DECLARATION

I, Candice Lee De Carvalho declare that this dissertation is my own work, unless otherwise stated in the text. It is being submitted for the degree of Bachelor of Science with Masters (MSc Med) at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other university.

……

CANDICE LEE DE CARVALHO

…….DAY OF……. [MONTH], 2008

ACKNOWLEDGEMENTS

I would like to thank the following for their support during this project:

- Dr. Klugman at the Johannesburg General Hospital and Dr. Westwood at the Red Cross Children's Hospital in Cape Town, as well as Dr Henderson for taking an interest in this project and assisting with the selection of patients
- Colleagues and staff at the Human Genetics Department at the NHLS
- Prof. Michele Ramsay for investing in me and providing me with endless support
- My family and friends
- God

ABSTRACT

INTRODUCTION: Cystic fibrosis (CF) is an autosomal recessive disease caused by mutations in the CFTR gene. The gene mutation profile is extremely heterogeneous and mutations show a variable distribution among population groups. In SA the 3120+1G->A splice site mutation has been found predominantly in Black and Coloured patients. It occurs in Black CF patients at an estimated frequency of 46%. The CF carrier frequency is estimated at 1/34 in Black and 1/55 in Coloured populations, and based on these rates, it is clear that a significant number of Black and Coloured patients remain undiagnosed.
Point mutations account for the majority of the mutations that have been found in the CFTR gene. Copy number mutations are, however, increasingly being detected in CF patients through the use of gene dosage-dependant assays. These mutations have been found to occur in the CFTR gene in various African American families and exon rearrangements are thought to account for 1.3% of all CF chromosomes across all populations. **AIMS:** To use haplotypes to analyse the origin(s) of the 3120+1G->A mutation and the likely frequencies of the remaining unknown mutations. To increase mutation detection in the SA Black and Coloured groups by searching for CFTR gene exons for copy number mutations. **METHODS**: In patients with at least one copy of the 3120+1G>A mutation haplotype studies will be used to elucidate the origin(s) of this mutation in SA Black and Coloured CF patients, by analyzing pyrosequencing SNP genotype data. In patients with at least one unknown mutation, haplotype studies will reveal the likely relative frequencies of the unknown mutations in these populations. In Black and Coloured CF patients with at least one unknown mutation, a multiplex ligation dependant probe amplification (MLPA) CF kit will be used for the detection of exon copy number mutations. **RESULTS:** The results of the haplotype data show that there is a G-G-C-G-T-A haplotype, for markers MetD-KM19-J44-T854T-Tub18-J32, associated with the 3120+1G->A mutation in both Black and Coloured patients. Unknown mutation-associated haplotypes indicate that there are two relatively common unknown mutations in each of these populations. MLPA results show that one patient is a carrier of an

exon 2 deletion. **CONCLUSION:** A single origin for the 3120+1G>A mutation in Black and Coloured CF patients is supported by the data. Exon copy number changes in the CFTR gene are not a major mutational mechanism leading to CF in SA Black and Coloured patients.

PAPER IN PREPARATION FOR PUBLICATION

Des Georges M, Guittard C, Templin C, Altiéri JP, **de Carvalho C**, Ramsay M, Claustres M. WGA Allows the Molecular Characterization of a Novel Large CFTR Rearrangement in a Black South African CF patient

POSTER PRESENTATIONS AT A CONFERENCE

CL de Carvalho, M Ramsay (2007) Molecular profiling of CFTR mutations in Black and Coloured South African Cystic Fibrosis patients
Southern African Society of Human Genetics (SASHG) Congress, Golden Gate, March 2007 (Joint prize for best poster presentation)

TABLE OF CONTENTS

DECLARATION I

ACKNOWLEDGEMENTS II

ABSTRACT III

PAPER IN PREPARATION FOR PUBLICATION V

POSTER PRESENTATIONS AT A CONFERENCE V

CHAPTER 1: INTRODUCTION - 1 -

1.1 CYSTIC FIBROSIS - 2 -

1.2. CF IN SOUTH AFRICA - 6 -

1.3. ORIGIN(S) OF MUTATIONS - 10 -

1.4. COPY NUMBER VARIANTS - 14 -

1.5. AIMS OF THE STUDY - 22 -

CHAPTER 2: SUBJECTS AND METHODS - 2 -

2.1. SUBJECTS - 3 -

2.2. GENOTYPING FOR HAPLOTYPE INFERENCE - 4 -

2.2.1 Carrier testing of controls and families - 4 -

2.2.2 Selection of markers for genotyping and haplotype study - 4 -

2.2.3 Pyrosequencing - 7 -

2.2.4 Data analysis - 35 -

2.3. DETECTION OF REARRANGEMENTS USING MLPA - 37 -

2.3.1 Principle of technique - 38 -

2.3.2 Principle of analysis of MLPA amplification products - 40 -

2.3.3 Laboratory protocol - 41 -

2.3.4 Analysis of MLPA products - 43 -
2.3.4.1 Analysis programs - 43 -

CHAPTER 3: RESULTS - 49 -

3.1. CARRIER TESTING - 50 -

3.2.1 PCR results - 51 -

3.2.2 Pyrosequencing results - 51 -

3.3. HAPLOTYPE RESULTS - 53 -

3.3.1 Control sample inferred haplotypes - 62 -

3.3.2 Genotype and allele frequencies - 64 -

3.4. MUTATION TESTING - 65 -
3.4.1 MLPA Results - 65 -

CHAPTER 4: DISCUSSION **- 68 -**

4.1 ORIGINS OF MUTATIONS - 87 -

4.1.1 Evolutionary origin of 3120+1G>A mutation in Black and Coloured SA populations - 87 -

4.1.2 ΔF508 associated haplotypes in Coloured CF patients - 89 -

4.1.3 Frequency of mutations among unidentified CF alleles - 90 -

4.1.4 CFTR haplotype inference in healthy controls - 91 -

4.1.5 Comparison between HapMap and SA population SNP genotype data - 94 -

4.2 EXON COPY NUMBER VARIATION IN BLACK AND COLOURED CF PATIENTS - 96 -

4.2.1 MLPA: technical discussion - 97 -

4.2.2 Analysis of MLPA products - 98 -

4.3 Future studies - 103 -

4.4 CONCLUSIONS - 104 -

APPENDICIES - 87 -

REFERENCES - 105 -

LIST OF FIGURES

Figure 1.1: Model of the CFTR protein in the cell membrane ……...…….2
Figure 1.2: Geographical distribution of 3120+1G>A mutation.…….…….8
Figure 1.3: D' plots for the YRI, CEU and CHB and JPT ……..14
Figure 1.4: Fully characterized gross CFTR rearrangements……..….7
Figure 1.5: Approaches used for the identification of CNVs ……..20

Figure 2.1: Outline of pyrosequencing method……..…..29
Figure 2.2: Genomic positioning of intragenic and extragenic SNPs …….…33
Figure 2.3: Systematic diagram of sample preparation time ……...……34
Figure 2.4: MLPA peak picture depicting a 35-55% ……..38
Figure 2.5: Model of a single probe pair for MLPA……39
Figure 2.6: Depiction of events during the MLPA technique……..…….……37

Figure 3.1: PCR result of rs1042077 (T845T)…….……51
Figure 3.2: Pyrogram showing passed PSQ results……..……52
Figure 3.3: Pyrogram of rs11770163 (Met D)…...……..…52
Figure 3.4: MLPA result……...……..….67

LIST OF TABLES

Table 1.1: Characterisation of CFTR mutational classes……..….....6
Table 1.2: Frequency of common CFTR mutations in Black ……..…..……...9
Table 1.3: Frequency of the 3120+1G>A mutation…….…...9
Table 1.4: Frequency of common CFTR mutations in Coloured …….……...10
Table 1.5: Summary of copy number variants in the CFTR gene……..…….....18

Table2.1: Standard colour scheme…….....……26
Table 2.2: General PCR mix protocol……...…….....31
Table 2.3: Primers used for amplification of products for pyrosequencing...….32
Table 2.4: Synopsis of CDC_Analysis macro calculations and worksheets..…46

Table 2.5: Assessment of the most effective way of using CDC_Analysis..…..48

Table 3.1: 3120+1G>A carrier test for Black unaffected controls……..……...50
Table 3.2: 3120+1G>A carrier test for Coloured unaffected controls……..….50
Table 3.3: 3120+1G>A carrier test result in relatives of CF patients……....50
Table 3.4: Coloured haplotypes of patient and family data…….…....….54
Table 3.5: Black haplotypes of patient and family data…….....….....56
Table 3.6: Inferred 3120+1G>A -associated haplotypes……..…..58
Table 3.7: Inferred ΔF508 -associated haplotypes…….…....59
Table 3.8: Inferred Coloured patient haplotypes……....60
Table 3.9: Inferred Black patient unknown mutation-associated haplotypes….61
Table 3.10: Postulated mutation detection rate……61
Table 3.11: Inferred results for Black and Coloured……...…...…63
Table 3.12: Genotype frequency and Hardy-Weinberg equilibrium…...….……62
T**able 3.13**: MLPA result for patients that passed the analysis quality control…66

Table 4.1: Summary of 3120+1G>A associated haplotypes…….…...…70
Table 4.2: Summary of ΔF508 associated haplotypes……...…..…72
Table 4.3: Allele frequencies for HapMap and SA populations…….…...…..77
Table 4.4: Significance testing for differences in allele frequency…….…….....78
Table 4.3: The advantages and disadvantages of four programs for MLPA……84

ABBREVIATIONS

A	adenine
ABCC7	ATP-binding cassette, subfamily C, member 7
AHR	allelic homologous recombination
AMP	adenosine monophosphate
bp	base pair
BSA	bovine serum albumin
BAC	bacterial artificial chromosome
C	cytosine
CBAVD	congenital bilateral absence of the vas deferens
CCD	charge coupled device
CEU	Centre d'Etude du Polymorphisme Humain
CGH	comparative genomic hybridisation
CHB	Han Chinese in Beijing
CI	confidence interval
CNV	copy number variation
CF	cystic fibrosis
CI	Confidence Interval
D'	D prime: pairwise measure of linkage disequilibrium
ddH_20	deionised double distilled water
DNA	deoxyribose nucleic acid
dNTPs	deoxynucleotide triphosphates
ddNTPs	dideoxynucleotide triphosphates
dbSNP	database of single nucleotide polymorphisms
DQ	dosage quotient
EDTA	ethylene diamine tetraacetic acid
f	frequency
FISH	fluorescent *in situ* hybridisation
G	guanine
GC	guanine and adenosine
HCl	hydrochloric acid
HR	homologous recombination

H-W	Hardy-Weinberg equilibrium
IC	internal control
IRT	immunoreactive trypsinogen
IUPAC	International Union of Pure and Applied Chemistry
JPT	Japanese in Tokyo, Japan
LD	linkage disequilibrium
LoH	loss-of-heterozygosity
MAF	minor allele frequency
Mb	megabase
MR	coalescent with recombination
MSD	membrane-spanning domains
$MgCl_2$	magnesium chloride
MI	meconium ileus
MLPA	multiplex ligation-dependant probe amplification
NaOH	sodium hydroxide
NAHR	non-allelic homologous recombination
NBD	nucleotide-binding domains
NCBI	National Center for Biotechnology Information
NEB	New England Biolabs
NHLS	National Health Laboratory Service
NPD	nasal potential difference
PCR	polymerase chain reaction
PSQ	pyrosequencing
P-value	probability value
QRT-PCR	quantitative real-time PCR
R	regulatory domain
RE	restriction enzyme
RFLP	restriction fragment length polymorphism
SA	South Africa
SD	standard deviation
SNP	single nucleotide polymorphism
T	thymine

TBE	tris borate EDTA
UCSC	University of California Santa Cruz
UK	United Kingdom
USA	United States of America
v	version
YRI	Yoruba in Ibadan, Nigeria

UNITS AND SYMBOLS

%	percentage
µl	micro litre
D’	D prime (LD unit)
mEq/L	milliequivalents of solute per litre of solvent
kb	kilobase
l	litre
mg	milligram
ml	millilitre
ng/µl	nanograms per micro litre
nm	nanometres
secs	seconds
™	trademark
V/cm	volts per centimetre
^{0}C	degrees Celsius

Chapter 1: Introduction

1.1 Cystic fibrosis

Cystic Fibrosis (CF) is an autosomal recessive genetic disorder caused by mutations in the Cystic Fibrosis Transmembrane Conductance Regulator gene (CFTR). The gene was identified in 1989 on the basis of its map location on chromosome 7, using a positional cloning approach (Kerem et al. 1989; Riordan et al. 1989). It is also known as the "ATP-binding cassette, subfamily C, member 7" (ABCC7). The CFTR gene product functions as a chloride channel and also controls the regulation of other transport pathways (Riordan et al.1989; Dörk et al. 1994). The gene is about 250kb in length, with 27 exons, and the transmembrane protein, which it encodes, is 1480 amino acids long. The CFTR protein consists of two membrane-spanning domains (MSD) (Figure 1.1). These are involved in the formation of the chloride ion channel. The MSD each contain six transmembrane segments. Two nucleotide-binding domains (NBD) and a regulatory domain (R) constitute the remainder of the protein. The NBD interact with cytosolic nucleotides to control channel activity. The R domain is phosphorylated by a cyclic-AMP-dependant protein kinase. Phosphorylation of this transmembrane channel causes it to open (Welsh et al. 1995).

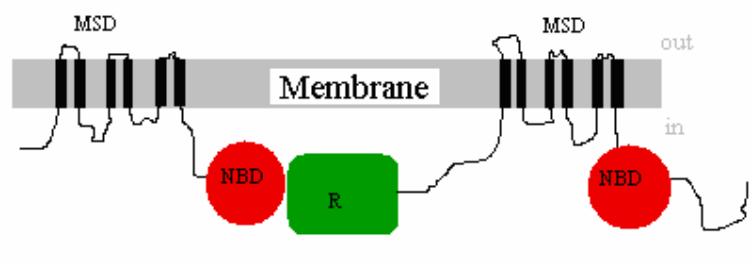


Figure 1.1: Model of the CFTR protein in the cell membrane (Adapted from Welsh et al. 1995)

Since this single-gene disorder is inherited as an autosomal recessive trait, carriers are asymptomatic and affected patients must have two mutant alleles in order to express the disease phenotype.

1.1.2 Diagnosis

CF may be difficult to diagnose, as patients show high clinical and molecular diversity. Since the most consistent CF feature is abnormal sweat chloride levels (Zielenski and Tsui 1995), the 'gold standard' for diagnosis of the disease is the sweat test: sweating is induced on the forearm of the patient, using pilocarpine nitrate and 100mg of sweat is collected. A digital chloridometer is used to determine the sweat chloride ion concentration (Gibson and Cooke 1959; Green et al. 1985). An increased concentration of electrolytes in the sweat result from mutations in the CFTR gene, which functions as a chloride channel on the apical membranes of epithelial cells lining the sweat ducts. Patients with severe CF usually have positive sweat test results. A positive result is characterised by a sweat chloride concentration of 60mEq/L or higher. These patients present with typical clinical features such as dysfunctional pancreatic exocrine secretions (due to blocked pancreatic ducts), progressive obstructive sino-pulmonary disease and meconium ileus (MI). MI is an obstruction of the distal ileum or proximal colon with inspissated meconium and this clinical feature is present in approximately 15% of neonatal patients with CF (Turcios 2005). Approximately 5% of CF patients have negative sweat test results (exhibit sweat chloride test results of less than 60mEq/L). These patients, or those with borderline/equivocal sweat test results often present with congenital bilateral absence of the vas deferens (CBAVD) and other less severe forms of the disease (Dörk et al. 2000; Mickle et al. 1998; Rosenstein and Cutting 1998). Transepithelial nasal potential difference (NPD) measurements may be taken in individuals over the age of six as an alternative to sweat testing (Knowles et al. 1995). Since respiratory epithelia regulate ion transport, the absence of a functional CFTR protein at this surface results in altered chloride and sodium transport. As a consequence, abnormal electrical potential across nasal epithelial surfaces occur, with raised (more negative) baseline NPD and a greater change in NPD during perfusion of the nasal mucosa with amiloride, an inhibitor of sodium channel activity. This measurement may only be performed using specialised equipment by qualified personnel, as is the case with sweat testing. In countries like the USA and UK, newborn screening using immunoreactive trypsinogen (IRT) assays, which are

performed on blood spots, have been implemented. Trypsinogen levels are elevated in CF patients. While the test is safe for newborns, it has a high false positive rate and is not appropriate for the diagnosis of CF in patients who do not have abnormal pancreatic functioning.

Molecular genetic testing is used for diagnosis in symptomatic individuals; for carrier testing and for prenatal diagnosis. This is done through targeted mutation analysis, sequence analysis and deletion or duplication analysis in the CFTR gene. The mutation detection rate depends on the number of identified CFTR mutations and the frequency of these mutations in a given population. Population-specific diagnostic mutation panels are used in a diagnostic setting to test patients that present with a clinical phenotype suggestive of CF, or who have at least one positive sweat test result.

The ethnicity, socioeconomic status and country of origin of a patient all have a large influence on whether or not they will receive a correct diagnosis. In SA, Black patients have a low chance of being correctly diagnosed with CF for two reasons: many healthcare workers believe that CF is extremely rare in Blacks and therefore do not test patients who present clinically with CF and have a high chance of misdiagnosing these patients. Secondly, because mutations in the CFTR gene are population-specific and very few 'Black' mutations have been identified there is a relatively low rate of molecular diagnosis of CF in this population. Those whose socio-economic status is particularly low, especially those in rural and disadvantaged urban communities, have limited or no access to specialist diagnostic centres. This lack of access can be traced back to their difficult living conditions, poverty and low levels of education. Country of origin also has an effect as in developing countries, the amount of money spent on research is relatively low and therefore the availability of knowledge, resources and skills is dependant on the economic status of the country.

1.1.3 Management

Symptomatic treatment of CF has resulted in an increase in lifespan and quality of life from 14 years in 1969 to 31.6 years in 2002 (Cystic Fibrosis Foundation), and is now even better. Treatment of individuals with respiratory dysfunction may include oral, inhaled or IV antibiotics, bronchodilators, anti-inflammatory agents, mucolytics and chest physiotherapy. Lung or heart/lung transplantation is an option for selected individuals with severe disease. New therapies include CFTR 'bypass' therapy to augment alternative chloride channels and 'protein assist' treatment to improve trafficking and function of defective CFTR protein (Gibson et al. 2003). Patients with gastrointestinal dysfunction undergo nutritional therapies whereby special formulas and supplemental feeding for infants are administered for increased intestinal absorption and increased caloric intake. Pancreatic insufficiency is treated with oral pancreatic enzyme replacement with meals. Some patients may undergo respiratory and gastrointestinal treatment, depending on the severity and range of symptoms.

1.1.4 Mutations in the CFTR gene

The CFTR locus is extremely heterogeneous and mutations in the CFTR gene result in a widely varied phenotype. The total number of mutations in the gene remains unclear, as mutations are sometimes restricted to a single family or a limited geographical region, while others account for a large proportion of CF alleles; 1556 mutations are currently listed in the Cystic Fibrosis Mutation Database (http://www.genet.sickkids.on.ca/cftr/). A major mutation that results in a single amino acid deletion (ΔF508) accounts for approximately 70% of the disease alleles in Caucasians globally (Cystic Fibrosis Genetic Analysis Consortium). Of the remainder, missense mutations have been detected in the gene more frequently than any other mutation type. Nonsense, frameshift, RNA splicing and copy number mutations have also been detected in the gene. Only four other mutations (G542X, N1303K, G551D, and W1282X) have overall frequencies >1% among the CF alleles in European populations (Dawson and Frossard 2000). The persistence of certain mutations throughout the ages, in various populations, for example the 3120+1G>A and ΔF508 mutations, provides

support for the concept that these mutations may have provided a selective advantage to the carrier. It has been proposed that heterozygotes for CF mutations may be protected against the dehydrating action of *Vibrio cholerae* (Romeo et al.1989; Gabriel et al. 1994) or *Salmonella typhi* (Pier et al. 1998; Pier 1999, 2000) enterotoxins, although this theory has been disputed (Hogenauer et al. 2000; Mateu et al. 2002).

The quantitative and qualitative effects of specific classes of mutations in the CFTR gene are outlined in Table 1.1.

Table 1.1: Characterisation of CFTR mutational classes (Zielenski and Tsui 1995)

Mutational Class	Effect on protein product	Mechanism
1	Defective protein production with premature termination of CFTR production	Non-sense, frameshift, splice site or rearrangement mutations
2	Defective trafficking of CFTR, protein does not reach apical surface membrane	Missense mutations, small amino acid deletions
3	Defective regulation of CFTR even though it is able to reach the apical cell surface	Missense mutations
4	CFTR reaches the apical surface but conduction through the channel is defective	Missense mutations

1.2. CF in South Africa

Globally, CF is the most common recessive disorder in populations of European origin but it is present in diverse populations from different continents (Dörk et al. 1998). CF was, at first, diagnosed sporadically in Black Africans, since the first record of a patient born to Bantu-speaking parents (Grove 1959). The second record of a Black African with CF was that of a Kenyan child born to unrelated parents (Mac Dougall 1962). Later, 'Bantu' twins were diagnosed in SA (Levin et al. 1967). Since these initial diagnoses, progressively more research has been done to ascertain the molecular basis and carrier frequency of CF in Black as well as other South African populations. It is likely that many CF patients in SA

remain undiagnosed and as a result, the infant mortality rate of untreated CF patients in SA is likely to be high. This is especially relevant in rural areas, where patients lack access to medical services. It is proposed that patients are frequently misdiagnosed with chronic pulmonary infection, malnutrition, TB, infantile diarrhea and failure to thrive (Padoa et al. 1999). Among non-CF specialists, CF is still thought to be uncommon in SA Coloureds and rare in SA Blacks. False negative sweat test results may also augment the problem of under-diagnosis of CF in SA (Goldman et al. 2001).

1.2.1 CF in South African Black populations

One of the most important discoveries concerning CF in SA Blacks was the discovery of the 3120+1G>A splice site mutation in exon 16 of the CFTR gene (Carles et al. 1996). This finding followed the initial discovery of this mutation in African Americans, in whom the CF carrier frequency is 1/61 (Macek et al. 1997; Hamosh et al. 1998). It was first discovered in three African American CF patients, two of whom were heterozygous and one of whom was homozygous for the mutation. The father of one of these African American patients originated in Cameroon and was found to be a carrier of the mutation. The 3120+1G>A mutation is found at a frequency of 12.3% in African American CF chromosomes and is the second most common mutation in this population, after ΔF508. Besides South African and African American populations, this mutation has since been found in Saudi Arabia (Banjar 1999) and Greece (Tzetis et al. 1997), as well as Brazil (Cabello et al. 2006; Giselda et al. 2006), the Reunion Islands (Bienvenu et al. 1996) and Bahrain (Eskandarani 2002) (Figure 1.2). Carles et al. (1996) show, through haplotype studies, that the 3120+1G>A mutation arose in Africa and proposed that the mutation has a single origin in African Blacks and that it probably arose before the Bantu expansion. Dörk et al. (1998) showed, again through haplotype studies, that there is a common origin of the 3120+1G>A mutation among African, Arab, Greek and African American populations and that the mutation is ancient and probably more common than is currently being detected, due to the under-diagnosis of CF in tropical and sub-tropical areas of the world. Haplotypes for these studies are shown in Chapter 4. The first report of

this mutation in Caucasian populations was in Greece and it has been found in the SA White population at very low frequencies, as it was probably introduced into this population through Black (or Coloured) admixture (Goldman et al. 2003). Figure 1.2 shows the geographical distribution of the 3120+1G>A mutation and a visualisation of the probable direction of gene flow between populations, as well as the frequency of the mutation among CF chromosomes in the different populations, as described in the literature.

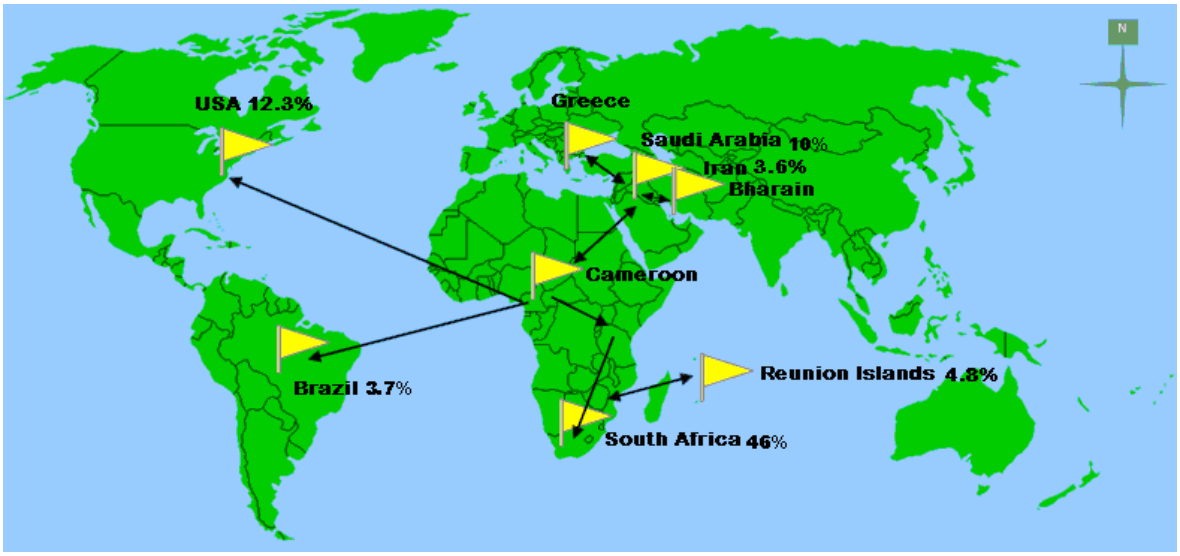


Figure 1.2: Geographical distribution of the 3120+1G>A mutation. Arrows indicate the probable direction of gene flow between populations. The frequency of this mutation among CF chromosomes, where available, is shown. (This frequency refers to African American populations in the USA and the central province of Saudi Arabia)

The carrier frequency of CF in the general Black SA population is approximately 1/34 [this is an adjusted frequency based on a 100% mutation detection rate; 95% CI: 1 in 14 to 1 in 59] (Padoa et al. 1999). The incidence of CF in SA Black populations is approximately 1/4624 (Goldman et al. 2003). Although a number of mutations have been detected in Southern African Black populations at low frequencies (Table 1.2), the 3120+1G>A mutation accounts for the largest number of CF chromosomes and is therefore the only mutation that is tested in a molecular diagnostic setting. As a result of this, only 21% of Black patients with

CF will be diagnosed with two pathogenic CFTR mutations in SA (Table 1.2 adapted from Goldman et al. (2003)).

Table 1.2: Frequency of common CFTR mutations in Black SA CF patients

Mutation	Black (N*=14)
3120+1G > A	0.464
G1249E	0.036
3196del54	0.036
-94G > T	0.036
2183delAA	0.036
Unknown	0.393
Total mutation detection	**0.607**

* N = number of patients studied

1.2.2 CF in South African Coloured populations

CF is more frequently diagnosed in SA Coloureds. More CFTR gene mutations have been identified in this population group and molecular genetic testing in Coloured CF patients yield a higher mutation detection rate than in Blacks. The most common mutations found in this group are the 3120+1G>A and the ΔF508 mutations. The ΔF508 mutation is the most common mutation in populations of European origin found worldwide and has likely been introduced into admixed African populations, such as the SA Coloured and African American, through ethnic admixture. The 3120+1G>A mutation may have been introduced into the Coloured population from the Black population, however, in a study done by Padoa et al. (1999), this mutation was not detected in 157 Nguni Black Africans (Table 1.3), with whom it is likely that Cape Coloureds would have admixed. It is possible that this mutation arose independently in the Cape Coloured population. Haplotype studies on the 3120+1G>A mutation in the Black and Coloured populations would reveal if this mutation arose once in the Black population and was subsequently passed into the Coloured population, or if the mutation arose independently in both.

Table 1.3: Distribution of the 3120+1G>A mutation in healthy Black Africans (adapted from Padoa et al. 1999)

Chiefdom	Subjects	Carriers
Nguni	157	0
Zulu	57	0
Xhosa	52	0
Ndebele	52	0
Swazi	25	0
Sotho/Tswana	372	6
Tsonga	76	1
Venda	45	1

Hill et al. 1988 found the incidence of CF in Coloured patients in Cape Town to be 1 in 12 000 and more recently Westwood (2007) found an incidence of 1 in 2853 (carrier frequency of 1 in 27) in the Western Cape. Both studies conclude that many cases of CF remain undiagnosed in rural areas. Approximately 55% of Coloured patients with a clinical phenotype suggestive of CF are diagnosed with two pathogenic CFTR mutations, using a diagnostic kit of multiple mutations, in SA Coloureds (Table 1.4).

Table 1.4: Frequency of common CFTR mutations in Coloured SA CF patients

Mutation	Coloured (N*=43)
ΔF508	0.500
3272-26A > G	0.012
3120+1G > A	0.174
G542X	0.023
G551D	0.023
R1162X	0.012
Unknown	0.256
Total mutation detection	**0.744**

* N = number of patients studied

1.3. Origin(s) of mutations

By tracing population genetic characteristics and the geographic distribution of population specific (and non-specific) mutations, gene flow, historical population migration and natural selection gradients, distinguishing features of human populations may be investigated. Clinically pertinent information that may be population specific, such as genotype-phenotype correlations, counseling options, prognosis and patient management may also be observed and evaluated (Dawson

and Frossard 2000). An example of how these population genetic characteristics can be used to give us this type of information is found in the studies that have been done on the ΔF508 mutation in populations of European descent. Stringer (1990), in a paper entitled 'The emergence of modern humans' showed through genetic marker studies that this mutation was present in the populations that entered Europe. After the initial introduction of this mutation into European populations, further population expansions occurred at different periods throughout Europe. This was supported by Morral et al. (1994) who demonstrated that different haplotypes were associated to the ΔF508 mutation and these haplotypes have different frequencies in different parts of Europe. Lucotte and Loirat (1993) and Lucotte et al. (1995) discovered that the ΔF508 mutation frequency differs across Europe in a south-easterly to north-westerly gradient, which correlates with the concept of humans spreading across Europe, starting from the Middle East. Lastly, through the analysis of microsatellite markers on CF and non-CF chromosomes, Morral et al. (1996) found that the mutation originated at least 52 000 years ago in an ancestral population that expanded and spread the mutation. The correlations between mutations and genetic markers therefore serve as tools for human genetic research. By studying the 3120+1G>A mutation in the context of Black and Coloured CF patients in SA, the origin of this mutation in these populations may be inferred.

1.3.1 Properties of markers

Studies of common genetic variants and the role that they may play in disease are possible because individuals who carry a particular SNP allele at one site, often predictably carry a specific allele or alleles at another variant site (The International HapMap Consortium). This correlation is known as linkage disequilibrium (LD). LD between alleles at specific loci within the genome may be particularly high, in which case the segment of DNA that contains these alleles is termed a haplotype block, where pairs of alleles segregate in a dependent manner. A pathogenic mutation may arise by chance within a specific haplotype block and is thereby linked to the polymorphic variants within the haplotype that are in LD with one another. This mutation will be non-randomly associated with

that particular haplotype and will thus segregate through a family with it. The LD structure of the genome may deteriorate through recombination and mutation over time. It is also important to consider the discontinuous nature of LD throughout the genome due to recombination and mutation hot and cold spots, which are regions of high and low recombination and mutation, respectively.

1.3.2 Using HapMap data

The DNA markers that are used for studying origin and frequency of mutations in SA populations will be selected by studying their allele and genotype frequencies and LD information contained in genotype data dumps from the HapMap project (www.hapmap.org). They will also be selected based on their presence in previously published haplotype studies for comparison of the haplotypes found in this study to those studies.

The haplotype map of the human genome is presented in a public database of common variation in the human genome. The project to compile these data was done in two phases. Phase I contained information on "more than one million SNPs for which accurate and complete genotypes have been obtained in 269 DNA samples from four populations including ten 500-kilobase regions in which essentially all information about common DNA variation has been extracted" (International HapMap consortium 2005) and served as a guide to genetic studies, as it was used in this project. Phase II of the project resulted in an additional 2.1 million SNPs that were genotyped in the same individuals. A minimum minor allele frequency (MAF) of 0.05 was targeted for the study and SNPs with a MAF greater than or equal to 0.05 are considered "common" by the HapMap database. MAFs and the minor alleles for a particular SNP as well as the LD between the markers may differ between the four populations that were genotyped for the HapMap project (Figure 1.3). These populations are: 1) 90 individuals (30 parent offspring trios) from the Yoruba in Ibadan, Nigeria (YRI); 2) 90 individuals (30 parent offspring trios) from Utah, United States of America (USA), from the Centre d'Etude du Polymorphisme Humain collection (CEU); 3) 45 Han Chinese in Beijing, China (CHB); 4) 44 Japanese in Tokyo, Japan (JPT). All markers for

which information is available in the database are polymorphic in at least one of the populations studied.

The marker information that is available for the CFTR gene may be obtained from the HapMap site and analysed in a program called HaploView. This is a program that was developed as a tool for the analysis and visualisation of LD and haplotype maps. "It provides computation of LD statistics and population haplotype patterns from primary genotype data" (Barrett et al. 2005) that is obtained from the HapMap project. A graphical presentation of the several pairwise measures of LD as well as haplotypes and their population frequencies are created by the software, with the display indicating the degree of LD between the bocks, as well as the haplotype block frequencies. Lastly, the software employs a suite of statistical tests to aid researchers in the selection of markers for disease studies, by exploiting the redundancy among SNPs (due to their non-random correlation), with the aim of maximising efficiency in the laboratory, while minimising the loss of information (Daly et al. 2001). The non-redundant markers that are selected by this software are called tag SNPs. Researchers may utilise this information to decrease the number of SNPs that they genotype for a particular study. Figure 1.3 shows the LD blocks in three main groups and clearly illustrates that these blocks tend to be much shorter in the African population compared to the European and Asian population. HaploView shows that there are two haplotype blocks for the European and Asian populations and seven for the YRI population. This indicates that there is less LD across the YRI CFTR gene than the other populations. These data indicate that many more tag SNPs will need to be used in African populations when compared to European and Asian populations than are found in African populations (HapMap consortium 2005).

1.3.3 Using pyrosequencing for marker genotyping

Pyrosequencing, a sequencing method based on sequencing by synthesis, through the detection of nucleotide incorporation, using a primer-directed polymerase extension is used in this project for marker genotyping (Ronaghi et al. 1996, 1998; Ronaghi 2000). The resultant genotypes for patient samples will be phased

manually and the control samples will be phased using a software program called PHASE (Stephens et al. 2001). The details of the genotyping technique and the haplotype inference are described in Chapter 2.

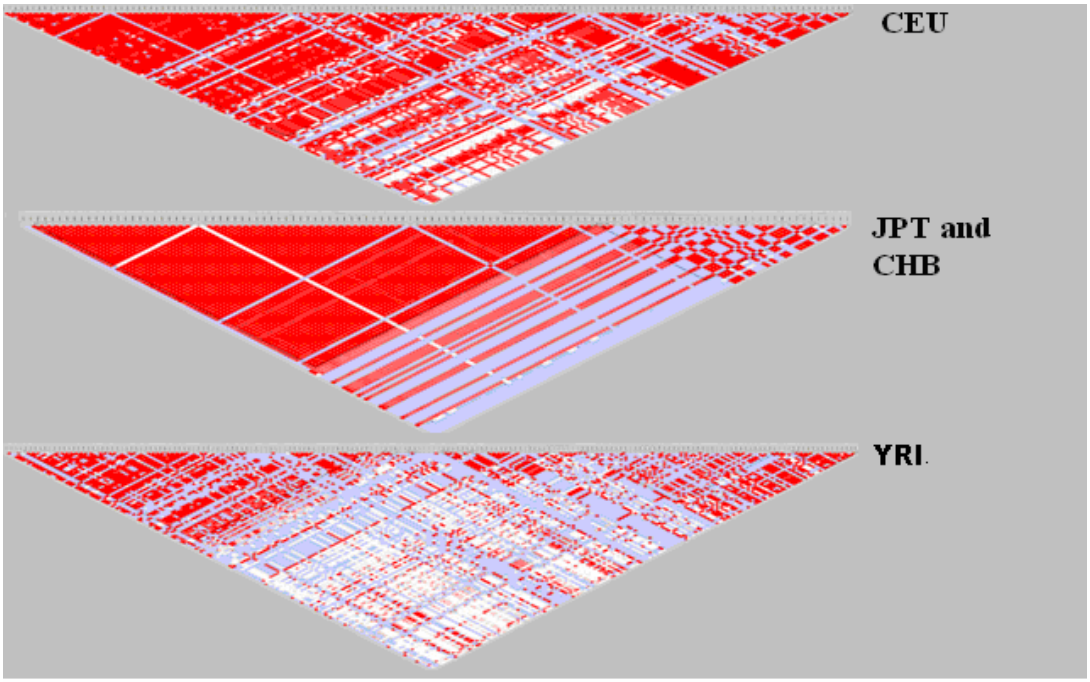


Figure 1.3: HapMap SNP genotype linkage data across CFTR gene (116 909 253 to 117 095 952, NCBI b36), as viewed in HaploView. D' plots for the YRI, CEU and CHB and JPT analysis panels are shown: white, D'<1 and LOD < 2; blue, D' = 1 and LOD < 2; pink, D' < 1 and LOD > 2; red, D' = 1 and LOD > 2. (D' is a pairwise measure of linkage disequilibrium)

1.4. Copy number variants

Copy number changes (also referred to as chromosomal rearrangements) are mutations in the genome involving new juxtapositions of chromosomal parts and are caused by unequal crossovers during replication as well as breakages (e.g.: translocation). The fragment parts of chromosomal regions may form a novel chromosomal arrangement of DNA sequence or they may rejoin in the same arrangement as before the breakage occurred, through the action of DNA repair mechanisms. There are two different categories of rearrangements: balanced and unbalanced, referring to the maintenance of the quantity of genetic material or a

change in copy number of genetic material, respectively. Balanced chromosomal rearrangements include inversions and reciprocal translocations and unless these involve breakages that disrupt normal gene function, they are not usually pathogenic. Unbalanced rearrangements are usually pathogenic as normal gene balance is disrupted through copy number changes that include insertions (duplications) or deletions. The effect of the mutation depends on the size and nature of the imbalance caused. Copy number variants (CNVs), which are structural variants in the genome that result in a gain or loss of genetic material, usually range from 1kb to several megabases in length and are currently being detected at a higher rate than ever before. Fifteen percent of the assembled human genome sequence is involved in CNVs. These types of variants are currently being included in many different types of studies and rare and common CNVs are being found to be involved in complex disorders (Database of genomic variants: http://projects.tcag.ca/variation/) (Estivill and Armengol 2007).

1.4.1 Origin of rearrangements

Allelic homologous recombination (AHR) is the normal crossing over that occurs during meiosis and gives rise to genetic diversity in the population. When repetitive segments within a single chromosome pair, or on different chromosomes act as substrates for illegitimate crossing over, non-allelic homologous recombination (NAHR) takes place. It is proposed that chromosomal rearrangements arise in the genome through these unequal cross-over events. The non-reciprocal exchange of DNA between duplicated sequences that have recombined is one of the mechanisms that can give rise to regions with a high level of sequence similarity and therefore perpetuate NAHR (Blanco et al. 2000; Papadakis and Patrinos 1999). Through this process, a cycle occurs, whereby paralogous sequences give rise to even more duplicated sequences with even higher sequence similarity, until the similarity between paralogs is higher than between orthologs. This homogenization process, which usually occurs within gene families, is called concerted evolution. Some examples of concerted evolution giving rise to more sequence similarity between paralogs, which may encourage further duplicated sequences within or between chromosomes and thus

increasing the incidence of chromosomal rearrangements are: NEMO and LARGE2 genes (Aradhya et al. 2001); on the Y-Chromosome AZFα region (human and ape genomes studied) (Hurles et al. 2004); CMTIA paralogus repeats (Hurles 2001) and many others. Hurles et al. 2004 confirms that the most likely passage of origin of chromosomal rearrangements begins with duplicated sequences in the genome that diverge through mutation and undergo gene conversion. Since homologous (allelic) recombination and non-homologous (non-allelic) recombination are determined by the presence of nucleotide sequence similarity between the parental sites of recombination, *Alu*, LINE-1 and SINE repeat sequences have been known to give rise to homologous recombination. Recombination can also occur in the absence of homologous sequences. This type of recombination can be produced by 'recombination-promoting motifs'. Alternating purine or pyrimidine tracts as well as polypyrimidine tracts are know to induce genomic instability and are often found in the vicinity of deletions (Abeysinghe et al. 2003; Bacolla et al. 2004; Feréc et al. 2006).

1.4.2 Rearrangements in the CFTR gene

Chromosomal rearrangements are often complex and exhibit extensive allelic heterogeneity (Audrezet et al. 2004). This is no different for the rearrangement mutations that have been found in the CFTR gene (Figure 1.4). In a study done by Audrezet et al. (2004) (Table 1.5) a frequency of 10-20% of deletions or insertions was found in CF chromosomes that previously had unknown mutations. It was concluded that "complex rearrangements [in CF] are far more frequent than previously expected". Similarly, Hantash et al. (2004) concludes that large chromosomal rearrangements may represent a gross underestimation of the total contribution of this type of mutation to the CF gene pool.

In a study of rearrangements in the CFTR gene, Audrezet et al. (2004) found that intron 3 and intron 18 of the CFTR gene are deletion breakpoint hotspots, but contain a low number of *Alu* repeats, which suggests that homologous recombination is not the predominant mutational mechanism giving rise to rearrangements in the CFTR gene. Similarly, Feréc et al. (2006) found that

homologous recombination does not appear to generate any of the deletions that have thus far been found in the CFTR gene (Table 1.5), as Alu, LINE-1 and SINE repeats are not commonly found in the vicinity of CFTR gene deletion breakpoints. This study also revealed that recombination-promoting motifs that give rise to non-homologous recombination (Abeysinghe et al. 2003), especially polypurine tracts, are the most frequently encountered motifs in the vicinity of CFTR deletion breakpoints. This study also revealed that CFTR deletion breakpoint junctions tend to occur in regions of low complexity (where sequence complexity was measured with reference to direct and inverted repeats as well as symmetric elements). Both studies conclude that diverse mutational mechanisms are implicated in the rearrangement mutations that have been found in the CFTR gene.

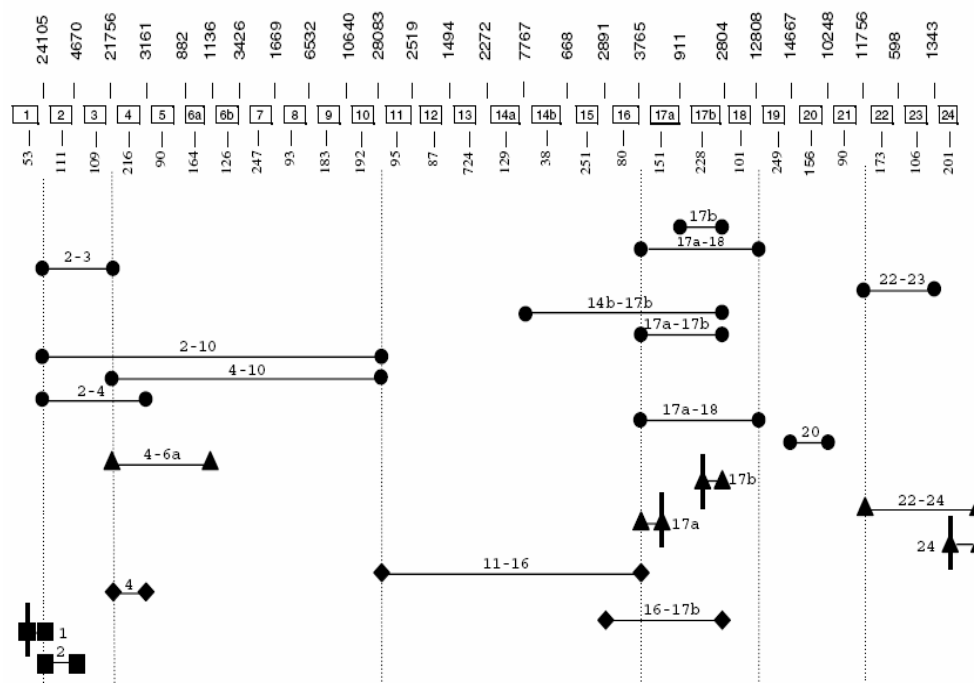


Figure 1.4: Schematic diagram of fully characterized gross CFTR genomic rearrangements involving deletions (Feréc et al. 2006)

Key:

● simple deletions with short direct repeats at 50 and 30 breakpoints
■ Complex deletions with large insertions of 4100 bp
▲ complex deletions with short insertions of 3–6 bp
◆ complex deletions with small insertions of 32–41 bp

Key for Figure 1.4 (continued):
Upper panel: Genomic structure of the CFTR gene
Numbers above and below denote the sizes (bp) of the introns and exons respectively
Lower panel: Fully characterized large genomic rearrangements involving deletions of the CFTR gene
Vertical bars: Breakpoints have occurred within coding sequences

Table 1.5: Summary of large chromosomal rearrangements in the CFTR gene

Deletion	Ethnicity	Reference
Exons 4-7 and 11-18	Spanish Caucasians	Morral et al. 1993
703bp deletion at the 5' end of exon 17b	Italian Caucasians	Mangani et al. 1996
Exons 14b-18	Caucasians	Shrimpton et al. 1997
Exons 4-10	Caucasians	Chevalier-Porst et al. 1998
86Kb spanning exons 17a, 17b and 18	Palastinian	Lerer et al. 1999
21kb spanning introns 1-3	Slavic Caucasians	Dörk et al. 2000
Introns 17b-18 with short Local duplications at the breakpoints	Unknown	Kilinic et al. 2002
Promoter, exons 1 and 2 Promoter and exon 1	African American African American	Hantash et al. 2004
Exon 4	French Caucasians	Audrezet et al. 2004
Exon 1	French Caucasians	Audrezet et al. 2004
Exons 4-6a	French Caucasians	Audrezet et al. 2004
Exons 11-16	French Caucasians	Audrezet et al. 2004
Exons 22-23	French Caucasians	Audrezet et al. 2004
3120+1kbdel8.6kb	Italian Caucasians	Bombieri et al. 2005
Exons 17a-18	Reunion Islands Caucasians	Nectoux et al. 2006
Exons 2-4	Irish Caucasians	Férec et al. 2006
Exons 17-18	French Caucasians	Férec et al. 2006
Exon 20	Spanish Caucasians	Férec et al. 2006
IVS16-908_c.3085del1005insGACAG	French Caucasians	Férec et al. 2006
c.4344_Stop486del585insTTG	Spanish Caucasians	Férec et al. 2006
Exon 1	Czech Caucasians	Férec et al. 2006
Complete deletion of CFTR gene	Italian Caucasians	Férec et al. 2006
Duplication	**Ethnicity**	**Reference**
Exons 7-10 Dup 26kb	American Caucasian	Hantash et al. 2007

The most frequently occurring CFTR chromosomal rearrangement worldwide is the 21kb deletion described by Dörk et al. (2000), which removes exons two and three. This mutation appears, on the basis of haplotype analysis, to have a common origin in Central and Eastern Europe. The remaining copy number mutations seem to be restricted to certain populations. Hantash et al. (2004) describe a deletion of the promoter and exons 1 and 2 of the CFTR gene in two African American families (Table 1.5). In the past, findings in African Americans have been extrapolated to SA Black and even Coloured patients with great success. In this study I aim to determine whether rearrangement mutations account for a proportion of the high number of unidentified alleles in Black and Coloured SA CF patients.

1.4.3 Detection of chromosomal rearrangements

Balanced chromosomal rearrangements may be detected through sequencing of breakpoints, if a particular region is implicated, as they do not cause a change in copy number of a particular sequence. Copy number changes are challenging to detect in a laboratory setting. This is because PCR-based methods amplify the normal allele, in the case of a heterozygous copy number change, which masks the effect of the mutated allele, and thus the mutated copy is simply invisible when using such techniques. Despite these difficulties, there are a number of techniques that may be used to detect copy number changes: comparative genomic hybridisation (CGH) (Kallioniemi et al. 1992); fluorescent *in situ* hybridisation (FISH) (Pinkel et al. 1986); BAC arrays (Snijders et al. 2001); Quantitative real-time PCR (ORT-PCR) (Casilli et al. 2002); loss-of-heterozygosity (LoH) assays (an assay that makes use of closely spaced markers to determine if an inordinate number of markers are homozygous, which would serve as an indication to investigate a region further for deletions) (Devilee et al. 2001) and multiplex ligation-dependant probe amplification (MLPA) (Schouten et al. 2002) (please refer to Figure 1.5 below for a summary of techniques that may be used to detect CNVs. The figure shows that MLPA has relatively low level of resolution when compared to other techniques). Once the rearrangements have

been found, their breakpoints are usually detected by long-range PCR (if the region is too big to amplify through traditional PCR) and sequencing.

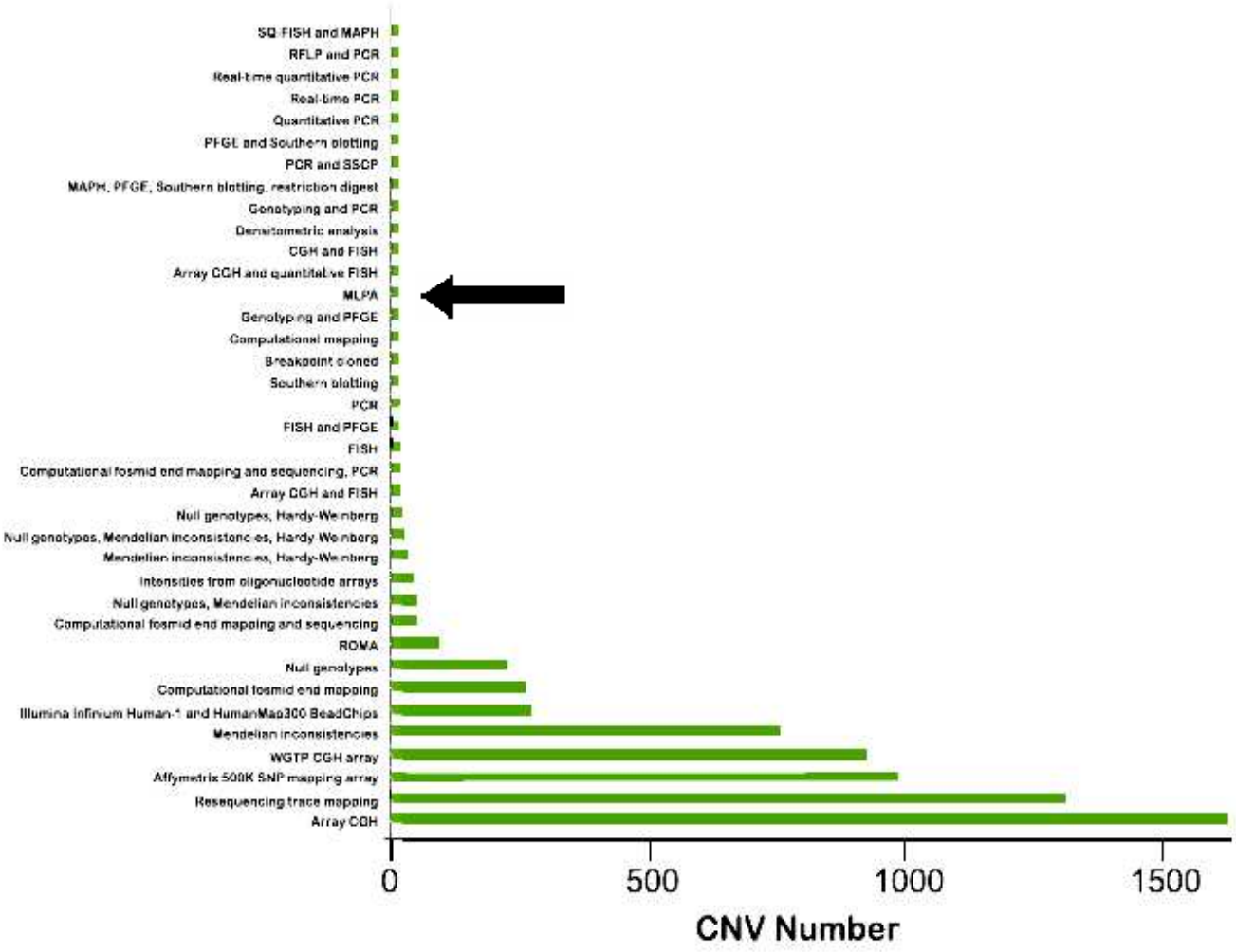


Figure 1.5: Approaches used for the identification of CNVs and other types of structural changes in the human genome. Many methods and technologies provide very different levels of resolution. The majority of findings (80%) are attributable to a restricted number of highthroughput experiments with a limited resolution (Estivill and Armengol 2007). (The arrow in the diagram points out MLPA)

1.4.4 Multiplex ligation-dependant probe amplification

MLPA is a multiplex, PCR-based mutation detection method, which effectively establishes exon copy number (Schouten et al. 2002). A kit for the detection of insertions and deletions in CFTR gene exons will be used in this project.

1.5. Aims of the study

It is proposed that South African Black and Coloured populations have a unique CFTR mutation profile. In order to investigate this hypothesis the following study objectives will be completed:

1.5.1 Use haplotypes to determine if the 3120+1G>A mutation has a single evolutionary origin in Black and Coloured populations

1.5.2 Use haplotypes to predict the frequency of mutations among unidentified CFTR alleles

1.5.3 Identify exon copy number variants in Black and Coloured SA patients with a clinical phenotype of CF and at least one unidentified CFTR mutation

In Chapter 2, the subjects and controls tested and the materials and methods used for genotyping and mutation detection will be discussed. The results of these tests are displayed in Chapter 3 and the implications of these results are discussed in Chapter 4.

Chapter 2: Subjects and Methods

2.1. Subjects

2.1.1 Patients and families

For copy number mutation detection, patients were chosen based on three selection criteria: they had to be Black or Coloured and South African; have a confirmed clinical diagnosis of CF (at least one positive sweat test or clinical features indicative of CF, as discussed with patient doctors: Dr Westwood, Dr Klugman and Dr Henderson) and a CFTR genotype with at least one unknown CFTR mutation. Out of approximately 1000 patients, for whom CFTR mutation detection had been requested, in the CF database at the National Health Laboratory Service (NHLS), 42 met all three selection criteria (24 Coloured CF patients and 18 Black CF patients). In order to study the origin of the 3120+1G>A mutation, patients with at least one copy of this mutation and family members were collected for the purpose of the haplotype phasing (seven Coloured family members and two Black family members). These family members were tested for their 3120+1G>A carrier status, in the cases where their offspring carried at least one copy of this mutation. Please refer to Appendix A for a summary of the infor124mation related to the subjects tested in this study as well as the ethics approval documentation.

2.1.2 Controls

The Black control samples that were used in this project had been collected before the commencement of this project and included samples from Johannesburg, Gauteng. These controls are representative of the Black population in Gauteng which is a mixture of different SA chiefdoms. The Coloured controls were selected from a Western Cape Coloured sample collection that was available in the laboratory, owing to the fact that a large proportion of the Coloured CF patient samples tested in this project originated from the Western Cape. Carrier testing of

controls for the 3120+1G>A mutation was done for both the Black and Coloured controls collected (51 Black and 50 Coloured controls were collected) (refer to Appendix A for ethics approval documentation).

2.2. Genotyping for haplotype inference

2.2.1 Carrier testing of controls and families

The 3120+1G>A mutation abolishes a B*st*NI (NEB) restriction enzyme recognition site, resulting in PCR fragment sizes of 537 and 33 bp in the presence of the mutation and 340, 197 and 33 bp fragments when the mutation is absent. The NHLS Division of Human Genetics: Molecular Genetics diagnostic protocol for testing of the 3120+1G>A mutation was used (Appendix B).

2.2.2 Selection of markers for genotyping and haplotype study

Markers were chosen that had previously been published in haplotype studies and would provide useful information for comparison. Additional markers were selected based on allele frequencies and LD relationships across the CFTR gene in the YRI population from the HapMap study.

2.2.2.1 Identification of 'rs' numbers for markers

It was necessary to identify the 'rs' numbers for markers in order to extract SNP genotype information from the HapMap database. Depending on what information was available for each of the markers, different avenues were explored to identify the correct 'rs' numbers for each of the SNPs. Previously published marker region and primer sequences were used for the markers Met D, XV-2C, KM 19, MP6-D9, J44, M470V, T854T and TUB 18 (Dörk et al. 1992; Horn et al. 1990) to run an *in silico* PCR reaction (www.genome.UCSC.org). This identified the genome co-ordinates for the sequence and was hyperlinked to an image of the region that showed all the 'rs' numbers for the polymorphic variants (SNPs, STRs, single base insertions and deletions) in that region. Detailed information and the sequences flanking the variant were identified. The restriction enzyme recognition sequence then allowed for the correct 'rs' number

to be matched to the marker names. In the absence of known primer sequences, the 'rs' number was found by searching ENSEMBL (www.ensembl.org/) for the particular genome co-ordinates for the region of interest, for example, 'intron nine of the CFTR gene', and these co-ordinates could either be viewed in UCSC or they could be searched in ENSEMBL.

2.2.2.2 Downloading information from HapMap

The genotype information that was downloaded from the HapMap site was obtained by searching for a particular 'rs' number and then downloading the SNP genotype data for defined populations (NCBI b36). SNP genotype information for YRI and the CEU populations was then saved as a text file. The following information was included: the marker ID (that is the 'rs' number); the observed as well as predicted heterozygosity; the Hardy-Weinberg (H-W) equilibrium p-value; the percentage of non-missing genotypes for a particular marker; the number of fully genotyped family trios for the marker; the number of observed Mendelian errors; the minor allele frequency (less frequent allele in a particular population) and the minor allele.

2.2.2.3. Analysis in HaploView

The marker genotype information obtained from HapMap for each SNP was consolidated into a single file that was uploaded into HaploView version 3.32. The program has thresholds for quality metrics so that it may filter out data that is deemed sub-standard quality. These thresholds may be adjusted by the user (default settings are also available): a H-W p-value cut off of 0.1 was set, meaning that the program should designate all markers as "bad" if the probability that their deviation from H-W could have been explained by chance was less than 10%. A minimum percentage of non-missing genotypes of 75% was set, as well as a maximum of 1 Mendelian error and a minimum minor allele frequency (MAF) of 0.10 (10%). HaploView generates blocks of LD whenever a file is uploaded but these blocks can be edited and redefined based on one of several block definitions and this partitions the genomic region of interest into segments of strong LD. The two most common pairwise measures of LD are D prime (D') and r^2. For this

study the D'/LOD score block definition (Gabriel et al. 2002) was used, which calculates a pairwise measure of LD (a block is created if 95% of informative [i.e. non-inconclusive] comparisons are in 'strong LD') and the log of the likelihood odds ratio (a measure of the confidence in the value of D'). This method allows the user to view the LD relationships between the SNPs in the LD plot in the "standard colour scheme" (Table 2.1).

Table 2.1: Standard colour scheme

	D' < 1	D' = 1
LOD < 2	white	blue
LOD >2	Shades of pink or red	bright red

This method, by default, ignores markers with a MAF of less than 0.05. Haplotypes for selected blocks are also calculated by the program. Haplotype phase as well as population frequency are inferred using an accelerated expectation-maximisation (EM) algorithm (Excoffier and Slatkin 1998; Long et al. 1995) with a partition-ligation approach (Qin et al. 2002) for blocks with more than 10 markers. Conformance with Hardy-Weinberg Equilibrium in the program is computed using an exact test. The haplotype display shows each haplotype in a block with its population frequency and connections from one block to the next. The level of recombination between the blocks is also shown. Markers with a high level of heterozygosity and a low level of redundance were selected by a program to which HaploView acts as an interface. Tagger, as the software for this program is called, combines the simplicity of pairwise r^2 methods (Chapman et al. 2003; Carlson et al. 2004) with the potential efficiency of multimarker haplotype approaches (HapMap consortium 2005). Tagger recognises which markers are in LD and aggressively searches for effective multimarker predictors to capture the alleles of interest, without selecting markers that are perfectly (the pairwise r^2 between such SNPs is 1), or highly correlated. Aggressive or pairwise tagging may be selected by the user. Aggressive tagging was used in this project to capture as many tags as possible (SNPs that could not be captured in the pairwise method are captured when aggressive tagging is used). There are fewer of these perfectly comparable (and therefore redundant) SNPs in YRI data as compared to

CEU data, adding weight to the hypothesis that African populations are the oldest extant populations on earth because of the increased variation induced through mutation and recombination that takes place over many generations. HaploView allows the user to view which SNPs are tagged within the blocks that are partitioned. In the cases where there were more than one tagged SNP in a particular block, redundance was further minimised by selecting the marker with the highest heterozygosity. YRI and CEU SNP genotype data were used as proxies for African and European populations. The CEU linkage information across the CFTR gene and across the entire genome shows fewer, but longer blocks of tightly linked variants (haplotype blocks) and less variation in general. For this reason, the YRI data was primarily used when choosing markers to include in this study. The markers chosen were: rs11770163, rs916727, rs43035, rs3802012, rs1042077 and rs213989.

2.2.3 Pyrosequencing

The markers that were chosen for genotyping for haplotype analysis were genotyped using Pyrosequencing. This is a technique based on sequencing by synthesis, through the detection of nucleotide incorporation, using a primer-directed polymerase extension. The laboratory protocol is outlined in Figure 2.1 below and the details of these methods are discussed.

Pyrosequencing: Flow diagram of method

Assay design

Pyrosequencing Assay Design Software was used to design the PCR primers and the PSQ sequencing primer.

⇩

PCR

Amplification of the target sequence was done by using the primer sequences determined above with the sequence containing a universal 'tail'. A universal biotinylated primer was included in the PCR assay so that the PCR product could be used for subsequent immobilisation of biotin-labelled DNA to sepharose beads.

⇩

Immobilisation of target strand

Immobilisation of the biotinylated PCR products onto the streptavidin-coated beads precedes the capture of the beads onto the probes of the vacuum prep tool for washing and neutralisation of the immobilised strand.

⇩

Annealing of sequencing primer

The vacuum prep tool probes, to which the immobilised biotinylated PCR product was attached, were exposed to sequencing primer and annealing buffer and the PCR product was released so that it binds the sequencing primer, which binds adjacently, or very near to the target variant sequence.

⇩

Sequencing by synthesis

The DNA template was incubated with the enzymes, DNA polymerase, ATP sulfurylase, luciferase and apyrase, and the substrates: adenosine 5´ phosphosulfate (APS) and luciferin. The first of four deoxynucleotide triphosphates (dNTPs) was added to the reaction. DNA polymerase catalyses the incorporation of the dNTP into the DNA strand, if it is complementary to the base in the template strand. Each incorporation event is accompanied by release of pyrophosphate (PPi) in a quantity equimolar to the amount of incorporated nucleotide. ATP sulfurylase quantitatively converts PPi to ATP in the presence of adenosine 5´ phosphosulfate. This ATP drives the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that are

proportional to the amount of ATP. The light produced in the luciferase-catalysed reaction was detected by a charge coupled device (CCD) camera and seen as a peak in a pyrogram™. Each light signal was proportional to the number of nucleotides incorporated. Apyrase, a nucleotide degrading enzyme, continuously degrades unincorporated dNTPs and excess ATP. When degradation is complete, another dNTP is added (Ronaghi 2000).

⇩

Haplotype analysis

Genotypes for patients were phased by hand according to homozygosity and family data.	Genotype data for controls was phased using the PHASE software program.

Figure 2.1: Outline of pyrosequencing and analysis

2.2.3.1 Assay design using the Pyrosequencing™ Assay Design Software

The target region sequence for each of the six markers was imported into the "Sequence Editor" tab in the "Assay" window. This sequence was obtained from dbSNP and contains the sequences flanking the variant. The assay type was selected as "genotyping" and therefore the analysis steps and the primer set scoring was automatically set according to this chosen application. The SNPs in the sequence, entered in FASTA format, were noted with IUPAC codes. The default target region of the sequence was then set to the first polymorphism in the sequence as there was only one polymorphism per sequence for all the SNPs genotyped in this project. Every primer set was assigned a score and quality index, which reflects the suitability of that primer set for use both in PCR and for pyrosequencing analysis. The primer sets were, by default sorted according to this rating, so that the best primer set was at the top of the list. The primers chosen were scored in blue indicating that they were of high quality and had a score greater than or equal to 88. Assay design settings were not modified for any of the markers used in this project. The assay report contains detailed information on the primers including: GC content (percentage); complementarity; duplex

formation; hairpin loops; melting temperature; mispriming; primer end stability; primer length as well as detailed information on the primer pairs, including: amplicon length; duplex formation; CG content difference; melting temperature difference.

2.2.3.2 Amplification of relevant region by PCR

Polymerase chain reaction (PCR) was used to amplify the target region for genotyping using pyrosequencing. The details of the PCR reactions are shown below. For immobilisation of single stranded PCR product, each reaction included a biotinylated universal primer. As a cost saving feature, rather than biotinylating one of the primers for each marker, one of the primers was synthesized with an additional sequence to which the biotinylated universal primer binds (Table 2.3). The reverse primer sequence (for rs11770163, rs916727, rs43035 and rs3802012) and the forward primer sequence (for rs1042077 and rs213989) included a complementary binding sequence for the universal primer. This universal primer was included in every reaction.

The products were amplified in the ABI GeneAmp® PCR system. PCR cycling conditions were as follows:

95°C; 5 minutes
95°C; 15 seconds
[Annealing temperature appropriate to SNP]; 30 seconds } 45 cycles
72°C; 15 seconds
72°C; 15 minutes

The general PCR mix protocol shown in Table 2.2 below was followed.

Table 2.2: General PCR mix protocol

PCR Components	Volume	Stock concentration	Final concentration
DNA	1µl	50ng/µl	2ng/µl
dNTPs	2.5µl	1.25mM	0.125mM
PCR Gold Buffer (Applied Biosystems, Roche)	2.5µl	10 x	1x
$MgCl_2$ (Applied Biosystems, Roche)	2.5µl	25mM	2.5mM
Universal biotinylated primer	0.5µl	10mM	0.2mM
Forward primer (Inquaba)	0.5µl	10mM	0.2mM
Reverse primer (Inquaba)	0.5µl	10mM	0.2mM
Amplitaq Gold (Inquaba)	0.2µl	5U/µl	0.04U/µl
Distilled H_2O	14.8µl		
TOTAL	25µl		

For rs1042077 and rs11770163, the above PCR mix was optimised to include 0.75µl for all of the primers. The PCR products were visualised and verified on a 3% agarose gel (run at 4 V/cm) and the successfully amplified samples were sequenced by using pyrosequencing as described below. Every PCR and pyrosequencing reaction included the following negative controls according to the manufacturers (Biotage) recommendations: i) PCR positive control (irrelevant genotype as no PSQ reagents are added and therefore this control is a PSQ negative control); ii) Sequencing primer only; iii) PCR negative control (to which sequencing primer is added); iv) PCR negative control (to which sequencing primer is not added); v) Annealing buffer only.

The genome positioning of the markers as well as the PCR primer information are shown in Figure 2.2.

Table 2.3: Primers used for amplification of products for pyrosequencing

SNP		Primers (5' - 3')	Score	Annealing temperature
Met D rs11770164	Forward Reverse Sequencing	TGT TTT GCA AGT GAC ATT TCT G GAC GGG ACA CCG CTG ATC GTT TAG CCT CTG GGT AAA ATG AGT CA ACT ATT GGA AGC ATG TTG	94	60
Km 19 rs916728	Forward Reverse Sequencing	TTT CGC TCG GTT CCC TAA A GAC GGG ACA CCG CTG ATC GTT TAA ACA TAC CAA GAA AGC CAA GAG A TGA CAT TTC ATA GAG CCT	92	51
J44 rs43035	Forward Reverse Sequencing	GAT TTT GCT TCA TTA GCA TGA TAT GAC GGG ACA CCG CTG ATC GTT TAT GGC AAA ATA CTT TGG AAC TCT TA AGC ATG ATA TAC TTT GTT CT	88	57
T854T rs1042077	Forward Reverse Sequencing	GAC GGG ACA CCG CTG ATC GTT TAA ACC ACA ATG GTG GCA TGA A CTC TGC CAG AAA AAT TAC TAA GCA AAA TTA AGC TCT TGT GGA C	91	53
TUB 18 rs213989	Forward Reverse Sequencing	GAC GGG ACA CCG CTG ATC GTT TAT AAG AAT GGC AGA CAA TTT CAC A GCC CGA CAA ATA ACC AAG TGA TTT ACA AGT TAT TTT TTA GG	89	57
J32 rs3802012	Forward Reverse Sequencing	GGA AAT TAT CTA GTA TGT CTT TCA GAC GGG ACA CCG CTG ATC GTT TAA TCA GCA CCT ACA CAG CAA TAA TCT TTC ACA AAA TTC TAT AA	72	55

Key

Red text — Indicates which primer contains the additional sequence complemetary to the universal primer

Score — PSQ Assay design software score (percentage)

Annealing temperature — Calculated by gradient PCR (°C) experiments

Universal primer sequence: 5' GAC GGG ACA CCG CTG ATC GTT TA 3'

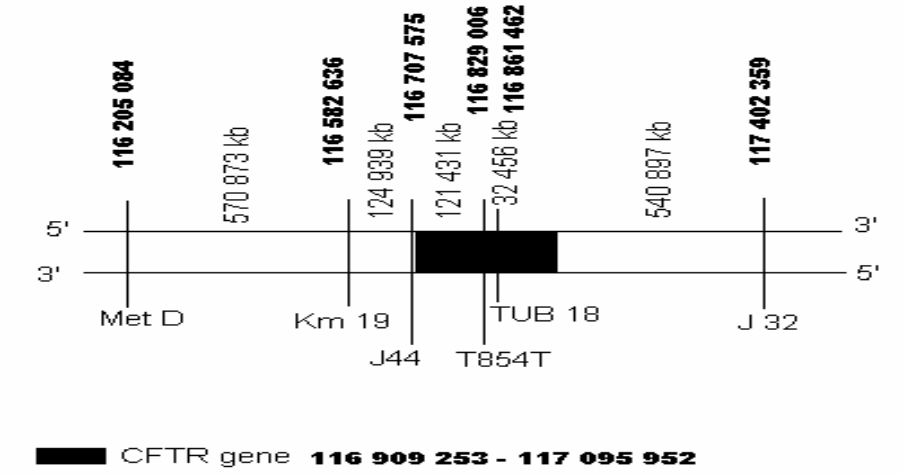


Figure 2.2: Genomic positioning of intragenic and extragenic SNPs genotyped in this project. Bolded numbers indicate genome co-ordinates (NCBI b36)

2.2.3.3 Immobilisation of biotinylated PCR products

The streptavidin-coated sepharose beads (streptavidin sepharose HP, Amersham Biosciences) were used for the immobilisation of PCR products at room temperature. The streptavidin sepharose beads were mixed with binding buffer to a volume equivalent to 40µl per sample and this was added to the amplified template that was mixed with ddH2O. The samples were then allowed to incubate by agitation for five to ten minutes.

2.2.3.4 Wash or neutralise the immobilised strand

The workstation troughs were filled with different solutions (Appendix C). Seventy percent ethanol (180ml) in trough 1; denaturation solution (120ml) in trough 2; washing buffer (180ml) in trough 3 and high purity water (180ml) in trough 4. Vacuum was applied to the Vacuum Prep Tool and the probes were washed by lowering it into the water trough for approximately 20 seconds. The Vacuum Prep Tool was slowly lowered into the PCR plate (or strip tubes) containing the PCR product. This process was performed within three minutes of immobilisation of the biotinylated PCR products onto the streptavidin-coated beads. When all the liquid in the wells was aspirated and the beads were captured

onto the probe tips they were slowly lowered into each of the troughs in numerical order as indicated above for five seconds (except for the water trough).

2.2.3.5 Anneal sequencing primer

The vacuum was released and the beads were released into the PSQ 96 Plate Low (Biotage) that was pre-filled with 0.4µM of sequencing primer and 40µl of Annealing buffer. This plate was heated to 80°C for two minutes (for all SNPs except Km 19, which was heated to 60°C for 5 minutes) on the PSQ 96 Sample Prep Thermoplate Low. Thereafter the samples were allowed to cool to room temperature. Refer to Figure 2.3 below for a summary of this process and an approximation of the time taken to complete these steps.

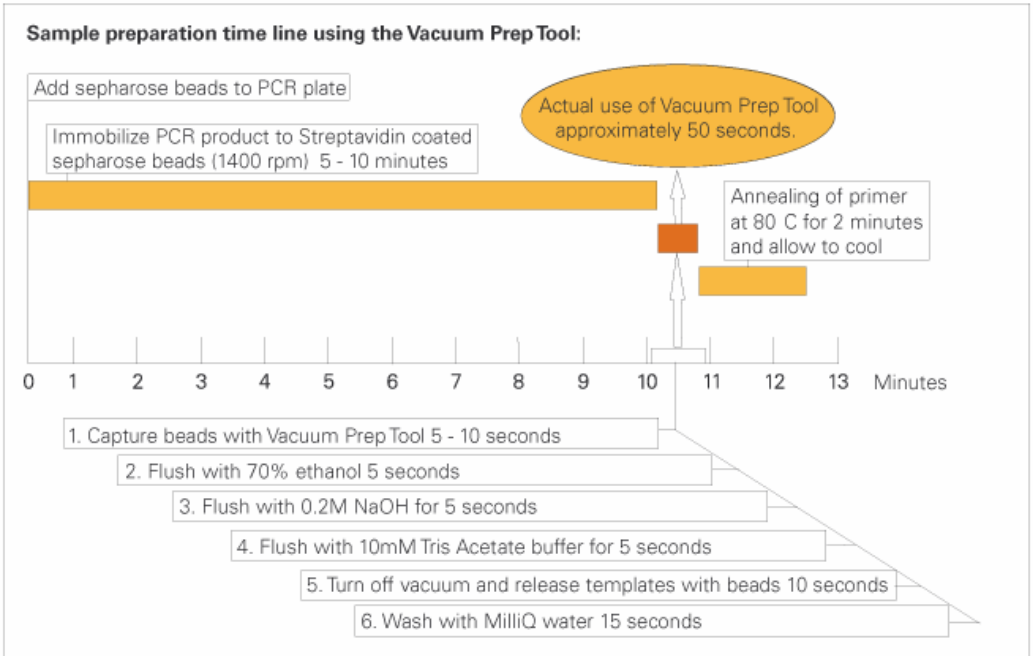


Figure 2.3: Systematic diagram of sample preparation time after PCR products have been immobilised to Sepharose beads http://www.pyrosequencing.com/graphics/2983.gif.

2.2.3.6 Setting up a run on the PSQ 96 Software

The run name, plate ID and instrument parameters as well as the sample information was entered into the Run setup windows of the software. In the "Web Browser" area, which is below the "Run Setup" window in the PSQ 96 Software, the run could be viewed. It is in this run-view that the recommended

volumes of substrate, enzyme and nucleotides are shown for pipetting into the reagent cartridge. These volumes are specific for each run.

2.2.3.7 Cartridge preparation

The clean and dry cartridge was filled with substrate and enzyme mixes as well as dNTPs, according to the volume recommendations which are specified for each run. The reagent cartridge was placed with the label facing forward and the reagents were pipetted according to the manufacturer's recommendations. Once the cartridge was prepared, the PSQ 96 run was started by loading the cartridge and plate into the instrument.

2.2.3.8 Post-run procedures

The well overview is a graphical representation of the activated runs and shows the progression of the runs by indicating what action is taking place in each well at a given time through the display of various colours. When the run ended, the "Close" button was selected in the instrument status window. This automatically caused the run to close and to be saved into the folder from which it was set up. To view the results of this run, "Analyse All" was selected once it had opened. This caused the PSQ 96 Software to analyse the run using default analysis settings and the results were saved.

2.2.4 Data analysis

2.2.4.1 Haplotype inference

2.2.4.1.1 Manual Phasing

The transmission of alleles from parents to offspring can sometimes be ambiguous in that the phase of the alleles is not always known. If all the loci were homozygous in an individual, or if only one of the loci was heterozygous, then the two haplotypes could be assigned without ambiguity. The remaining haplotypes were phased according to family data, where possible. The haplotypes that could not be phased according to homozygosity or family data were inferred: the most

likely phase was determined to be compatible with the known unambiguous mutation associated haplotype.

2.2.4.1.2 PHASE

The haplotype structure in the SA Black and Coloured non-CF populations, representatives of which were genotyped in this study, was compared using a cohort of 51 Black and 50 Coloured unaffected individuals. The inferred allelic phase haplotype of these unrelated controls was deduced through a Bayesian statistical approach using the software package PHASE v 2.1.1 (I would like to acknowledge Dries Oelefse for running the data through this software on my behalf). This method approximates the coalescent (evolutionary history) of the alleles through specialised algorithms. The genotype data were converted into the format required by the program and were run using default settings. Genotype data sets that contained gaps (markers that did not have a genotype result) were included in each run and the results of these analyses were used. The software was used to test datasets with zero gaps only (control samples with gaps were removed) and datasets with control samples with two or more gaps were removed. Analysis of these results in comparison to the analysis results of the datasets with the gaps showed similar results, and for this reason the datasets with gaps were used, so as to minimise the loss of information that could occur through the exclusion of samples.

2.2.4.2 Statistical analysis of data

The Hardy-Weinberg law is a fundamental principle in population genetics stating that the genotype frequencies and allele frequencies of a large, randomly mating population remain constant provided that gene flow, mutation and selection do not take place. The Hardy-Weinberg (H-W) Principle can be used to calculate the expected frequency of genotypes. Testing deviation from H-W equilibrium is generally performed using Pearson's chi-squared test (Pearson 1903), using the observed genotype frequencies obtained from the data and the expected genotype frequencies obtained using the H-W Principle. In some cases, for instance where there are a large number of alleles and not enough individuals in the population

(small sample size) to adequately represent all the genotype classes, it may be necessary to test if the population is in H-W equilibrium by using Fisher's exact test (probability test), which requires a computer to solve (Guo and Thompson 1992; Wigginton et al. 2005). In this project the Fisher's exact test was used to test if the control populations were in H-W Equilibrium. The exact test was calculated using Powermarker v3.25 (Liu and Muse 2005). Please refer to Appendix B for further information on the calculations used for these tests.

2.3. Detection of rearrangements using MLPA

MLPA is a multiplex, PCR-based mutation detection method, which effectively establishes exon copy number. The technique is divided into three parts: Hybridisation of probe pairs to target sequences; ligation of probe pairs and finally amplification of the ligated probe. Up to 45 pairs of probes are hybridised to target DNA, ligated and amplified in a single reaction using universal primers, which are fluorescently labelled. The fluorescence of each band is detected and quantified and compared to controls. The CFTR P091 kit (MRC-Holland; www.mrc-holland.com) contains 44 different probes with amplification products between 130 and 463 base pairs (Schouten et al. 2002) (Appendix E). The use of the kit requires a thermocycler with a heated lid as well as sequencer-type electrophoresis equipment. Heterozygous deletions or duplications of probe recognition sequences will result in an apparent 35-50% reduced or increased relative peak area of the amplification product of that probe (Figure 2.4). The relative amount of probe hybridisation and ligation is dependant on the copy number of the target sequence. For example, in the case of a duplication, where there is increased gene dosage, there will be an increase in target sequence and consequently more amplification of products. This will be reflected in a higher probe peak relative to the peak for a control with a diploid number of copies of the gene (usually two copies).

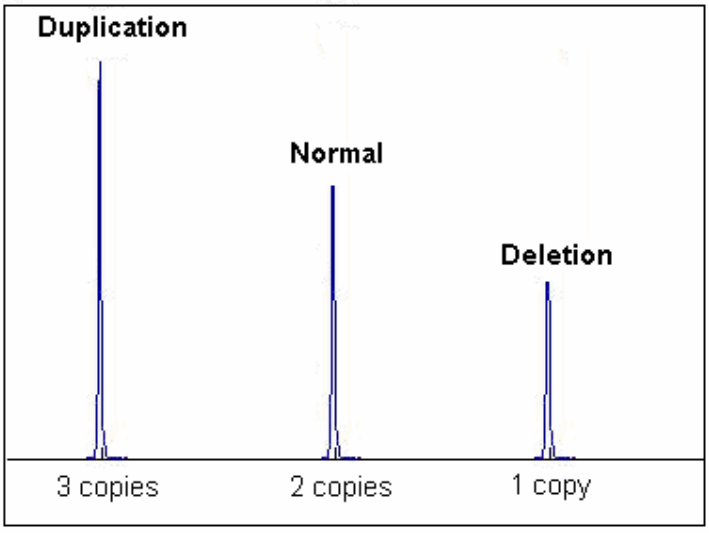


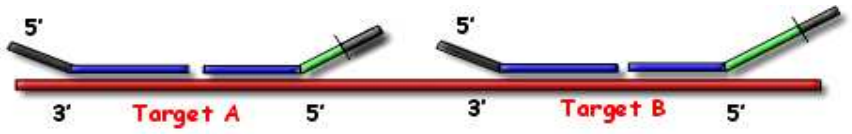
Figure 2.4: MLPA peak picture depicting a 35-55% increase and decrease of a single target exon in a heterozygous state as well as the normal diploid dosage (2 copies). The relative amount of amplification depends on the copy number of the target region in the sample

2.3.1 Principle of technique

Many pairs of probes hybridise, in a multiplex reaction, to a unique region of the genome and will detect the copy number thereof. The probe pair is made up of a synthetic probe, consisting of a primer sequence which is complementary to the universal primer and a hybridisation sequence which is complementary to a target sequence on the genomic DNA. The probe pair is also made up of a single stranded M13-derived probe, which also contains a primer sequence that is complementary to the universal primer; a stuffer sequence as well as a hybridisation sequence (Figure 2.5).

Figure 2.5: Model of a single probe pair for MLPA) (http://www.mlpa.com/pages/support_mlpa_infopag.html)

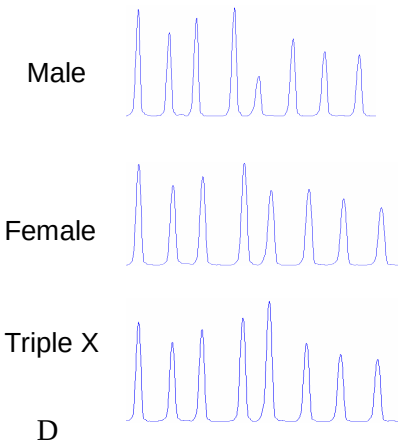
The M13 derived probes in the multiplex reaction differ in terms of their hybridisation and stuffer sequences (shown in green in Figure 2.6). The stuffer sequence serves only to lengthen the M13 derived probe so that the amplification products may be distinguishable on a polyacrylamide gel through size separation. Many target regions of DNA may be investigated simultaneously, using the universal PCR primer pair once the probe pairs in the mix have been ligated. All resultant fragments differ by six to eight base pairs.



A

B

C



D

Figure 2.6: Depiction of events on a molecular level, during the MLPA technique. A) Hybridisation of probe, each with a longer stuffer sequence to target sequences in the genome. B) Ligation of the directly adjacent probe pair. C) PCR products are separated due to their size differences. D) Products are visualised as peaks (http://www.mlpa.com/pages/support_mlpa_infopag.html)

2.3.2 Principle of analysis of MLPA amplification products

Dosage data are quantitative and can therefore be difficult to analyse. It is for this reason that dosage quotients and statistics are used. Before the details of the analysis are discussed, it is important to understand why MLPA data is a challenge in terms of analysis: Firstly, because of inter-run variation. This means that the peak pattern of the same sample may change slightly from run to run. Secondly, longer fragment peaks may wane off, that is, fragments longer than 300-400bp may suffer from decreased peak signals and lastly, MLPA is susceptible to contaminants and therefore there may be preferential amplification of certain probes, which will result in an artificial result for that sample. These caveats will be discussed in more detail in Chapter 4. The analysis is done by comparing relative peak area ratios between patient samples and control samples. First, each peak is compared to the sum of the internal control peaks within the same sample and this is referred to as normalisation. This normalised peak is compared to the same normalised peak in the control samples; this is called a dosage quotient (DQ). A normal, deleted and duplicated range of values is specified for the DQs.

2.3.2.1 Control fragments

In each kit probe mix, DNA quantity control fragments are included (Appendix E). These fragments generate amplification products that are smaller in length than the probe amplification products (64, 70, 76 and 82bp). Even if the technique fails these fragments will be present and will indicate when the sample DNA was less than the required 20ng of DNA (Schouten et al. 2002). In addition to these, a 92bp amplification product is present which is sample-DNA and ligation-dependant. This reflects the copy number of a 2q14 DNA sequence. The purpose of this probe is to warn the user when ligation of probe pairs in the mix did not take place or was sub-optimal. The peak area and height of this probe is usually 60-80% of the peak dimensions of the other peaks. These DNA quality control probes are not the only non-CFTR probes in the kit to serve as controls for the technique. There are thirteen other probes in the kit that hybridise to loci throughout the genome that will be referred to in this dissertation as internal

control (IC) probes. The purpose of these probes is for relative peak ratio analysis. If the entire gene was deleted, these probes would indicate that the technique had not failed.

2.3.2.3 Filtering raw data

"Raw data" refers to the peak signals (both height and area) that are generated by a capillary genetic analyser. These are scrutinised in GeneMapper. Filtering is needed because there is always a certain amount of "noise" in the sample, which is not relevant to the result of that sample. These are small peaks that lie on either side of the larger, relevant peaks that correspond to regions on the genome, as specified by the kit manufacturers. These small peaks are manually removed from the data.

2.3.2.4 Identification of homozygous deletions

The first goal of the analysis is to identify homozygous deletions in the samples. These will be obvious to the user, as missing peaks can be identified from the peak picture. An astute user may be able to eyeball homozygous duplications; however these will have to be confirmed by the analysis.

2.3.3 Laboratory protocol

2.3.3.1 Preparation of sample and control DNA

DNA quality and quantity of sample and control DNA along with dilutions of DNA were used to select and prepare samples for rearrangement detection using MLPA. Quality assessment of samples and controls was done by running 1µl of genomic DNA on a 0.8% agarose gel for 30 minutes. Samples that showed exaggerated smearing due to degradation of DNA or excess salt concentrations as a result of the salting out DNA extraction method, or those that showed very low concentrations were excluded from the analysis. Samples were quantified by analysing 1µl of genomic DNA on the *ND-1000 Spectrophotometer* nanodrop spectrophotometer. These results were used to dilute the samples and controls to 75ng/µl, in 1xTE buffer (pH 8). Please refer to Appendix D for dilution

calculation equations that were used. Diluted samples were re-quantified in the nanodrop spectrophotometer to check that the DNA was correctly diluted.

2.3.3.2 MLPA protocol

A CFTR Kit (P091 from MRC-Holland) was used to test patients who present clinically with CF, and have at least one unknown CFTR mutation. The technique is divided into three steps and requires a minimum of 25 hours to complete (please refer to Appendix E for a detailed protocol of these steps).

2.3.3.2.1 The Hybridisation reaction

The hybridisation takes place over a period of 16 hours. Two probe pairs bind adjacently on their single stranded DNA target sequences after denaturation of the DNA for 5 minutes at 98°C. The probes will only bind perfectly when 100% complementarity exists between them and the target DNA. The thermal cycler lid was opened once the products had cooled to 25°C. The hybridisation probes and buffer were added to the denatured DNA and this was incubated for 1 minute at 98°C and then 60°C overnight (16-18 hours).

2.3.3.2.2 The Ligation reaction

The adjacent probes must be ligated in order for amplification to take place. A thermostable ligase requires 15 minutes to ligate the two ends of a single probe pair. The ligase mix was added to the hybridisation products once they were cooled to 54°C. The ligation reaction was stopped by incubating at 98°C for 5 minutes.

2.3.3.2.3 Amplification

SALSA (kit) buffers and water were added to the ligation products and heated to 60°C. At this temperature polymerase mix, which includes primers, one of which is fluorescently labelled (FAM), is added and the PCR reaction is immediately started. The PCR conditions used were: 30 seconds 95°C; 30 seconds 60°C; 60 seconds 72°C for 25 cycles and then a final incubation at 72°C for final extension.

Exponential amplification of the probes takes place if two probe pairs were successfully ligated.

2.3.3.2.4 Visualisation of MLPA products

The separation of MLPA products takes place in a capillary genetic analyser (3130xl, ABI) and the fluorescence of the fragments is quantified by the software and visualised in the form of peaks, the height and area of which is proportional to the amount of fluorescence. The fragment size of the MLPA amplification products is determined by the software (GeneMapper) that compares the products to an internal size standard (ROX 500). Please refer to Appendix E for the details of the protocol for visualisation of MLPA amplification products on a capillary genetic analyser.

2.3.4 Analysis of MLPA products

2.3.4.1 Analysis programs

The analysis of MLPA PCR fragments may be done using various macros and software analysis packages. A macro is a series of commands and functions that are stored in Microsoft Visual Basic and it is automated in Microsoft Excel. The macro created in this project was recorded, whereby Microsoft Excel stores information about each step of calculations as they are performed. Each macro is given a specific identifier (name) and is stored in a workbook. Many macros and software programs for the analysis of MLPA dosage data are commercially available but MRC-Holland has developed an analysis package for public use called Coffalyser (http://www.mrc-holland.com/pages/support_mlpa_analysis_coffalyserpag.html). The national genetics reference laboratory at Manchester University (http://www.ngrl.org.uk/Manchester/Informaticspubs.htm#MLPA) has also made analysis sheets freely available. Softgenetics have produced a program called Genemarker (version 5.1), which was assessed on a 30 day trial basis. By making use of some of the guidelines for analysis specified by the authors at MRC-Holland as well as the University of Manchester, an analysis macro in Microsoft Excel for the analysis of filtered MLPA data from Genotyper was also developed

in this project (CDC_AnalysisMacro). These programs were assessed and compared for future use in the laboratory. A short synopsis of these analysis programs and macros is given below:

A) Coffalyser

Principles of analysis are as mentioned above. Some of the differences include intra-normalisation, where each peak signal is divided by each internal control peak signal and the median of these is calculated for each peak. The analysis offers a median of all peak signals as an alternative to using internal control peaks for normalisation. Regression analysis is also done by the software to correct for falling peak signals in longer probes. It is calculated through the internal control probes and a line is drawn through these values. Every probe signal is then calculated in relation to this line and all subsequent normalisation are one minus the distance of every probe signal to the regression line. Internal control probes that fall between 70 and 130% of the median of all the internal control probes are used to build the regression line. A second regression line is calculated so that internal control probes that fall outside of the first regression line are not unfairly excluded from the analysis (a program-defined number of standard deviations is used to define how far from the regression line probes may fall, and still be included in the analysis). The internal control peak normalisation is actually done by comparing probes to one minus the value to the second regression line.

B) Manchester analysis sheets

This macro uses a maximum of five controls to calculate dosage quotients (as described above) and graphically displays the mean dosage quotient for each ligation product, as well as the likelihood probability and odds for each ligation product calculated for one of three hypotheses: normal dosage, deleted and duplicated. The mean and standard deviations of each product are also calculated for normalised control and test samples. The controls give a measure of variability and the probability of deviation is estimated using a t-statistic. The relative likelihood of each of the three competing hypotheses is calculated for each ligation product as an odds ratio, to indicate which hypothesis is more likely. A quality control calculation is included by calculating DQ's for all the internal control fragments and measuring the standard deviation (SD) of these within each

sample. A SD of greater that 0.1 for a sample deems that sample as poor quality and warns the user that the result obtained for that sample is not reproducible.

C) Genemarker v 1.5

The user may choose their preferred analysis method when using this program: either the internal control probe method or the population method. The normalisation in this program increases the signal intensities of the longer fragments. The two different methods use different probes to create an intensity ratio that is comparable; for both, the square root of this value is calculated and plotted to a linear regression, using the control probes as reference points. The population normalisation calculates this ratio by comparing peak intensities of the amplified probes between all peaks in the lane. Regression analysis uses a t-distribution to form the regression line between the square root of the sample and of the control. It is iteration based which means that it is computationally repeated until the regression line has reached a certain amount of confidence (greater than 99%).

D) CDC_AnalysisMacro v 2

This macro normalises patient as well as control samples by using the sum of internal control peaks, and compares these normalised peaks to one another. The result of this comparison is called a dosage quotient. The macro allows the user to visualise which samples have erratic internal control peak values and flags samples that are of poor quality, that is, their results may be un-reproducible. The macro includes a one tailed t-test whereby a z-score for single samples is calculated. Individual data values are compared to the mean of control sample data. The z-score is a measure of the distance of each data point from the mean, in standard deviations. To test if the difference between data points and the control population mean is statistically significant, a p-value can be determined from the z-score. For significance testing, the null hypothesis is that there are no rearrangements. This hypothesis can then by rejected at the 0.05 significance level. The lower the p-value, the more likely that the null hypothesis is true by chance. The macro displays both the DQ and p-value for each data point. A short synopsis of the analysis macro is given in the Table 2.4 below.

Table 2.4: Synopsis of CDC_AnalysisMacro calculations and worksheets

Worksheet	Action	Calculations
Raw Data	1. Past filtered peak areas of up to 15 controls and 20 patients according to P091 kit order of exons	
Normalisation	1. Exons arranged in CFTR gene order 2. Internal control (IC) areas are summed 3. Normalisation (N) : Every exon peak (PA) is divided by the sum of IC 4. Internal control peak areas are also normalised	$\sum(IC_i...IC_p)$ $N = PA / \sum(IC_i...IC_p)$
z-test	Normalised peak values for control samples are: 1. Averaged¥ 2. Standard deviation is calculated¥ 3. z is calculated¥ 4. p values are calculated (NORMSDIST(z)¥) where NORMSDIST is defined by the positive value of z	$\mu = \sum(N_i...N_p) / p$ $S.D. = \sqrt{\frac{\sum_{s=1}^{m}\sum_{i=1}^{n}(y_{is} - M)^2}{(n_y - 1)}}$ $M = \frac{\sum_{s=1}^{m}\sum_{i=1}^{n} y_{is}}{n_y}$ $z = X - \mu / SD$ $p = 1 - f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$
Quality control	1. Dosage quotients (DQ) for IC peaks are calculated by dividing the normalised IC peak value by the average of the IC peak areas for the same IC peak in the control samples 2. Standard deviation of DQ's is calculated¥ 3. Bar graph¥ representing SD and p-values for IC peaks is shown	$DQ = N_{(IC)} / \mu_{(IC)}$ $S.D. = \sqrt{\frac{\sum_{s=1}^{m}\sum_{i=1}^{n}(y_{is} - M)^2}{(n_y - 1)}}$ $M = \frac{\sum_{s=1}^{m}\sum_{i=1}^{n} y_{is}}{n_y}$
Results	1. DQ for exon peaks are calculated DQ ranges define if a samples is deleted, duplicated or normal 2. Scatter plot¥ of DQ values is shown 3. Bar graph¥ representing SD and p-values for PA is shown	$DQ = N_{(PA)} / \mu_{(PA)}$

¥ Calculation or plot was done using EXCEL functions and formulae

IC Internal control peak
PA Exon peak
N Normalised peak
SD Standard deviation
μ Sample mean
n Sample size
Σ Sum of

In order to assess the most effective way of using the analysis macro, various analysis experiments were done. Through this, some conclusions could be drawn and these guided the way that the data were analysed. Some of the questions that were answered concerned the use of controls samples, internal control peaks as well as exon peaks. This analysis experimentation was possible because of the availability of two known deletion controls. These were good controls because they were heterozygous for the deletion; they were verified in other laboratories and contained no other unknown rearrangements in the CFTR gene. The one patient originated form South Korea and the other was a Black South African (this deletion was identified both in this project and by unpublished data of Prof M Claustres in France). The questions asked, and the ways that they were answered are covered in the Table 2.5.

Table 2.5: Assessment of the most effective way of using CDC_Analysis Macro

Question	Action	Method
Control samples		
1. How can control samples be evaluated?	Check samples for homozygous deletions of IC and exon peaks Evaluate 64-84bp peaks to check DNA concentration and 92bp peak to check ligation Manually fail samples that wane off	Manual (visual) analysis
	Analyse each control against all other controls from the same and different laboratory runs	Use analysis program(s)
2. What is the effect of comparing samples to unmatched controls?	Controls samples that were unmatched (for ethnicity) were used to analyse the deletion controls	The Black SA deletion control and the Korean deletion control were compared to Black, Coloured and a mixture of both controls
3. What is the effect of comparing samples to controls that were generated in different laboratory runs?	Control samples that were generated in different laboratory runs were used to analyse the deletion controls	Deletion controls were compared to controls generated in the same run, different runs and a mixture of both
Internal control peaks		
1. Should certain IC peaks be removed from the analysis?	Samples with many IC peak aberrations and samples with two or less IC peak aberrations were tested	The peaks that were deleted or duplicated were removed from the analysis
Exon peaks		
1. Should certain exon peaks be removed from the analysis?	Check deletion control samples for rearrangement of exon 6b	Test deletion controls, that should not show a deletion or duplication of exon 6b

Chapter 3: Results

3.1. Carrier testing

The Black and Coloured unaffected controls, that were representative of these general SA populations, were tested to determine if they were carriers of the 3120+1G>A CFTR mutation. The results of this RFLP test are shown below:

Table 3.1: 3120+1G>A carrier test for Black unaffected controls

Control	RFLP result for 3120+1G>A
RB 1 - RB 51	Non-carrier
RB 52	Carrier

Table 3.2: 3120+1G>A carrier test for Coloured unaffected controls

Control	RFLP result for 3120+1G>A
RC 1 - RC 50	Non-carrier

The control sample that was determined to be a carrier was excluded from haplotype testing. The objective of haplotype testing of controls was to calculate the frequency of the 3120+1G>A mutation-associated haplotype in the general population. The introduction of controls that carried this mutation would have biased the calculation, albeit very marginally. The result in the general Black SA population can be translated into a 3120+1G>A carrier frequency of 1 in 52.

Patient families were genotyped for the purpose of phasing haplotype(s) in this study, therefore the available family members were also tested for the presence of the 3120+1G>A mutation. The results of this test are shown below:

Table 3.3: 3120+1G>A carrier test result in relatives of CF patients

Parent	Ethnicity	RFLP result for 3120+1G>A
CF 225 M	Black	Carrier
CF 225 F	Black	Non-carrier
CF 620 M	Coloured	Carrier
CF 672 M	Coloured	Non-carrier
CF 788 M	Coloured	Carrier

Key

M	Mother
F	Father

3.2. Genotype results

3.2.1 PCR results

The six markers genotyped for haplotype analysis by using pyrosequencing were amplified first by PCR. Below is a representative gel picture of the results of one these amplification experiments.

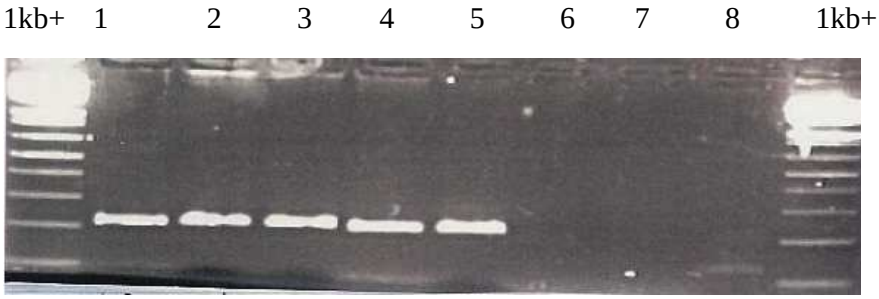


Figure 3.1: PCR result of rs1042077 (T845T). Lane 1) RB21; 2) RB8; 3)RB11; 4)RB22; 5)PCR positive control; 6) and 7) PCR negative control (no DNA added). PCR products run on a 3% Agarose gel. Expected fragment size of 225bp was found (please refer to Appendix B for molecular marker fragment sizes)

3.2.2 Pyrosequencing results

Pyrosequencing software was used to determine genotype results for each sample. A representative of these is shown below in Figure 3.2. For the purpose of this dissertation, the genotype results for all six markers for all patient and control samples are not shown.

Samples that are designated as “checks” by the PSQ verification software may be re-tested or they may be “manually called” by the user. In the Figure 3.3, B is an example of a sample that needed to be checked, and was manually determined to have a G / C genotype. This manual calling is done by comparing sample pyrograms to hypothetical bar graphs that are shown by the PSQ software. These graphs may show all three genotypes of a particular marker. The peaks that are highlighted in yellow are representative of sample genotype. The user makes a

comparison of these and the proceeding peak patterns to the hypothetical graph patterns.

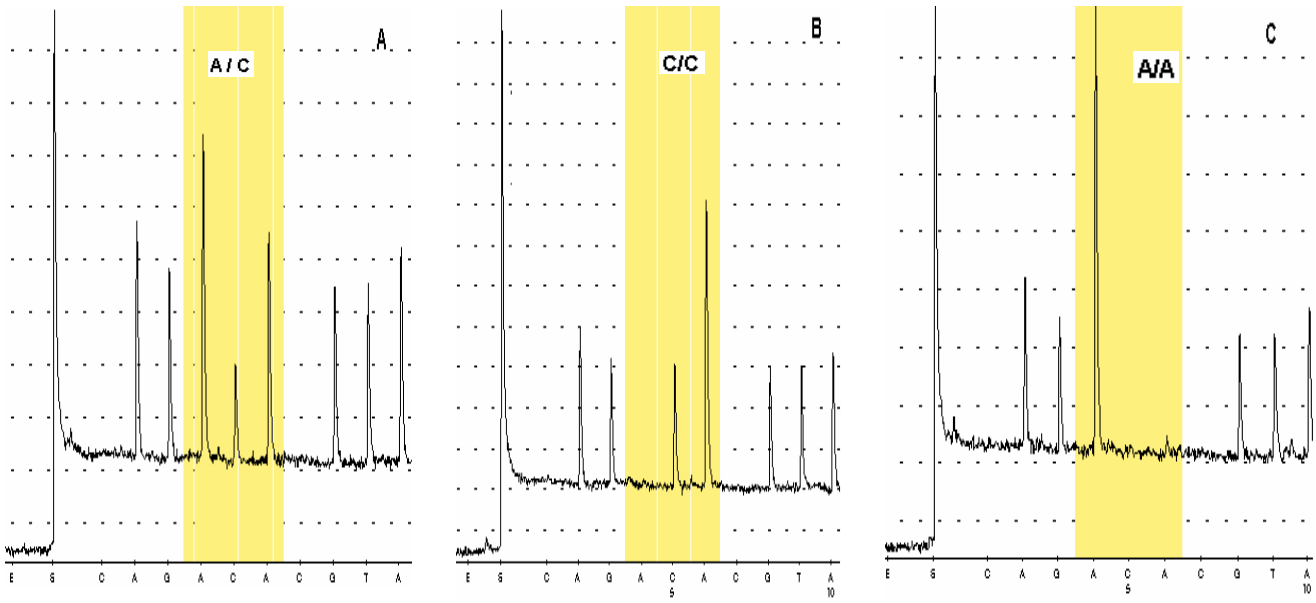


Figure 3.2: Pyrogram showing passed PSQ results for rs43035 (J44); **A)** CF 638 genotype A / C; **B)** CF 736 genotype C / C; **C)** CF 800 genotype A / A; Sequence to analyse: AG(A/C)AGTA

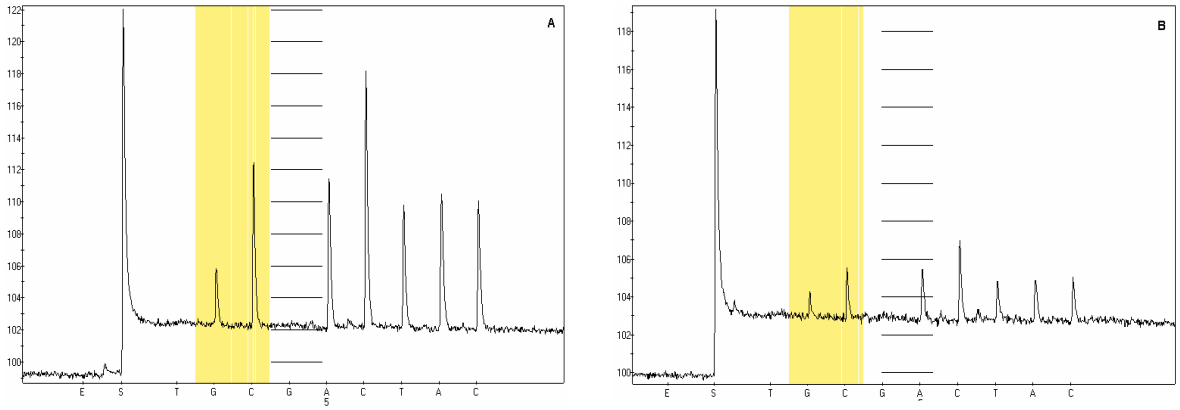


Figure 3.3: Pyrogram of rs11770163 (Met D) showing a G / C genotype. **A)** Sample that passed the software verification for a G/C genotype. **B)** Sample that was designated as a "check" by the PSQ verification software and that was passed manually for a G/C genotype. Sequence to analyse: (G/C)ACTAC

3.3. Haplotype results

Please refer to Appendix F for summary tables of the genotype results for Black and Coloured CF patients and Black and Coloured unaffected controls. Tables 3.4 and 3.5 show inferred haplotypes of patient genotype data. The phasing was manually determined by analysis of homozygosity as well as family data, where available. Family data is shown in a shade of grey as the CF chromosomes represented in these parents are identical to those that present in their offspring and are thus duplicated in this table. They are included for the convenience of the reader. Inference is indicated in the table by the use of brackets.

In accordance with the aims of this study, the 3120+1G>A mutation-associated haplotype(s) were inferred. Table 3.6, pg 56 is a summary of all the Black and Coloured 3120+G>A mutation-associated haplotypes. By analysing the inferred haplotypes in Tables 3.4 and 3.5 a frequent, 3120+1G>A compatible haplotype was found that is common to both Black and Coloured patient samples. There is little deviation from this haplotype, with only two markers, TUB 18 and J32 presenting with a different allele on only one CFTR haplotype each. This result indicates that there is a common origin for 3120+1G>A in Black and Coloured SA CF patients. The results of this analysis are further discussed in Chapter 4.

Analysis of the ΔF508-associated haplotypes in the Coloured patients was not a major aim of this project; however, as the data were available, the haplotypes were phased. The selected, compatible haplotypes are shown below in Table 3.7, pg 57. These results show a relatively common ΔF508-associated haplotype. These results indicate that there is a common origin for this mutation in the Coloured population of the Western Cape. These results will be further discussed in Chapter 4.

Table 3.4: Coloured haplotypes of patient and family data with inferred genotype phases shown

Subjects	CFTR genotype	Met D rs11770163	Km 19 rs916727	J44 rs43035	T845T rs1042077	TUB 18 rs213989	J32 rs3802012
CF 32	ΔF508 U	(C) (G)	(G) (A)	C C	(T) (G)	G G	(G) (A)
CF 120	ΔF508 ΔF508	(C) (G)	Nr Nr	C C	T T	G G	G A
CF 120 M	ΔF508 N	(G) (C)	Nr Nr	C A	T T	G G	G A
CF 120 F	ΔF508 N	(G) (C)	Nr Nr	C C	T T	G G	A A
CF 217	3120+1G>A R1162X	(G) (C)	(G) (A)	C C	G G	(T) (G)	A A
CF 433	U U	G C	G G	C A	T T	G T	A A
CF 546	U U	(G) (C)	G G	(A) (C)	G G	(G) (T)	A A
CF 620	3120+1G>A U	G G	(G) (A)	C C	G G	(T) (G)	A A
CF 620 M	3120+1G>A N	G G	(G) (A)	C A	G T	(T) (G)	A A
CF 626	ΔF508 3120+1G>A	(C) (G)	nr nr	C C	T G	g (t)	(g) a
CF 626 M	3120+1G>A N	(G) (C)	(G) (A)	C A	G T	T G	A A
CF 626 F	ΔF508 N	(C) (G)	(G) (A)	C A	T T	G G	(G) (A)
CF 640	U U	G G	A A	(A) (C)	(G) (T)	G G	(A) (G)
CF 651	3120+1G>A 3120+1G>A	G G	G G	C C	G G	T T	(A) (G)
CF 662	ΔF508 U	(C) (G)	(G) (A)	C C	(T) (G)	G G	(G) (A)

Table 3.4: Continued

Subjects	CFTR genotype	Met D rs11770163	Km 19 rs916727	J44 rs43035	T845T rs1042077	TUB 18 rs213989	J32 rs3802012
CF 664	ΔF508 U	nr nr	(G) (A)	C C	(T) (G)	G G	(G) (A)
CF 665	U U	G G	(A) (G)	A A	(G) (T)	(G) (T)	A A
CF 665 M	N U	G G	(A) (G)	(A) (C)	(G) (T)	(T) (G)	A A
CF 666	3120+1G>A U	G G	G G	C C	(G) (T)	(T) (G)	A A
CF 672	ΔF508 3120+1G>A	(C) (G)	G G	C C	T G	G T	(G) (A)
CF 672 M	ΔF508 N	(C) (G)	G A	nr nr	T T	G G	(G) (A)
CF 675	ΔF508 3120+1G>A	(C) (G)	G G	C C	(T) (G)	(G) (T)	A A
CF 685	U U	(G) (C)	G G	(C) (A)	(T) (G)	(G) (T)	(A) (G)
CF 741	U U	(G) (C)	A A	A A	(G) (T)	G G	A A
CF 766	3120+1G>A 3120+1G>A	G G	G G	C C	G G	T T	A A
CF 788	ΔF508 3120+1G>A	C G	G G	C C	T G	(G) (T)	G A
CF 788 M	3120+1G>A N	G G	G G	C C	G G	(T) (G)	A A
CF 917	U U	G G	A A	(A) (C)	nr nr	(G) (T)	(A) (G)
CF 947	U U	C C	G G	C C	(G) (T)	G G	G G
CF 1008	U U	G G	G G	nr nr	G G	T T	A A
CF 1014	U U	(G) (C)	G G	nr nr	G G	(T) (G)	(A) (G)

Key:

( )	Inferred phase
g	lower case indicates that although there was no result, a genotype may be determined
(g)	lower case, bracketed indicates that the determined result could not be phased
nr	No result
U	Unknown CFTR mutation
N	A non-CF allele
	Grey shading: unaffected family members
F	Father
M	Mother

Table 3.5: Black haplotypes of patient and family data with inferred allele phases shown

Subjects	CFTR genotype	Met D rs11770163	Km 19 rs916727	J44 rs43035	T845T rs1042077	TUB 18 rs213989	J32 rs3802012
CF 225	3120+1G>A	G	G	C	(G)	T	A
	G1249E	G	G	C	(T)	G	A
CF 225 M	3120+1G>A	G	G	C	(G)	T	A
	N	G	G	A	(T)	T	G
CF 225 F	G1249E	G	G	A	(G)	G	A
	N	C	nr	C	(T)	T	G
CF 387	U	G	G	A	T	G	(A)
	U	G	G	A	T	G	(G)
CF 389	U	(G)	G	A	(T)	G	A
	U	(C)	G	A	(G)	G	A
CF 390	3120+1G>A	G	G	C	(G)	T	nr
	U	G	G	C	(T)	G	nr
CF 395	U	C	(A)	A	(G)	(G)	A
	U	C	(G)	A	(T)	(T)	A
CF 462	U	G	(G)	(A)	T	G	A
	U	G	(A)	(C)	T	G	A
CF 526	3120+1G>A	G	nr	C	G	T	A
	3120+1G>A	G	nr	C	G	T	A
CF 534	3120+1G>A	G	(G)	C	(G)	(T)	A
	U	G	(A)	C	(T)	(G)	A
CF 538	3120+1G>A	(G)	(G)	C	(G)	G	A
	U	(C)	(A)	C	(T)	G	A
CF 603	3120+1G>A	G	G	C	G	T	A
	3120+1G>A	G	G	C	G	T	A
CF 638	U	G	(G)	(C)	(G)	G	(A)
	U	G	(A)	(A)	(T)	G	(G)
CF 652	3120+1G>A	G	nr	C	G	T	A
	3120+1G>A	G	nr	C	G	T	A
CF 736	U	G	G	C	(G)	(G)	A
	U	G	G	C	(T)	(T)	A
CF 782	U	C	G	A	T	(T)	(A)
	U	C	G	A	T	(G)	(G)
CF 800	U	G	(G)	A	(G)	G	A
	U	G	(A)	A	(T)	G	A
CF 801	U	(G)	G	C	(G)	G	A
	U	(C)	G	C	(T)	G	A
CF 826	U	C	A	(C)	G	G	G
	U	C	A	(A)	G	G	G
CF 849	3120+1G>A	G	G	C	G	T	A
	U	G	G	A	G	T	G

Key:

( )	Inferred phase
g	lower case indicates that although there was no result, a genotype may be determined
(g)	lower case, bracketed indicates that the determined result could not be phased
nr	No result
N	A non-CF allele
	Grey shading: unaffected family members
F	Father
M	Mother

Many of the patient samples that were genotyped did not have a known CFTR mutation. These patient genotypes were analysed and phased according to homozygosity (Table 3.8). One patient had family data that was largely uninformative for the purpose of haplotype phasing. The results summarised in Tables 3.8 and 3.9, show that there may be one mutation of appreciable frequency in each of the groups, but that the majority of the remaining unknown mutations are likely to be low frequency alleles in this group. On six of the 27 (22%) alleles in the Coloured samples the following haplotype was inferred: G-A-C-G-G-A. If there is an unknown allele in the Coloured population that is associated with this haplotype, then this allele may account for approximately 22% of the 25.6% of mutations that are not currently detected through molecular diagnostic mutation testing. Similarly, there is a common inferred haplotype (G-G-A-T-G-A) in the Black samples that accounts for 7 of the 23 (30.4%) alleles genotyped. These data indicate that there may be a mutation in the population that accounts for 5.63% and 16.3% of all the mutations in the Coloured and Black populations respectively. [The total mutation detection in the Black population is 60.3%, but the only mutation that is tested for in a molecular diagnostic setting is the 3120+1G>A mutation, and therefore the mutation detection rate when testing for this mutation is used]. Table 3.10 shows the total increase in mutation detection that could occur if the mutations that are associated with these haplotypes are detected and included in molecular diagnostic testing in these populations. (Please refer to Table 1.2 and 1.3 on page 9 for a summary of mutation detection in Coloured and Black populations).

Table 3.6: Inferred 3120+1G>A -associated haplotypes for Black and Coloured CF patients

Subjects	Ethnicity	Met D rs11770163	Km 19 rs916727	J44 rs43035	CFTR genotype	T845T rs1042077	TUB 18 rs213989	J32 rs3802012
CF 526	Black	G	nr	C	3120+1G>A	G	T	A
		G	nr	C	3120+1G>A	G	T	A
CF 603	Black	G	G	C	3120+1G>A	G	T	A
		G	G	C	3120+1G>A	G	T	A
CF 652	Black	G	nr	C	3120+1G>A	G	T	A
		G	nr	C	3120+1G>A	G	T	A
CF 225	Black	G	G	C	3120+1G>A	G	T	A
CF 390	Black	G	G	C	3120+1G>A	G	T	nr
CF 534	Black	G	G	C	3120+1G>A	G	T	A
CF 849	Black	G	G	C	3120+1G>A	G	T	A
CF 538	Black	G	G	C	3120+1G>A	G	G	A
CF 766	Coloured	G	G	C	3120+1G>A	G	T	A
		G	G	C	3120+1G>A	G	T	A
CF 788	Coloured	G	G	C	3120+1G>A	G	T	A
CF 672	Coloured	G	G	C	3120+1G>A	G	T	A
CF 626	Coloured	G	nr	C	3120+1G>A	G	t	a
CF 620	Coloured	G	G	C	3120+1G>A	G	T	A
CF 675	Coloured	G	G	C	3120+1G>A	G	T	A
CF 666	Coloured	G	G	C	3120+1G>A	G	T	A
CF 217	Coloured	G	G	C	3120+1G>A	G	T	A
CF 651	Coloured	G	G	C	3120+1G>A	G	T	A
		G	G	C	3120+1G>A	G	T	G

Key

nr	Indicates that there is no result
	Colour indicates 3120+1G>A -associated haplotype
	Colour indicates deviation from the 3120+1G>A -associated haplotype
A	Red letter indicates that phase is inferred for that particular marker

Table 3.7: Inferred ΔF508 -associated haplotypes for Coloured CF patients

Subjects	Ethnicity	Met D Rs11770163	Km 19 rs916727	CFTR genotype	J44 rs43035	T845T rs1042077	TUB 18 rs213989	J32 rs3802012
CF 672	Coloured	C	G	ΔF508	C	T	G	G
CF 675	Coloured	C	G	ΔF508	C	T	G	A
CF 788	Coloured	C	G	ΔF508	C	T	G	G
CF 626	Coloured	C	nr	ΔF508	C	T	nr	nr
CF 662	Coloured	C	G	ΔF508	C	T	G	G
CF 32	Coloured	C	G	ΔF508	C	T	G	G
CF 664	Coloured	nr	G	ΔF508	C	T	G	G
CF 120	Coloured	C	nr	ΔF508	C	T	G	G
		G	nr	ΔF509	C	T	G	A

Key

nr	Indicates that there is no result
	Colour indicates ΔF508-associated haplotype
	Colour indicates deviation from the ΔF508-associated haplotype
A	Red letter indicates that phase is inferred for that particular marker

Table 3.8: Inferred Coloured patient unknown mutation-associated haplotypes

Subjects	Ethnicity	CFTR genotype	Met D rs11770163	Km 19 rs916727	J44 rs43035	T845T rs1042077	TUB 18 rs213989	J32 rs3802012
CF 32	Coloured	U	G	A	C	G	G	A
CF 620	Coloured	U	G	A	C	G	G	A
CF 662	Coloured	U	G	A	C	G	G	A
CF 664	Coloured	U	nr	A	C	G	G	A
CF 666	Coloured	U	G	G	C	T	G	A
CF 433	Coloured	U	G	G	C	T	G	A
		U	C	G	A	T	T	A
CF 546	Coloured	U	G	G	A	G	G	A
		U	C	G	C	G	T	A
CF 640	Coloured	U	G	A	A	T	G	G
		U	G	A	C	G	G	A
CF 665	Coloured	U	G	A	A	G	G	A
		U	G	G	A	T	T	A
CF 685	Coloured	U	G	G	C	T	G	A
		U	C	G	A	G	T	G
CF 741	Coloured	U	G	A	A	G	G	A
		U	C	A	A	T	G	A
CF 741	Coloured	U	G	A	A	G	G	A
		U	C	A	A	T	G	A
CF 917	Coloured	U	G	A	C	nr	G	A
		U	G	A	A	nr	T	G
CF 947	Coloured	U	C	G	C	G	G	G
		U	C	G	C	T	G	G
CF 1008	Coloured	U	G	G	nr	G	T	A
		U	G	G	nr	G	T	A
CF 1014	Coloured	U	G	G	nr	G	T	A
		U	C	G	nr	G	G	G

Key

nr	Indicates that there is no result
A	Red letter indicates that allele phase is inferred for that particular marker
U	Indicates an unknown CFTR genotype
	Most common inferred Coloured haplotype
	Second most common inferred Coloured haplotype

Table 3.9: Inferred Black patient unknown mutation-associated haplotypes

Subjects	Ethnicity	CFTR genotype	Met D rs11770163	Km 19 rs916727	J44 rs43035	T845T rs1042077	TUB 18 rs213989	J32 rs3802012
CF 534	Black	U	G	A	C	T	G	A
CF 538	Black	U	C	A	C	T	G	A
CF 849	Black	U	G	G	A	G	T	G
CF 387	Black	U	G	G	A	T	G	A
		U	G	G	A	T	G	G
CF 389	Black	U	G	G	A	T	G	A
		U	C	G	A	G	G	A
CF 395	Black	U	C	A	A	G	T	A
		U	C	G	A	T	G	A
CF 462	Black	U	G	G	A	T	G	A
		U	G	A	C	T	G	A
CF 638	Black	U	G	G	A	T	G	A
		U	G	A	C	G	G	G
CF 736	Black	U	G	G	C	G	G	A
		U	G	G	C	T	T	A
CF 782	Black	U	C	G	A	T	T	A
		U	C	G	A	T	G	G
CF 800	Black	U	G	G	A	T	G	A
		U	G	A	A	G	G	A
CF 801	Black	U	G	G	C	G	G	A
		U	C	G	C	T	G	A
CF 826	Black	U	C	A	C	G	G	G
		U	C	A	A	G	G	G

Key

nr	Indicates that there is no result
A	Blue letter indicates that allele phase is inferred for that particular marker
U	Indicates an unknown CFTR genotype
	Most common inferred Black haplotype

Table 3.10: Postulated mutation detection rate if unknown alleles that are associated with the most common haplotypes in Coloured and Black populations are tested for in a molecular diagnostic setting

Mutation detection		Population
Current	Postulated*	
74.40%	80.03%	**Coloured**
46.40%	69.90%	**Black**

* Calculated by studying the frequency of unknown-associated haplotypes

It should be noted that the highlighted haplotype in Table 3.9 occurs at a relatively high frequency in the normal Black and Coloured control groups (9% and 8%) respectively. From these findings it is not explicitly clear if there is a single or multiple mutations associated to this relatively common haplotype, and therefore the detection of this haplotype in the control group weakens the evidence that there may be another single mutation that accounts for a high proportion of Black CF alleles in whom no mutation is currently detected.

3.3.1 Control sample inferred haplotypes

In order to infer haplotypes for Black and Coloured control samples, a statistical program, PHASE, was used. The results are shown in Table 3.11 and are discussed in Chapter 4. The results show that the 3120+1G>A mutation-associated haplotype identified in Black and Coloured SA healthy controls was inferred in only one of 102 Black chromosomes in the controls samples, which are representative of the general population. This haplotype was not inferred in the general Coloured population. These results confirm that the G-G-C-G-T-A haplotype is highly correlated with 3120+1G>A and suggests that the mutation did not arise independently in the Coloured population.

Table 3.11: Inferred results for Black and Coloured controls, using PHASE

Inferred haplotype*	Black	Coloured
GGCGGA	21	3
GGCGGG	8	4
GGATGA	9	8
GGAGGA	11	13
GGAGTG	1	1
GACTGA	0	4
GACGGA	0	12
GACGTA	2	6
GACGTG	0	1
GAATGA	1	3
GAAGGA	0	10
GAAGGG	0	8
GAAGTA	0	2
GGATTA	2	0
GGAGTA	5	0
GGCTGA	4	0
GGCGTA♦	1	0
GGCGTG	1	0
GACTGG	1	0
CGCGGA	9	1
CGCGTA	3	1
CGATGA	0	3
CGAGGA	4	5
CGAGGG	0	2
CACTTA	0	1
CACGGA	0	6
CACGTA	1	4
CAAGGA	2	2
CGAGTA	4	0
CGCTGA	4	0
CGCGGG	5	0
CGCGTG	2	0
CAATTA	1	0
Total alleles	102	100

Key

Light grey shading indicates the commonest three inferred haplotypes in Blacks and Coloureds

* Markers in chromosomal order: rs1177016; rs916727; rs43035; rs1042077; rs213989; rs3802012

♦ 3120+1G>A-associated haplotype in Black and Coloured CF patients

3.3.2 Genotype and allele frequencies

In order to test weather the control populations are in Hardy-Weinberg equilibrium for each of the markers, a Fisher's exact test was used. This was calculated using PowerMarker v3.25. The results are shown in Table 3.12.

Table 3.12: Genotype frequency and Hardy-Weinberg equilibrium for control samples form two populations

		Black			Coloured		
Marker	**Genotype**	***f***	**n**	**H-W**	***f***	**n**	**H-W**
rs11770163	G / G	0.3913	18		0.5435	25	
	G / C	0.4783	22		0.3478	16	
	C / C	0.1304	6		0.1087	5	
total			**46**	**Yes (1.00)**		**46**	**Yes (0.47)**
rs916727	G / G	0.7857	22		0.1778	16	
	G / A	0.1429	4		0.4667	21	
	A / A	0.0714	2		0.3556	8	
total			**28**	**Yes (0.07)**		**45**	**Yes (0.78)**
rs43035	C / C	0.3469	17		0.2245	16	
	A / C	0.5306	26		0.4490	22	
	A / A	0.1224	6		0.3265	11	
total			**49**	**Yes (0.57)**		**49**	**Yes (0.57)**
rs1042077	G / G	0.5349	23		0.6744	29	
	G / T	0.3953	17		0.1627	7	
	T / T	0.0698	3		0.1627	7	
total			**43**	**Yes (1.00)**		**43**	**No (<0.001)**
rs213989	G / G	0.6200	31		0.7200	36	
	G / T	0.2600	13		0.2200	11	
	T / T	0.1200	6		0.0600	3	
total			**50**	**Yes (0.05)**		**50**	**Yes (0.13)**
rs3802012	A / A	0.6327	31		0.6522	30	
	A / G	0.3673	18		0.3478	16	
	G / G	0	0		0	0	
total			**49**	**Yes (0.36)**		**46**	**Yes (0.32)**

Key

f	Genotype frequency
n	number of genotypes
	Sample result indicates that the population is in Hardy-
H-W	Whether the sample is in Hardy-Weinberg equilibrium
Brackets	p-value

The control samples were found to be in H-W equilibrium at all but one locus in the Coloured samples. This may be a chance event since this locus is closely

linked to the other loci and one would not expect that drift or selection are acting to cause the disequilibrium. However, it is possible that there is linkage disequilibrium, due to hitch-hiking, with another nearby locus which has been selected for or against. On the other hand, since multiple testing was done and a 5% significance level was chosen, the application of the Bonferroni correction may render this finding non-significant.

3.4. Mutation testing

3.4.1 MLPA Results

MLPA was used to test for rearrangements in the CFTR gene using a CFTR kit from MRC-Holland. As MLPA can be sensitive to certain types of contaminants, this was resolved through the use of low concentrations (30-75ng/μl) of genomic DNA. Decreased signal strength of long fragments is not entirely avoidable and most analysis programs have built-in regression analysis to prevent longer probes from appearing to be deleted, however the peak signal strengths are greatly improved by increased mixing of MLPA reagents and DNA at every step of the protocol. There should not appear to be deletions in the regions spanned by the probes that yield longer peak fragments if the controls to which they are compared suffer from the same effect, however due to the normal inter-run variation, it is not always prudent to rely on this fact. Empirically, peak areas are generally more robust than peak heights for analysis. Table 3.13 shows that one copy number mutation was found: a deletion of exon two of the CFTR gene. This deletion was detected in the heterozygous state in one black patient (CF390). Figure 3.4 shows the Genemapper peak picture of this patient and a control sample. This figure clearly demonstrates a reduction in peak height and area for this exon.

T**able 3.13**: MLPA result for patients that passed the analysis quality control

Ethnicity	Sample	Genotype	Detection of exon copy number variation
Black	CF 389	U / U	No
Black	CF 390	3120+1G>A / U	Yes*
Black	CF 395	U / U	No
Black	CF 462	U / U	No
Black	CF 534	3120+1G>A / U	No
Black	CF 538	3120+1G>A / U	No
Black	CF 638	U / U	No
Black	CF 736	U / U	No
Black	CF 782	U / U	No
Black	CF 801	U / U	No
Black	CF 849	3120+1G>A / U	No
Coloured	CF 32	ΔF508 / U	No
Coloured	CF 433	U / U	No
Coloured	CF 620	3120+1G>A / U	No
Coloured	CF 664	ΔF508 / U	No
Coloured	CF 665	U / U	No
Coloured	CF 666	3120+1G>A / U	No
Coloured	CF 685	U / U	No
Coloured	CF 1016	U / U	No
Coloured	CF 1131	U / U	No

Key

* Heterozygous deletion of exon 2

U Unknown CFTR mutation

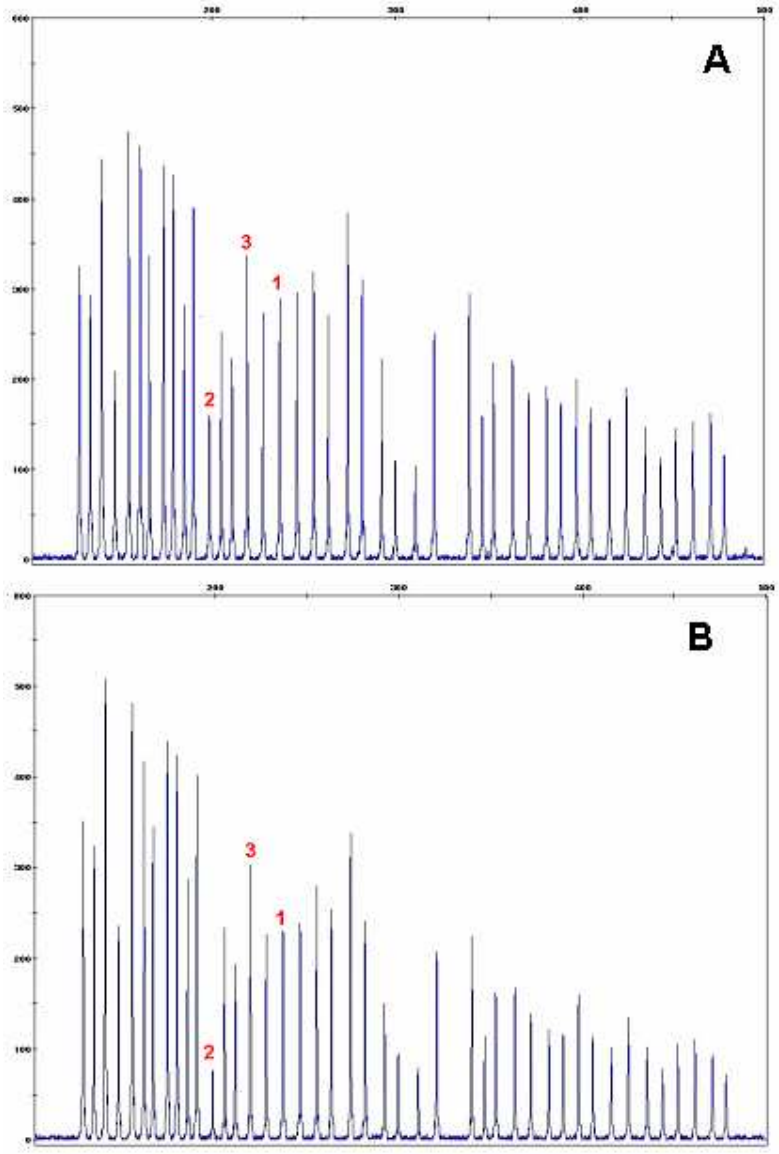


Figure 3.4: MLPA result (Genemapper) confirming the presence of an exon 2 deletion of the CFTR gene. The numbers above the peaks refer to the exon number of the gene. Exons 1-3 are shown for comparison. A: Peak profile of a control showing no rearrangements of the CFTR gene. B: Peak profile of the SA black CF patient with a heterozygous deletion of exon 2

As this deletion is present in only one of the subjects tested in this study, it would seem that copy number changes of CFTR gene exons are not a critical mechanism for mutations in the SA population.

Chapter 4: Discussion

In this project we propose that Black and Coloured SA populations have unique CFTR mutation profiles. Since a significant proportion of mutations remain unknown, we aimed to search for copy number variants in CFTR gene exons, with a view to increasing molecular diagnostic mutation detection in these populations. Another aim was to do a haplotype analysis across the CFTR gene with the primary objective of shedding light on the origin of the 3120+1G>A mutation in Black and Coloured SA populations. The significance of our findings and our conclusions are discussed below.

4.1 Origins of mutations

4.1.1 Evolutionary origin of 3120+1G>A mutation in Black and Coloured SA populations

In this study I found that the Met D-Km 19-J44-T854T-TUB18-J32 CFTR haplotype for 10 Black 3120+1G>A chromosomes was compatible with the G-G-C-G-T-A haplotype. Of these, three were inferred based on this common haplotype, the others were determined without ambiguity. Three Black patients were homozygous for the mutation and every marker genotyped. This haplotype was also inferred on, and compatible with, 10 Coloured CFTR alleles that carried the 3120+1G>A mutation. Four of these were phased without inference and one Coloured sample was homozygous for the mutation and every marker genotyped. The shared haplotype identity for these intra- and extragenic markers for Black and Coloured SA CF patients suggests that the mutation arose once in the Black population and was then introduced into the Coloured population through genetic admixture.

This haplotype was compared to haplotypes found for the same markers in previous studies. As shown in Table 4.1, the haplotype associated with the 3120+1G>A mutation found in this study correlates perfectly with those found in two previous studies.

Table 4.1: Summary of 3120+1G>A associated haplotypes

Met D	Km 19	J44	T845T	TUB 18	J32	
rs11770163	rs916727	rs43035	rx1042077	rs213989	rs3802012	
G	G	C	G	T	A	This study
	G	C	G	T		Dörk et al. 1998*
G	G	C				Carles et al. 1996*

*NB the markers shown in this table are not fully representative of all the markers that were genotyped in these studies

The haplotypes found in the Carles et al. (1996) study (three SA families and one Cameroonian family were studied) suggest that this mutation has a single origin and that it is an old mutation that accounts for many of the CFTR mutations in African Blacks, as it had been found associated with the same haplotype in Southern Africa and Cameroon. The haplotype data described in the Dörk et al. (1998) study, on 17 unrelated CF patients of African American, Arabic, African (8 samples) and Greek origin strongly suggested that this mutation has a common origin in all but the Greek population, which differed only at a single microsatellite locus by one repeat unit. The authors conclude: 'Greek and Arab/African haplotypes of the 3120+1G>A mutation [thus] may have diverged from a common ancestor and then evolved separately in the respective populations'. In a study done by Padoa et al. (1999) to identify the frequency of the 3120+G>A mutation in African populations, the 3120+1G>A mutation was found in several of the Sotho-Tswana-speaking populations, but not in the Xhosa (Nguni). This led us to postulate that this mutation may have arisen independently in the Cape Coloured population, rather than through Xhosa-speaking Black admixture, as they are the predominant Black group in the Cape and the 3120+1G>A mutation was not found in them. The present study shows that this mutation was most likely passed from the Black to the Coloured population, indicating that the sample size studied in the Padoa et al. (1999) (n= 52 Xhosa out of n=157 Nguni) study was too small to detect this mutation. The results, therefore mitigate against the possibility that the 3120+1G>A mutation arose independently in the Coloured population.

4.1.2 ΔF508 associated haplotypes in Coloured CF patients

Analysis of the **Δ**F508 associated haplotypes revealed a C-G-C-T-G-G haplotype in Coloured samples genotyped for Met D-Km 19-J44-T854T-TUB18-J32. These samples were genotyped as they carried one copy of the 3120+1G>A mutation or one copy of an unknown CFTR gene mutation. Seven of the ΔF508 alleles out of a total of nine appeared to carry this haplotype. The two that did not carry this haplotype showed deterioration of the haplotype on the extremities, those being the furthest extragenic markers which lie 571Mb upstream and 541Mb downstream of their nearest neighbours. This pattern of decay usually reflects that the haplotype has existed for a long time and that recombination has eroded its ends.

The ΔF508 mutation has been extensively studied and has been shown to have a single origin in many different populations, as discussed in Chapter 1 (Stringer et al. 1990, Morral et al. 1994 and 1996, Lucotte et al. 1995 and 1993). Table 4.2 below highlights this, as there is a shared haplotype between Coloured SA CF patients; admixed African American population (Dörk et al. 1998) and German Caucasians (this haplotype is also found in many other European countries, specifically France) (Kerem et al. 1989; Dörk et al. 1992) indicating that there is a common origin for this mutation in these populations. The study done on the German population supports the theory that the ΔF508 mutation arose from a 'single mutational event that spread through European populations according to a south-east to north-west gradient'. The ΔF508 mutation is assumed to have entered the African American population through genetic admixture because known European ΔF508 associated haplotypes were found in these patients in the Dörk et al. (1998) study. In the present study, the ΔF508-associated haplotype indicates that this mutation did not arise independently in the SA Coloured population but was introduced from the White SA population. This haplotype has been found in French populations and now has been found in the Western Cape Coloured population, which is likely to be admixed with descendants of French Huguenots, as predicted by Westwood et al. (2007).

Table 4.2: Summary of ΔF508 associated haplotypes

Met D	Km 19	J44	T845T	TUB 18	J32	
rs11770163	rs916727	rs43035	rx1042077	rs213989	rs3802012	
C	G	C	T	G	G	This study
	G	C	T	G		Dörk et al. 1998*
	G	C	T			Kerem et al. 1989*, Dörk et al. 1992*

*NB the markers shown in this table are not fully representative of all the markers that were genotyped in these studies

4.1.3 Frequency of mutations among unidentified CF alleles

Haplotype inference of unknown mutation-associated haplotypes showed a common haplotype for six and seven of the Coloured and Black samples respectively. If this is calculated as a percentage of the total number of alleles genotyped (27 Coloured and 23 Black), then we find that the mutation detection rate increases by 5.63% and 16.3%. When adding this to the total current mutation detection rate, we find that it is increased to 80.03% in the Coloured and 69.9% in the Black population. It is important to note that these samples were phased manually with very little homozygosity and no informative family data and therefore the allele phase was rarely assigned without ambiguity.

The remaining unknown mutations are likely low frequency mutations, as no other dominating haplotypes were found, and therefore will each account for a low proportion of mutant CF alleles in these populations. This result provides evidence for high heterogeneity in the CFTR gene in SA Black and Coloured CF patients. The high amount of genetic heterogeneity in these populations suggests that there has been an accumulation of many different mutations and, other than possibly 3120+1G>A and ΔF508, selective pressure has not driven specific alleles to high frequencies in these populations.

In conclusion, the inferred unknown mutation-associated haplotypes show that there may be two mutations that account for an appreciable proportion of the currently undetected CF alleles in these populations. If these mutations may be detected, then it may be worth including these in the molecular diagnostic mutation testing. The remaining unknown mutation-associated haplotypes indicate that up to 20% and 30% of currently undetected mutations in the Coloured and Black mutations, respectively, are highly likely to be low-frequency mutations that would probably not be included in diagnostic mutation detection as they are probably very rare and may be confined to specific families.

4.1.4 CFTR haplotype inference in healthy controls

Carrier testing of healthy controls revealed a 1 in 52 carrier frequency for the 3120+1G>A mutation in the SA Black population. This is comparable to a study done by Padoa et al. (1999), that recorded a 3120+1G>A carrier frequency of 1 in 91 (8/728) in South African blacks with a 95% confidence interval of 1 in 46 to 1 in 197.

PHASE software was used to predict haplotypes in the healthy controls genotyped in this project. In healthy Black control samples, the G-G-C-G-T-A haplotype (the haplotype associated with the 3120+1G>A mutation in both Black and Coloured populations) was predicted only once out of a total of 102 Black chromosomes and was not predicted in 100 Coloured chromosomes. The C-G-C-T-G-G haplotype associated with the ΔF508 mutation was not predicted by the software to be present in the samples genotyped for this project. A total of 23 different Black and 22 different Coloured haplotypes were predicted in these controls. Eleven exclusively in the Black samples and 10 exclusively in the Coloured samples. Twelve haplotypes were found in both, but not at the same frequencies (five of these were found at similar but low frequencies). The results of this analysis show that there is a clear distinction in haplotype distribution between non-CF chromosomes and CF chromosomes, with regard to the markers studied. The haplotypes found associated with the 3120+1G>A and ΔF508 mutations are underrepresented in healthy controls. This finding is in line with

studies in European populations that have compared CF to non-CF chromosomes (Estivill et al. 1997; Morral et al. 1996).

The PHASE version used takes into consideration the spacing of the markers and the decay of linkage disequilibrium with distance. In this way, the pattern of LD among multiple loci, which depends on the underlying recombination rate in the region, is estimated by the software. This prior (probability distribution of parameter values before observing the data) is referred to as the "coalescent with recombination" (MR) which is a setting that can be selected by the user, and is computationally more intensive than settings that do not take recombination rate into account. Studies, that have used this setting, show that there are benefits to taking distance into account, even over short distances (Stephens and Donnelly 2003).

The accuracy of this software for inferring missing genotypes (gaps) has also been shown in previous studies (Stephens and Scheet 2005). This is important as there are few studies, including this one, in which gaps are not found. PHASE uses the same algorithms to infer haplotypes as it does missing genotypes, and therefore infers these with comparable accuracy. The utility of this feature was confirmed in this study, whereby two datasets were compared, where the one contained missing genotypes and the other did not. This comparison showed that both data sets resulted in similar haplotype distributions, albeit occurring at different frequencies, with the dataset that did not contain gaps having three additional haplotypes. By removing samples with missing genotypes for one or more of the markers genotyped, an unnecessary loss of information occurs. The overall loss of information and reduced sample size that occurred had a greater effect on the results than the presence of gaps. Because the haplotype results for datasets that contain gaps and those that do not contain gaps are highly comparable, and in order to retain as much information as possible, it makes sense to run the PHASE program with datasets that contain gaps. This recommendation, however, is subject to the sample size and the number of samples that contain gaps as well as how many gaps there are in the samples on average.

This haplotype inference program does not rely strictly on the assumption of H-W equilibrium. This was an important factor for the choice of this software for inference of haplotypes in Black and Coloured populations as population stratification, gene flow, drift, mutation and selection are all expected to be acting on these populations. This even more so in the Black samples genotyped in this project, as they represent many different chiefdoms (sub-populations) in Gauteng and in the Coloured population, that is expected to show high levels of population-specific variation and heterogeneity. In fact, the H-W tests (exact testing done by PowerMarker software) on these populations show that they are in H-W equilibrium at all but one locus (KM 19 in the Coloured population). As a result, the PHASE program was expected to offer a fairly accurate estimation of haplotypes for these populations due to its relative leniency when it comes to H-W equilibrium.

The number of haplotypes that may theoretically be expected when studying six biallelic markers is 2^6=64. Approximately 35% of these potential haplotypes was detected in both populations, which is expected as the sample size is not large enough to detect extremely rare haplotypes. The results of the program for these populations may be overestimating the number of haplotypes found and the differences between the two populations. This is because the program allows for recombination, which results in a higher number of 'plausible' haplotypes for each individual; and because analysing samples from different genetic backgrounds separately would tend to consistently overestimate the differences between them and by the same token, analysing samples from different genetic backgrounds together would consistently underestimate the differences between them. This is because of the assumption in the software that haplotypes tend to cluster together due to their shared ancestry and therefore the prediction of the relationship between genotypes at specific loci is dependant on the prediction of haplotypes at other loci. It may be worthwhile analysing populations that are not highly diverged in the same run, for the sake of sample size and to maximise genotype information per run (Stephens and Scheet 2005). These types of

statistical analyses are more accurate with increased sample sizes, number of markers used and the uniformity of their position in the genome (evenly spaced) and by analysing smaller regions and lastly, by doing many computational iterations. In studies like this one, where statistically reconstructed haplotypes are not critically important on an individual basis, the inaccuracies which are expected may be accepted, as long as the researchers bear them in mind.

In conclusion, the most reliable methods for haplotype inference among unrelated samples are Bayesian approaches which take into account recombination and which do not rely heavily on the assumptions of H-W equilibrium (Stephens and Scheet 2005). These approaches are all exercised in the PHASE run that was used for this project. The results of the haplotype study in Black and Coloured healthy controls show that the two populations have dissimilar haplotype structures for the markers genotyped, highlighting the genetic differences that exist between Western Cape Coloured and Gauteng Black populations, despite the Black admixture in the Coloured samples. The 3120+1G>A mutation-associated haplotype in the CF patient samples is found at a very low frequency in the healthy Black population, which either indicates that the mutation arose on a rare haplotype in the ancient population, that is still rare today, or it arose on a haplotype that was common but is no longer common, as a result of genetic drift. Because this haplotype is extremely rare in the healthy Coloured population (was not detected in this dataset by PHASE) but is found at a very high frequency in the patient group (Black and Coloured) associated to the 3120+1G>A mutation, this mutation was probably introduced into the Coloured population from the Black population and therefore did not arise independently in the Coloured population. This conclusion also makes sense in light of the fact that the Coloured population is much younger than the Black population, which contributed to its gene pool.

4.1.5 Comparison between HapMap and SA population SNP genotype data

A comparison of SNP genotype information between SA and HapMap populations is highly relevant to genetic studies on SA populations that may

require the use of HapMap genotype information. The comparison may reveal that the HapMap data for specific populations may be extrapolated to certain SA populations, or that the population differences in allele frequencies between them are dissimilar and therefore, HapMap population data would not be relevant as proxies for SA data. The allele frequencies in the Black and Coloured SA populations (computed by studying Black and Coloured controls genotyped in this project), and the YRI and CEU populations (information obtained from the HapMap database) were compared for the six markers studied. Table 4.3 below shows allele frequencies for these datasets. Significance testing was done for the YRI and SA Black allele data in order to test whether the differences between them were significant of not. These results are shown in Table 4.4. (An example of the calculations that were done are shown in Appendix D).

Table 4.3: Allele frequencies for HapMap and SA populations

Marker	**Population**			
	HapMap: YRI	**HapMap: CEU**	**SA: Black (Gauteng)**	**SA: Coloured (Western Cape)**
Met D rs11770163	0.48 (G)	NA	0.63 (G)	0.72 (G)
Km 19 rs916727	0.48 (A)	0.66 (A)	0.14 (A)	0.59 (A)
J44 rs43035	0.45 (C)	0.52 (C)	0.61 (C)	0.45 (C)
T854T rs1042077	0.254 (T)	0.57 (T)	0.27 (T)	0.24 (T)
TUB 18 rs213989	0.12 (T)	0.23 (T)	0.25 (T)	0.17 (T)
J32 rs3802012	0.19 (G)	NA	0.18 (G)	0.17 (G)

Key

( )	Allele is indicated in brackets
NA	No SNP genotype information available

The significance testing (Table 4.4) showed that the differences in allele frequencies between the HapMap YRI population and the SA Black population were statistically significant at the 5% confidence level at all the marker sites.

Table 4.4: Significance testing for differences in allele frequency between YRI and SA Black populations

Marker	P-value
Met D	0.03
rs11770163	
Km 19	6.03
rs916727	
J44	0.017
rs43035	
T854T	0.83
rs1042077	
TUB 18	0.011
rs213989	
J32	0.95
rs3802012	

* df = 1
p = significance level = 0.05
Ho: Allele frequencies are not significantly different

For markers Km 18, T854T and J32, the null hypothesis was accepted and the allele frequencies between the populations were not significantly different and therefore, YRI SNP genotype information may be used as proxies for SA Gauteng Black SNP genotype information. For markers Met D, J44 and TUB 18, this result indicates that, because of their significantly different allele frequencies, HapMap YRI SNP genotype data and SA Black SNP genotype data are not comparable. As this work represents only a tiny fraction of the genome, this mini-study is probably best viewed as preliminary data for a larger, more comprehensive comparison. Despite this, this result serves as a potential warning, as it indicates that up to half of HapMap data may not be appropriate as proxies for the SA Black population.

4.2 Exon copy number variation in Black and Coloured CF patients

In this study SA Black and Coloured CF patients were investigated for exon copy number changes in the CFTR gene, using MLPA. The PO91 kit developed by MRC-Holland was designed to detect exon deletions or duplications. Although most of the mutations that have been detected in this gene have been point mutations and small insertions and deletions, there is growing evidence in the literature of copy number variants in this gene, which are listed in the CF

mutation database (http://www.genet.sickkids.on.ca/cftr/StatisticsPage.html) as large in/dels and account for 2.83% of the 1556 mutations listed in the database.

4.2.1 MLPA: technical discussion

MLPA is a quantitative, reliable and robust method of detecting copy number changes in the genome. When compared to Southern Blotting, the technique is faster; less genomic DNA is used and large numbers of samples can be analysed simultaneously without a corresponding proportional increase in man-labour. There is also no threat of incomplete digestion and a lower risk of false positives. When compared to other techniques such as comparative genomic hybridization (CGH) and fluorescence *in situ* hybridization (FISH) as well as quantitative-real time PCR (QRT-PCR), many would argue that the techniques are comparable. In the past five years, numerous researchers have made use of QRT-PCR, or modifications of this technique, to detect copy number changes in many different genes, including the CFTR gene. The main benefits of this technique are that it is also multiplex (albeit not with universal primers, but usually multiple primers, which decrease the robustness of the technique) and quantitative measures of DNA copy number are ascertained using light-cycler technology, whereby the DNA of interest is compared to a reference (control) sample. MLPA is also rapid, reliable and sensitive and does not require the intense computational analysis that, for example CGH requires.

Schouten et al. (2002) proposes that Southern blotting and long range PCR be used to verify the result found by using the MLPA kit. In a study by Meuller et al. (2004), genomic deletions in the APC gene were found, first by performing MLPA and then quantitative real-time PCR (QRT-PCR) was used to verify the MLPA result. These are both quantitative methods and were followed by subsequent verification using fluorescent *in situ* hybridisation (FISH). The deletion was further verified and characterised with long-range PCR and sequencing. The authors concluded that MLPA "ensures a sensitive high-throughput screening for large deletions of the APC gene". This study highlights

the importance of verification of MLPA results. The MLPA technique is only diagnostic in cases where a mutation in a family is known.

MLPA is relatively labor intensive but it is extremely sensitive and is easily performed with equipment that is readily available in most research laboratories. It is very effective for the detection of deletions or duplications in genes, even of single exons in a multiplex-type reaction, which allows for relatively rapid and high-throughput mutation detection.

4.2.2 Analysis of MLPA products

Once quantitative MLPA data were generated, the analysis of the data was done by using a macro that was developed specifically for this project. The most effective way of using this CDC_AnalysisMacro for MLPA data analysis was investigated (the aims of this are outlined in Table 2.5). The following are some conclusions that were drawn. These conclusions are applicable not only to the analysis of the data using this particular macro, but also addresses some of the principles of analysing MLPA data in general: It was found that control samples do not necessarily yield good quality products, and they must be evaluated as controls before they are used as a standard to which sample data is compared. Controls with homozygous exon copy number variants and internal control peak copy number variants as well as controls that fail quality control evaluation, when compared to other controls, should be eliminated from the analysis. When comparing the deletion control samples to unmatched controls (for ethnicity) it was found that the result was the same or very similar when comparing to the results of data that was analysed against matched controls. Interestingly, when comparing the Korean deletion control to unmatched controls (no matched controls were available), the use of Black controls yielded better results (no other exon copy number variants were found, other than the deletion that is present in that sample). This may be because there were a higher number of Black controls than Coloured controls and therefore the standard deviation of normal inter-run variation was found to be higher in Black controls. This testing showed that it is

discretionary to compare Black and Coloured samples to unmatched controls (for ethnicity), so long as the controls were checked and deemed reliable.

When comparing the Black deletion control to controls that were generated in different laboratory runs, it was found that the analysis showed false duplications or deletions if those controls were not checked. For any statistical analysis, it is always better to have as many control samples as possible. For this reason, the question of whether samples may be compared to control samples that were generated in other runs is important, so that the number of controls tested can be balanced against cost. If samples are to be compared only to controls that are generated in the same run, then at least five good quality controls must be generated in each run. This could be very expensive, therefore it would be advantageous if samples could be compared to controls from other runs; and if this were possible, then control sample sizes could accumulate, until fewer numbers of controls would have to be tested per run. The evaluation showed that samples can be compared to a mixture of control samples: some that were generated in the same run and some that were generated in other runs, so long as the controls were checked and deemed reliable.

Internal control peaks are used in the analysis to calculate the quality of the samples used. When testing whether internal control peaks can be removed from the analysis of MLPA products, it was found that the quality control calculation showed less accurate results, the fewer IC peaks were used to calculate it. The kit that was used in this project was produced for Caucasian European or North American populations. This means that it was not created with SA population-specific SNP and STR variants in mind. These variants may have an effect on probe binding, even if they are near to, and not at the probe ligation site (Schouten et al. 2002). There is therefore, a possibility that IC peak signals may be affected by variants that are located in or near the probe binding sites, in SA populations. If this is found, should these peaks be removed from the analysis? These peaks are meant to give an indication of the quality of that sample. The question of whether IC peaks should be removed is important because if one or more of the

peaks had low or high peak signal due to population specific variants, then the peak(s) that are affected should be removed rather than discarding (or optimising) the sample. The evaluation of the analysis showed that these peaks can be removed only if one of these peaks is deleted or duplicated. If there was a common variant in a population, the control and the sample peaks would both have it and therefore it is unlikely that the analysis would pick it up. Samples with IC peak aberrations should be optimised and re-run and if still present, the mutations should be validated and if the copy number changes are not verified (shown to be a false artefact) then the sample should be discarded.

Users are warned, in the P091 kit specifications, that exon 6b shows more inter-run variability than other exons, and this was confirmed in this project. Removing this exon from the analysis had no effect on the analysis of other exons. By removing exon peaks from the analysis, one runs the risk of missing true deletions or duplications in those exons, however by keeping such peaks in, one runs the risk of publicising exon copy number variations that are not real. This problem emphasises the importance of validation of copy number changes that are found using MLPA, in the research setting.

The evaluation of the MLPA data analysis assessed how control samples, internal control peaks and exon peaks should be evaluated and applied for the analysis of MLPA data. In conclusion: Control samples should always be checked before they are used in the analysis. Control samples of poor DNA quality (degradation or contamination), those that show copy number changes or severe waning off of peaks signals of longer fragments should not be used. Black and Coloured samples may be analysed using a mixture of Black and Coloured controls. Samples may be compared to controls that are generated both in the same and in different runs, but not exclusively to controls generated in different runs. Removing internal control peaks and exon peaks from the analysis is not recommended.

The analysis macro that was made for use in this project was assessed in the ways discussed above and was also contrasted against some of the other software that is currently available (all of these are available for free download except Genemarker). Table 4.5 below highlights only the major advantages and disadvantages of these programs. The finer points that could be discussed are beyond the scope and the aims of this project. The major reasons that these other programs were not used are outlined in this table. The macro that was developed in this project was based largely on the calculations that are used in Coffalyser and the Manchester analysis sheets. The only program that I would not advise for the analysis of MLPA data is Genemarker (version 1.5 was used in this project). The major disadvantage of this program is that it selects only one control sample to which the patient samples are compared, and while the user may change this control at their own discretion, the software tended to select a patient sample rather than a control sample, to use as a control. This is very disconcerting as, if there is a common aberration in the patient samples, it would not be detected and would appear normal; and as there are usually more patient samples than healthy controls in a typical run, the controls would 'seem' mutated and the patients samples normal, if they all contained the same mutation. Caution should be exercised when using the Manchester analysis sheets, as the macro tends to overestimate the number of mutations in a sample. This means that the macro calculations are too 'lenient' or, more likely, that the ranges that are set to define the parameters of their hypotheses (this being normal, deleted or duplicated) are too wide. This could, in my opinion be avoided in these sheets, by simply allowing the user to analyse their data against more controls (if need be, the SD range would therefore be widened). This problem highlights the importance of allowing the data to dictate the ranges rather than the researcher, and this can only be done if enough data is available.

Table 4.5: The advantages and disadvantages of four programs for the analysis of MLPA data

Software	Advantages	Disadvantages
Coffalyser	• Regression analysis • Option of filtering data for the user	• Program contains many bugs and could not be used
Genemarker (30 day trial version used)	• User may choose between two normalisation methods • Regression analysis • Performs many iterations	• A major disadvantage is that while the user inputs both controls and sample data, the program selects which it considers to be the control and analyses all samples against this control. The software tended to choose a patient sample as a control
Manchester analysis sheets	• Calculates DQs, and likelihood of three hypotheses: normal, deleted or duplicated dosage • Sample quality is assessed	• Maximum of five controls and ten samples can be analysed at once • Sheets showed the incorrect deletion when tested with the known deletion control
CDC_AnalysisMacro	• Sample quality is assessed • Can analyse up to 15 controls per run • Calculates DQs as well as significance testing	• Requires manual filtering • No regression analysis

4.2.3 MLPA findings

In this project a deletion of exon two of the CFTR gene was found in one Black patient (CF 390). This deletion was not confirmed by alternative techniques and the breakpoints were not determined, however, it was also found in this patient independently in a French laboratory to whom the DNA sample had been sent many years ago for mutation detection. The mutation was verified by the French group using semi-quantitative fluorescent PCR. Once the deletion was detected, whole genome amplification was used to increase the amount of DNA that was available (as only minute quantities remained) and the intronic mutation (c.54-1161_c.164+1603del2875) was characterised through mapping and sequencing. The study revealed that the mutational mechanism could be explained by a classical model of replication slippage due to direct repeats present at the 5' and 3' breakpoints (des Georges et al. 2007). As the mutation was

concurrently detected and verified in the French group, it was used in this project, after it had been independently identified, as a positive control to test the analysis macros and programs for MLPA, as discussed in section 4.2.2.

This mutation removes residues 91 to 163 of the amino acid chain and the effect of this mutation on the protein is the removal of the first two transmembrane regions within the first membrane spanning domain (which forms the chloride channel). Two cytoplasmic and one extracellular topological domains (topological domains are the regions that connect the transmembrane domains, either on the extracellular or on the cytoplasmic sides of the channel) are also removed (http://www.expasy.org, http://www.genet.sickkids.on.ca/cftr/). This patient has one positive sweat test and other symptoms suggestive of CF, including respiratory failure. The seemingly classical symptoms of CF combined with the mutation that results in a defective channel that will hinder conduction through the membrane indicates that this is a severe mutation and could probably be classified as a class 4 mutation (according to the classification of Zielenski and Tsui 1995). The classification of a mutation is important for treatment and counselling purposes, but since this mutation was not found in any other patients, it is probably isolated to this family and is present at very low frequencies amongst CF alleles. This implies that the mutation will not be included in the diagnostic panel of mutations that are tested for in Black patients with symptoms suggestive of CF. This result suggests that copy number variants of CFTR gene exons is not a major mutational mechanism giving rise to CF in Black and Coloured SA CF patients. It must be noted that, had the sample size been larger, a more accurate frequency of the mutation in these populations may have been calculated.

4.3 Future studies

Some follow-up studies that may be undertaken as a result of the conclusions drawn from this project for mutation detection in the gene include: a) Test Black patient samples with at least one unknown mutation for the low frequency mutations that have been detected in this population (G1249E, 3196del54, -94G >

T, 2183delAA); b) Increase sample size of patients with at least one unknown CFTR mutation and test for copy number changes to get a more accurate frequency of the exon 2 deletion and to further detect CNVs in SA Black and Coloureds CF patients c) Denaturing gradient get electrophoresis or single stranded conformation polymorphism (SSCP) and sequencing of products that that showed a different mobility from those in a reference control, of coding and flanking intronic sequences of the CFTR gene, starting with the samples that showed a common unknown mutation-associated haplotype. In these samples, the exons or introns that harbour the highest number of mutations worldwide should be studied first; d) Copy number mutation detection in non-coding sequences as well as upstream and downstream regulatory sequences using QRT-PCR.

4.4 Conclusions

The MLPA results obtained in this study indicate that exon copy number variants in the CFTR gene do not contribute significantly to mutations in Black and Coloured SA CF patients. The detection of a heterozygous exon 2 deletion in a single Black patient is probably isolated to that family.

Haplotype analysis of patients with at least one 3120+1G>A mutation shows a common origin of the 3120+1G>A mutation in Black and Coloured South African CF patients. These data provide further evidence that CF has a common origin in many populations and that there is probably an ancient, singular, Black African origin for the 3120+1G>A mutation.

Haplotype analysis of patients with at least one unknown mutation shows that there is likely one mutation each in the Coloured and Black population that may account for an appreciable fraction of the unknown CFTR mutations in this population. The remaining 20% of Coloured and 30% of Black mutations are likely low-frequency. These unidentified mutations were shown in this project not to be deletions or duplications of CFTR gene exons.

APPENDICIES

APPENDICIES

APPENDIX A: SUBJECTS AND ETHICS APPROVAL

Table A.1: Coloured patients and family members: phenotype and genotype information

CF Code	Molecular tests	Genotype	Selection Criteria
CF032	Kit12, 3120+1G>A	ΔF508/U	One positive sweat test Patient in ICU Pseudomonas
CF120 CF120 M CF120 F	Kit12, 3120+1G>A ΔF508508	ΔF508/ΔF508 ΔF508/N ΔF508/N	Two CFTR mutations Haplotype phasing Haplotype phasing
CF217	Kit22,3120+1G>A	3120+1G>A /R1162X	Two positive sweat tests Two CFTR mutations
CF433	3120, 394 delTT	U/U	Personal communication with Dr Westwood
CF546	Kit22,3120+1G>A	U/U	Two positive sweat tests
CF620 CF 620 M	Kit20, 3120+1G>A 3120+1G>A	3120+1G>A/U 3120+1G>A/N	Personal communication with Dr Westwood Haplotype phasing
CF626 CF626 M CF 626F	Kit22,3120+1G>A Kit22,3120+1G>A Kit22,3120+1G>A	ΔF508/3120+1G>A ΔF508/N 3120+1G>A/N	Two CFTR mutations Haplotype phasing Haplotype phasing
CF640	Kit22,3120+1G>A	U/U	Personal communication with Dr Klugman
CF651	Kit22,3120+1G>A	3120+1G>A / 3120+1G>A	Two CFTR mutations
CF662	Kit22,3120+1G>A	ΔF508508 /U	Personal communication with Dr Westwood
CF664	Kit20, 3120+1G>A	ΔF508508 /N	Personal communication with Dr Westwood
CF665 CF665 M	Kit20, 3120+1G>A 3120+1G>A	U/U U/N	Personal communication with Dr Westwood Haplotype phasing
CF666	3120+1G>A	3120+1G>A/N	Personal communication with Dr Westwood
CF672 CF672M	Kit22,3120+1G>A Kit22,3120+1G>A	ΔF508/3120+1G>A ΔF508/N	Two CFTR mutations Haplotype phasing
CF675	Kit22,3120+1G>A	ΔF508/3120+1G>A	Two CFTR mutations
CF685	Kit22,3120+1G>A	ΔF508508 /U	Personal communication with Dr Westwood
CF741	Kit22,3120+1G>A	N/N F508or 3120	Personal communication with Dr Klugman
CF766	Kit22,3120+1G>A	3120+1G>A / 3120+1G>A	Two CFTR mutations

Table A1: Continued

CF788 CF788 M	Kit22,3120+1G>A 3120+1G>A	ΔF508/3120+1G>A 3120+1G>A/N	Two CFTR mutations Haplotype phasing
CF917	Kit31	U/U	Recurrent wheezing symptoms suggestive of CF
CF947	Kit31	U/U	One sibling died of CF
CF1008	Kit32	U/U	One positive sweat test
CF1014	Kit33	U/U	One positive sweat test

Key

M	Mother
F	Father
U	Unknown CFTR genotype
N	Non-CF allele

Table A.2: Black patients and family members: phenotype and genotype information

CF Code	Molecular tests	Genotype	Selection Criteria
CF225	ΔF508, G1249E, 3120+1G>A	G1249E / 3120+1G>A	Three positive sweat tests Fatty stool Failure to thrive
CF225 M	3120+1G>A	3120+1G>A / N	Haplotype phasing
CF225 F		G1249E / N	Haplotype phasing
CF387	3120+1G>A, ΔF508	U/U	Two positive sweat tests Chronic or recurrent pulmonary infections Parents are related
CF389	3120+1G>A, ΔF508	U/U	Symptoms strongly suggestive of CF Recurrent respiratory infections Chronic lung disease
CF390	3120+1G>A	3120+1G>A/U	One positive sweat test Symptoms suggestive of CF One sibling died of respiratory problems
CF395	3120+1G>A, ΔF508	N/N	One positive sweat test Severe asthmatic Pseudomonas infection Failure to thrive Kwashiorkor/Marasmus Chronic or recurrent pulmonary infections
CF462	Kit12, 3120+1G>A	U/U	Severe inspissated meconium Patient on respirator
CF526	3120+1G>A	3120+1G>A/3120+1G>A	Diarrhoea, Kwashiorkor Bowel obstruction Meconium Ileus

Table A.2: Continued

CF534	3120+1G>A	3120+1G>A/U	Personal communication with Dr Henderson
CF538	3120+1G>A	3120+1G>A/U	Meconium Ileus Bowel obstruction
CF603	3120+1G>A	3120+1G>A/3120+1G>A	Failure to thrive
CF638	3120+1G>A	U/U	Meconium Ileus Constipation Chronic or recurrent chest infections
CF652	3120+1G>A	3120+1G>A/3120+1G>A	Two CFTR mutations
CF736	3120+1G>A	U/U	Recurrent chest infections Failure to thrive Unsuccessful sweat test
CF782	Kit22,3120+1G>A	U/U	One positive, two negative sweat tests Recurrent chest infection Lobectomy Personal communication with Dr Klugman
CF800	3120+1G>A	U/U	One positve and one negative sweat test Lower respiratory infection Patient treated as CF and recovered Personal communication with Dr Klugman
CF801	3120+1G>A	U/U	Two positive sweat tests Severe failure to thrive
CF826	3120+1G>A	U/U	Oxygen Needed Low Chymotrypsin Failure to thrive Respiratory difficulty
CF849	Kit29, 3120+1G>A	3120+1G>A/U	Two positive sweat tests Symptoms strongly suggestive of CF

Key

M Mother

F Father

U Unknown CFTR genotype

N Non-CF allele

University
of the Witwatersrand,
Johannesburg

Human Research Ethics Committee (Medical)
(formerly Committee for Research on Human Subjects (Medical)

Secretariat: Research Office, Room SH10005, 10th floor, Senate House • Telephone: +27 11 717-1234 • Fax: +27 11 339-5708
Private Bag 3, Wits 2050, South Africa

18 December 2006

Ms CL De Carvalho
MSC (Med) Students
Department of Human Genetics
NHLS

Dear Ms De Carvalho

RE: Ethics for Masters Project (M050533)

This letter serves to confirm that the Chairman of the Human Ethics Research Committee (Medical) has approved your request to use the abovementioned protocol number for your MSc (Med) research project. Copy attached

Yours sincerely,

Anisa Keshav (Ms)
Secretary
Human Research Ethics Committee (Medical)

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Ramsay/Soodyal et al

CLEARANCE CERTIFICATE	**PROTOCOL NUMBER M050706**
PROJECT	Genetic Disease Related Studies in Southern African Population: Control Samples
INVESTIGATORS	Professors M/H Ramsay/Soodyal et al
DEPARTMENT	School of Pathology/Human Genetics
DATE CONSIDERED	05.07.29
DECISION OF THE COMMITTEE*	Approved unconditionally

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE 05.08.01 **CHAIRPERSON** .....
(Professor PE Cleaton-Jones)

*Guidelines for written 'informed consent' attached where applicable

cc: Supervisor : Prof A Krause

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10005, 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. **I agree to a completion of a yearly progress report.**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

M. Ramsay 29/8/05
H Soodyal 29/8/05
alcraus 29/8/05

APPENDIX B: RFLP protocol

Diagnostic protocol for CF 3120+1G>A mutation

1. PCR reaction

PCR Components	Stock concentration	Final concentration	Volume (µl)
DNA			1
dNTPs	1.25mM (10x)	1x	2.5
Amplitaq polymerase	2.5 U/µl	1U	0.2
$MgCl_2$	25mM	2.5mM	2.5
Spermidine	1/40 dilution of 0.1M		2.5
Forward primer (16bi-5)	10pmol/µl	5pmol/µl	0.5
Reverse primer (16bi-3)	10pmol/µl	5pmol/µl	0.5
Amplitaq polymerase buffer	10 x	1 x	2.5
Distilled H_2O			12.3
TOTAL			25

Include a normal and a heterozygote control for each PCR run.
Include an additional sample in the PCR reaction to be used as an uncut control

PCR conditions: (2.5 hours)
94℃; 5 minutes
94℃; 1 minute}
51℃; 1minute} 30 cycles
72℃; 1 minute}
72℃; 10 minutes

Check 5µ PCR product on a 3% agarose gel. PCR product size: 570bp

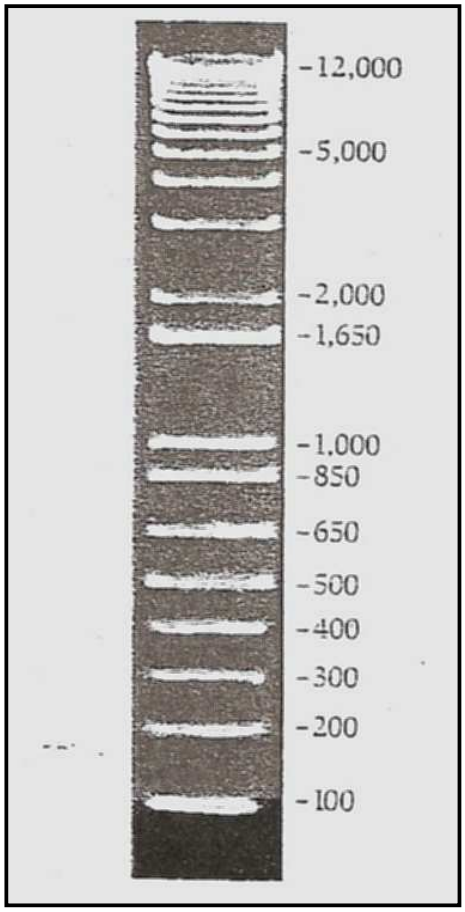
2. Restriction enzyme digestion

Reagent	Stock concentration	Final concentration	Volume (µl)
PCR product			20
BstN1 (NEB)	10 U/µl	10 U	1
NEB buffer 2	10x	1x	3.5
Distilled H_2O			10.5
Total			35

Incubate at 60℃ for 2 hours
Run 35µl of digestion product on a 3% agarose gel
Expected fragment sizes: Mutation: 537bp, 33bp; no mutation: 340bp, 197bp, 33bp

APPENDIX C: SOLUTIONS AND BUFFERS

Expected fragment sizes of 1kb+ ladder (Invirogen)



0.5M EDTA (pH 8)

93.06g	EDTA
500ml	dH_2O

Adjust pH by adding concentrated 5M NaOH
Adjust the volume of the solution to 1 liter using dH_2O

10x TBE Buffer

108g	Tris base
55g	Boric acid
7.4g	EDTA

Adjust the volume of the solution to 1 liter using dH_2O

1x TBE Buffer

1ml	10x TBE buffer
9ml	dH_2O

1x TE Buffer pH 8.0
5ml 10mM Tris
1ml 1mM EDTA
Make up to 500ml with dH_2O
Adjust pH by adding concentrated HCl

3% Agarose Gel
300ml 1x TBE
9g Agarose
Mix agarose and 1x TBE
Heat until all agarose is dissolved
Add 3µl ethidium bromide (10mg/µl) and mix

0.8% Agarose Gel
300ml 1x TBE
2.4g Agarose
Mix agarose and 1x TBE
Heat until all agarose is dissolved
Add 3µl ethidium bromide (10mg/µl) and mix

dNTPs (1.25mM stock) (Invitrogen)
12.5µl 100mM dGTP
12.5µl 100mM dATP
12.5µl 100mM dCTP
12.5µl 100mM dTTP
950µl ddH_2O

Ficol-bromophenol blue loading dye (100ml)
25g sucrose
5ml 0.5M EDTA (pH7)
0.05g Bromophenol blue dye
5g Ficoll
Make up to 50ml with dH_2O

1kb/ 1kb+ Ladder
109µl 1kb/1kb+ ladder
50µl Ficoll dye
8.4µl 1xTE

PCR primers
For a 100µM stock solution, 1xTE (pH8) buffer was added to the primer according to the primer manufacturers specifications.
To prepare working stock: Add 10µl of the 100µM solution to 90µl of 1xTE buffer for a final volume of 100µl

NaOH (10N)

800ml	distilled water
400g	NaOH pellets

Adjust the volume of the solution to 1 liter using dH_2O

Tris-Cl (1M) pH 8.0

121.1g	Tris base
800ml	ditilled water

Adjust pH by adding concentrated HCl
Adjust the volume of the solution to 1 liter using dH_2O

Binding buffer pH 7.6

10mM Tris-HCl	1.21g
2M NaCl	117g
1mM EDTA	0.292g
Tween 20	1ml

Adjust pH using HCL
Adjust the volume of the solution to 1 liter using dH_2O

Annealing buffer pH 7.6

20mM Tris-Acetate	2.42g
5mM Mg-Acetate	1.07g

Adjust pH using acetic acid
Adjust the volume of the solution to 1 liter using dH_2O

0.5M Denaturation solution

NaOH	20g

Dissolve in 1 liter dH_2O

Washing buffer pH7.6

10mM Tris-acetate	1.21g

Adjust pH by adding acetic acid
Adjust the volume of the solution to 1 liter using dH_2O

70% Ethanol

100% ethanol	70ml
dH_2O	30ml

APPENDIX D: EQUATIONS AND CALCULATIONS

A) PowerMarker

H-W equilibrium is calculated; for a single locus, the MLE of the disequilibrium coefficient D_{uv} for alleles A_u and A_v is:

$$\hat{D}_{uv} = \begin{cases} \tilde{P}_{uv} - \tilde{p}_u \tilde{p}_v, & u = v \\ \tilde{p}_u \tilde{p}_v - \frac{1}{2}\tilde{P}_{uv}, & u \neq v \end{cases},$$

and the variance is estimated using the follow formulas:

$$\mathrm{Var}(\hat{D}_{uu}) \triangleq \frac{1}{n}\left[\tilde{p}_u^2(1-\tilde{p}_u)^2 + (1-2\tilde{p}_u)^2\hat{D}_{uu} - \hat{D}_{uu}^2\right]$$

$$\mathrm{Var}(\hat{D}_{uv}) \triangleq \frac{1}{2n}\left\{\tilde{p}_u\tilde{p}_v(1-\tilde{p}_u)(1-\tilde{p}_v) + \sum_{w\neq u,v}\left(\tilde{p}_u^2\hat{D}_{uw} + \tilde{p}_v^2\hat{D}_{vw}\right)\right.$$
$$\left. -\left[(1-\tilde{p}_u-\tilde{p}_v)^2 - 2(\tilde{p}_u-\tilde{p}_v)^2\right]\hat{D}_{uv} + \tilde{p}_u^2\tilde{p}_v^2 - 2\hat{D}_{uv}^2\right\}.$$

The chi-square goodness-of-fit test is formed by calculating the chi-square statistic:

$$X_T^2 = \sum_u \frac{(n_{uu} - n\tilde{p}_u^2)^2}{n\tilde{p}_u^2} + \sum_u\sum_{v\neq u}\frac{(n_{uv} - 2n\tilde{p}_u\tilde{p}_v)^2}{2n\tilde{p}_u\tilde{p}_v}$$

B) Significance testing

Allele:		G	C	TOTAL
	YRI	**57**	**61**	**118**
	Expected	64.6190476	53.38095	
	Chi Squared	0.89834018	1.087464	
	SA Black	**58**	**34**	**92**
	Chi Squared	1.15221892	1.394791	
	Expected	50.3809524	41.61905	
	TOTAL	115	95	210

To calculate chi squared: Sum all Chi Squared values in table
Therefore, Chi Squared = 4.532
To calculate p: Use CHIDIST Excel function
Therefore p = 0.03325
Significance level: 0.05

Therefore, the null hypothesis is rejected at the 5% significance level

APPENDIX E: MLPA INFORMATION

Length (nt)	SALSA Probe #	Chromosomal position
64-70-76-82	DQ-control bands*	
94	Synthetic Control probe	2q14
130	Control probe 0797-L0463	5q31
136	GASZ probe 3571-L3264	58 Kb before the CFTR gene
142	CFTR probe 2956-L2388	Exon 13
148	CFTR probe 3842-L3315	Exon 23
154	CFTR probe 2944-L2376	Exon 1 promoter region
160	CFTR probe 2957-L2389	Exon 14
166	Control probe 2881-L2348	19q12
172	CFTR probe 3578-L2939	Exon 24
178	CFTR probe 2958-L2390	Exon 14
184	Control probe 2882-L2349	19q13
190	CFTR probe 3574-L2400	Exon 24
198	CFTR probe 2946-L3265	Exon 2
204	CFTR probe 2959-L3266	Exon 15
211	Control probe 0472-L0088	12q14
220	CFTR probe 2947-L2379	Exon 3
229	CFTR probe 2960-L2392	Exon 16
238	CFTR probe 3839-L3312	Exon 1
247	CFTR probe 2948-L2380	Exon 4
256	CFTR probe 2961-L2393	Exon 17
265	Control probe 2318-L1809	19p13
274	CFTR probe 2949-L2381	Exon 6
283	CFTR probe 2962-L2394	Exon 17
292	CFTR probe 3841-L3314	Exon 12
301	CFTR probe 2950-L2382	Exon 6
310	CFTR probe 3576-L3179	Exon 18
320	Control probe 1866-L1425	1p34
328	CFTR probe 3102-L2510	DF508 mutation specific. Not yet present
337	CFTR probe 2951-L2383	Exon 7
346	CFTR probe 3840-L3313	Exon 5
353	CFTR probe 3577-L2396	Exon 19
364	CFTR probe 2952-L2384	Exon 8
373	Control probe 1589-L1161	13q14
382	CFTR probe 2965-L2397	Exon 20
391	CFTR probe 2953-L2385	Exon 9
400	Control probe 2598-L2069	5q35
409	CFTR probe 2966-L2398	Exon 21
418	CFTR probe 2955-L2387	Exon 11
427	Control probe 0680-L0121	7q35
436	CFTR probe 2967-L2399	Exon 22
445	CORTBP2 probe 3572-L3267	58 Kb after the CFTR gene
454	Control probe 0605-L0018	15q26
463	CFTR probe 2954-L2386	Exon 10
472	Control probe 1032-L0604	1p13.2
481	Control probe 1060-L0628	17q21

Table E.1: Chromosomal position of probes in MLPA kit (P091)

Table E.2: Laboratory protocol for MLPA

MLPA Kit-Based PROTOCOL

Reagents		**Amount (µl)**
1. HYBRIDISATION PROTOCOL		
SALSA PROBE-MIX	(black cap)	1,5
MLPA BUFFER	(yellow cap)	1,5
DNA	(100ng/ul	5
Total		**8**
2. LIGATION PROTOCOL		
LIGASE-65 BUFFER A	(transparent cap)	3
LIGASE-65 BUFFER B	(white cap)	3
WATER		25
(mix thoroughly)		
LIGASE-65	(brown cap)	1
(mix thoroughly)		
Hybridisation product		8
Total		**40**
3. PCR PROTOCOL		
MLPA LIGATION REACTION		10
SALSA PCR BUFFER	(red cap)	4
WATER		26
(heat to 60 C)		
POLYMERASE MIX	(to be prepared)	10
TOTAL		**50**
POLYMERASE MIX		
SALSA PCR-PRIMERS	(purple cap)	2
SALSA ENZYME DILUTION	(blue cap)	2
WATER		5.5
SALSA POLYMERASE	(orange cap)	0.5
TOTAL		**10**

Thermocycler protocol

Denaturation:
98˚C 5 mins
25˚C open lid

Hybridisation:
95˚C 1min
60˚C 16hrs
54˚C hold

Ligation:
54˚C 15mins
98˚C 5mins

PCR:
60˚C hold
95˚C 30s }
60˚C 30s } 35cycles
72˚C 1min }

72˚C 20min

Notes on the MLPA laboratory protocol

1. Centrifuge reagents before first use only. Thereafter, resuspend by repeated pipetting to prevent the disintegration of long probes.
2. Buffers are extremely viscous; therefore all mixes should be thoroughly resuspended.
3. To avoid contamination in the PCR reaction, be sure to use dedicated pipettes and to clean bench surfaces with 1% hypochlorite solution. Ideally work in another room, and remember to wipe thermocycler. Always use a PCR negative control.
4. Ligation products may be stored up to one week at 4°C or longer at -20°C.
5. MLPA is extremely sensitive to contaminants; therefore low concentrations (30-74ng/µl) of genomic DNA should be used.

Protocol for visualisation of products on capillary genetic analyser

Sample preparation

1. Prepare master mix: HiDi formamide (8.7µl) and ROX 500 (0.3µl)
2. Aliquot 9µl of this mix into wells of an ABI plate and add 1µl of MLPA amplification product
3. "Blank" empty wells with 10µl of HiDi formamide
4. Denature products by heating plate in thermalcycler for 2 minutes at 95°C

3120xl software protocol

In plate manager:

1. Choose a run-specific name
2. Application: GeneMapper 3130xl

In plate editor

Priority: 100
Size standard: GS 500 (-250)
Panel: None
Analysis method: Microsatellite default
Results group: create
Instrument protocol: pp16_36_pop7_ANY4DYE

Load plate onto the try and link the plate in the software
Click PLAY

Analysing products in GeneMapper

1. After uploading results from the results group into GeneMapper, select the analysis settings tab
2. Select specific settings, or use default settings and analyse by pressing play
3. View the product peak pictures and export peak size, height and area information

APPENDIX F: GENOTYPE RESULTS

Table F.1: Genotype results for Coloured patient samples including family results where available

Subjects	CFTR genotype	Met D rs11770163	Km 19 rs916727	J44 rs43035	T845T rs1042077	TUB 18 rs213989	J32 rs3802012
CF 32	ΔF508 / U	G / C	G / A	C / C	G / T	G / G	A / G
CF 120	ΔF508 / ΔF508	G / C	nr	C / C	T / T	G / G	A / G
CF 120 M	ΔF508 / N	G / C	nr	A / C	T / T	G / G	A / G
CF 120 F	ΔF508 / N	G / C	nr	C / C	T / T	G / G	A / G
CF 217	R1162X / 3120+1G>A	G / C	A / G	C / C	G / G	G / T	A / A
CF 433	U / U	G / C	G / G	A / C	T / T	G / T	A / A
CF 546	U / U	G / C	G / G	A / C	G / G	G / T	A / A
CF 620	3120+1G>A / U	G / G	A / G	C / C	G / G	G / T	A / A
CF 620 M	3120+1G>A / N	G / G	A / G	A / C	G / T	G / T	A / A
CF 626	ΔF508 / 3120+1G>A	G / C	nr	C / C	G / T	nr	nr
CF 626 M	3120+1G>A / N	G / C	A / G	A / C	G / T	G / T	A / A
CF 626 F	ΔF508 / N	G / C	A / G	A / C	T / T	G / G	A / G
CF 640	U / U	G / G	A / A	A / C	G / T	G / G	A / G
CF 651	3120+1G>A / 3120+1G>A	G / G	G / G	C / C	G / G	T / T	A / G
CF 662	ΔF508 / U	G / C	A / G	C / C	G / T	G / G	A / G
CF 664	ΔF508 / U	nr	A / G	C / C	G / T	G / G	A / G
CF 665	U / U	G / G	A / G	A / A	G / T	T / G	A / A
CF 665 M	U / N	G / G	A / G	A / C	T / G	G / T	A / A
CF 666	3120+1G>A / U	G / G	G / G	C / C	G / T	G / T	A / A
CF 672	ΔF508 / 3120+1G>A	G / C	G / G	C / C	G / T	G / T	A / G
CF 672 M	ΔF508 / N	G / C	A / G	nr	T / T	G / G	A / G
CF 675	ΔF508 / 3120+1G>A	G / C	G / G	C / C	G / T	G / T	A / A
CF 685	U / U	G / C	G / G	A / C	G / T	G / T	A / G
CF 741	U / U	G / C	A / A	A / A	G / T	G / G	A / A
CF 766	3120+1G>A / 3120+1G>A	G / G	G / G	C / C	G / G	T / T	A / A
CF 788	ΔF508 / 3120+1G>A	G / C	G / G	C / C	G / T	G / T	A / G
CF 788 M	3120+1G>A / N	G / G	G / G	C / C	G / G	G / T	A / A
CF 917	U / U	G / G	A / A	A / C	nr	G / T	A / G
CF 947	U / U	C / C	G / G	C / C	G / T	G / G	G / G
CF 1008	U / U	G / G	G / G	nr	G / G	T / T	A / A
CF 1014	U / U	G / C	G / G	nr	G / G	G / T	A / G

Key

nr	Indicates that there is no result
U	Indicates an unknown CFTR genotype
N	In obligate carriers, indicates where there is no CFTR mutation genotype
M	Mother
F	Father

Table F.2: Genotype results for Black patient samples including family results where available

Subjects	CFTR genotype	Met D rs11770163	Km 19 rs916727	J44 rs43035	T845T rs1042077	TUB 18 rs213989	J32 rs3802012
CF 225	G1249E / 3120+1G>A	G / G	G / G	C / C	T / G	G / T	A / A
CF 225 M	3120+1G>A / N	G / C	G / G	A / C	G / T	T / T	A / G
Cf 225 F	G1249E / N	G / C	nr	A / C	G / T	G / T	A / G
CF 387	U / U	G / G	G / G	A / A	T / T	G / G	A / G
CF 389	U / U	G / C	G / G	A / A	T / G	G / G	A / A
CF 390	3120+1G>A / del 2	G / G	G / G	C / C	T / G	G / T	nr
CF 395	U / U	C / C	A / G	A / A	G / T	G / T	A / A
CF 462	U / U	G / G	A / G	A / C	T / T	G / G	A / A
CF 526	3120+1G>A / 3120+1G>A	G / G	nr	C / C	G / G	T / T	A / A
CF 534	3120+1G>A / U	G / G	A / G	C / C	G / T	G / T	A / A
CF 538	3120+1G>A / U	G / C	A / G	C / C	G / T	G / G	A / A
CF 603	3120+1G>A / 3120+1G>A	G / G	G / G	C / C	G / G	T / T	A / A
CF 638	U / U	G / G	A / G	A / C	G / T	G / G	A / G
CF 652	3120+1G>A / 3120+1G>A	G / G	nr	C / C	G / G	T / T	A / A
CF 736	U / U	G / G	G / G	C / C	G / T	G / T	A / A
CF 782	U / U	C / C	G / G	A / A	T / T	G / T	A / G
CF 800	U / U	G / G	A / G	A / A	G / T	G / G	A / A
CF 801	U / U	G / C	G / G	C / C	G / T	G / G	A / A
CF 826	U / U	C / C	A / A	A / C	G / G	G / G	G / G
CF 849	3120+1G>A / U	G / G	G / G	A / C	G / G	T / T	A / G

Key

nr	Indicates that there is no result
U	Indicates an unknown CFTR genotype
N	In obligate carriers, indicates where there is no CFTR mutation genotype
M	Mother
F	Father

Table F.3: Genotype results for Black unaffected controls

Controls	Met D	KM 19	J44	T845T	Tub 18	J32
	rs11770163	rs916727	rs43035	rs1042077	rs213989	rs3802012
RB1	G / G	G / G	C / C	nr	G / G	A / A
RB2	G / C	A / A	A / C	T / T	G / T	A / G
RB3	G / G	G / G	A / A	G / G	G / G	A / A
RB4	G / G	A / G	C / C	G / G	G / T	A / A
RB5	G / C	nr	A / C	G / G	G / G	A / A
RB6	nr	G / G	A / A	G / T	G / T	A / A
RB7	G / G	A / G	A / C	G / T	G / G	A / G
RB8	G / G	G / G	A / A	nr	T / T	A / G
RB9	G / C	G / G	A / C	T / T	G / G	A / A
RB10	G / C	A / A	C / C	G / G	T / T	A / A
RB11	G / G	G / G	A / C	G / G	G / G	A / G
RB12	G / C	nr	A / C	G / G	G / T	A / G
RB13	G / C	nr	A / C	G / T	T / T	A / A
RB14	C / C	A / G	A / C	G / G	G / G	A / G
RB15	G / C	nr	A / C	G / G	G / G	A / G
RB16	G / C	nr	C / C	G / G	G / G	A / G
RB17	C / C	G / G	C / C	G / T	G / T	A / G
RB18	G / C	G / G	C / C	G / G	G / T	A / G
RB19	nr	G / G	nr	G / G	G / G	A / A
RB20	G / C	A / G	A / C	G / G	G / G	A / A
RB21	G / C	nr	nr	G / T	nr	nr
RB22	G / C	G / G	A / C	G / G	G / T	A / A
RB23	nr	G / G	A / C	G / T	G / G	A / A
RB24	G / G	nr	C / C	nr	G / G	A / A
RB25	G / G	nr	A / C	G / T	G / G	nr
RB26	G / C	nr	A / A	G / T	T / T	A / A
RB27	G / C	nr	C / C	nr	G / G	A / G
RB28	G / C	nr	A / C	G / G	G / G	A / A
RB29	G / C	nr	A / C	G / G	G / T	A / A
RB30	G / G	nr	A / C	G / G	G / T	A / A
RB31	G / G	G / G	C / C	G / T	G / G	A / A
RB32	G / G	nr	A / C	G / T	G / G	A / A
RB33	G / G	G / G	C / C	G / T	G / G	A / G
RB34	G / G	G / G	A / A	T / T	G / G	A / A
RB35	G / C	nr	A / C	G / T	G / G	A / A
RB36	G / G	nr	C / C	G / T	G / G	A / G
RB37	G / C	nr	C / C	G / G	G / T	A / A
RB38	nr	nr	C / C	nr	G / G	A / A
RB39	C / C	nr	A / A	G / T	G / G	A / A
RB40	G / C	G / G	A / C	nr	T / T	A / A

Table F.3: Continued

Controls	Met D	KM 19	J44	T845T	Tub 18	J32
	rs11770163	rs916727	rs43035	rs1042077	rs213989	rs3802012
RB41	G / C		A / C	nr	G / G	A / A
RB42	G / G	G / G	C / C	G / G	G / T	A / G
RB43	C / C	G / G	A / C	G / G	T / T	A / A
RB44	G / G		A / C	G / T	G / G	A / A
RB45	G / C	G / G	C / C	G / T	G / G	A / G
RB46	G / C	nr	A / C	G / G	G / T	A / A
RB47	nr	nr	A / C	nr	G / G	A / G
RB48	C / C	G / G	A / C	G / G	G / G	A / A
RB49	C / C	G / G	C / C	G / G	G / G	A / G
RB50	G / G	G / G	A / C	G / G	G / T	A / G
RB51	G / G	G / G	C / C	G / T	G / G	A / A

Key

nr No result

RB Random Black

Table F.4: Genotype results for Coloured unaffected controls

Controls	Met D	Km 19	J44	T845T	TUB 18	J32
	rs11770163	rs916727	rs43035	rs1042077	rs213989	rs3802012
RC 1	G / C	A / G	A / C	T / T	G / G	A / A
RC 2	nr	A / G	A / C	G / T	G / G	A / A
RC 3	G / C	G / G	A / A	nr	G / G	A / G
RC 4	G / G	A / G	nr	G / T	G / G	A / G
RC 5	G / G	A / A	C / C	G / G	T / G	A / A
RC 6	G / C	A / A	A / C	G / G	G / G	A / G
RC 7	G / C	nr	A / A	G / G	G / G	A / G
RC 8	G / C	A / A	A / C	G / G	G / G	A / G
RC 9	G / C	A / G	A / C	G / G	T / G	A / A
RC 10	G / G	A / G	A / C	G / G	T / G	A / A
RC 11	G / G	A / A	A / A	G / G	G / G	A / G
RC 12	G / G	A / G	C / C	G / G	G / G	A / G
RC 13	G / G	A / G	A / A	G / G	G / G	A / A
RC 14	G / G	A / G	A / C	G / G	G / G	A / A
RC 15	G / G	A / G	C / C	G / G	G / G	A / G
RC 16	C / C	A / G	A / C	G / G	G / G	A / G
RC 17	G / C	A / A	A / C	G / G	G / G	A / A
RC 18	C / C	A / A	C / C	G / G	T / G	A / A
RC 19	G / C	nr	C / C	G / G	T / T	A / G
RC 20	G / C	G / G	A / C	nr	G / G	A / G
RC 21	G / C	A / A	A / C	G / G	T / G	A / G

Table F.4: Continued

Controls	Met D	Km 19	J44	T845T	TUB 18	J32
	rs11770163	rs916727	rs43035	rs1042077	rs213989	rs3802012
RC 22	C / C	A / A	A / C	G / G	T / G	A / A
RC 23	G / G	A / G	A / A	G / T	G / G	A / G
RC 24	nr	A / A	A / A	G / T	G / G	A / A
RC 25	G / G	A / A	C / C	G / T	T / G	A / A
RC 26	nr	A / G	A / A	G / G	G / G	A / G
RC 27	C / C	A / G	C / C	nr	T / T	nr
RC 28	G / G	nr	A / C	G / T	G / G	nr
RC 29	G / G	A / G	C / C	G / G	G / G	A / A
RC 30	G / G	A / A	A / C	T / T	G / G	A / A
RC 31	nr	nr	A / A	T / T	G / G	A / A
RC 32	G / G	A / G	A / C	T / T	G / G	A / A
RC 33	G / G	A / G	A / A	nr	G / G	A / A
RC 34	G / G	A / G	A / C	nr	G / G	A / A
RC 35	G / G	G / G	A / A	G / T	G / G	nr
RC 36	G / G	A / A	A / C	G / G	T / G	A / A
RC 37	G / C	G / G	C / C	G / G	G / G	A / A
RC 38	G / G	G / G	A / A	T / T	G / G	A / A
RC 39	G / G	A / G	A / A	G / G	T / T	A / G
RC 40	C / C	A / G	A / C	T / T	T / G	A / A
RC 41	G / C	G / G	A / A	T / T	G / G	A / A
RC 42	G / C	A / G	A / A	G / G	G / G	A / A
RC 43	G / C	G / G	A / C	G / G	G / G	A / G
RC 44	G / G	A / A	A / A	nr	T / G	nr
RC 45	G / G	A / A	C / C	G / G	G / G	A / A
RC 46	G / C	A / A	A / C	G / G	G / G	A / A
RC 47	G / G	A / A	C / C	G / G	G / G	A / A
RC 48	G / G	A / G	A / C	G / G	T / G	A / A
RC 49	G / C	G / G	A / C	G / G	G / G	A / A
RC 50	G / G	nr	A / A	nr	G / G	A / A

Key

nr	No Result
RC	Random Coloured

REFERENCES

1. Abeysinghe, S.S., Chuzhanova, N., Krawczak, M., Ball, E.V., Cooper, D.N. (2003) Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination associated motifs. *Hum Mutat.* **22**: 229– 244
2. Aradhya, S., Bardaro, T., Galgóczy, P., Yamagata, T., Esposito, T., Patlan, H., Ciccodicola, A., Munnich, A., Kenwrick, S., Platzer, M., D'Urso, M., Nelson, D.L. (2001) Multiple pathogenic and benign genomic rearrangements occur at a 35 kb duplication involving the NEMO and LAGE2 genes. *Hum Mol Genet.* **10**:2557-2567
3. Audrezet, M.P., Chen, J.M., Ranguenes, O., Chuzhanova, N., Giteau, K., Le Marechal, C., Quere, I., Cooper, D.N., Feréc, C (2004) Genomic Rearrangements in the CFTR Gene: Extensive Allelic Herterogeneity and Diverse Mutational Mechanisms. *Hum Mutation.* **23**:343-357
4. Bacolla, A., Jaworski, A., Larson, J.E. et al (2004) Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc Natl Acad Sci USA.* **101**:14162 –14167
5. Banjar, H. (1999) Geographic distribution of cystic fibrosis transmembrane regulator gene mutations in Saudi Arabia. *East Mediterr Health J.* **6**: 1230-1235
6. Barrett, J.C., Fry, J., Malleri, Daly, M.J. (2005) Haploview: Analysis and visualisation of LD and haplotype maps. *Bioinfo.* **21**:263-265
7. Bienvenu, T., Cartault, F., Lesure, F., Renouil, M., Beldjord, C., Kaplan, J.C. (1996) A splicing mutation in intron 16 of the cystic fibrosis transmembrane conductance regulator gene, associated with severe disease, is common on Reunion Island. *Hum Hered.* **46**:168-171
8. Blanco, P., Shlumukova, M., Sargent, C.A., Jobling, M.A., Affara, N., Hurles, ME. (2000) Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. *J Med Genet.* **37**:752-758

9. Bombieri, C., Bonizzato, A., Castellani, C., Assael, B.M., Pignatti, P.F. (2005) Frequency of large CFTR gene rearrangements in Italian CF patients. *Eur J Hum Genet.* **13**:687-689
10. Cabello, M. K. Pedro, H. Cabello, Juan, C. Llerena, J.R., Fernandes, O. (2006) Polymorphic Markers Suggest a Gene Flow of CFTR Gene from Sub-Saharan/Arabian and Mediterranean to Brazilian Population. *Journal of Heredity.* **97**:313-317
11. Carles, S., Desgeorges, M., Goldman, A., Thiart, R., Guittard, C., Kitazos, C.A., de Ravel, T.J., Westwood, A.T., Claustres, M., Ramsay, M. (1996) First report of CFTR mutations in black cystic fibrosis patients of Southern African origin. *J Med Genet.* **33**: 802-804
12. Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., Nickerson, D.A. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet.* **74**:106-120
13. Casilli, F., Di, Rocco, Z.C., Gad, S., Tournier, I., Stoppa-Lyonnet, D., Frebourg, T., Tosi, M. (2002) Rapid detection of novel BRCA1 rearrangements in high-risk breast-ovarian cancer families using multiplex PCR of short fluorescent fragments. *Hum Mutat.* **20**:218-226
14. Chapman, J. M., Cooper, J. D., Todd, J. A. & Clayton, D. G. (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered.* **56:**18-31
15. Chevalier-Porst, F., Bonardot, A.M., Chazalette, J.P., Mathieu, M., Bozon, D. (1998) 40 kilobase deletion (CF 40kb del 4-10) removes exons 4 to 10 of the Cystic Firbrosis Transmembrane Conductance Regulator gene. *Hum Mutat Suppl.* **1**:291-294
16. Cystic Fibrosis Genetic Analysis Consortium (1990) Worldwide survey of the ΔF508 mutation: report from the Cystic Fibrosis Genetic Analysis Consortium. *Am J Hum Genet.* **47:** 354-359

17. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., Lander, E. S. (2001) High-resolution haplotype structure in the human genome. *Na Genet.* **29**:229–232
18. Dawson, K.P., Frossard, P.M. (2000) The geographic distribution of cystic fibrosis mutations gives clues about population origins. *Eur J Pediatr.* **159**:496–9
19. des Georges, M., Ramsay, M., Guittard, C., Altieri,J., Templin,C., Rene, C., Beroud, C., Claustres, M. (2007) Spectrum of CFTR mutations in a group of black patients from South Africa: Identification of a novel large rearrangement. *J of cystic fibr.* suppl 1
20. Devilee P, Cleton-Jansen AM, Cornelisse, C.J. (2001) Ever since Knudson. *Trends Genet.* **17**:569-73
21. Dörk, T., El-Harith, E-H., Stuhrmann, M., Macek, M Jr., Egan, M., Cutting, G., Tzetis, M., Kanavakis, E., Carles, S., Claustres, M., Padoa, C., Ramsay, M., Schmidtke, J. (1998) Evidence for a common ethnic origin of cystic fibrosis mutation 3120+1G-to-A in diverse populations. (Letter) *Am J Hum Genet.* **63**: 656-662
22. Dörk, T., Macek, M. Jr., Mekus, F., Tummler, B., Tzountzouris, J., Casals, T., Krebsova, A., Koudova, M., Sakmaryova, I., Macek, M. Sr., Vavrova, V., Zemkova, D., Ginter, E., Petrova, N.V., Ivaschenko, T., Baranov, V., Witt, M., Pogorzelski, A., Bal, J., Zekanowsky, C., Wagner, K., Stuhrmann, M., Bauer, I., Seydewitz, H.H., Neumann, T., Jakubiczka, S. (2000) Characterization of a novel 21-kb deletion, CFTRdele2,3 (21 kb), in the CFTR gene: a cystic fibrosis mutation of Slavic origin common in Central and East Europe. *Hum Genet.* **3**:259-268
23. Dörk, T., Mekus, F., Schmidt, K., Bosbhammer, J., Fislage, R., Heuer, T., Dziadek, V., Neuman, T., Kalin, N., Wulbran, U., Wulf, B., von der Hardt, H., Maab, G., Tummler, B. (1994) Detection of more than 50 different CFTR mutations in a large group of German cystic fibrosis patients. *Hum Genet.* **94**:533-542
24. Dörk, T., Neumann, T., Wulbrand, U., Wulf, B., Kälin, N., Maass, G., Krawczakm, M., Guillermit, H., Feréc, C., Horn, G. (1992) Intra- and

extragenic marker haplotypes of CFTR mutations in cystic fibrosis families. *Hum Genet.* **88**:417-25

25. Eskandarani, H.A. (2002) Cystic fibrosis transmembrane regulator gene mutations in Bahrain. *J Trop Pediatr.* **48**:348-350
26. Estivill, X., Bancells, C., Ramos, C. (1997) Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. *Hum Mutat.* **10**:135-154
27. Estivill,X. Armengol,L. (2007) Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet.* **3**:1787-1799
28. Excoffier, L., Slatkin, M. (1998) Incorporating genotypes of relatives into a test of linkage disequilibrium. *Am J Hum Genet.* **62**:171-80
29. Férec, C., Casals, T., Chuzhanova, N., Macek, M. Jr., Bienvenu, T., Holubova, A., King, C., McDevitt, T., Castellani, C., Farrell, P.M., Sheridan, M., Pantaleo, S.J., Loumi, O., Messaoud, T., Cuppens, H., Torricelli, F., Cutting, G.R., Williamson, R., Ramos, M.J., Pignatti, P.F., Raguénès, O., Cooper, D.N., Audrézet, M.P., Chen, J.M. (2006) Gross genomic rearrangements involving deletions in the CFTR gene: characterization of six new events from a large cohort of hitherto unidentified cystic fibrosis chromosomes and meta-analysis of the underlying mechanisms. *Eur J Hum Genet.* **14**:567-576
30. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., Altshuler, D. (2002) The structure of haplotype blocks in the human genome. *Science.* **296:**2225-2229
31. Gabriel, S.E., Brigman, K.N., Koller, B.H., Boucher, R.C., Stutts, M.J. (1994) Cystic fibrosis heterozygote resistance to cholera toxinin the cystic fibrosis mouse model. *Science.* **266**:107–109
32. Gibson, L., Cooke, R. (1959) A test for concentration of electrolytes in sweat in cystic fibrosis of the pancreas utilizing pilocarpine by iontophoresis. *Pediatrics.* **23**:545-549

33. Gibson, R., Burns, J., Ramsey, B. (2003) Pathophysiology and management of pulmonary infections in cystic fibrosis. *Am J Respir Crit Care Med.* **168**:918-951
34. Giselda, M.K., Cabello, Pedro, H., Cabello, Juan, C., Llerena Jr., Octavio, Fernandes. (2006) Polymorphic Markers Suggest a Gene Flow of CFTR Gene from Sub-Saharan/Arabian and Mediterranean to Brazilian Population. *J Heredity*. **97**:313-317
35. Goldman, A., Graf, C., Ramsay, M. (2003) Molecular diagnosis of cystic fibrosis in South African populations. *S Afr Med J.* **93**:518-519
36. Goldman, A., Labrum, R., Claustres, M., Desgeorges, M., Guittard, C., Wallace, A., Ramsay, M. (2001) The molecular basis of Cystic Fibrosis in South Africa. *Clin Genet.* **59**:37-41
37. Green, A., Dodds, P., Pennock, C. (1985) A study of sweat sodium and chloride; criteria for the diagnosis of cystic fibrosis. *Ann Clin Biochem.* **2**:171-174
38. Grove, S. (1959) Fibrocystic disease of the pancreas in the Bantu. *S Afr J Lab Clin Med.* **2**:113-119
39. Guo, S., Thompson, E. (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics.* **48**: 361-372.
40. Hamosh, A., FitzSimmons, S.C., Macek, M. Jr., Knowles, M.R., Rosenstein, B.J., Cutting, G.R. (1988) Comparison of the clinical manifestations of cystic fibrosis in black and white patients. *J Pediatr.* **132**:255-259
41. Hantash, F.M., Redman, J.B., Anderson, B., Tubman, C., Starn, K., Buller, A., McGinniss, M., Quan, F., Sun, W., Strom, C.M. (2004) Frequent Ocurrance of DNA Rearrangements in Cystic Fibrosis Transmembrane Conductance Regulator Gene *ASHG* Abstract
42. Hantash, F.M., Redman, J.B., Goos, D., Kammesheidt, A., McGinniss, M.J., Sun, W., Strom, C.M. (2007) Characterization of a recurrent novel large duplication in the cystic fibrosis transmembrane conductance regulator gene. *J Mol Diagn.* **9**:556-560

43. Hill, I.D., MacDonald, W.B., Bowie, M.D., Ireland, J.D. (1988) Cystic fibrosis in Cape Town. *S Afr Med J.* **73**:147-149
44. Hogenauer, C., Santa, Ana, C.A., Porter, J.L., Millard, M., Gelfand, A., Rosenblatt, R.L., Prestidge, C.B., Fordtran, J.S. (2000) Active intestinal chloride secretion in human carriers of cystic fibrosis mutations: an evaluation of the hypothesis that heterozygotes have subnormal active intestinal chloride secretion. *Am J Hum Genet.* **67**:1422–1427
45. Horn, G.T., Richards, B., Merrill, J.J., Klinger, K.W. (1990) Characterization and rapid diagnostic analysis of DNA polymorphisms closely linked to the cystic fibrosis locus. *Clin Chem.* **36**:1614-1619
46. Hurles, M.E. (2001) Gene conversion homogenizes the CMT1A paralogous repeats. *BMC Genomics.* **2**:11-20
47. Hurles, M.E., Willey, D., Matthews, L., Hussain, S.S. (2004) Origins of chromosomal rearrangement hotspots in the human genome: evidence from the AZFa deletion hotspots. *Genome Biol*. **5**:55
48. Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F., Pinkel, D. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science.* **30**:818-821
49. Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., Tsui, L.C. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science.* **245**:1073-1080
50. Kilinic, M.O., Ninis, V.N., Dagli, E., Demirkol, M., Ozkinay, F., Arikan, Z., Cogulu, O., Hunter, G., Karakoc, F., Tolun, A. (2002). Highest heterogeneity for cystic fibrosis: 36 mutations account for 75% of all CF chromosomes in Turkish patients. *Am J Med Genet.* **113**:250-257
51. Knowles, M., Hohneker, K., Zhou, Z., Olsen, J., Noah, T., Hu, P., Leigh, M., Engelhardt, J., Edwards, L., Jones, K. (1995) A controlled study of adenoviral-vector-mediated gene transfer in the nasal epithelium of patients with cystic fibrosis. *N Engl J Med.* **13**:823-831
52. Lerer, I., Laufer-Chana, A., Rivlin, J.R., Augarten, A., Abeliovich, D. (1999) A large deletion mutation in the CFTR gen (3120+1Kbdel 8.6Kb):

a founder mutation in the Palestinian Arabs. Mutation in brief no. 231. Online. *Hum Mutat.* **13**:337

53. Levin, S.E., Blumberg, H., Zamit, R., Schmaman, A., Wagstaff, L. (1967) Mucoviscidosis (cystic fibrosis of the pancreas) in Bantu twin neonates. *S Afr Med J.* **19**: 482-485
54. Liu, K., Muse, S.V. (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics.* **21**:2128-2129
55. Long, J.C., Williams, R.C., Urbanek, M. (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet.* **56**:799-810
56. Lucotte, G., Hazout, S., De Braekeleer, M. (1995) Complete map of cystic Fibrosis mutation DF508 frequencies in Western Europe and correlation between mutation frequencies and incidence of the disease. *Hum Biol.* **67**: 797-803
57. Lucotte, G., Loirat, F. (1993) A more detailed map of the cystic Fibrosis mutation DF508 frequencies in Europe. *Hum Biol.* **65**: 503-507
58. MacDougall, L.G. (1962) Fibrocystic disease of the pancreas in African children. *Lancet.* **II:**409-410
59. Macek, M. Jr., Mackova, A., Hamosh, A., Hilman, B.C., Selden, R.F., Lucotte, G., Friedman, K.J., Knowles, M.R., Rosenstein, B.J., Cutting, G.R. (1997) Identification of common cystic fibrosis mutations in African-Americans with cystic fibrosis increases the detection rate to 75%. *Am J Hum Genet.* **60**:1122-1127
60. Magnani, C., Cremonesi, L., Giunta, A., Magnaghi, P., Taramelli, R., Ferrari, M. (1996) Short direct repeats at the breakpoints of a novel large deletion in the CFTR gene suggest a likely slipped mispairing mechanism. *Hum Genet.* **98**:102-108
61. Mateu, E., Calafell, F., Dolors, Ramos, M., Casals, T., Bertranpetit, J. (2002) Can a Place of Origin of the Main Cystic Fibrosis Mutations Be Identified? *Am J Hum Genet.* **70**:257–264
62. Meuller, J., Kanter-Smoler, G., Nygren, A.O., Errami, A., Grönberg, H., Holmberg, E., Björk, J., Wahlström, J., Nordling, M. (2004) Identification

of genomic deletions of the APC gene in familial adenomatous polyposis by two independent quantitative techniques. *Genet Test.* **8**:248-56

63. Mickle, J., Macek, M Jr., Fulmer-Smentek, S., Egan, M., Schwiebert, E., Guggino, W., Moss, R., Cutting, G. (1998) A mutation in the cystic fibrosis transmembrane conductance regulator gene associated with elevated sweat chloride concentrations in the absence of cystic fibrosis. *Hum Mol Genet.* **7**:729-735
64. Morral, N., Bertranpetit, J., Estivill, X., Nunes, V., Casals, T., Gimenez, J., Reis, A., Varon-Mateeva, R., Macek, M., Kalaydjieva, L., Angelicheva, D., Dancheva, R., Romeo, G., Russo, M.P., Garnerone, S., Restagno, G., Ferrari, M., Magnani, C., Claustres, M., Desgeorges, M., Schwartz, M., Dallapiccola, B., Novelli, G., Feréc, C., de Arce, M., Kere, J., Anvret, M., Dahl, N., Kadasi, L. (1994) The origin of the major cystic Fibrosis mutation (DF508) in European populations. *Nat Genet.* **7**: 169-175
65. Morral, N., Dörk, T., Llevadot, R., Dziadek, V., Mercier, B., Férec, C., Costes, B., Girodon, E., Zielenski, J., Tsui, L.C., Tümmler, B., Estivill, X. (1996) Haplotype analysis of 94 cystic fibrosis mutations with seven polymorphic CFTR DNA markers. *Hum Mutat.* **8**:149-159
66. Morral, N., Nunes, V., Casalas, T., Cobos, N., Aseniso, O., Dapena, J., Estivill, X. (1993) Uniparental inheritance of microsatellite alleles of the cystic fibrosis gene (CFTR): identification of a 50 kilobase deletion. *Hum Mol Genet.* **2**:677-681
67. Nectoux, J., Audrezet, M.P., Viel, M., Leroy, C., Raguenes, O., Feréc, C., Lesure, J.F., Davy, N., Renouil, M., Cartault, F., Bienvenu, T. (2006) A frequent large rearrangement in the CFTR gene in cystic fibrosis patients from Reunion Island. *Genet Test.* **10**:208-214
68. Padoa, C., Goldman, A., Jenkins, T., Ramsay, M. (1999) Cystic fibrosis carrier frequencies in populations of African origin. *J Med Genet.* **36**:41-44
69. Papadakis, M.N., Patrinos, G.P. (1999) Contribution of gene conversion in the evolution of the human beta-like globin gene family. *Hum Genet.* **104**: 117-125

70. Pearson, K. (1903) Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Phil Transact Royal Soc of L.* Ser A **200**: 1-66
71. Pier, G.B. (1999) Evolution of the DF508 CFTR mutation: response. *Trends Microbio.* **7**:56–58
72. Pier, G.B. (2000) Role of the cystic fibrosis transmembrane conductance regulator in innate immunity to *Pseudomonas aeruginosa* infections. *Proc Natl Acad Sci USA.* **97**:8822–8828
73. Pier, G.B., Grout, M., Zaidi, T., Meluleni, G., Mueschenborn, S.S., Banting, G., Ratcliff, R., Evans, M.J., Colledge, W.H. (1998) *Salmonella typhi* uses CFTR to enter intestinal epithelial cells. *Nature.* **393**:79–82
74. Pinkel, D., Straume, T., Gray, J.W. (1986) Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proc Natl Acad Sci U S A.* **83**:2934-2938
75. Qin, Z.S., Niu, T., Liu, J.S. (2002) Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet.* **71:**1242-1247
76. Riordan, J., Rommens, J., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J., Drumm, M., Iannuzzi, M., Collins, F., Tsui, L. (1989) Identification of the cystic Fibrosis gene: chromosome walking and jumping. *Science.* **245**:1059-1065
77. Romeo, G., Devoto, M., Galietta, L.J.V. (1989) Why is the cystic fibrosis gene so frequent? *Hum Genet.* **84**:1–5
78. Ronaghi, M. (1998) Pyrosequencing: A tool for sequence-based DNA analysis. Doctoral thesis, The Royal Institute of Technology, Stockholm, Sweden.
79. Ronaghi, M. (2000) Improved performance of Pyrosequencing using single-stranded DNA-binding protein. *Anal Biochem.***286:** 282–288
80. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M., and Nyren, P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Ana Biochem.* **242:** 84–89

81. Rosenstein, B., Cutting, G. (1998) The diagnosis of cystic fibrosis: a consensus statement. *J Pediatr.* **132**:589-595
82. Schouten, J.P., McElgunn, C.J., Waaijer, R., Zwijnenburg, D., Diepvens, F and Pals, G. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acid Research.* **30**: e57
83. Shrimpton, A.E., Borowitz, D., Swender, P. (1997) Cystic Fibrosis mutation frequencies in upstate New York. *Hum Mut.* **10**: 436-442
84. Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D., Albertson, D.G. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet.* **29**:263-264
85. Stephens, M., Donnelly, P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet.* **73**:1162–1169
86. Stephens, M., Scheet, P. (2005) Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation. *Am J Hum Genet.* **76**:449–462
87. Stephens, M., Smith, N.J., Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* **68**:978–989
88. Stringer, C.B. (1990) The emergence of modern humans. *Sci Am.* **263**: 98-104
89. The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature.* **437**:1299-1320
90. Turcios, N. (2005) Cystic Fibrosis An overview. *J Clin Gastroenterol.* **39**:307-317
91. Tzetis, M., Kanavakis, E., Antoniadi, T., Doudounakis, S., Adam, G., Kattamis, C. (1997) Characterization of more than 85% of cystic fibrosis alleles in the Greek population, including five novel mutations. *Hum Genet.* **99**:121-125

92. Welsh, M., Tsui, L., Boat, T., Beaudet, A. (1995) Cystic Fibrosis. In: Scriver CR, Beaudet AL, Sly WS, Valle D (eds) The metabolic and molecular basis of inherited disease 7th edition McGraw-Hill New York: 3799-3940
93. Westwood, T. (2007) The epidemiology of cystic fibrosis in the Western Cape province. *S Afr Med J.* **1**: 78-81
94. Wigginton, J., Cutler, D., Agbecasis, G.R. (2005) A note on exact tests of Hardy-Weinberg equilibrium. *Am J H Genet.* **76**: 887-893
95. Zielenski, J., Tsui, L. (1995) Cystic Fibrosis: Genotypic and Phenotypic variations. *Ann Rev Genetics.* **29**:777-807

ONLINE REFERENCES

The following websites were accessed between February and November 2007:

Cystic Fibrosis Foundation.: http://www.cysticfibrosis.ca/page.asp?id=1
CF mutation database http://www.genet.sickkids.on.ca/cftr/
HapMap database: www.hapmap.org
Database of genomic variants: http://projects.tcag.ca/variation
Protein database: http://www.expasy.org
Genome browser at UCSC: www.genome.UCSC.org
Genome browser at ENSEMBL: www.ensembl.org/
PSQ figures: http://www.pyrosequencing.com/graphics/2983.gif
MLPA figures: http://www.mlpa.com