    GameticPhase= 0
```

```
    RecessiveData= 0
    RecessiveAllele= null
    MissingData= '?'
# Some advenced settings the experienced user can unkomment
#     Frequency= ABS            # - {ABS, REL}
#     CompDistMatrix= 0         # - {0, 1}
#     FrequencyThreshold= 1.0e-5 # - (Any real number between 1.0e-7 and 1.e-2)
#     EpsilonValue= 1.0e-7      # - (Any real number between 1.0e-12 and 1.0e-7)

[Data]

[[Samples]]

        SampleName="Random White Population"
        SampleSize= 30
        SampleData= {
 #Sample of genotypic data (4 loci):
    RW251    1        18 12 18 18
                      18 12 19 20
    RW252    1        17 12 18 19
                      22 20 23 19
    RW253    1        16 12 18 20
                      17 19 21 22
    RW254    1        17 19 19 18
                      20 20 23 20
    RW255    1        17 12 21 20
                      22 17 23 22
    RW256    1        18 19 18 16
                      23 24 23 16
    RW257    1        17 12 18 19
                      22 21 19 20
    RW258    1        17 19 23 18
                      22 19 23 20
    RW259    1        18 20 18 20
                      18 21 19 22
    RW260    1        23 20 2  20
                      23 25 26 20
    RW261    1        16 19 18 18
                      17 20 24 21
    RW262    1        17 12 19 18
                      17 19 21 22
    RW263    1        17 12 18 18
                      21 19 19 21
    RW264    1        17 20 18 19
                      18 22 23 19
    RW265    1        17 12 18 16
```

```
                    20 19 23 18
    RW266    1      15 17 18 18
                    17 19 22 20
    RW267    1      18 18 23 17
                    18 19 23 19
    RW268    1      17 12 21 16
                    17 19 24 16
    RW269    1      18 19 18 16
                    22 20 21 18
    RW270    1      15 19 21 19
                    18 19 23 19
    RW271    1      20 12 18 19
                    22 19 23 21
    RW272    1      17 12 21 20
                    22 20 23 22
    RW273    1      17 12 18 20
                    17 21 19 22
    RW274    1      15 21 19 16
                    18 23 21 16
    RW275    1      18 12 18 20
                    18 20 21 22
    RW276    1      17 19 18 18
                    18 21 21 20
    RW277    1      18 19 22 16
                    21 19 25 18
    RW278    1      15 12 19 18
                    17 19 21 20
    RW279    1      17 17 18 19
                    17 19 23 20
    RW280    1      17 12 19 20
                    17 19 21 22
}
```

5.6.3 Random Coloured population haplotypic LD estimation project file

```
[Profile]

Title="Linkage disequilibrium estimation for the random Coloured Namaqualand population"
    NbSamples= 1        #Number of sample in the Project.
    DataType= MICROSAT
    GenotypicData= 0
    LocusSeparator= WHITESPACE
    GameticPhase= 0
    RecessiveData= 0
    RecessiveAllele= null
```

```
    MissingData= '?'
    Frequency= REL
#    CompDistMatrix= 0         # - {0, 1}
#    FrequencyThreshold= 1.0e-5 # - (Any real number between 1.0e-7 and 1.e-2)
#    EpsilonValue= 1.0e-7      # - (Any real number between 1.0e-12 and 1.0e-7)

[Data]

[[Samples]]

        SampleName="Random Namaqualand Estimated Haplotypes"
        SampleSize= 81
        SampleData= {
RC1     0.01042         15 16 21 19
RC2     0.01042         15 19 19 21
RC3     0.01042         15 20 23 12
RC4     0.02083         15 21 19 18
RC5     0.01042         15 22 26 16
RC6     0.01042         15 23 24 22
RC7     0.01042         15 26 19 20
RC8     0.01042         16 18 23 19
RC9     0.01042         16 20 24 20
RC10    0.01042         17 12 19 17
RC11    0.01042         17 12 23 16
RC12    0.03125         17 12 23 18
RC13    0.01042         17 16 19 24
RC14    0.01042         17 17 18 21
RC15    0.01042         17 18 18 18
RC16    0.01042         17 18 18 20
RC17    0.03125         17 18 19 21
RC18    0.01042         17 18 24 19
RC19    0.01042         17 19 19 18
RC20    0.02083         17 19 20 21
RC21    0.02083         17 19 23 22
RC22    0.01042         17 19 25 22
RC23    0.01042         17 19 26 22
RC24    0.01042         17 19 32 19
RC25    0.01042         17 20 18 20
RC26    0.01042         17 20 26 19
RC27    0.01042         17 21 19 20
RC28    0.01042         17 21 21 22
RC29    0.01042         17 22 21 17
RC30    0.02083         17 22 23 20
RC31    0.01042         17 22 25 21
RC32    0.01042         17 24 27 16
RC33    0.01042         17 26 26 21
```

RC34	0.01042	18 17 21 16
RC35	0.01042	18 17 23 13
RC36	0.01042	18 17 25 22
RC37	0.01042	18 17 27 18
RC38	0.01042	18 18 23 18
RC39	0.01042	18 18 23 21
RC40	0.01042	18 19 21 20
RC41	0.01042	18 19 23 21
RC42	0.01042	18 19 25 20
RC43	0.02083	18 20 23 19
RC44	0.01042	18 20 24 20
RC45	0.01042	18 20 25 19
RC46	0.01042	18 21 26 22
RC47	0.01042	18 22 21 19
RC48	0.01042	18 22 24 13
RC49	0.01042	18 23 23 20
RC50	0.01042	19 12 23 23
RC51	0.01042	19 12 25 22
RC52	0.01042	19 17 23 16
RC53	0.01042	19 17 23 18
RC54	0.02083	19 18 19 19
RC55	0.01042	19 19 18 18
RC56	0.01042	19 19 27 21
RC57	0.01042	19 20 27 21
RC58	0.03125	20 12 21 22
RC59	0.01042	20 12 25 24
RC60	0.01042	20 12 26 19
RC61	0.01042	20 17 22 21
RC62	0.01042	20 17 23 23
RC63	0.01042	20 17 27 17
RC64	0.01042	20 18 19 15
RC65	0.01042	20 18 27 19
RC66	0.01042	20 21 26 19
RC67	0.01042	20 22 20 18
RC68	0.01042	20 22 27 16
RC69	0.03125	21 18 18 19
RC70	0.01042	21 18 19 19
RC71	0.01042	21 18 24 19
RC72	0.01042	21 19 19 20
RC73	0.01042	21 19 28 20
RC74	0.01042	21 20 21 16
RC75	0.01042	21 22 20 20
RC76	0.01042	21 22 23 19
RC77	0.01042	21 22 26 22
RC78	0.01042	22 20 24 20
RC79	0.02083	23 18 21 20

```
RC80  0.01042        23 19 19 19
RC81  0.01042        24 20 22 21
 }
```

5.6.4 Random White population haplotypic LD estimation project file

```
[Profile]

Title="Linkage disequilibrium estimation using random White haplotypes"
    NbSamples= 1       #Number of sample in the Project.
    DataType= MICROSAT
    GenotypicData= 0
    LocusSeparator= WHITESPACE
    GameticPhase= 0
    RecessiveData= 0
    RecessiveAllele= null
    MissingData= '?'
    Frequency= REL
#    CompDistMatrix= 0        # - {0, 1}
#    FrequencyThreshold= 1.0e-5 # - (Any real number between 1.0e-7 and 1.e-2)
#    EpsilonValue= 1.0e-7      # - (Any real number between 1.0e-12 and 1.0e-7)

[Data]

[[Samples]]

        SampleName="Random White estimated haplotypes"
        SampleSize= 43
        SampleData= {
        RW1   0.01667        15 12 21 18
        RW2   0.01667        15 17 22 20
        RW3   0.01667        15 19 21 19
        RW4   0.01667        15 23 21 16
        RW5   0.01667        16 19 18 20
        RW6   0.01667        16 20 24 21
        RW7   0.01667        17 12 18 22
        RW8   0.08333        17 12 21 22
        RW9   0.01667        17 12 24 16
        RW10  0.01667        17 17 18 19
        RW11  0.08333        17 19 18 18
        RW12  0.01667        17 19 19 18
        RW13  0.03333        17 19 19 20
        RW14  0.01667        17 19 21 16
        RW15  0.05000        17 19 23 20
        RW16  0.03333        17 20 23 19
        RW17  0.03333        17 21 19 20
```

```
        RW18 0.03333        18 12 18 20
        RW19 0.01667        18 12 19 18
        RW20 0.01667        18 18 23 17
        RW21 0.01667        18 19 23 16
        RW22 0.03333        18 19 23 19
        RW23 0.01667        18 19 25 16
        RW24 0.01667        18 20 18 16
        RW25 0.01667        18 20 19 22
        RW26 0.01667        18 20 21 22
        RW27 0.01667        18 21 18 20
        RW28 0.01667        18 21 19 16
        RW29 0.01667        18 21 21 20
        RW30 0.01667        18 22 18 19
        RW31 0.01667        20 12 23 16
        RW32 0.01667        20 19 23 21
        RW33 0.01667        20 20 19 18
        RW34 0.01667        21 12 19 21
        RW35 0.01667        21 19 22 18
        RW36 0.05000        22 12 18 19
        RW37 0.01667        22 17 23 20
        RW38 0.01667        22 19 21 18
        RW39 0.01667        22 19 23 18
        RW40 0.01667        22 20 23 20
        RW41 0.01667        23 20 2 20
        RW42 0.01667        23 24 18 16
        RW43 0.01667        23 25 26 20
}
```

5.6.5 Coloured LP haplotypic LD project file

```
[Profile]

Title="Namaqualand LP Linkage disequilibrium estimate"
    NbSamples= 1       #Number of sample in the Project.
    DataType= MICROSAT
    GenotypicData= 1
    LocusSeparator= WHITESPACE
    GameticPhase= 1
    RecessiveData= 0
    RecessiveAllele= null
    MissingData= '?'
# Some advenced settings the experienced user can unkomment
#    Frequency= ABS           # - {ABS, REL}
#    CompDistMatrix= 0        # - {0, 1}
#    FrequencyThreshold= 1.0e-5 # - (Any real number between 1.0e-7 and 1.e-2)
#    EpsilonValue= 1.0e-7     # - (Any real number between 1.0e-12 and 1.0e-7)
```

```
[Data]

[[Samples]]

        SampleName="Namaqualand LP Haplotypes"
        SampleSize= 48
        SampleData= {
LP1     34      19 12 23 22
LP2     6       19 12 23 23
LP3     2       19 12 23 20
LP4     2       19 12 18 22
LP5     1       19 12 23 18
LP6     1       19 12 19 22
LP7     1       19 12 26 23
LP8     1       14 12 23 22
}
```

5.6.6 White LP haplotypic LD project file

```
[Profile]

Title="White LP Linkage disequilibrium Estimation"
    NbSamples= 1        #Number of sample in the Project.
    DataType= MICROSAT
    GenotypicData= 1
    LocusSeparator= WHITESPACE
    GameticPhase= 1
    RecessiveData= 0
    RecessiveAllele= null
    MissingData= '?'
# Some advenced settings the experienced user can unkomment
#     Frequency= ABS            # - {ABS, REL}
#     CompDistMatrix= 0         # - {0, 1}
#     FrequencyThreshold= 1.0e-5 # - (Any real number between 1.0e-7 and 1.e-2)
#     EpsilonValue= 1.0e-7      # - (Any real number between 1.0e-12 and 1.0e-7)

[Data]

[[Samples]]

        SampleName="White LP Population"
        SampleSize= 4
        SampleData= {
LP1     8       19 12 23 22
LP9     2       17 12 23 18
```

LP10 1 15 12 23 20
LP11 1 22 12 23 22
}

5.7 Arlequin result files

5.7.1 Random Coloured population genotypic LD result file

//
RUN NUMBER 1 (08/01/1904 at 15:39:39))
//

Project information:

NbSamples = 1
DataType = MICROSAT
GenotypicData = 1
GameticPhase = 0
RecessiveData = 0

Sample: Random Namaqualand Population

Pairwise linkage disequilibrium: (Random Namaqualand Population)

Test of linkage disequilibrium for all pairs of loci:

Permutation test using the EM algorithm
Number of permutations: 100000
Number of initial conditions for EM: 10

Pair(0, 1) : Exact P= 0.0255261 +- 0.000479137 (100172 permutations done)
Chi-square test value=89.5712 (P = 1, 819 d.f.)
Pair(0, 2) : Exact P= 0.0707982 +- 0.000819767 (100172 permutations done)
Chi-square test value=91.5793 (P = 1, 783 d.f.)
Pair(1, 2) : Exact P= 0.118516 +- 0.00103656 (100172 permutations done)
Chi-square test value=100.527 (P = 1, 890 d.f.)
Pair(0, 3) : Exact P= 0.918042 +- 0.000833909 (100172 permutations done)
Chi-square test value=64.5359 (P = 1, 819 d.f.)
Pair(1, 3) : Exact P= 0.675538 +- 0.00139915 (100172 permutations done)
Chi-square test value=81.5169 (P = 1, 950 d.f.)
Pair(2, 3) : Exact P= 0.558679 +- 0.0016182 (100172 permutations done)
Chi-square test value=91.3574 (P = 1, 1023 d.f.)

Histogram of the number of linked loci per locus

```
---------------------------------------------------------------------
Locus:    0   1   2   3
---------------------------------------------------------------------
          1   1   0   0
---------------------------------------------------------------------
```

Table of significant linkage disequilibrium (significance level=0.05):

```
       ----------------
Locus  # |  0|  1|  2|  3|
       ----------------
     0 |  *   +   -   -
     1 |  +   *   -   -
     2 |  -   -   *   -
     3 |  -   -   -   *
```

//
END OF RUN NUMBER 1 (08/01/1904 at 20:49:26))
Total ellapsed time for this run : 5h 9m 47s 257 ms
//

5.7.2 Random Coloured population haplotype estimation result file

//
RUN NUMBER 1 (08/01/1904 at 12:00:04))
//

Project information:

```
--------------------
```

NbSamples = 1
DataType = MICROSAT
GenotypicData = 1
GameticPhase = 0
RecessiveData = 0

Sample: Random Namaqualand Population

Haplotype frequencies estimation: (Random Namaqualand Population)

No. of gene copies in sample: 96
No. of random initial conditions for EM: 50
No. of different maximum likelihoods found: 5

Epsilon value for stopping iterations: 1.000000e-07
Logarithm of the sample maximum-likelihood: -382.018951
Standard deviations not computed

Maximum-likelihood haplotype frequencies:

Total number of possible haplotypes: 495
Minimum frequency to reach for output: 1.000e-05

#	Haplotype	Freq.	s.d.	
1	UNKNOWN	0.01042	0.00000	15 16 21 19
2	UNKNOWN	0.01042	0.00000	15 19 19 21
3	UNKNOWN	0.01042	0.00000	15 20 23 12
4	UNKNOWN	0.02083	0.00000	15 21 19 18
5	UNKNOWN	0.01042	0.00000	15 22 26 16
6	UNKNOWN	0.01042	0.00000	15 23 24 22
7	UNKNOWN	0.01042	0.00000	15 26 19 20
8	UNKNOWN	0.01042	0.00000	16 18 23 19
9	UNKNOWN	0.01042	0.00000	16 20 24 20
10	UNKNOWN	0.01042	0.00000	17 12 19 17
11	UNKNOWN	0.01042	0.00000	17 12 23 16
12	UNKNOWN	0.03125	0.00000	17 12 23 18
13	UNKNOWN	0.01042	0.00000	17 16 19 24
14	UNKNOWN	0.01042	0.00000	17 17 18 21
15	UNKNOWN	0.01042	0.00000	17 18 18 18
16	UNKNOWN	0.01042	0.00000	17 18 18 20
17	UNKNOWN	0.03125	0.00000	17 18 19 21
18	UNKNOWN	0.01042	0.00000	17 18 24 19
19	UNKNOWN	0.01042	0.00000	17 19 19 18
20	UNKNOWN	0.02083	0.00000	17 19 20 21
21	UNKNOWN	0.02083	0.00000	17 19 23 22
22	UNKNOWN	0.01042	0.00000	17 19 25 22
23	UNKNOWN	0.01042	0.00000	17 19 26 22
24	UNKNOWN	0.01042	0.00000	17 19 32 19
25	UNKNOWN	0.01042	0.00000	17 20 18 20
26	UNKNOWN	0.01042	0.00000	17 20 26 19
27	UNKNOWN	0.01042	0.00000	17 21 19 20
28	UNKNOWN	0.01042	0.00000	17 21 21 22
29	UNKNOWN	0.01042	0.00000	17 22 21 17
30	UNKNOWN	0.02083	0.00000	17 22 23 20
31	UNKNOWN	0.01042	0.00000	17 22 25 21
32	UNKNOWN	0.01042	0.00000	17 24 27 16
33	UNKNOWN	0.01042	0.00000	17 26 26 21
34	UNKNOWN	0.01042	0.00000	18 17 21 16
35	UNKNOWN	0.01042	0.00000	18 17 23 13
36	UNKNOWN	0.01042	0.00000	18 17 25 22

37	UNKNOWN	0.01042	0.00000	18 17 27 18
38	UNKNOWN	0.01042	0.00000	18 18 23 18
39	UNKNOWN	0.01042	0.00000	18 18 23 21
40	UNKNOWN	0.01042	0.00000	18 19 21 20
41	UNKNOWN	0.01042	0.00000	18 19 23 21
42	UNKNOWN	0.01042	0.00000	18 19 25 20
43	UNKNOWN	0.02083	0.00000	18 20 23 19
44	UNKNOWN	0.01042	0.00000	18 20 24 20
45	UNKNOWN	0.01042	0.00000	18 20 25 19
46	UNKNOWN	0.01042	0.00000	18 21 26 22
47	UNKNOWN	0.01042	0.00000	18 22 21 19
48	UNKNOWN	0.01042	0.00000	18 22 24 13
49	UNKNOWN	0.01042	0.00000	18 23 23 20
50	UNKNOWN	0.01042	0.00000	19 12 23 23
51	UNKNOWN	0.01042	0.00000	19 12 25 22
52	UNKNOWN	0.01042	0.00000	19 17 23 16
53	UNKNOWN	0.01042	0.00000	19 17 23 18
54	UNKNOWN	0.02083	0.00000	19 18 19 19
55	UNKNOWN	0.01042	0.00000	19 19 18 18
56	UNKNOWN	0.01042	0.00000	19 19 27 21
57	UNKNOWN	0.01042	0.00000	19 20 27 21
58	UNKNOWN	0.03125	0.00000	20 12 21 22
59	UNKNOWN	0.01042	0.00000	20 12 25 24
60	UNKNOWN	0.01042	0.00000	20 12 26 19
61	UNKNOWN	0.01042	0.00000	20 17 22 21
62	UNKNOWN	0.01042	0.00000	20 17 23 23
63	UNKNOWN	0.01042	0.00000	20 17 27 17
64	UNKNOWN	0.01042	0.00000	20 18 19 15
65	UNKNOWN	0.01042	0.00000	20 18 27 19
66	UNKNOWN	0.01042	0.00000	20 21 26 19
67	UNKNOWN	0.01042	0.00000	20 22 20 18
68	UNKNOWN	0.01042	0.00000	20 22 27 16
69	UNKNOWN	0.03125	0.00000	21 18 18 19
70	UNKNOWN	0.01042	0.00000	21 18 19 19
71	UNKNOWN	0.01042	0.00000	21 18 24 19
72	UNKNOWN	0.01042	0.00000	21 19 19 20
73	UNKNOWN	0.01042	0.00000	21 19 28 20
74	UNKNOWN	0.01042	0.00000	21 20 21 16
75	UNKNOWN	0.01042	0.00000	21 22 20 20
76	UNKNOWN	0.01042	0.00000	21 22 23 19
77	UNKNOWN	0.01042	0.00000	21 22 26 22
78	UNKNOWN	0.01042	0.00000	22 20 24 20
79	UNKNOWN	0.02083	0.00000	23 18 21 20
80	UNKNOWN	0.01042	0.00000	23 19 19 19
81	UNKNOWN	0.01042	0.00000	24 20 22 21

Sum of all 495 haplotype frequencies: 1.00000
Sum of 81 listed frequencies: 1.00000

//
END OF RUN NUMBER 1 (08/01/1904 at 12:00:04))
Total ellapsed time for this run : 0h 0m 0s 361 ms
//

5.7.3 Random White population genotypic LD result file

//
RUN NUMBER 1 (09/01/1904 at 12:54:25))
//

Project information:

NbSamples = 1
DataType = MICROSAT
GenotypicData = 1
GameticPhase = 0
RecessiveData = 0

Sample: Random White Population
Pairwise linkage disequilibrium: (Random White Population)
Test of linkage disequilibrium for all pairs of loci:

Permutation test using the EM algorithm
Number of permutations: 100000
Number of initial conditions for EM: 10

Pair(0, 1) : Exact P= 0.534011 +- 0.00153022 (100172 permutations done)
Chi-square test value=43.9476 (P = 1, 371 d.f.)
Pair(0, 2) : Exact P= 0.159086 +- 0.000991606 (100172 permutations done)
Chi-square test value=47.3711 (P = 1, 371 d.f.)
Pair(1, 2) : Exact P= 0.0911332 +- 0.000865628 (100172 permutations done)
Chi-square test value=52.4577 (P = 1, 459 d.f.)
Pair(0, 3) : Exact P= 0.739618 +- 0.00130345 (100172 permutations done)
Chi-square test value=35.2883 (P = 1, 357 d.f.)
Pair(1, 3) : Exact P= 0.15686 +- 0.00101697 (100172 permutations done)
Chi-square test value=48.032 (P = 1, 495 d.f.)
Pair(2, 3) : Exact P= 0.285958 +- 0.00140217 (100172 permutations done)
Chi-square test value=43.6017 (P = 1, 408 d.f.)

Histogram of the number of linked loci per locus

```
------------------------------------------------------------------------
Locus:    0   1   2   3
------------------------------------------------------------------------
          0   0   0   0
------------------------------------------------------------------------
```

Table of significant linkage disequilibrium (significance level=0.05):

```
       ----------------
Locus  # |  0|  1|  2|  3|
       ----------------
     0 | *   -   -   -
     1 | -   *   -   -
     2 | -   -   *   -
     3 | -   -   -   *
```

//
END OF RUN NUMBER 1 (09/01/1904 at 15:21:41))
Total ellapsed time for this run : 2h 27m 16s 35 ms
//

5.7.4 Random White population haplotype estimation result file

//
RUN NUMBER 1 (08/01/1904 at 11:57:59))
//

Project information:

```
--------------------
        NbSamples = 1
        DataType = MICROSAT
        GenotypicData = 1
        GameticPhase = 0
        RecessiveData = 0
```

Sample: Random White Population
Haplotype frequencies estimation: (Random White Population)

No. of gene copies in sample: 60
No. of random initial conditions for EM: 50
No. of different maximum likelihoods found: 2
Epsilon value for stopping iterations: 1.000000e-07
Logarithm of the sample maximum-likelihood: -204.682220

Standard deviations not computed

Maximum-likelihood haplotype frequencies:

Total number of possible haplotypes: 248
Minimum frequency to reach for output: 1.000e-05

#	Haplotype	Freq.	s.d.	
1	UNKNOWN	0.01667	0.00000	15 12 21 18
2	UNKNOWN	0.01667	0.00000	15 17 22 20
3	UNKNOWN	0.01667	0.00000	15 19 21 19
4	UNKNOWN	0.01667	0.00000	15 23 21 16
5	UNKNOWN	0.01667	0.00000	16 19 18 20
6	UNKNOWN	0.01667	0.00000	16 20 24 21
7	UNKNOWN	0.01667	0.00000	17 12 18 22
8	UNKNOWN	0.08333	0.00000	17 12 21 22
9	UNKNOWN	0.01667	0.00000	17 12 24 16
10	UNKNOWN	0.01667	0.00000	17 17 18 19
11	UNKNOWN	0.08333	0.00000	17 19 18 18
12	UNKNOWN	0.01667	0.00000	17 19 19 18
13	UNKNOWN	0.03333	0.00000	17 19 19 20
14	UNKNOWN	0.01667	0.00000	17 19 21 16
15	UNKNOWN	0.05000	0.00000	17 19 23 20
16	UNKNOWN	0.03333	0.00000	17 20 23 19
17	UNKNOWN	0.03333	0.00000	17 21 19 20
18	UNKNOWN	0.03333	0.00000	18 12 18 20
19	UNKNOWN	0.01667	0.00000	18 12 19 18
20	UNKNOWN	0.01667	0.00000	18 18 23 17
21	UNKNOWN	0.01667	0.00000	18 19 23 16
22	UNKNOWN	0.03333	0.00000	18 19 23 19
23	UNKNOWN	0.01667	0.00000	18 19 25 16
24	UNKNOWN	0.01667	0.00000	18 20 18 16
25	UNKNOWN	0.01667	0.00000	18 20 19 22
26	UNKNOWN	0.01667	0.00000	18 20 21 22
27	UNKNOWN	0.01667	0.00000	18 21 18 20
28	UNKNOWN	0.01667	0.00000	18 21 19 16
29	UNKNOWN	0.01667	0.00000	18 21 21 20
30	UNKNOWN	0.01667	0.00000	18 22 18 19
31	UNKNOWN	0.01667	0.00000	20 12 23 16
32	UNKNOWN	0.01667	0.00000	20 19 23 21
33	UNKNOWN	0.01667	0.00000	20 20 19 18
34	UNKNOWN	0.01667	0.00000	21 12 19 21
35	UNKNOWN	0.01667	0.00000	21 19 22 18
36	UNKNOWN	0.05000	0.00000	22 12 18 19

37	UNKNOWN	0.01667	0.00000	22 17 23 20
38	UNKNOWN	0.01667	0.00000	22 19 21 18
39	UNKNOWN	0.01667	0.00000	22 19 23 18
40	UNKNOWN	0.01667	0.00000	22 20 23 20
41	UNKNOWN	0.01667	0.00000	23 20 2 20
42	UNKNOWN	0.01667	0.00000	23 24 18 16
43	UNKNOWN	0.01667	0.00000	23 25 26 20

Sum of all 248 haplotype frequencies: 1.00000
Sum of 43 listed frequencies: 1.00000

//
END OF RUN NUMBER 1 (08/01/1904 at 11:57:59))
Total ellapsed time for this run : 0h 0m 0s 270 ms
//

5.7.5 Truncated random Coloured haplotypic LD estimation result file

//
RUN NUMBER 1 (09/01/1904 at 12:38:42))
//

Project information:

NbSamples = 1
DataType = MICROSAT
GenotypicData = 0

Sample: Random Namaqualand Estimated Haplotypes
Pairwise linkage disequilibrium: (Random Namaqualand Estimated Haplotypes)

Test of linkage disequilibrium for all pairs of loci:

Exact test using a Markov chain:
Chain length: 100000
Dememorization: 1000
Note: The test is only performed for polymorphic pairs of loci

Pair (0, 1) Exact P=0.105419 +- 0.00093619 (100172 Steps done)
Pair (0, 2) Exact P=0.0943877 +- 0.000474807 (100172 Steps done)
Pair (1, 2) Exact P=0.0638602 +- 0.000483281 (100172 Steps done)
Pair (0, 3) Exact P=0.646069 +- 0.00080982 (100172 Steps done)

Pair (1, 3) Exact P=0.00175698 +- 9.79536e-05 (100172 Steps done)
Pair (2, 3) Exact P=0.272012 +- 0.00125002 (100172 Steps done)

Table of significant linkage disequilibrium (significance level=0.05):

```
        ----------------
Locus  # |  0|  1|  2|  3|
        ----------------
      0 |  *   -   -   -
      1 |  -   *   -   +
      2 |  -   -   *   -
      3 |  -   +   -   *
```

Number of linked loci per polymorphic locus (significance level=0.05):

y\Locus:	0	1	2	3	No. of Loci
1	0	1	0	1	4
2	0	1	0	1	4
3	0	1	0	1	4
4	0	1	0	1	4
5	0	1	0	1	4

//
END OF RUN NUMBER 1 (09/01/1904 at 12:38:43))
Total ellapsed time for this run : 0h 0m 0s 580 ms
//

5.7.6 Truncated random White population haplotypic LD estimation result file

//
RUN NUMBER 1 (09/01/1904 at 12:46:07))
//

Project information:

```
--------------------
        NbSamples  = 1
        DataType = MICROSAT
        GenotypicData = 0
```

Sample: Random White estimated haplotypes
Pairwise linkage disequilibrium: (Random White estimated haplotypes)

Test of linkage disequilibrium for all pairs of loci:

Exact test using a Markov chain:
Chain length: 100000
Dememorization: 1000
Note: The test is only performed for polymorphic pairs of loci

Pair (0, 1) Exact P=0.405622 +- 0.00133903 (100172 Steps done)
Pair (0, 2) Exact P=0.0662061 +- 0.000692723 (100172 Steps done)
Pair (1, 2) Exact P=0.541089 +- 0.00103682 (100172 Steps done)
Pair (0, 3) Exact P=0.498972 +- 0.0015881 (100172 Steps done)
Pair (1, 3) Exact P=0.0693308 +- 0.000531422 (100172 Steps done)
Pair (2, 3) Exact P=0.41256 +- 0.00168096 (100172 Steps done)

Table of significant linkage disequilibrium (significance level=0.05):

```
        ----------------
Locus  # |  0|  1|  2|  3|
        ----------------
      0 |  *   -   -   -
      1 |  -   *   -   -
      2 |  -   -   *   -
      3 |  -   -   -   *
```

Number of linked loci per polymorphic locus (significance level=0.05):

y\Locus:	0	1	2	3	No. of Loci
1	0	0	0	0	4
2	0	0	0	0	4
3	0	0	0	0	4
4	0	0	0	0	4
5	0	0	0	0	4

//
END OF RUN NUMBER 1 (09/01/1904 at 12:46:07))
Total ellapsed time for this run : 0h 0m 0s 561 ms

//

5.7.7 Truncated Coloured LP population haplotypic LD estimation results file

//
RUN NUMBER 1 (08/01/1904 at 12:50:25))
//

Project information:

NbSamples = 1
DataType = MICROSAT
GenotypicData = 1
GameticPhase = 1
RecessiveData = 0

Sample: Namaqualand LP Haplotypes
Pairwise linkage disequilibrium: (Namaqualand LP Haplotypes)

Test of linkage disequilibrium for all pairs of loci:

Exact test using a Markov chain:
Chain length: 100000
Dememorization: 1000
Note: The test is only performed for polymorphic pairs of loci

Pair (0, 1) Exact P=0 +- 0 (100172 Steps done)
Pair (0, 2) Exact P=0 +- 0 (100172 Steps done)
Pair (1, 2) Exact P=0 +- 0 (100172 Steps done)
Pair (0, 3) Exact P=0 +- 0 (100172 Steps done)
Pair (1, 3) Exact P=0 +- 0 (100172 Steps done)
Pair (2, 3) Exact P=0 +- 0 (100172 Steps done)

Table of significant linkage disequilibrium (significance level=0.05):

Locus #	0	1	2	3
0	*	+	+	+
1	+	*	+	+
2	+	+	*	+
3	+	+	+	*

Number of linked loci per polymorphic locus (significance level=0.05):

y\Locus:	0	1	2	3	No. of Loci
1	3	3	3	3	4
2	3	3	3	3	4
3	3	3	3	3	4
4	3	3	3	3	4
5	3	3	3	3	4

//
END OF RUN NUMBER 1 (08/01/1904 at 12:50:26))
Total ellapsed time for this run : 0h 0m 0s 591 ms
//

5.7.8 Truncated White LP population haplotypic LD estimation results file

//
RUN NUMBER 1 (08/01/1904 at 12:54:42))
//

Project information:

NbSamples = 1
DataType = MICROSAT
GenotypicData = 1
GameticPhase = 1
RecessiveData = 0

Sample: White LP Population
Pairwise linkage disequilibrium: (White LP Population)

Test of linkage disequilibrium for all pairs of loci:

Exact test using a Markov chain:
Chain length: 100000
Dememorization: 1000
Note: The test is only performed for polymorphic pairs of loci

Pair (0, 1) Exact P=0 +- 0 (100172 Steps done)
Pair (0, 2) Exact P=0 +- 0 (100172 Steps done)
Pair (1, 2) Exact P=0 +- 0 (100172 Steps done)
Pair (0, 3) Exact P=0 +- 0 (100172 Steps done)
Pair (1, 3) Exact P=0 +- 0 (100172 Steps done)
Pair (2, 3) Exact P=0 +- 0 (100172 Steps done)

Table of significant linkage disequilibrium (significance level=0.05):

```
        ----------------
Locus  # |  0|  1|  2|  3|
        ----------------
      0 |  *   +   +   +
      1 |  +   *   +   +
      2 |  +   +   *   +
      3 |  +   +   +   *
```

Number of linked loci per polymorphic locus (significance level=0.05):

y\Locus:	0	1	2	3	No. of Loci
1	3	3	3	3	4
2	3	3	3	3	4
3	3	3	3	3	4
4	3	3	3	3	4
5	3	3	3	3	4

//
END OF RUN NUMBER 1 (08/01/1904 at 12:54:43))
Total ellapsed time for this run : 0h 0m 0s 621 ms
//

5.8 dHPLC Results

Table 5-2: Random Black South African dHPLC results

Exon 1	Exon 2	Exon 3	Exon 4	Exon 5	Exon 6pI	Exon 6pII	Exon 7pI	Exon 7pII	Exon 8pI	Exon 8pII	Exon 9	Exon 10
Y	N	N	N	Y	N	N	N	Y	N	N	N	N
Y	N	N	N	N	Y	N	N	Y	N	Y	Y	N
N	N	N	Y	Y	Y	N	N	Y	N	N	N	N
N	N	N	N	N	Y	N	Y	Y	N	N	N	N
N	Y	N	N	N	N	N	Y	Y	N	Y	N	N
Y	N	N	N	N	N	Y	N	N	N	N	Y	N
N	N	N	N	N	Y	N	N	Y	N	Y	N	N
Y	N	N	N	N	N	N	N	Y	N	Y	N	N
Y	N	N	N	N	N	N	N	N	N	N	N	N
N	N	N	N	N	N	N	Y	N	N	N	Y	N
N	N	N	Y	N	N	N	Y	Y	N	N	N	N
N	N	N	N	N	Y	N	N	Y	N	N	N	N
N	N	N	N	N	N	N	N	N	N	N	N	N
N	N	N	N	N	N	N	Y	Y	N	Y	Y	N
Y	N	N	N	N	N	N	N	Y	N	N	N	N
N	N	N	N	N	Y	Y	Y	Y	N	N	N	N
Y	N	N	N	N	N	Y	N	N	N	Y	N	N
N	N	N	N	Y	Y	Y	Y	N	N	Y	N	N
N	N	N	N	Y	N	Y	N	N	N	Y	N	N
Y	N	N	N	N	N	Y	N	N	N	Y	N	N

Y indicates that a dHPLC shift was seen in a sample, while N indicates that there was no dHPLC shift seen.

Table 5-3: Random Coloured Namaqualand dHPLC results

Exon 1	Exon 2	Exon 3	Exon 4	Exon 5	Exon 6pI	Exon 6pII	Exon 7pI	Exon 7pII	Exon 8pI	Exon 8pII	Exon 9	Exon 10
N	Y	N	N	Y	N	N	N	Y	Y	N	N	N
N	N	N	N	N	Y	N	Y	Y	N	N	N	N
N	N	N	N	N	Y	N	Y	Y	N	N	N	N
N	N	N	N	N	N	Y	N	Y	N	N	N	N
N	Y	N	N	Y	N	N	Y	N	N	Y	Y	N
N	N	N	N	N	Y	N	Y	Y	N	N	N	N
Y	N	N	N	N	N	N	N	N	N	N	N	N
Y	N	N	N	N	N	N	N	N	N	N	N	N
N	N	N	N	N	Y	N	Y	Y	N	Y	N	N
N	N	Y	N	N	N	N	N	N	N	N	N	N
N	N	N	N	N	Y	N	Y	Y	Y	N	N	N
N	N	N	N	N	Y	N	Y	Y	Y	Y	N	N
N	Y	N	N	N	Y	N	Y	N	N	N	N	N
N	N	N	N	N	N	N	N	N	N	N	N	N
N	Y	N	N	N	N	N	N	N	N	N	N	N
N	N	N	N	N	Y	Y	Y	Y	N	N	N	N
N	N	N	N	N	N	N	N	N	N	N	N	N
N	N	N	N	N	N	N	N	N	N	N	N	N
N	N	N	N	N	Y	N	N	Y	N	Y	N	N
Y	N	N	N	N	Y	N	N	N	N	N	N	N

Y indicates that a dHPLC shift was seen in a sample, while N indicates that there was no dHPLC shift seen.

Table 5-4: Random White South African dHPLC results

Exon 1	Exon 2	Exon 3	Exon 4	Exon 5	Exon 6pI	Exon 6pII	Exon 7pI	Exon 7pII	Exon 8pI	Exon 8pII	Exon 9	Exon 10
N	N	N	N	Y	Y	N	N	N	Y	N	N	N
N	N	Y	Y	N	Y	N	N	Y	N	Y	N	N
N	N	Y	N	N	Y	N	Y	Y	N	Y	N	N
N	N	Y	N	N	Y	N	Y	Y	N	Y	Y	N
Y	N	N	N	N	Y	N	Y	Y	N	N	N	N
Y	N	N	N	N	N	N	Y	Y	N	N	N	N
N	N	Y	N	N	N	N	Y	Y	N	N	N	N
N	N	Y	N	N	N	N	N	N	N	N	N	N
N	N	N	N	N	N	N	N	N	N	N	Y	N
N	N	Y	N	N	Y	N	N	N	N	N	N	N
N	N	Y	N	N	Y	N	N	N	N	N	N	N
Y	N	Y	N	N	N	N	Y	Y	N	N	N	N
N	N	Y	N	N	N	N	Y	Y	N	N	N	N
N	N	Y	N	Y	N	N	N	Y	N	N	N	N
N	N	Y	N	N	Y	N	N	N	N	N	N	N
Y	N	N	N	N	Y	N	N	N	N	N	N	N
N	N	N	N	N	Y	N	Y	Y	N	N	N	N
N	N	N	Y	N	N	N	N	Y	N	N	N	N
N	N	N	N	N	Y	N	Y	Y	N	N	N	N
N	N	N	N	N	Y	N	Y	N	N	N	N	N

Y indicates that a dHPLC shift was seen in a sample, while N indicates that there was no dHPLC shift seen.

5.9 Reagents used during this study

0.025µm Filter Paper
Available from Millipore

0.5M EDTA (pH 8)
93.06g EDTA (available from Merck)
500ml ddH_2O

pH to 8 with 5M NaOH.

1Kb Ladder
250µl 1 Kb ladder (available from Invitrogen)
125µl Ficoll Dye
2.1ml TE buffer

1Kb plus Ladder
250µl 1 Kb plus ladder (available from Invitrogen)
125µl Ficoll Dye (available from Sigma)
2.1ml TE buffer

1% Alcanox
1g Alcanox (available from Aldrich Chemical company, Inc)
100ml ddH_2O

1x TBE Buffer
1ml 10x TBE buffer
9ml ddH_2O

1x TE Buffer pH 8.0
5ml 10mM Tris (available from Merck)
1ml 1mM EDTA

Make up to 500ml with ddH_2O. pH with HCl if necessary. Autoclave.

10% BSA
1µl 100% BSA (available from Promega)

3% Agarose Gel
300ml 1x TBE
9g Agarose (available from Hispangar)

Mix agarose and 1x TBE.
Heat until all agarose is dissolved.
Allow to cool, but mixture must still be liquid.

Add 9µl ethidium bromide and mix.
Pour into gel tray and allow to set.

4.3% Acrylamide Solution
10.6ml Acrylamide/Bis acrylamide (available from Promega)
10ml 10x TBE
36g Urea (available from Promega)
50ml ddH_2O

4.3% Polyacrylamide Gel
40ml 4.3% acrylamide solution
200µl APS (available from Analar)
24µl Temed (available from Promega)

10%SDS
10g SDS (available from BDH Laboratory Supplies)
100ml ddH_2O

Weigh SDS in a fume hood.

10x TBE Buffer
108g Tris base
55g Boric acid (available from Sigma)
7.4g EDTA

Make up to 1l with ddH_2O. Autoclave.

70% Ethanol
70ml Ethanol (available from BDH Laboratory Supplies)
30ml ddH_2O

Acetonitrile
Available from Riedel-de Haen

Big Dye v3.1
Available from Applied Biosystems.

Big Dye v3.1 Buffer
Available from Applied Biosystems.

***Bfa*I**
Available from New England Biosystems.

ddH_2O
H_2O is distilled and then passed through a Millipore filtration system. Autoclave if using in a PCR reaction.

Dextran-Formamide Dye
10ml Formamide (available from Fluka)
10mg Bromophenol Blue (available from Merck)
200µl 0.5M EDTA pH 8.0

dNTP Mix (125mM)
125µl 10mM dATP (available from Invitrogen)
125µl 10mM dCTP (available from Invitrogen)
125µl 10mM dGTP (available from Invitrogen)
125µl 10mM dTTP (available from Invitrogen)

Make up to 1ml with ddH_2O.

Ethidium Bromide
Available from Sigma.

Ficoll Dye
5g sucrose (available from Gibco BRL)
1ml 0.5M EDTA (pH 7.0)
0.01g bromophenol blue
1g ficoll (available from Sigma)

Make up to 10ml with ddH_2O.

NEBuffer 4
Available from New England Biosystems

Nucleospin® Extract columns
Available from Mackerey-Nagel

PCR Buffer
Available from Roche

Primers
Available from Inqaba Biotec

Proteinase K Solution
100mg Proteinase K
10ml ddH_2O

Rox Size Standard
Available from Applied Biosystems

Saturated NaCl
Add NaCl to ddH_2O until the NaCl no longer dissolves. (available from South African Point Instruments)

Sigma Spin™ Post Reaction Purification columns
Available from Sigma

Sucrose-Triton X Lysing Buffer
10ml 1M Tris-HCl pH 8.0 (available from Merck)
5ml 1M $MgCl_2$ (available from Merck)
10ml Triton X (available from BDH Laboratory Supplies)
109.5g Sucrose (available from Gibco)

Make up to 1l with ddH_2O. Autoclave.

T20E5
20mM Tris pH 8.0
5mM EDTA

Make up to volume with ddH_2O. Autoclave.

Taq
Available from Roche.

5.10 Ethics clearance certificates

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

COMMITTEE FOR RESEARCH ON HUMAN SUBJECTS (MEDICAL)
Ref: R14/49 Jenkins/Ramsay

CLEARANCE CERTIFICATE **PROTOCOL NUMBER** M01-01-19

PROJECT Localising, Cloning & Characterising The Gene For Lipoid Proteinosis

INVESTIGATORS Prof's T/M Jenkins/Ramsay

DEPARTMENT Dept of Human genetics, SAIMR

DATE CONSIDERED 01-01-26

DECISION OF THE COMMITTEE *

Approved unconditionally

DATE 01-02-26 **CHAIRMAN** .....(Professor P E Cleaton-Jones)

* Guidelines for written "informed consent" attached where applicable.

c c Supervisor: Prof T Jenkins
Dept of Dept of Human Genetics, SAIMR

Words2/ann0015/humEth97.wdb M 01-01-19

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10001, 10th Floor, Senate House, University.

I/we fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee.

DATE 6-4-01 SIGNATURE .....

PROTOCOL NO.: M 01-01-19

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

COMMITTEE FOR RESEARCH ON HUMAN SUBJECTS (MEDICAL)
Ref: R14/49 Van Hougenhouck-Tulleken

CLEARANCE CERTIFICATE **PROTOCOL NUMBER** M02-04-31

PROJECT Characterization of The Role of The Extra-Cellular Matrix Protein 1 Gene In Lipoid Proteinosis

INVESTIGATORS Mr W Van Hougenhouck-Tulleken

DEPARTMENT School of Pathology, NHLS

DATE CONSIDERED 02-04-05

DECISION OF THE COMMITTEE *

Approved unconditionally

DATE 02-04-08 **CHAIRMAN** ....(Professor P E Cleaton-Jones)

* Guidelines for written "informed consent" attached where applicable.

c c Supervisor: Prof M Ramsay
Dept of School of Pathology, NHLS

M. Ramsay
29/4/02

Works2\lain0015\humEth97.wdb\M 02-04-31

===

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10001, 10th Floor, Senate House, University.

I/we fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

CHAPTER 6 Reference List

1. Hamada T, McLean WH, Ramsay M *et al*. Lipoid proteinosis maps to 1q21 and is caused by mutations in the extracellular matrix protein 1 gene (ECM1). *Hum Mol Genet* 2002; **11**: 833-40.

2. Urbach E, Wiethe C. Lipoidosis cutis et mucosae. *Virchows Arch. Path. Anat.* 1929; **273**: 285-319.

3. Hamada T, Wessagowit V, South AP *et al*. Extracellular matrix protein 1 gene (ECM1) mutations in lipoid proteinosis and genotype-phenotype correlation. *J Invest Dermatol* 2003; **120**: 345-50.

4. van Hougenhouck-Tulleken W, Chan I, Hamada T *et al*. Clinical And Molecular Characterization Of Lipoid Proteinosis In Namaqualand, South Africa. *Br J Dermatol* 2004; **151**: 413-23.

5. Dinakaran S, Desai SP, Palmer IR *et al*. Lipoid proteinosis: clinical features and electron microscopic study. *Eye* 2001; **15**: 666-8.

6. Freedberg I, Eisen A, Wolff K *et al*. *Dermatology in general medicine 5th edition* 1999; 1825-9.

7. Muda AO, Paradisi M, Angelo C *et al*. Lipoid proteinosis: clinical, histologic, and ultrastructural investigations. *Cutis* 1995; **56**: 220-4.

8. Hamada T. Lipoid proteinosis. *Clin Exp Dermatol* 2002; **27**: 624-9.

9. Kleinert R, Cervos-Navarro J, Kleinert G *et al*. Predominantly cerebral manifestation in Urbach-Wiethe's syndrome (lipoid proteinosis cutis et mucosae): a clinical and pathomorphological study. *Clin Neuropathol* 1987; **6**: 43-5.

10. Heyl T. Lipoid proteinosis in South Africa. *Dermatologica* 1971; **142**: 129-32.

11. Heyl T. Genealogical study of lipoid proteinosis in South Africa. *Br J Dermatol* 1970; **83**: 338-40.

12. Heyl T. Genealogy of lipoid proteinosis. *Lancet* 1969; **2**: 162-3.

13. Gordon H, Gordon W, Botha V. Lipoid proteinosis in an inbred Namaqualand community. *Lancet* 1969; **1**: 1032-5.

14. Emsley RA, Paster L. Lipoid proteinosis presenting with neuropsychiatric manifestations. *J Neurol Neurosurg Psychiatry* 1985; **48**: 1290-2.

15. Moschella S, Hurley H. DERMATOLOGY 2nd edition. *W. B. Saunders Company* 1985.

16. Sams WM, Lynch PJ. Principals and Practice of Dermatology 2nd edition. *Churchill Livingstone Inc.* 1996.

17. Nanda A, Alsaleh QA, Al-Sabah H *et al*. Lipoid proteinosis: report of four siblings and brief review of the literature. *Pediatr Dermatol* 2001; **18**: 21-6.

18. Harper JI, Duance VC, Sims TJ *et al*. Lipoid proteinosis: an inherited disorder of collagen metabolism? *Br J Dermatol* 1985; **113**: 145-51.

19. Bauer EA, Santa-Cruz DJ, Eisen AZ. Lipoid proteinosis: in vivo and in vitro evidence for a lysosomal storage disease. *J Invest Dermatol* 1981; **76**: 119-25.

20. Gordon H, Gordon W, Botha V *et al*. Lipoid proteinosis. *Birth Defects Orig Artic Ser* 1971; **7**: 164-77.

21. Strassberger E. The Rhenish Mission Society in South Africa 1830-1950. *C. Struik (Pty.) LTD.* 1969.

22. Warnich L, Kotze MJ, Groenewald IM *et al*. Identification of three mutations and associated haplotypes in the protoporphyrinogen oxidase gene in South African families with variegate porphyria. *Hum. Mol. Genet.* 1996; **5**: 981-4.

23. Jenkins T. The molecular basis of South African genetic porphyria established at last! *S. Afr. Med. J.* 1997; **87**: 733-5.

24. Defesche JC, van Diermen DE, Lansberg PJ *et al*. South African founder mutations in the low-density lipoprotein receptor gene causing familial hypercholesterolemia in the Dutch population. *Hum. Genet.* 1993; **92**: 567-70.

25. Le Saux O, Beck K, Sachsinger C *et al*. Evidence for a founder effect for pseudoxanthoma elasticum in the Afrikaner population of South Africa. *Hum. Genet.* 2002; **111**: 331-8.

26. Torrington M, Viljoen D. Founder effect in 20 Afrikaner kindreds with pseudoxanthoma elasticum. *S. Afr. Med. J.* 1991; **79**: 7-11.

27. Tipping AJ, Pearson T, Morgan NV *et al*. Molecular and genealogical evidence for a founder effect in Fanconi anaemia families of the Afrikaner population of South Africa. *PNAS* 2001; **98**: 5734-9.

28. Christiano A, Amano S, Eichenfield L *et al*. Premature Termination Codon Mutations in the Type VII Collagen Gene in Recessive Dystrophic Epidermolysis Bullosa Result in Nonsense-Mediated mRNA Decay and Absence of Functional Protein. *J. Invest. Dermatol.* 1997; **109**: 390-4.

29. Pulkkinen L, Christiano A, Gerecke D *et al*. A Homozygous Nonsense Mutation in the *B3* Chain Gene of Laminin 5 (LAMB3) in Herlitz Junctional Epidermolysis Bullosa. *Genomics* 1994; **24**: 357-60.

30. Smits P, Ni J, Feng P *et al*. The human extracellular matrix gene 1 (ECM1): genomic structure, cDNA cloning, expression pattern, and chromosomal localization. *Genomics* 1997; **45**: 487-95.

31. Mongiat M, Fu J, Oldershaw R *et al*. Perlecan protein core interacts with extracellular matrix protein 1 (ECM1), a glycoprotein involved in bone formation and angiogenesis. *J Biol Chem* 2003; **278**: 17491-9.

32. Smits P, Poumay Y, Karperien M *et al*. Differentiation-dependent alternative splicing and expression of the extracellular matrix protein 1 gene in human keratinocytes. *J Invest Dermatol* 2000; **114**: 718-24.

33. Bhalerao J, Tylzanowski P, Filie JD *et al*. Molecular cloning, characterization, and genetic mapping of the cDNA coding for a novel secretory protein of mouse. Demonstration of alternative splicing in skin and cartilage. *J Biol Chem* 1995; **270**: 16385-94.

34. Han Z, Ni J, Smits P *et al*. Extracellular matrix protein 1 (ECM1) has angiogenic properties and is expressed by breast tumor cells. *FASEB J* 2001; **15**: 988-94.

35. Deckers MM, Smits P, Karperien M *et al*. Recombinant human extracellular matrix protein 1 inhibits alkaline phosphatase activity and mineralization of mouse embryonic metatarsals *in vitro*. *Bone* 2001; **28**: 14-20.

36. Sekiya I, Vuoristo J, Larson B *et al*. *In vitro* cartilage formation by human adult stem cells from bone marrow defines the sequence of cellular and molecular events during chondrogenesis. *PNAS* 2002; **99**: 4397-402.

37. Le Naour F, Hohenkirk L, Grolleau A *et al*. Profiling Changes in Gene Expression during Differentiation and Maturation of Monocyte-derived Dendritic Cells Using Both Oligonucleotide Microarrays and Proteomics. *J. Biol. Chem.* 2001; **276**: 17920-31.

38. Rickman D, Bobek M, Misek D *et al*. Distinctive Molecular Profiles of High-Grade and Low-Grade Gliomas Based on Oligonucleotide Microarray Analysis. *Cancer Research* 2001; **61**: 6885-91.

39. Nurse GT, Weiner JS, Jenkins T. The Peoples of Southern Africa and their Affinities. *Oxford Science Publications, Research Monographs on Human Population Biology No 3* 1985.

40. Miller S, Dykes D, Polesky H. A simple salting out procedure for extracting DNA from nucleated cells. *Nucleic Acid Research* 1988; **16**: 1215.

41. Brown TA. Gene cloning, an introduction. 3rd edition. 1997.

42. Saiki R, Gelfand D, Stoffel S *et al*. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 1988; **239**: 487-91.

43. Sambrook, Russell. Molecular Cloning 3rd edition, volumes 1-3. *Cold Spring Harbor Laboratory Press* 2001.

44. Nasri M, Thomas D. Relaxation of recognition sequence of specific endonuclease HindIII. *Nucleic Acids Research* 1986; **14**: 811-21.

45. Bitinaite J, Schidkraut I. Self Generated DNA termini relax the specificity of SgrAI restriction endonuclease. *PNAS* 2002; **99**: 1164-9.

46. NEW ENGLAND BIOLABS. *Bfa*I Certificate of analysis (R0568S). *NEW ENGLAND BIOLABS Certificate of analysis* 1.

47. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids research* 1999; **27**: 573-80.

48. Transgenomic Incorporated. An Introduction to the Transgenomic WAVE DNA Fragment Analysis System. *Transgenomic WAVE DNA Fragment Analysis System user manual* 2001.

49. Strachan T, Read A. Human Molecular Genetics. *BIOS Scientific Publishers Limited* 1998.

50. Klug W, Cummings M. *Concepts of Genetics 3rd edition* 1991.

51. Harlt D, Clark A. Principles of Population Genetics 2nd edition. 1989.

52. Schneider S, Roessli D, Excoffier L. Arlequin ver. 2.000: A software for population genetics data analysis. *Genetics and Biometry Laboratory, University of Geneva, Switzerland* 2000.

53. Devlin B, Risch R. A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics* 1995; **29**: 311-22.

54. Zar JH. Biostatistical Analysis. *Prentice-Hall, Inc., Engelwood Cliffs, New Jersey 07632* 1984.

55. Rice W. Analyzing Tables of Statistical Tests. *Evolution* 1989; **43**: 223-5.

56. Whittaker JC, Harbord RM, Boxall N *et al*. Liklihood-based Estimation of Microsatellite Mutation Rates. *Genetics* 2003; **164**: 781-7.

57. Latchman D. Gene Regulation, A eukaryotic perspective. *Chapman & Hall* 1996.

58. Oyama N, Chan I, Neill SM *et al*. Autoantibodies to extracellular matrix protein 1 in lichen sclerosus. *Lancet* 2003; **362**: 118-23.