Rarefaction as a tool to determine variant diversity in monogenetic disorders

By

Jeanne van Rensburg

26043484

Submitted in fulfilment of the requirements for the degree

Masters of Science (Immunology)

In the Faculty of Health Sciences

Department of Immunology

University of Pretoria

30 April 2015

Supervisor: **Prof. M.S. Pepper**

Co-supervisor: **Dr. M. Alessandrini**

Ethics Statement

The author, whose name appears on the title page of this dissertation, has obtained, for the research described in this work, the applicable research ethics approval.

The author declares that she has observed the ethical standards required in terms of the University of Pretoria's *Code of ethics for researcher* and the *Policy guidelines for responsible research*.

Summary

Genetic diversity is a well-described concept within many biological disciplines. However, mathematical models determining genetic diversity are often applied within ecological disciplines and are rarely explored within the medical field. Given that genetically associated disorders and complications can occur at high frequency in developing countries, the primary aim of this study was to determine whether or not diversity theory could be applied to disease-associated variants. Two monogenic disorders were selected for this purpose – one commonly observed disorder known as cystic fibrosis (CF), and one rare disorder known as metachromatic leukodystrophy (MLD). Despite being a common monogenic disorder, the clinical and molecular presentation of CF in the different population groups of South Africa is largely unknown. Thus, the medical records of 45 CF patients attending the Steve Biko Academic Hospital CF clinic were investigated to better understand the manifestation of this disorder in these patients. Additionally, molecular data was collected for both CF and MLD through published reports and analysed via the Shannon-Weaver, Simpson, Simpson Diversity, and rarefaction diversity methods. The rarefaction method was found to be the most informative measure of diversity and a potentially powerful tool to employ in the development and/or refinement of population-specific screening panels.

Key words: Genetic diversity, South Africa, monogenic, Cystic fibrosis, Metachromatic leukodystrophy, Shannon index, Simpson index, Simpson Diversity index, effective species numbers, rarefaction.

Acknowledgements

My Lord and Saviour, my Healer, Deliverer and Redeemer! All praise be unto You - Your mercy is great and Your faithfulness endures forever. I am humbled by the grace I am under. Thank you Jesus!

To my family: Dad, Mom, Ouma and Boet – You may not always understand the nature of my work, but your patience and understanding through this process has been undeniable and is sincerely appreciated. Thank you for all your love and support.

To Prof Michael Pepper and Dr Marco Alessandrini: It's finally finished! I do not have words to describe my gratitude for the roles you have played in this project. Thank you for your belief in my abilities, your support and encouragement from start to end, your guidance, corrections and patience. I doubt I would have been able to achieve even half of this without your respective inputs – thank you a million times over!

To Prof Refiloe Masekela and Dr Cheryl Stewart: Thank you for all your help in the data collection process. To Prof Mark Robertson: Thank you for your time in helping us fully understand the outputs from our diversity analyses. To Prof Robert Colwell: Thank you for your willingness to aid us with our data processing challenges. There would not have been much of a project without any of your efforts and I am truly grateful for the contribution each of you made.

To all my colleagues at the ICMM: Thanks for all the laughs, providing tissues and hugs for all the tears, for the food (notably in the form chocolate, coffee, and cupcakes [Danielle]), and for all your encouragement. What an exceptional privilege to be a part of this awesome team of people!

Finally, to Pastors Bernard & Belinda Mulder, Philip & Anrie le Roux, the Musto family (Uncle David, Aunty Wendy, Neil and Lesley), Ivan Kruger, Ryan Rothschild, Olivier Zablocki, Wolfgang Wehrmeyer, Febé Meyer, and James Mehl: Thank you for your prayers, your love, your patience, your mentorship and friendship. It truly means the world to me.

THE BEST IS YET TO COME!

Table of Contents

Ethics Statement i

Summary ii

Acknowledgements iii

List of Figures vii

List of Tables viii

List of Equations x

List of Abbreviations and Symbols xi

Chapter 1 – Introduction 1

Chapter 2 – Reviewing Two Monogenic Disorders in the South African Context: Historical and Statistical Perspectives 5

2.1 Introduction 5

2.2 Population bottlenecks, founder effects and the generation of mutational diversity 6

2.3 Diversity indices: The new, the old and the permanent debates 14

2.4 Final remarks: The past paves the way for the future 33

Chapter 3 – Investigating the presence of cystic fibrosis in patients attending Steve Biko Academic Hospital 37

3.1 Introduction 37

3.1.1 Biochemistry of Cystic Fibrosis 37

3.1.2 Clinical presentation 41

3.1.3 Molecular diagnostic aspects 42

3.1.4 Diagnostic capacity in South Africa 44

3.2 Methods 45

3.2.1 Study population and ethical considerations 45

3.2.2 Clinical, laboratory and molecular presentation 46

3.2.3 Data collection 47

3.2.4 Statistical analysis 48

3.3 Results 49

3.3.1 Demographics 49
3.3.2 Clinical presentation 50
3.3.3 Biochemical and Molecular test results 54
3.3.4 CFTR molecular data 58
3.4 Discussion and conclusion 62
3.4.1 Demographic and clinical presentation of patients attending the SBAH CF clinic 62
3.4.2 Biochemical and Molecular test results 64
3.4.3 CFTR molecular data 65
Chapter 4 – Cystic Fibrosis and diversity: Unification through variation 68
4.1 Introduction 68
4.2 Methods 71
4.2.1 Data assimilation 72
4.2.2. Diversity analysis 73
4.3 Results 75
4.3.1 Data assimilation 75
4.3.2 Diversity analysis 85
4.3.2.1 Matrix 1 85
4.3.2.2 Matrix 2 89
4.3.2.3 Matrix subset 3 94
4.4 Discussion 101
4.4.1 Matrix 1 101
4.4.2 Matrix 2 103
4.4.3 Matrix 3 104
4.5 Conclusions 105
Chapter 5 – A global perspective of MLD: A review of incidence, prevalence and mutation data 107
5.1 Introduction 107
5.2 Materials and Methods 113
5.2.1 Data assimilation 113
5.2.2 Diversity analysis 114
5.3 Results 117

5.3.1 Variant comparison 117
5.3.2 Variant Diversity 123
5.4 Discussion and Conclusion 127
5.4.2 Variant comparison 127
5.4.3 Variant Diversity comparison 129
Chapter 6 – Concluding Discussion 130
References 134
Appendix A: Informed consent for SBAH patients 143
Appendix B: Ethical Approval Certificate – 40-2014 148
Appendix C: Ethical Approval Certificate – 4-2013 149

List of Figures

Figure 1: Map detailing the territories of the Gcaleka Xhosa, Shaka Zulu and Mzilikatzi 10

Figure 2: Graphical description of how population bottlenecks and founder effects are established 12

Figure 3: Hypothetical example of a rarefaction curve 23

Figure 4: Summary of the CFTR system 39

Figure 5: Graphical representation of South African CF patients attending SBAH according to ethnicity and gender 50

Figure 6: Graphical representation of the number of South African patients attending the CF clinic at SBAH in relation to the number of patients who underwent sweat electrolyte and PFE-1 tests 56

Figure 7: Frequency distribution (%) of the five positively identified mutational types present in 42 CF patients attending the SBAH CF clinic 58

Figure 8: Hypothetical example of gene diversity for a high-cholesterol gene in three different populations ... 69

Figure 9: Algorithm used to determine world-wide CF variant diversity 74

Figure 10: Shannon and Simpson Diversity Effective Species Number analysis of Matrix 1 87

Figure 11: Rarefaction analysis of Matrix 1 88

Figure 12: Shannon and Simpson Diversity Effective Species Number analysis of Matrix 2 91

Figure 13: Rarefaction analysis of Matrix 2 93

Figure 14: Shannon, Simpson and Simpson Diversity ESN analysis of Matrix 3 97

Figure 15: Rarefaction analysis of adjusted Matrix 3 variant data 99

Figure 16: Schematic representation of cerebroside sulfate hydrolysis 108

Figure 17: Algorithm used to determine MLD variant diversity 116

Figure 18: Combined frequency of the c.459+1G>A, P426L, and I179S variants 120

Figure 19: Frequency of the c.459+1G>A variant (%) reported in various countries 121

Figure 20: Frequency comparison of the P426L variant (%) 122

Figure 21: Shannon and Simpson ESNs for MLD variant subset data 126

Figure 22: Rarefaction analysis of MLD variant subset data 127

List of Tables

Table 1: Summary of commonly used diversity indices 32

Table 2: Summary of the number of CF and genetics clinics available in South Africa in 2012 (20) 44

Table 3: Average weight, height and BMI values in Black and White South African CF patients attending the SBAH CF clinic 51

Table 4: Liver damage and pancreatic insufficiency associated with South African CF patients attending SBAH 52

Table 5: Illustration of South African patients attending the SBAH CF clinic who were diagnosed with any combination of MIE/DIOS, GORD, osteoporosis, infertility and/or CF-associated diabetes 54

Table 6: Average sweat electrolyte of select South African CF patients attending SBAH 55

Table 7: Summary of the number of infections with various bacterial species identified in the lungs of South African CF patients attending SBAH CF clinic 57

Table 8: Summary of South African CF variants reported in literature 60

Table 9: Distribution of CF variants in the South African population according to literature 62

Table 10: World-wide frequency (%) occurrence of 17 CF variants 77

Table 11: Variant subset data - Matrix 1 80

Table 12: Variant subset data - Matrix 2 82

Table 13: Variant subset data - Matrix 3 84

Table 14: Variant values used for Shannon and Simpson Diversity analysis in CF Matrix 1 through simple random sampling 85

Table 15: Results from Shannon-Weaver and Simpson Diversity analysis of Matrix 1 86

Table 16: Relative values used for Shannon and Simpson Diversity analysis in CF Matrix 2 89

Table 17: Results from Shannon-Weaver and Simpson Diversity analysis of Matrix 2 90

Table 18: Adjusted Matrix 2 variant data 92

Table 19: Comparison of the significance of variant diversity between countries represented in Matrix 2 94

Table 20: Relative values used for Shannon and Simpson Diversity analysis in CF Matrix 3 95

Table 21: Results from Shannon-Weaver and Simpson Diversity analysis of Matrix 3 96

Table 22: Adjusted Matrix 3 variant data 98

Table 23: Comparison of the significance of variant diversity between countries represented in Matrix 3 100

Table 24: Frequency of the most commonly reported allelic variants of MLD in 18 countries 119

Table 25: MLD variant subset used in diversity analysis 124

Table 26: Shannon, Simpson and Simpson Diversity analysis of MLD variant subset data 125

List of Equations

Equation 1: Effective population equation when differences in population sizes exist between generations 16

Equation 2: Change in hetezygotic allele frequency between generations 16

Equation 3: Allelic richness equation developed by Hulbert in 1971 18

Equation 4: Kalinowski's 2004 rarefaction equation determining the expected number of private alleles in a population 19

Equation 5: Kalinowski's 2004 rarefaction equation determining the expected number of regionally private alleles present in a population 19

Equation 6: Smith and Grassle's 1977 equation describing allele richness 20

Equation 7: Smith and Grassle's 1977 equation determining private allele richness 20

Equation 8: Individual-based rarefaction calculation as developed by Colwell *et al*. in 2012 22

Equation 9: Shannon-Weaver Diversity index 25

Equation 10: Shannon-Weaver Diversity index sample evenness equation 26

Equation 11: Shannon-Weaver Effective Species Number equation 26

Equation 12: Simpson index equation in populations of infinite size 27

Equation 13: Simpson index equation in populations of finite size 27

Equation 14: Simpson Diversity/Gini-Simpson Diversity index in populations of infinite size 28

Equation 15: Simpson Diversity/Gini-Simpson Diversity index in populations of finite size 29

Equation 16: Simpson Reciprocal index 29

Equation 17: Simpson index effective species number equation 29

Equation 18: Simpson Diversity index effective species number equation 29

Equation 19: BMI equation for adults who are 21 or older (117) 46

List of Abbreviations and Symbols

ARSA	Arylsulfatase A gene
ASA	Arylsulfatase A
ATPase	Adenosinetriphosphatase
BMI	Body mass index
c.	chromosomal
cAMP	cyclic Adenosine monophosphate
CDC	Centers for disease control and prevention
CF	Cystic Fibrosis
CFTR	Cystic Fibrosis transmembrane conductance regulator gene
CI	Confidence interval
CIs	Confidence Intervals
cm	centimetre
CRC	Convention on the Rights of the Child
CT	Computerized Tomography
CVS	Chorionic Villus sampling
ELISAs	Enzyme-linked immunosorbent assays
ENaC	Epithelial Sodium Channel
ER	Endoplasmic reticulum
ESNs	Effective Species Numbers
et al.	*et alia*
GA	Golgi Apparatus
GORD	Gastro-oesophageal reflux disease
H.	*Haemophilus*
HIV/AIDS	Human Immunodeficiency Virus/Autoimmune deficiency syndrome
i.e.	*id est*
Indel	Insertion/deletion
ISI	Institute for Scientific Information
K^+ channel	Potassium ion channel
kb	kilobase
kg	kilogram
L	Litre
m	meter
Max	Maximum
MIE/DIOS	Meconium ileus/Distal intestinal obstruction syndrome
Min	Minimum
MLD	Metachromatic Leukodystrophy
mmol	micromoles
mmol/L	micromoles/Litre
MRI	Magnetic Resonance Imaging
mRNA	Messenger Ribonucleic Acid
MRSA	Methicillin resistant *Staphylococcus aureus*
N	Sample size
Na^+	Sodium ion
Na^+/K^+ ATPase	Sodium Potassium Adenosine Triphosphatase
$Na^+/K^+/2Cl^-$	Sodium/Potassium/Chloride co-transporter system
NBFs/NBDs	Nucleic Binding Folds/Nucleic Binding Domains
NHLS	National Health Laboratory Services
No.	Number of
OMIM	Online Mendelian Inheritance of Man
ORCC	Outwardly rectifying chloride channel
P.	*Pseudomonas*
P_2Y_2	Purinergic receptor
PEM	Protein Energy Malnutrition
PFE-1	Pancreatic Faecal Elastase-1

PKA	Protein kinase A
PSAP	Prosaposin gene
R-domain	Receptor-domain
ROMK	Renal rectifying outer medullary potassium channel
RSA	Republic of South Africa
S.	*Staphylococcus*
SAPB	Saposin B
SBAH	Steve Biko Academic Hospital
SD	Standard Deviation
SE	Standard Error
TB	Tuberculosis
UAE	United Arab Emirates
UK	United Kingdom
UN	United Nations
USA	United States of America
v.	Version
x.	Gene-coverage factor
µg/g	microgram/gram
I	Roman numeral for 1
II	Roman numeral for 2
III	Roman numeral for 3
IV	Roman numeral for 4
V	Roman numeral for 5
%	Percentage
Δ	Delta
Σ	Summation
/ OR −	Division
-	Subtraction/negative value
×	Multiplication
+	Addition
±	Plus/minus
′	Derivative of
Π	Product of
∈	Element of
$\hat{\pi}$	Allele richness
λ	Species diversity
α	Allelic richness
Cr	Total number of combinations in which r can be sampled from R
D	Simpson effective species number
$\widetilde{D}$	Simpson Diversity index
$\overline{D}$	Average duration of disease
${}^{1}D$	Shannon-Weaver effective species number
${}^{2}D$	Simpson Reciprocal OR Simpson Diversity effective species number
$E_{H'}$	Shannon-Weaver diversity evenness
Exp	Exponent
f_k	Total number of alleles found k times within randomly selected individuals
g	gene
H'	Shannon-Weaver diversity
I	Incidence
i	i^{th} element
j	j^{th} element
k	k^{th} region OR Total number of times a randomly selected allele is found within randomly selected individuals
m	distinct alleles
N or n	Sample size
Ne	Effective population size

Ni OR ni	Size of population in the i^{th} generation
Nj	Number of genes in the j^{th}population
Nij	Number of copies of the i^{th} allele in the j^{th}population
P	Prevalence
pi	Proportion of Ni as a function of N
$Pijg$	Probability of finding the i^{th} allele in the j^{th}population from a sample of g genes
$Qijg$	Probability of not finding the i^{th} allele in the j^{th}population from a sample of g genes
r	Sub-region that is studied
R	Total number of regions studied
S	Number of Variants
$S_{(est)}$	Estimated number of variants
$\tilde{S}_{ind}$	Expected number of alleles within randomly selected individuals
S_{obs}	Total number of alleles found in a studied population
t	Time
u	Number of i^{th} alleles present in the sample
X_k	Set of populations from region k
Y_{kcr}	C^{th} set of combinations r can be sampled from region k

Chapter 1 – Introduction

The South African population is both genetically and culturally diverse, a fact that has garnered South Africa the title of "The rainbow nation". However, in order to gain an understanding of how such diversity could have originated within South Africa, it is necessary to study historical events and to adopt a multi-disciplinary approach.

Used extensively within ecological disciplines, three frequency-based methods, namely, the Shannon-Weaver index (1), the Simpson index (2), the Simpson Diversity/Gini-Simpson index (3-5), are typically applied in diversity studies. However, many have realised the potential short-comings of these three methods and have subsequently suggested improved ways through which diversity can be determined. These methods include the Shannon and Simpson effective species numbers (ESNs) (6-11), as well as rarefaction (12-16).

Shannon and Simpson ESNs were developed to indicate the relative number of variants that would need to be sampled in a country, region or population in order to generate the observed level of diversity. Although described as being more informative than their complementing equations, both Shannon and Simpson ESNs are dependent on frequency-based results derived from the Shannon-Weaver, Simpson and Simpson Diversity indices (6-8). Rarefaction, contrary to these methods, has been favoured in recent years due to its unique approach to diversity determination – diversity is determined as a function of all

variants that are observed in a sampled country, region or population. Rarefaction additionally determines diversity taking variations in sample size into account (12-16).

The application of diversity theories in the medical disciplines is scant. More particularly, diversity theories have to our knowledge not been investigated with regard to monogenic disorders in humans, let alone their associated disease-causing alleles. The key aim of this project was therefore to determine whether or not diversity theories could be applied in the medical context of monogenic disorders, and if so, ascertain the best method to use in determining disease-causing variant diversity.

In order to determine whether or not diversity theory could be applied effectively in this context, it was decided that it would be necessary to investigate its application in both a commonly observed monogenic disorder (cystic fibrosis) as well as comparatively rare monogenic disorder (metachromatic leukodystrophy). It was thus important to determine 1) how incidence, prevalence and variant data for these disorders compare to other countries, either within or outside of African borders, and 2) if successfully applied, the way(s) in which each of the diversity methods meaningfully contribute to existing knowledge of our already genetically heterogeneous population.

Cystic fibrosis (CF) is one of the most well studied monogenic disorders globally, including South Africa, and has an approximate incidence of 1/2,000 – 1/3,000 in White South Africans (17, 18), 1/4,600 Black South Africans (19), and 1/10,300 – 1/12,000 in Mixed race South

Africans (17, 18). This autosomal-recessive disorder is caused by mutations in the cystic fibrosis transmembrane conductance regulator gene (*CFTR*). In South Africa, the two most frequently observed CF variants, ΔF508 and c.3120+1G>A, are respectively observed in 76% and 46% of White and Black South African CF populations (17). Nevertheless, the clinical presentation of CF can vary significantly between patients. Typically affecting children, but also present in adults, approximately 85-90% of all CF patients will suffer from pancreatic insufficiency, while all patients will experience pulmonary complications and have decreased sweat chloride concentrations (20).

As a stark contrast to CF, metachromatic leukodystrophy (MLD) is a comparatively rare monogenic disorder - not only in South Africa but in other countries too. To date, only one published report of MLD in South Africa exists (21). MLD is an autosomal recessive disorder, neurodegenerative in nature, and is associated with the arylsulphatase A (*ARSA*) gene. It is reported to occur in every 1/40,000 to 1/100,000 live births in patients of European descent (22). Three variants (c.549+1G>A, P426L and I179S) are commonly observed in Caucasian patients of European descent, having been described in 45-60% of this population group (23, 24). Despite being a comparatively rare disorder, the clinical presentation of MLD has been very well documented and is dependent on the age at which it manifests. Although disease severity decreases with increased age at diagnosis, MLD is a severely debilitating disorder and is fatal in nature (25-27).

We were thus able to formulate the following hypothesis for the present study: diversity theories which are commonly applied in ecological studies can successfully be applied in the medical field in order to investigate the molecular diversity of disease-associated variants. In order to do this, it was necessary to consider historical events that could have contributed to the known genetic diversity within South Africa. In order to determine which diversity method would be the best to use for future studies of this nature, it was also necessary to test the applicability of several diversity methods. This dissertation therefore concludes with a summary of the outcomes of the study and discusses the implications thereof.

Chapter 2 – Reviewing Two Monogenic Disorders in the South African Context: Historical and Statistical Perspectives

2.1 Introduction

"Population genetics is frequently considered an abstract and theoretical subject, of no relevance to the real world. The truth is that population genetics is an increasingly important component of many areas of mainstream biology." As stated in 2004 by Halliburton in the "Introduction to population genetics", one cannot overlook the use and application of many population genetics theories and theorems in a variety of biological fields (28). In the medical disciplines, one of the most common applications of population genetics principles can be observed when studying sickle cell anaemia. Described in 1910, this disorder has been used to illustrate not only the predictive power that population genetics can have in the field of medicine, but also how genetic variation can be beneficial to an entire population (28, 29). Thus, the concept of genetic variation/diversity is neither new nor difficult to understand, with Nei (30) defining gene variation as being "the probability that two randomly chosen copies of a gene will be different alleles". However, in order to gain an appreciable idea of the extent to which genetic diversity/genetic variation can be described in the medical disciplines, one need only initially focus on Mendelian disorders.

Mendelian disorders, also known as monogenic disorders, are found the world over. As many as 6,000 monogenic disorders have been described to date, while countless more are likely to be described in the future (31, 32). The number of Mendelian disorders and the associated

causative mutations in each country, region or population can vary significantly. Additionally, many different theories regarding the observed variability in the occurrence of these disorders and their associated disease alleles within and between populations/geographical regions have been developed. Some of these theories include the "selfish-gene theory", the "theory of natural selection and mutational fitness", and "Muller's ratchet" to name but a few (33-36). As epitomized by Jost (6-8) and irrespective of which theory holds more credit, variant and disease diversity is an obvious and somewhat intuitive concept. Here we will investigate the pros and cons of four different measures of diversity and seek to introduce these concepts in the context of two monogenic disorders – one commonly observed and well-studied disorder, namely cystic fibrosis, and one rarely observed but well-studied disorder, namely metachromatic leukodystrophy.

2.2 Population bottlenecks, founder effects and the generation of mutational diversity

In order to understand the methods that are used to determine genetic diversity, it is important to understand some of the events generating diversity in the first place. Notably, diversity can be affected by many different external influences. Factors such as war, migration and natural disasters are known to influence diversity, often resulting in genetic drift through population bottlenecks and founder effects (28, 37-40). South African populations have not been immune to such changes or influences. Two clear-cut examples of such changes include 1) the Xhosa cattle-killing event (41) and 2) the loss of millions of lives at the hands of Zulu warriors (42, 43).

Many South Africans, and even more Xhosa individuals, would be able to explain what the cattle-killing event entailed and how it affected the Gcaleka Xhosa people at the time. Occurring between April 1856 and May 1857, it has been estimated that over 400,000 cattle were slaughtered by roughly 85% of all adult Gcaleka Xhosa men. It is further speculated that 40,000 Gcaleka Xhosas died during this period and that 40,000 left their homes in search of sustenance and new areas in which to settle (41, 44). However, as described by Peires (41, 44), the cattle-killing event occurred as a consequence to a prophesy given to the Gcaleka by the Xhosa prophetess, Nongqawuse. Extracted from Peires' paper of 1987 (41), the prophesy reads as follows:

"*It happened in one of the minor chiefdoms among the Gcaleka Xhosa, that of Mnazabele, in the year 1856. Two girls went out to guard the fields against birds. One was named Nongqawuse, the daughter of Mhlkaza, and the other was very young. At the river known as the place of the Strelitzia, they saw two men arriving. These men said to the girls – Give our greetings to your homes. Tell them we are So-and-so… and they told their names, those of people who had died long ago. Tell them that the whole nation will rise from the dead if all living cattle are slaughtered because these have been reared with defiled hands, since there are people about who have been practising witchcraft. There should be no cultivation. Great new corn pits must be dug and new houses built. Lay out great big cattle-folds, cut out new milk-sacks, and weave doors from buka roots, many of them. So say the chiefs, Napakade, the son of Sifuba-sibunzi. The people must abandon their witchcraft, for it will soon be revealed by diviners.*"

After already having been involved in eight Frontier wars between 1779 and 1853, the cattle-killing event of 1856, as well as the ninth and final Frontier war that ended in 1879 would have a major effect on the Gcaleka population (Figure 1). Although the exact number of Xhosa deaths, and ultimately the number of Gcaleka Xhosa that remained alive during and after the nine Frontier wars is unknown, when considered in parallel with the loss of lives and large-scale migration of survivors during the cattle-killing event, a picture of significant population decline starts to develop.

Following the final Frontier war in the early 19th century, Zulu tribes had started their own wars. One of the most notable characters appearing in these wars was Shaka Zulu (42). Having been rejected by his father Senzangakhona, Shaka and his mother Nandi found refuge in the presence of king Dingiswayo. Under the guidance and training of Dingiswayo's leadership and army, Shaka became one of the most revered warriors in Zulu history. The death of Dingiswayo through the hands of his nemesis, Zidwe, as well as the death of Shaka's half-brother Sigujana at Shaka's own hand, afforded Shaka the opportunity to firmly establish his reign over the Zulu peoples. After acquiring a vast following, Shaka specialised in expertly training all men fit for battle in specific military formations, strategies and close-range combat (42, 45, 46).

Mzilikatzi, although trained through Zidwe's armies, was among those who had sworn allegiance to Shaka Zulu. However, Mzilikatzi's personal desires for wealth and power resulted

in him fleeing from Shaka's control in 1822 while later conquering most of the Northern Transvaal and Zimbabwean (formally Rhodesian) territories (Figure 1) (43). Mzilikatzi also ventured into Botswana, splitting his 15,000 supporters into two groups in order to control the Botswana region too. Unfortunately for Mzilikatzi, tsetse fly infestations in Botswana caused the death of many of his people's cattle and as a result drove Mzilikatzi to travel north and eastwards. He eventually moved into Zimbabwean territories where he died on the 22 September 1868 and was laid to rest at the Matopo Hills (43).

It is known that Shaka Zulu's reign only lasted 10 years (1818-1828), but in that time is has been written that his wars were "accompanied by great slaughter and caused many migrations. Their effects were felt even far north of the Zambezi River" (42). As shown in Figure 1, the territories under Shaka's control were so immense that it has been estimated that at least 1-2 million people met their fate at the hands of Shaka's warriors (42, 45). However, after having all women bearing his children put to death, Shaka died without an heir and was killed by his bodyguard, Mbopha, and his two half-brother's, Dingane and Mhlanga in 1828. Nevertheless, the mass murders and migration caused by Zidwe, Dingiswayo, Shaka Zulu and Mzilikatzi, as well as the death and migration of over 80,000 Gcaleka Xhosa affords the opportunity to explore the concept of population bottlenecks, founder effects and their influence on genetic diversity.

Figure 1: Map detailing the territories of the Gcaleka Xhosa, Shaka Zulu and Mzilikatzi

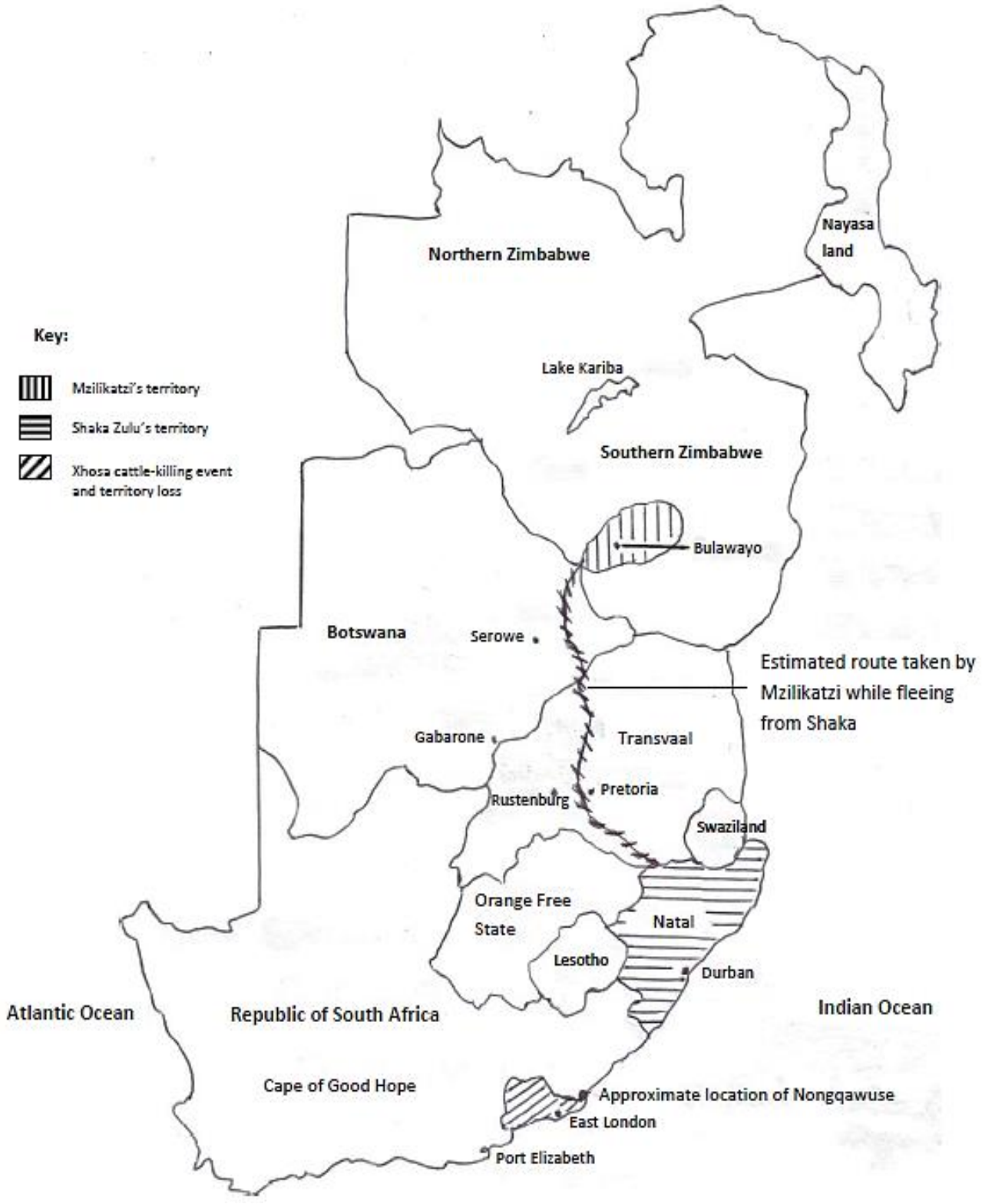
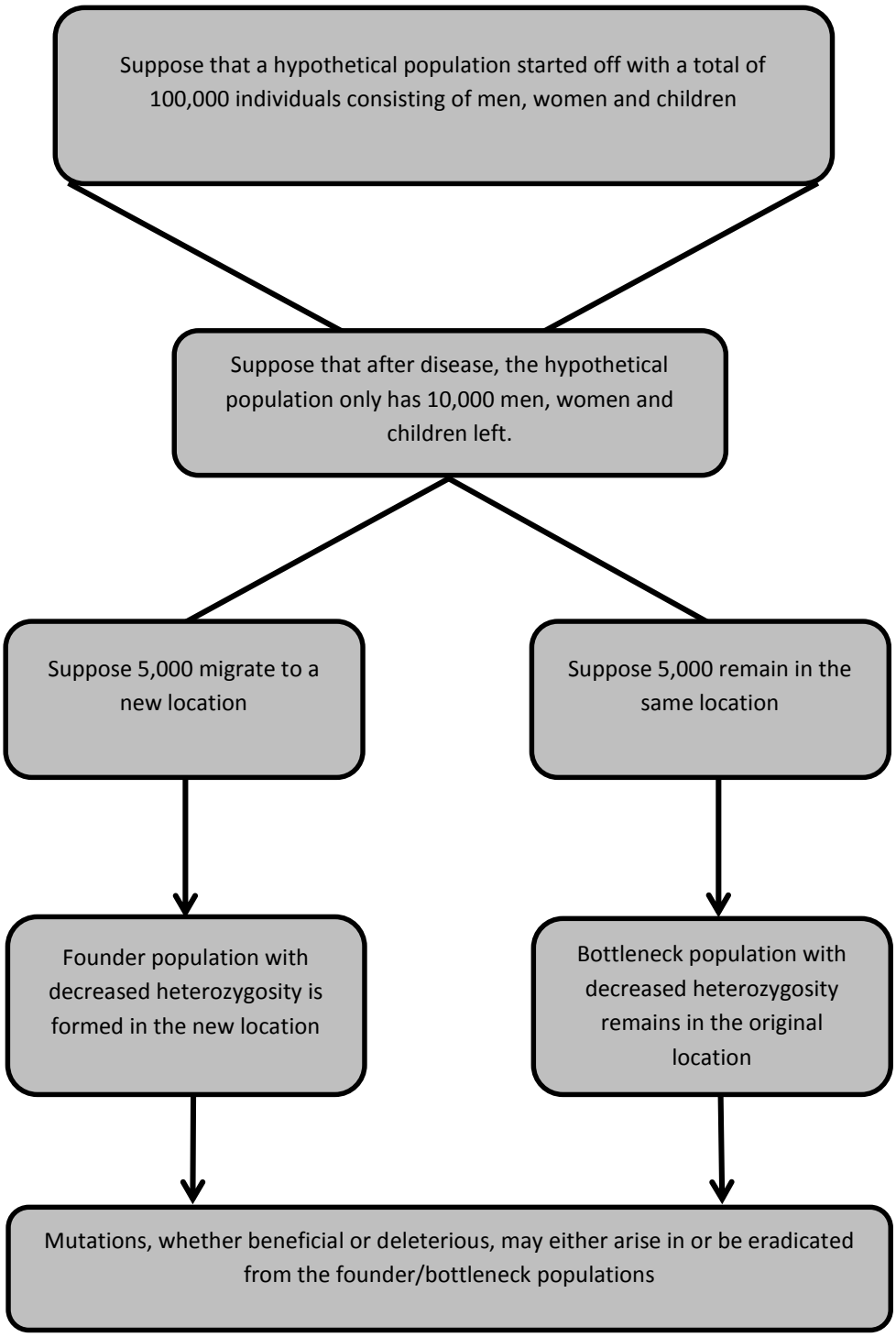


Figure 1 shows a map indicating the territories occupied by Mzilikatzi and Shaka Zulu during the 19th century. The Xhosa territory affected by some of the Frontier wars, as well as the cattle-killing event is also indicated. The map has been adapted from The Reader's Digest Great World Atlas (47), Peires (41), and several websites (48-50).

Population bottlenecks and founder effects are often inseparable terms and tend to occur in parallel with each other. Best illustrated in Figure 2, population bottlenecks occur when a significant reduction in the size of a population occurs over a period of time (28), while Mayr (40) describes a founder event as being "the establishment of a new population by a few original founders (in an extreme case, by a single fertilized female) which carry only a small fraction of the total genetic variation of the parental population". Such reductions in populations can be caused by many different factors: migration (as seen in the Xhosa population) and wide-spread death (such as those causes by Shaka Zulu) are just some examples. Thus, the period of time in which the bottleneck or founder event may occur may either be extensive or short-lived.

However, from the perspective of genetic variation, the effects that bottleneck and founder events have on populations has been thoroughly studied – both processes result in the inevitable loss of rare gene variants within the population(s) and, ultimately, a reduction in genetic heterozygosity takes place. More importantly though, it is known that population bottleneck and founder events are associated with altered allele frequencies. Additionally, given that genetic drift is described as being the "random variation in allele frequencies from one generation to the next" (28), population bottleneck and founder events are important measures in the study of genetic drift (28, 40, 51). It is thus an unsurprising fact that the observed effects of genetic drift are greater the smaller founder/bottleneck populations are.

Figure 2: Graphical description of how population bottlenecks and founder effects are established



Genetic drift has the power to "cause beneficial variations to be eliminated or deleterious variants to become common in a population" (28) and explains why some disorders, such as familial hypercholesterolaemia and sickle-cell anaemia, are observed more frequently in some populations or regions than in others (52, 53). Familial hypercholesterolaemia is commonly associated with Afrikaner populations in South Africa and results in elevated low-density lipoprotein cholesterol concentrations in affected individuals. This has been associated with increased risk of cardio-vascular diseases and complications and is thus noteworthy within this particular population group (53). Similarly, sickle-cell anaemia, although not specific to any one population, tends to be observed at increased frequencies in regions that experience high levels of malaria infections. Although compound heterozygous states result in sickle-shaped red blood cells that impair oxygen uptake within these cells, heterozygous states provide a buffering effect against the malaria parasite thereby resulting in an adaptive immunity to malaria infections (52). Nevertheless, each bottleneck and/or founder event will result in genetic drift which ultimately creates measureable and distinct genetic variation between populations despite reducing genetic variation within populations (37-39, 54, 55).

With the existence of events such as the Xhosa cattle-killing and frontier wars, as well as the devastation caused by Shaka Zulu and Mzilikatzi, it can clearly be shown how easily population bottlenecks could be formed and how founder populations could have arisen in some of South Africa's African populations. However, population bottleneck events have occurred in other South African populations and have been well-described in the Afrikaner peoples of South Africa. "Die Groot Trek", which occurred during the 1830's and 1840's, has been used as an

example of such events (56). South Africa is also host to Mixed race populations. These populations are generally considered to be genetically heterogeneous (57-59).

More importantly, these happenings provide an example of how bottleneck/founder events could have contributed to the establishment or eradication of beneficial and/or deleterious variants within these populations. Furthermore, given that approximately 200 years have passed since these events occurred and by considering the substantial changes that transpired in several of the populations' structures since then, two pivotal questions can be asked. Firstly, to what extent have distinct and harmful variants become established within modern day South African populations; and secondly, is it possible to determine the genetic diversity within and between modern day South African populations?

2.3 Diversity indices: The new, the old and the permanent debates

The effects that genetic drift can have in a collective population can be somewhat misleading. Wright (38) realised this conundrum – from a genetic perspective, genetic drift causes populations to act as if they are smaller than what they truly are. In Layman's terms, the genetic variation of a population of 100,000 people could effectively be represented by only 10,000 people. Irrespective of the numbers, the point is that the actual population, when it has experienced genetic drift, will behave as if it is smaller than it truly is. The smaller "population" was referred to by Wright (38) as being the effective population and several

equations exist through which the size of such a population can be mathematically determined (37-39, 60, 61).

Effective population sizes can be determined under many different conditions (28). One such condition is when the size of the studied population(s) between successive generations is uneven (as will be observed under population bottleneck or founder events). Thus, as shown in Equation 1 and using Mzilikatzi's story as an example, the effective population size of Mzilikatzi's following can be determined. It is known that upon entering into Botswana, Mzilikatzi split his 15,000 supporters into two groups of 7,500 people each (43). Let us hypothetically assume that the population that remained with Mzilikatzi increased to 8,500, 13,000 and 20,000. According to Wright's equation (38), the effective population size from when Mzilikatzi's 15,000 supporters split until the population had increased to 20,000 individuals, can be determined as being approximately 10,600 . We can thus see that the effective population is much smaller than the actual population size. This method can be applied to any population in a practical manner, as long as population growth or decline data is available.

It is known that genetic drift results in decreased heterozygotic states and altered allele frequencies (28). If the population size(s) is/are known, it is possible to determine the change in the frequency of heterozygotic states in the affected population (Equation 2). If we furthermore assume that heterozygous allelic states were in Hardy-Weinberg equilibrium and therefore present at a frequency of 0.5 in the example of Mzilikatzi's party, we can calculate

that the change in the frequency of heterozygotic allele states at the time of the party's split into two groups of 7,500 people was 0.00008333 (0.0083% reduction). As an arbitrary and more extreme comparison, had 10 of the 15,000 individuals branched off, the change in heterozygotic allele frequency could be calculated as 0.0951 (9.51% reduction).

Equation 1: Effective population equation when differences in population sizes exist between generations

$$Ne = t/\sum_{i=1}^{t} \frac{1}{Ni}$$

Ne represents the effective population size, while t indicates the total number of generations under observation. Ni represents the actual population size in the i^{th} generation (38).

Equation 2: Change in hetezygotic allele frequency between generations

$$H_t = H_0\left(1 - \frac{1}{2N}\right)^t$$

H_t represents the frequency of the heterozygous allelic state in the t^{th} generation, while H_0 indicates the initial frequency of the heterozygous allelic state within the sampled population. N represents the size of the population in each of the generations under investigation, while t indicates time in respect to the number of generations studied.

In terms of understanding the genetic makeup of a particular population, it is important to consider effective population size as it will "describe the drift experienced by the actual population" (28, 38). Nevertheless, one might (rightfully) be tempted to ask the question "so what"? Recall that although genetic drift reduces genetic heterozygosity, it also has the power to "cause beneficial variations to be eliminated or deleterious variants to become common in

a population" (28). This is important to note due to the fact that mutations are involved in the process of generating genetic diversity between populations, despite reducing genetic diversity within affected populations.

Nei (30) describes gene diversity as being "the probability that two randomly chosen copies of a gene will be different alleles". Traditionally, genetic diversity has usually been well studied and described in instances where variants are neutral in effect, but what of deleterious/disease-causing variants? We can furthermore ask how genetic diversity is determined, and to what extent is it relevant in the medical disciplines.

In order to answer these questions, we must first understand that diversity can be ascribed to species on either a genomic/genotypic or phenotypic level. This has, in the past, caused so much outrage that Hurlbert rather infamously stated "The term 'species diversity' has been defined in such various and disparate ways that it now conveys no information other than 'something to do with community structure'; species diversity has become a nonconcept" (12). Although viewed as a radical statement at the time, his intention to overcome the obstacle created in defining "species diversity" was validated with the development of the diversity method now known as rarefaction (Equation 3) (12). Predecessors to this method, Hill (9) and MacArthur (11), both understood the need to take a multi-level approach to determine species/variant diversity within and between selected populations.

Equation 3: Allelic richness equation developed by Hulbert in 1971

$$a_g^{(j)} = \sum_{i=1}^{m} Pijg \qquad \text{(a)}$$

Where

$$Pijg = 1 - Qijg \qquad \text{(b)}$$

And

$$Qijg = \frac{\binom{Nj - Nij}{g}}{\binom{Nj}{g}} = \prod_{u=0}^{g-1} \frac{Nj - Nij - u}{Nj - u} \qquad \text{(c)}$$

Hurlbert (1971) showed that it is possible to calculate the allelic richness (denoted by a_g) from randomly selected genes (g) in the j^{th} population. $Pijg$ indicates the probability that such a sample of g genes will contain the i^{th} allele in the j^{th} population from a total of m distinct alleles observed at a selected locus/gene region (3a). Hurlbert (12) further showed that $Qijg$ measures the probability that a sample of g genes will not contain the i^{th} allele in the j^{th} population. Thus, if $Qijg$ is known, $Pijg$ can be calculated and vice versa (3b). Solving for Q, Hurlbert (12) showed that Nj indicates the total number of genes sampled from the j^{th} population, while Nij indicates the number of copies of the i^{th} allele from the j^{th} population. $Qijg$ is thus calculated by considering the number of g genes that do not include the i^{th} allele in the j^{th} population, without replacement $\binom{Nj - Nij}{g}$ as a factor of the total number of combinations of g genes that can be made from the genes present in the j^{th} population $\binom{Nj}{g}$, without replacement. The total number of i^{th} alleles found in the sampled region/population is represented by u (3c).

Hurlbert (12) originally derived this method to compensate for uneven sample sizes, but also to compensate for the presence rare alleles/variants present within a population. By taking into consideration that two types of rare alleles (private and regionally private alleles) can be present in a population, Kalinowski expanded on Hurlbert's 1971 equation (13). Private alleles, as described by Kalinowski (13) simply refer to "alleles that are observed in only one population", while regionally private alleles refer to "alleles that are observed in only one region". Due to the difference in definition, Kalinowski developed two separate equations, respectively shown in Equation 4 and Equation 5, to calculate the presence and impact of private and regionally private alleles in a sampled population (13). However, in the derivation

of his equation, Kalinowski incorporated methods and nomenclature developed by Smith and Grassle (62). Thus, as shown in Equation 6 and Equation 7, the approaches developed by Smith and Grassle (1977) are capable of determining the allele richness of and the number of private allele expected in a selected population, thereby generating unbiased results as well as providing the minimum expected variance in variant diversity found in the studied population(s).

Equation 4: Kalinowski's 2004 rarefaction equation determining the expected number of private alleles in a population

$$\prod_{r,g}^{\widehat{(k)}} = \sum_{i=1}^{m}\left\{\frac{1}{Cr}\sum_{c_{1=1}}^{S_1}\sum_{c_{2=1}}^{S_2}\dots\sum_{c_{R=1}}^{S_R}\left[\left(\sum_{j\in Y_{kc_k}r}\left(P_{ijg}\prod_{\substack{j\prime\in Y_{kc_k}r\\ J'\neq j}}Q_{ij'g}\right)\right)\prod_{\substack{k'=1\\ k'\neq 1}}\prod_{j\in Y_{kc_{k\prime}r}}Q_{ijg}\right]\right\}$$

Equation 5 indicates the method developed by Kalinowski (13) in determining the expected number of private alleles that could be found in region k. The notation for both equations 4 and 5 is the same, where S_k represents the number of sampled populations within region k. R indicates the number of regions that were studied, while r is indicative of the standardised number of populations studied per region. C_r calculates the total number of ways in which the standardised number of populations per region (r) can be sampled in the number of regions studied (R). X_k represents the set of populations from region k, while Y_{kcr} represents the cth set of the different combinations from which the standardised number of mutations can be made from the number of sampled populations within region k $\binom{Sk}{r}$. Y_{kcr} thus also represents the r subset of X_k.

Equation 5: Kalinowski's 2004 rarefaction equation determining the expected number of regionally private alleles present in a population

$$\hat{\rho}_{r,g}^{(k)} = \sum_{i=1}^{m}\left\{\frac{1}{C_r}\sum_{c_1=1}^{S_1}\sum_{c_2=1}^{S_2}\dots\sum_{c_R=1}^{S_R}\left[\left(1-\prod_{j\in Y_{kc_kr}}Q_{ijg}\right)\prod_{\substack{k'=1\\ k'\neq k}}^{R}\prod_{j\prime\in X_{k\prime c_{k\prime}r}}Q_{ijg}\right]\right\}$$

As developed by Kalinowski (13), Equation 6 indicates the number of regionally private alleles that is expected to be found in region k. Notation used in Equation 5 and Equation 6 are identical.

Equation 6: Smith and Grassle's 1977 equation describing allele richness

$$\hat{\alpha}_g^{(j)} = a_g^{(j)} = \sum_{i=1}^{m} Pijg$$

Equation 6 indicates the equation developed by Smith and Grassle (62) and the concept and nomenclature which Kalinowski incorporated into his rarefaction equation. Allelic richness $\left(\hat{\alpha}_g^{(j)}\right)$ in the j[th] population is approximately equal to the "minimum variance, unbiased estimate of the allelic richness" (13) as determined by $\left(a_g^{(j)}\right)$.

Equation 7: Smith and Grassle's 1977 equation determining private allele richness

$$\hat{\pi}_g^{(i)} = \sum_{i=1}^{m}\left[Pijg\left(\prod_{\substack{j'=1 \\ j'\neq j}}^{J} Qij'g\right)\right]$$

Equation 7 shows the method developed by Smith and Grassle (62) used to calculate private allele richness (π). The number of private alleles that can be expected in the i^{th} $\left(\hat{\pi}_g^{(i)}\right)$ sampled population, given that g genes have been sampled from all populations, is denoted by J. Just as Hurlbert's 1971 equation, $Pijg$ calculates the probability that at least one i allele will be found in the j^{th} population if g genes are studied from the j^{th} population (12). The probability that not a single i allele will not be found in the j^{th} population if g genes are sampled, is denoted by $\left(\prod_{\substack{j'=1 \\ j'\neq j}}^{J} Qij'g\right)$. According to Kalinowski (13) this method is also a "minimum variance, unbiased estimator" measuring private allele richness.

Although crediting both Smith and Grassle (62), as well as Burnham and Overton (63) for the approach followed in the development of their rarefaction and variance estimation equations, Colwell *et al.* (14) realised, like Kalinowski, that there were limitations to the traditional rarefaction approach. However, unlike Kalinowski, Colwell *et al.* do not distinguish between private alleles and regionally private alleles, but rather between the type of "sample unit" used in the rarefaction analysis - the sampling unit undergoing rarefaction analysis could be either individual-based or sample-based.

Colwell *et al.* (14) defined an individual-based sampling unit as being "individuals, ideally sampled randomly and independently" which were then "counted and identified to species", while "a trap net, quadrat, plot or a fixed period of survey time … that are sampled randomly and independently" was used to define a sample-based sampling unit. However, Colwell *et al.* (14) further subdivided sample-based sampling units as representing either abundance or incidence data. Sample-based abundance data is used "if the number of individuals for each species appearing within each [sample-based] sampling unit can be measured or approximated", while Sample-based incidence data is used if only the "presence or absence of each species in each [sample-based] sampling unit can be accurately recorded". Nevertheless, given that individual-based sampling units would be used within the context of monogenic disorders, sample-based sampling units will not be discussed any further.

As shown in Equation 8, Colwell *et al.* not only took a simplistic view in the development of their individual-based rarefaction equation, but incorporated this approach in the statistical program known as EstimateS (14, 64). One of the greatest advantages of EstimateS, is that, it not only determines the expected number of alleles within a selected region/population, but is simultaneously able to calculate and describe all supporting statistics. These supporting statistics have been constructed to provide an unbiased estimate of variance, as developed by Colwell et al. (14, 15), and are an in-built function of EstimateS from version 9 upwards (64).

Equation 8: Individual-based rarefaction calculation as developed by Colwell *et al*. in 2012

$$\tilde{S}_{ind}(m) = S_{obs} - \sum_{k=1}^{n} \left[\frac{\binom{n-k}{m}}{\binom{n}{m}} \right] f_k$$

As developed by Colwell *et al.* (14), $\tilde{S}_{ind}(m)$ indicates the expected number of alleles within m randomly selected individuals, where S_{obs} represents the total number of alleles found within the studied population, n represents the sampled number of individuals within m randomly selected individuals, k represents the total number of times a randomly selected allele is found within m randomly selected individuals, and f_k represents the total number of alleles that are found k times within m randomly selected individuals.

Ultimately, the rarefaction method determines diversity as a function of species abundance – results are therefore indicative of the expected number of species that could be found in a sampled region/population (12-15). Moreover, the rarefaction method, through the construction of rarefaction curves, indicates 1) the level of diversity (either high or low), 2) how many disease chromosomes are present in the sampled region(s), as well as 3) the rate of diversification. As shown in the hypothetical example of a rarefaction curve in

Figure 3, the y-axis is indicative of how high or low diversity in each sampled region/population is – the higher the value along this axis, the greater the diversity. The number of disease chromosomes identified in a selected country, region or population is indicated by the x-axis – the further right along this axis, the greater the number of sampled chromosomes. The gradient of the rarefaction curve indicates the rate of diversification – the steeper the gradient of the rarefaction curve, the higher the rate of diversification. Conversely, the more gradual the curve's gradient, the slower the rate of diversification is.

Figure 3: Hypothetical example of a rarefaction curve

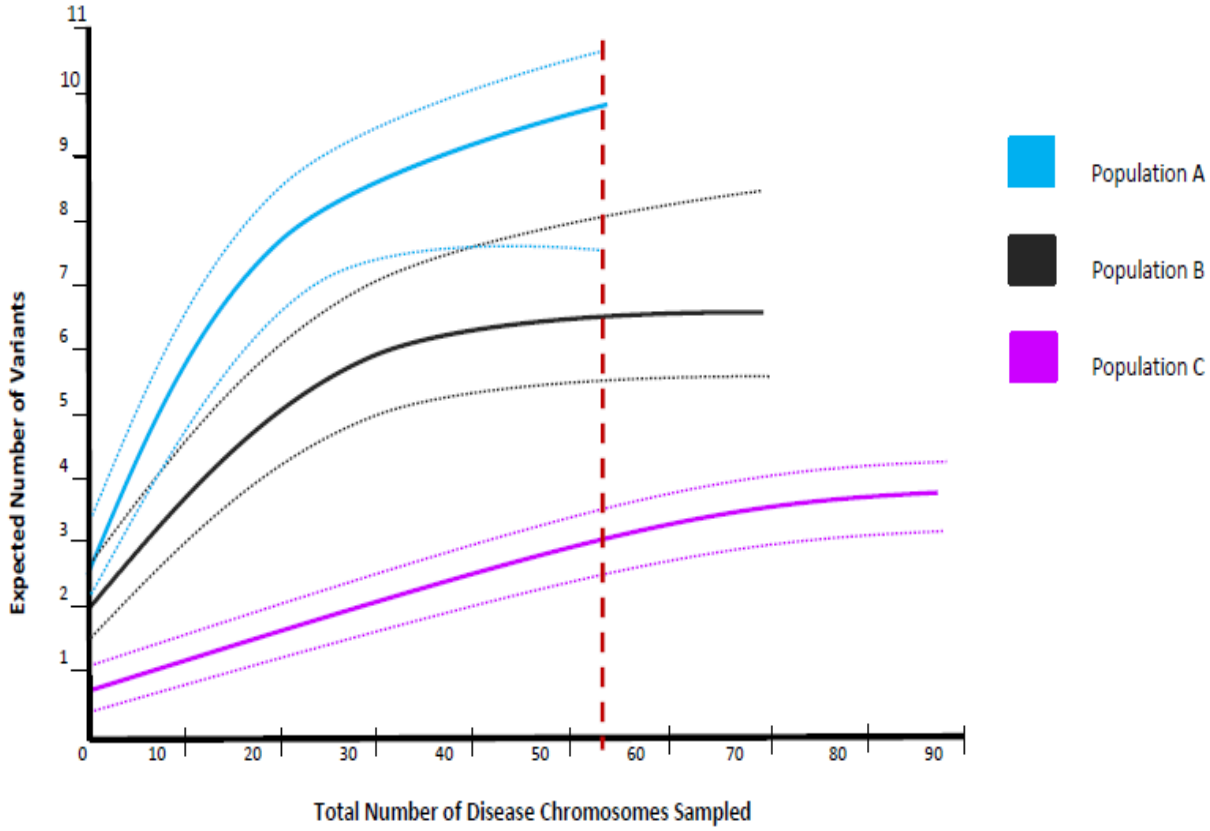


Figure 3 indicates an example of rarefaction curves. Solid lines indicate the calculated variant diversity per population, while dotted lines indicate the CI values. Diversity results between populations A and B are observed as not significant, but the difference in diversity between populations A and B in comparison to population C is shown to be statistically significant. The dashed red line indicates the point at which the comparison took place.

Furthermore, upper and lower 95% confidence intervals (CIs) are used to indicate whether the diversity between various sampled regions/populations is statistically significant. Diversity differs significantly if CI values between sampled regions/populations do not overlap. Diversity is deemed to not differ significantly if the CIs between sampled regions/populations do overlap (14, 15, 65, 66). Lastly, rarefaction curves that reach a plateau

are indicative of sampling saturation. This implies that, for a given set of variants investigated, the highest fraction of variants that could realistically be found within the sampled region, were actually found (14, 65).

Regardless, the rarefaction method is not the only way in which diversity can be measured. The Shannon diversity index (also known as the Shannon-Weaver or Shannon-Weiner diversity index), the Simpson index, the Simpson Diversity index, and the Simpson Reciprocal index, are just four more examples of methods that can be used to calculate diversity. However, an important aspect of each of these mathematical models is the concept of "richness", and "evenness", where Peet describes richness as being "an indicator of the relative wealth of species in a community" and evenness as the "concept concerning the evenness with which importance is divided among species" (67). Thus, in simplifying the definition for "evenness", we can say that evenness is a measure of how equal each component/species of a studied population is in terms of sample size. Importantly, although the description of richness as given by Peet is the most accepted form of the term "richness", one of the biggest challenges with the concept of richness, is that it is dependent on the sample sizes (9, 67).

Nonetheless, in transposing this information from an ecological to medical perspective, we can use patient ethnicity as an example. If in region "A" we find that 40 Black South African, 20 Mixed race and 70 White South African patients were examined, then we have not only described the sample distribution (i.e. sample evenness) by evaluating the number of patients

studied per ethnic group in region "A", but also how many ethnic groups were studied in region "A" (i.e. species richness). In this example, the sample sizes in the three studied ethnic groups are uneven.

Possibly the most well-known of the various indices, the Shannon-Weaver index was developed in 1949 by Shannon and Weaver and popularised through Margalef's application of this index to the field of ecology in 1957 (1, 67, 68). Determining both species richness (Equation 9) and sample evenness (Equation 10), this index assumes that the greater the area(s) that is sampled, the greater the diversity is likely to be. It is however important to recognise that although this method determines diversity, it does so by calculating the frequencies of the selected entity (n_i) as a function of the total population (N). This ultimately leads to the development of biases towards variants that occur at higher frequencies when compared to rarer alleles occurring at low frequencies within the sampled populations/regions (6, 7, 9, 10, 67).

Equation 9: Shannon-Weaver Diversity index

$$H' = -\sum_{i=1}^{S} pi \log pi \qquad \text{(a)}$$

Where

$$pi = \left(\frac{n_i}{N}\right) \qquad \text{(b)}$$

Equation 9 described the diversity method developed by Shannon and Weaver in 1949 (1) and examined in depth by Jost (6-8). The number of species observed in the sampled population is denoted by S. In some instances, R (representing species richness) will be used instead of S (a). pi is represented by the number of times a particular species/entity (n_i) is found relative to the size of the total population (N) in the region under study (b).

Equation 10: Shannon-Weaver Diversity index sample evenness equation

$$E_{H'} = \frac{H'}{H'_{max}} = \frac{H'}{logS}$$

Equation 10 indicates the method utilised in determining sample evenness ($E_{H'}$) when using the Shannon-Weaver index to determine diversity. H' represents the frequency obtained from the Shannon Diversity index while S represents the total number of species within the total sampled population (1, 9, 67).

Due to the frequency-based nature of the Shannon index, results will always fall between one and zero. However, in order to interpret results obtained from using the Shannon-Weaver index, one must bear in mind that outcomes are not only indicative of high or low species diversity, but also of how even sample sizes are. This therefore makes diversity comparison between populations fairly challenging, and for that reason Hill developed a technique known as "Effective Species Numbers", "Hill numbers", or "Hill ratios" (6-9). As shown in Equation 11, effective species numbers allow the frequency results generated by the Shannon-Weaver diversity index to be converted into the "equivalent number of equally abundant species that would be needed to give the same value of the diversity measure" within the selected population/region (69). When comparing results within and between populations, this method becomes noticeably valuable and useful.

Equation 11: Shannon-Weaver Effective Species Number equation

$${}^{1}D = \ Exp\ (H')$$

As seen in Equation 11 and as developed by Hill, the Shannon effective species number is determined by calculating the exponent of H' (6-9).

Also in 1949, Simpson developed a different method by which diversity could be measured. Simply termed as the Simpson index, his method determines the likelihood that, within a specific region/population, two randomly selected individuals/entities will be identical (2). Additionally, Simpson had the foresight to consider the fact that samples selected from a region/population are usually finite in size (2, 67). He thus proposed two equations – one able to calculate the diversity in populations of infinite size (Equation 12), and the other to determine diversity in populations with finite sizes (Equation 13).

Equation 12: Simpson index equation in populations of infinite size

$$\lambda = \sum_{i=1}^{s} p_i \qquad \text{(a)}$$

Where

$$pi = \left(\frac{n_i}{N}\right) \qquad \text{(b)}$$

As developed by Simpson, Equation 12 describes the Simpson index for populations of infinite size (2). λ represents the species diversity, while the total number of species in the sampled population(s) is represented by (s). The frequency of the i^{th} species in the sampled population is represented by p_i (a). , p_i represents the number of the selected species/entity (n_i) in a sampled population as a function of the total size of the sampled population(s) (N) (b).

Equation 13: Simpson index equation in populations of finite size

$$\lambda = \sum \left(\frac{n_i\,(n_i - 1)}{N\,(N - 1)}\right)$$

Equation 13 represents the way in which diversity is determined using the Simpson index in a population of a finite size, where n_i is the number of times the selected species/entity is observed in the sampled population/region, and N represents the total size of the sampled population.

The Simpson index is similar to the Shannon-Weaver diversity index in that it 1) calculates the frequency of identified species as a function of the total population from which samples were taken, and 2) calculates the evenness of sample sizes and species numbers per studied population. This method, like the Shannon-Weaver index, thus tends to lead to biases in favour of species found at higher frequency within sampled regions/populations. However, this particular model, unlike the Shannon-Weaver diversity index, indicates high levels of diversity as results tend towards zero and low diversity as results tend towards one.

Recognising this problem and by including methods devised by Gini (3), Greenberg (4) and Pielou (5) modified Simpson's original index and developed what is now known as the Simpson Diversity/Gini-Simpson index (Equation 14 and Equation15), while Williams (70) and MacArthur (11), respectively, established the Simpson Reciprocal index (Equation16) and Simpson Effective Species Number method (Equation17) (6, 7, 67). Using the methods proposed by MacArthur, Jost (6, 7) developed the Simpson Diversity Effective Species Number equation in order to convert diversity results obtained from using the Simpson Diversity/Gini-Simpson index (Equation18) into the relative species richness within the sampled region/population.

Equation 14: Simpson Diversity/Gini-Simpson Diversity index in populations of infinite size

$$\widetilde{D} = 1 - \sum p_i$$

Equation 14 indicates Simpson Diversity index (also referred to as the Gini-Simpson Diversity index) describing sampled populations with infinite sample sizes. Nomenclature used here is the same as those used in equations 13 – 15, where the frequency of the i^{th} species in the sampled population is represented by p_i (2-5).

Equation 15: Simpson Diversity/Gini-Simpson Diversity index in populations of finite size

$$\widetilde{D} = 1 - \sum \left\{ \frac{[n_i\ (n_i - 1)]}{[N\ (N - 1)]} \right\}$$

Equation 15 describes the Simpson Diversity index for sampled populations of finite sizes. The number of times the selected species/entity is observed in the sampled population/region is represented by n_i, while N represents the total size of the sampled population (2-5).

Equation 16: Simpson Reciprocal index

$$^2D = \frac{1}{\sum p_i}$$

As developed by Williams (70), Equation 16 shows the Simpson Reciprocal index calculation $\left(^2D\right)$ for populations of infinite size. The frequency of the i^{th} species in the sampled population is represented by p_i .

Equation 17: Simpson index effective species number equation

$$D = \frac{1}{\lambda}$$

Equation 17 shows the Simpson Effective Population Size method developed by MacArthur (11), where λ represents the diversity value calculated by using the Simpson index method.

Equation 18: Simpson Diversity index effective species number equation

$$^2D = \frac{1}{(1 - \widetilde{D})}$$

As derived by Jost (6, 7), Equation 18 shows the Simpson Diversity Effective Population Size, where $\widetilde{D}$ represents the diversity value calculated by using the Simpson Diversity index method.

The Simpson reciprocal index, contrary to the Simpson index and Simpson diversity index, outputs results that are indicative of the species richness found within a sampled

population/region (Equation 16). Additionally, both the Simpson and Simpson Diversity effective species number equations (Equation 17 and Equation 18) alleviate the challenges associated with comparing Shannon diversity index results to those respectively derived from the Simpson and Simpson Diversity index methods, and were developed in order to indicate the "equivalent number of equally abundant species that would be needed to give the same value of the diversity measure" (6-9, 11, 67, 69, 70).

However, in order to compare the three frequency-based diversity methods, namely the Shannon, Simpson and Simpson Diversity indices, it must be assumed that equal sampling effort has occurred within sampled regions. By "equal sampling effort", it is implied that the same amount of time and effort has been exerted in generating sampling results from sampling region "A" as was exerted in sampling region "B". In relaying this information from an ecological point of view to a medical one, equal sampling effort would imply that an equal number of patients had been tested for the same disease-causing variants in each sampled population/region. Nevertheless, as these methods are frequency-based, results will range between zero and one only. The Shannon-Weaver and Simpson Diversity indices indicate high levels of diversity when frequency results tend towards one. Conversely, high levels of diversity are ascribed to frequency results that tend towards zero when using the Simpson index (1, 2, 6, 7, 11).

Finally, as summarised in Table 1, it is important to note that when diversity equations calculate allele richness, the generated results do not require conversion into their effective

species while allele frequency results, for comparative purposed, must be converted into their effective species numbers/counts found per sampled region/population (6-9, 11, 67, 69).

Table 1: Summary of commonly used diversity indices

Diversity Method	Method function?	Result conversion required?	Conversion method available?	References
Rarefaction	Measures allele richness, true measure of diversity. Takes into consideration variation is sample size a. Unbiased measure of diversity.	No	Not Applicable	(12-15, 67, 71, 72)
Shannon Diversity index	Measures allele frequency and sample evenness. Biases are created towards the specie(s) with the highest frequency. Low and high diversity tend towards zero and one respectively.	Yes	Yes	(1, 6-8, 67, 73)
Shannon effective species number	Converts Shannon index allele frequency results obtained from Shannon Diversity index results into corresponding allele richness values.	No	Not Applicable	(6-9, 67, 74)
Simpson index	Measures allele frequency and sample evenness. Biases are created towards the species with the highest frequency. However, diversity can be measured from samples that are either infinite or finite in size. Low and high diversity tend towards one and zero respectively.	Yes	Yes	(2, 6-8, 67, 75)
Simpson Diversity index	Measures allele frequency. Also known as the Gini-Simpson Diversity index. Biases are created towards species with the highest frequency. Two equations exist; one to calculate diversity in samples of infinite size, and the other in samples of finite sample size. Low and high diversity respectively tend towards zero and one.	Yes	Yes	(3-8, 67, 76)
Simpson reciprocal index	Measures allele/species richness, considers evenness and is a true measure of diversity. Biases may however result in sampled populations with uneven species numbers or sample sizes.	No	Not Applicable	(6-8, 67, 70, 77)
Simpson Effective species number	Converts Simpson index allele frequency results into corresponding allele richness values.	No	Not Applicable	(6-8, 11, 67, 75)
Simpson Diversity Effective species number	Converts Simpson Diversity index allele frequency results into corresponding allele richness values.	No	Not Applicable	(6-9)

2.4 Final remarks: The past paves the way for the future

Reflecting back on the history of the Zulu nation, it can be seen that although the Zulu nation expanded immensely under Shaka Zulu's reign, it also resulted in several population bottlenecks. This occurred not only through Mzilikatzi fleeing from Shaka's kingdom, but also through others who either fled fearing death, or were killed at the hands of Shaka's armies. Through population bottlenecks and founder effects, those who escaped during Shaka's period of influence would undoubtedly have generated novel founder mutations if they remained an isolated population. If such founder populations had integrated with other peoples in search of safety, then genetic diversity would most certainly have arisen both within, and between different populations, and such genetic diversity can be determined and compared within sampled countries, regions, and populations.

However, in order to illustrate the potential that diversity theory can have in the medically-orientated fields, first consider a common monogenic disorder such as CF. Secondly, bear in mind that several different variants may exist at a given locus for any one monogenic disorder (31, 32). Although it is known that monogenic disorders are relatively rare if considered individually, it is estimated that they are collectively responsible for the manifestation of a genetic disorder in at least one in every 200 live births world-wide (78). Of the estimated 6,000 recognized monogenic disorders, CF, an autosomal recessive disorder, has been most widely recognised and studied globally – including South Africa. With nearly 2,000 mutations having been identified in CF patients (79, 80), it is not difficult to see the extent to which diversity theory can be applied.

Metachromatic leukodystrophy (MLD), in stark contrast to CF, is considered to be a rare, pan-ethnic, autosomal recessive monogenic disorder. Despite being described both biochemically and molecularly in 1980 and 1990, respectively (81, 82), less than 200 variants are known to be associated with the onset of this particular disorder (23). However, even though only one case of MLD has been reported within the South African population (21), this disorder has been well studied world-wide. Nevertheless, if one is to investigate the applicability of diversity theory in a common monogenic disorder, such as CF, it stands to reason that the same theory can be applied to a comparatively rare monogenic disorder, such as MLD. This is an important consideration given that many population-specific disorders are present within South Africa (52, 53, 83).

Interestingly, the Department of Health (84) has stated that in developing countries such as South Africa, up to 8% of children aged five are likely to present with either a severe birth defect or a genetic disorder. According to Statistics South Africa (85), a total of 6,039,160 children, between the ages of birth and five, resided in South Africa in 2013. If the 8% is modestly applied to this number, it can be estimated that in 2013 alone, approximately 483,000 children would have manifested with either a severe birth defect or genetic disorder. Additionally, it is expected that approximately 60-65% of a third world country's population is likely to present with a genetic disorder of late-onset, such as hypertension, diabetes, certain cancers and psychoses, in their lifetime (84). According to the mid-year population estimates of 2013 (86), the South African population consisted of approximately 32 million individuals that were at least 20 years or older. We can thus estimate that approximately 19,415,000 - 21,032,000 South African adults may have been affected by a genetically-linked

disorder in 2013. These estimates show the staggering number of people that would need health and welfare support in South Africa for genetically related disorders or birth defects.

This does not however reflect on the health and welfare costs experienced by patients and their family members as well as the costs experienced by the state. Statistics South Africa (87) has estimated that the average monthly income for White South Africans in 2011 was R30,400, while Indian and Mixed race South Africans, in the same year, received an average monthly income of R20,900 and R20,960, respectively. The Black South African population, according to the same report, showed an average monthly income of approximately R5,000 (87). Given that the required medication for South African CF patients can cost as much as R30,000 a month, the average South African cannot afford to pay for the required CF medication (88). This estimate does not include hospital costs, molecular/biochemical testing fees, physiotherapeutic costs and the costs associated with healthy living lifestyles required by CF patients. This also only reflects the costs associated with one example of approximately 6,000 known monogenic disorders. Furthermore, costs can be considerably higher in cases where patients harbour variants associated with exceptionally rare disorders, such as MLD - multiple hospital visits and biochemical tests are often required before a diagnosis can be made (25, 89).

Therefore, if the UN CRC requirements concerning health and welfare in South Africa are to be met, a significant effort must be made by several different sectors in order to achieve this. This study merely presents one way through which a contribution to this plight might be made

– if we know the relative contribution of disease variants in terms of diversity between different regions or populations within South Africa, then steps can be taken to tailor health systems accordingly.

Chapter 3 – Investigating the presence of cystic fibrosis in patients attending Steve Biko Academic Hospital

3.1 Introduction

3.1.1 Biochemistry of Cystic Fibrosis

Described in 1989, the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene is found at chromosomal position 7q31.3 (90-92). Associated with the autosomal recessive disorder known as cystic fibrosis (CF), the *CFTR* gene has been well studied, is approximately 250kb in length, and encompasses 24 exonic regions (90-93). The CFTR protein, synthesized in the endoplasmic reticulum (ER) of epithelial cells, is transported from the ER into the Golgi apparatus (GA) for further processing. Once processed, the CFTR protein is transported into the cell's cytoplasm where it initiates a signalling cascade. The presence of the CFTR protein in the cell's cytoplasm is important for two reasons: 1) it activates the purinergic receptor P_2Y_2, and 2) it activates the $Na^+/K^+/2Cl^-$ co-transporter system. As shown in Figure 4, the purinergic receptor is ultimately responsible for the transport of chloride and sodium ions into and out of the epithelium, while the $Na^+/K^+/2Cl^-$ co-transporter system ensures that sodium and potassium ions are pumped back into the cell while facilitating the expulsion of chloride ions (91, 94, 95).

Once activated, P_2Y_2 signals the initiation of the cyclic AMP (cAMP) mechanism. The cAMP mechanism in turn activates protein kinase A (PKA) synthesis as well as the outwardly rectifying chloride channel (ORCC). The ORCC mediates the direct transportation of chloride

ions to the apical surface of airway cells, while PKA further activates the CFTR ion channel. The CFTR ion channel consists of a highly charged receptor (R) domain, two nucleotide binding folds/ domains (NBFs/NBDs), and two membrane-associated regions. Of the two membrane-associated regions of the CFTR channel, one acts as a cationic node, while the other acts as an anionic node (Figure 4) (91, 94, 95).

The distinction between the two membrane-associated regions of the CFTR channel is an important one to make as the cationic and anionic nodes, respectively, facilitate the movement of negatively charged chloride and positively charged sodium ions out of the apical surface of epithelial cells. The movement of chloride ions through the CFTR channel is referred to as electrolyte secretion and results in chloride ions being deposited on the surface of airways. Sodium ions that are secreted via the CFTR channel are reabsorbed into the cytosolic region of the epithelial cell. The reabsorption of sodium ions is facilitated through the epithelial sodium channel (ENaC), the action of which is down regulated by the CFTR channel (91, 94, 95).

Figure 4: Summary of the CFTR system

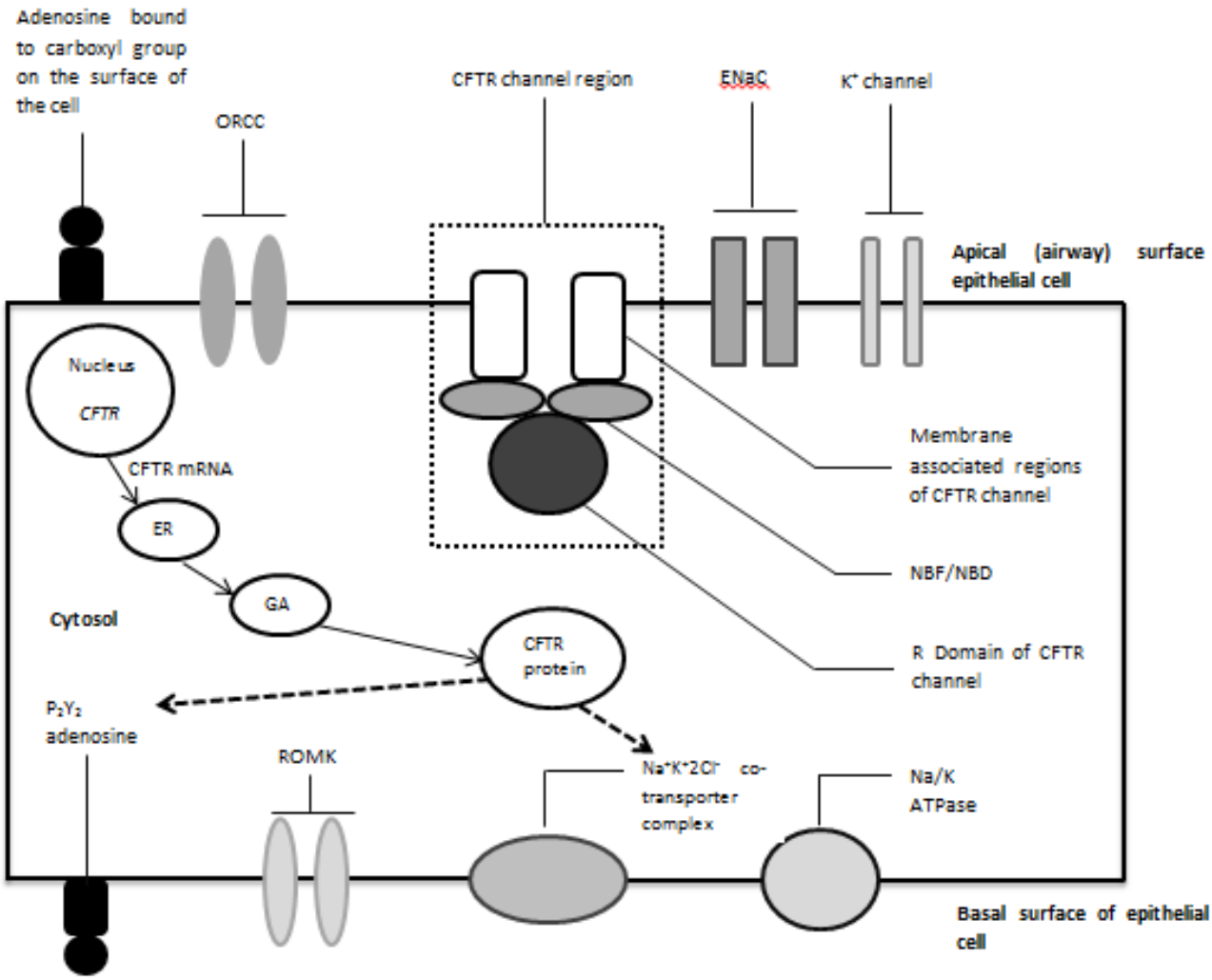


Figure 4 shows the functional units of CFTR protein synthesis and regulation in epithelial cells as well as those responsible for chloride and sodium ion transportation. CFTR mRNA is synthesised in the nucleus and then transported to the endoplasmic reticulum (ER) where the protein is produced. The protein is subsequently transported to the Golgi apparatus (GA) for further processing, after which it is released into the cytosol of the cell. Once released in the cytosol, the CFTR protein activates 1) the adenosine-bound purinergic receptor (P_2Y_2) region, 2) the $Na^+K^+2Cl^-$ co-transporter complex, and 3) the epithelial sodium channel (ENaC). Activation of the P_2Y_2 region results in the initiation of cyclic AMP (cAMP) release. This stimulation of cAMP results in the activation of 1) the outwardly rectifying chloride channel (ORCC), and 2) the release of protein kinase A (PKA). Activation of the ORCC allows chloride ions to be transported to the airway on the apical surface of the epithelial cell. PKA stimulates the activation of the CFTR channel (comprised of one highly charged receptor (R) domain, two nucleotide ATP-binding folds/domains (NFB/NFD) and two membrane associated regions.) One membrane associated region is positively charged; allowing the transportation of chloride ions, while the other has a negative charge and facilitates the movement of sodium ions. Sodium ions transported out of the apical surface of the epithelial cell are re-absorbed into the cell's cytosol via the ENaC. Sodium ions (Na^+) present in the $Na^+K^+2Cl^-$ co-transporter complex are basally pumped out of the cell via the Na/K -ATPase, while potassium ions (K^+) are reabsorbed into the epithelial cell. The reabsorption of potassium into the epithelial cell is facilitated through the rectifying renal outer medullary potassium channel (ROMK). Chloride ions present in the $Na^+K^+2Cl^-$ co-transporter complex are cycled out of the epithelial cell via the CFTR channel. Potassium released into the cell's cytosolic region via the Na/K – ATPase is pumped out of the apical surface of the cell via the K^+ channel. Figure 4 has been adapted from several sources (91, 94, 95).

Finally, recall that the $Na^+/K^+/2Cl^-$ co-transporter system is activated by the CFTR protein and that its primary function is to facilitate the 1) reabsorption of sodium and potassium ions, and 2) secretion of chloride ions. Sodium ions are pumped back into the cytosolic region of the epithelial cell through the Na^+/K^+ ATPase system, while potassium ions are reabsorbed into the cytosolic region with aid of the renal rectifying outer medullary potassium channel (ROMK). However, as shown in Figure 4, the Na^+/K^+ ATPase system releases potassium ions into the cytosolic region during its expulsion of sodium ions. The potassium "waste products" are relieved from the cell through the potassium (K^+) channel and secreted to the airway surface of epithelial cells. Chloride ions originating from the $Na^+/K^+/2Cl^-$ co-transporter system are directly secreted through the CFTR channel (91, 94, 95). Thus, failures occurring during any one of these processes ultimately results in decreased electrolyte secretory ability and the consequent decrease in sodium ion concentrations in the cytosol of epithelial cells (95).

The nearly 2,000 mutations occurring in the *CFTR* gene can be divided into five categories/classes (79, 80). Mutations that affect protein production in the ER as a result of defective mRNA fragments are referred to as class I mutations. Class I mutations are often the result of nonsense, insertion/deletion (indel), frameshift or splice site mutations. Class II mutations occur as a result of faulty protein processing between the ER and GA and result in the incorrect or incomplete folding of the CFTR protein. Class III mutations are the result of incorrect protein regulation occurring on the CFTR channel's R-domain ATP/phosphorylation site. Class III mutations thus affect the activation of the CFTR channel complex. Mutations which decrease the capacity of the CFTR channel to secrete chloride and reabsorb sodium ions are referred to as class IV mutations, while class V mutations are responsible for

decreasing the regulatory capacity of the ORCC and ENaC to secrete chloride ions and absorb sodium ions, respectively (91, 94, 96-98).

3.1.2 Clinical presentation

The phenotypic expression of CF is often varied among patients, but pancreatic insufficiency occurs in an estimated 85-90% of all patients, which results in poor weight gain, malnutrition/decreased nutrient absorption, and anaemia. Rectal prolapse has also been noted to occur in CF patients. CF patients frequently present with respiratory *pseudomonas* infections, ultimately resulting in obstructed airways and often causing recurrent or persistent complications in the respiratory systems as well as bronchiectasis if untreated (18, 20, 94, 99). Nasal polyposis and chronic sinupulmonary disease may also present in CF patients. Intestinal obstructions/meconium ileus has also been found in CF patients, while male CF patients may additionally present with infertility due to congenital bilateral absence of the vas deferens. However, it has been shown that CF patients will show a significantly reduced concentration in sweat chloride levels – an important diagnostic characteristic in all CF patients (17, 18, 20, 80, 94, 99, 100). Finally, liver damage has also been reported as a feature of CF patients (20), which is diagnosed through an ultrasound scoring system (known as the Westaby score) developed by Westaby *et al.* in 1995 (101). Westaby scores of 3 indicate normal liver function with no liver damage, while scores ranging between 4 and 7 indicate that liver damage occurred to some extent with potential cirrhosis. Westaby scores of either 8 or 9 indicate advanced liver damage with confirmed cirrhosis.

3.1.3 Molecular diagnostic aspects

Prenatal diagnosis for CF can be performed through either amniocentesis or chorionic villus sampling (CVS), where genetic testing of the foetus will determine CF status. However, two additional screening options exist for parents if performing an amniocentesis or CVS is not desired - 1) genetic carrier screening, and 2) new-born genetic screening. However, genetic carrier screening requires that both parents be genetically tested and provides only the probability that parents will give birth to a CF positive child (20). Regardless, when a familial history of CF is known to exist, carrier screening is advised (18, 100, 102-109).

Additionally, although genetic screening has the potential to identify causative *CFTR* variants, Padoa *et al*. (19) outlined some of the difficulties associated with detecting many of these variants in the South African population. Nevertheless, their research revealed that the c.3120+1G>A variant is the most commonly observed CF variant in the Black South African CF population, occurring in nearly 50% of all such patients, while the ΔF508 variant occurs in nearly 80% of all White South African CF patients and is consequently also the most frequently observed variant in this population group (17, 19, 110). Despite this, Goldman *et al.* (17, 110), in their genetic investigation of 14 Black, 43 Mixed race, and 201 White South African CF patients, have found that up to 79%, 45% and 17% of all variants causative of CF remain unidentified in each of these population groups, respectively.

The sweat test is still considered to be the “gold standard” on which a CF diagnosis can be made. Two different types of sweat tests exist: 1) the sweat electrolyte test, and 2) the sweat

conductivity test. Sweat electrolyte tests are preferentially performed over sweat conductivity tests and determine the concentration of chloride ions (electrolytes) present in the sweat of CF patients (102, 103, 107, 108, 111). Furthermore, Shwachman and Mahmoodian (102) suggested that sweat chloride levels exceeding 60 mmol per litre is abnormally high and potentially indicative of CF. This limit is considered to form part of standard diagnostic test for CF with tests usually being performed in duplicate, and sometimes triplicate, in order to ensure the accuracy of the results. Sweat electrolyte tests revealing sweat chloride concentrations between the "grey zone" of 30 and 60 mmol/L indicate the potential presence of CF, while sweat chloride concentrations below 30 mmol/L are considered to be normal (20). Sweat conductivity tests may also be performed in order to diagnose patients with CF but have been reported to result in high numbers of false-negative results. Sweat conductivity tests are thus considered to be more of a screening test used to suggest the presence of CF (112). In South Africa, a sweat conductivity value of 90 mmol/L is considered to be indicative of CF in circumstances where performing a sweat electrolyte test is not an available option (20).

Additionally, given that an estimated 85-90% of all CF patients will suffer from pancreatic insufficiency, pancreatic faecal elastase-1 enzyme (PFE-1) testing may also be conducted in order to determine CF status (18, 20, 94, 100, 104, 105). Determined through enzyme-linked immunosorbent assays (ELISAs), PFE-1 enzyme concentrations less than 100µg PFE-1 per gram of faecal matter is indicative of severe pancreatic insufficiency, while PFE-1 concentrations between 100-200 µg/g is indicative of mild pancreatic insufficiency. However,

due to the fact that decreased PFE-1 concentrations are not uniquely observed in patients with CF, it is advised that PFE-1 tests only be suggestive, not indicative, of CF (100).

3.1.4 Diagnostic capacity in South Africa

Despite the usefulness of sweat electrolyte, sweat conductivity and PFE-1 tests, genetic testing is currently the best CF diagnostic method available in the South Africa. Table 2 shows the number of CF and genetic clinics available in South Africa in 2012 (20). According to Kromberg et al. (89), there were only five genetic service facilities in South Africa in 2008, while Jenkins (113) states that 14 genetic counselling clinics were available in South Africa in 1990. Nevertheless, it is known that the National Health Laboratory Services (NHLS) in Johannesburg is the largest referral centre in South Africa having performed nearly 12,400 biochemical, molecular, cytogenetic and genetic ancestry tests in 2008 alone (89). These tests were not limited to CF determination, but consisted of approximately 9,100 biochemical and molecular, 2,900 cytogenetic, and 350 genetic ancestry tests.

Table 2: Summary of the number of CF and genetics clinics available in South Africa in 2012 (20)

Region in South Africa	Number of CF clinics	Number of Genetics clinics
Johannesburg	5	4
Pretoria	1	1
Bloemfontein	2	1
Durban	2	1
Port Elizabeth	2	0
Cape Town/Stellenbosch	3	2
Total	**15**	**9**

Henley and Hill (114), Duff (115), Mutesa and Bours (95), and Masekela et al. (99) have indicated that diagnosing CF in African and South African Black populations can be complicated due to the presence of HIV/AIDS, tuberculosis (TB), protein energy malnutrition (PEM), and chronic pulmonary infections. Additionally, it is generally regarded that CF is a rare monogenic disorder in Black/African populations. Not only does this influence a medical practitioner's diagnostic considerations, but it also has a detrimental effect on the patient's health (19, 95, 99).

From a global perspective, the general clinical presentation, molecular and diagnostic features of CF have been well-described in most populations. Unfortunately, this does not hold true for the CF population(s) in South Africa. In order to address this, clinical presentation and molecular findings of CF patients attending the Steve Biko Academic Hospital (SBAH) CF clinic were thoroughly investigated. Medical records of these patients were accessed, the information databased and analysed, and finally compared to published reports.

3.2 Methods

3.2.1 Study population and ethical considerations

Ethical approval was obtained from the Research Ethics Committee of the Faculty of Health Sciences at the University of Pretoria under application numbers 4-2013 and 40-2014 (Appendix A – C). Informed consent was collected from 45 CF patients attending SBAH CF

clinic and patient confidentiality ensured through the assignment of randomly generated alphanumeric codes. Demographic data extracted from each patient's file included the patient's age, ethnicity, gender and current geographical location.

3.2.2 Clinical, laboratory and molecular presentation

Each patient's most recent weight and height measurements were also documented. This was collected in order to determine the BMI. In accordance with the standards established by the Centers for Disease Control and Prevention (CDC) (116), BMI values were calculated using Equation 19, as determined by the age of each respective patient at the point when the last weight and height measurement was made. Weight status was determined by comparing each patient's age, height and BMI values to growth charts developed by the CDC (116).

Equation 19: BMI equation for adults who are 21 or older (117)

$$BMI = \left(\frac{weight\ (kg)}{height\ (m)^2}\right)$$

Where determined and described in the medical records, the status of liver function as measured by the Westaby score, pancreatic insufficiency, whether meconium ileus/distal intestinal obstruction syndrome had ever been present (MIE/DIOS), whether the patient had ever been diagnosed with gastro-oesophageal reflux disease (GORD), and whether a family history of CF existed was additionally recorded. The presence of any other clinical feature, such as CF-associated diabetes, osteoporosis and infertility, were also recorded. Furthermore,

where present within patient files, results obtained from biochemical and molecular tests, which included sweat electrolyte, PFE-1 test results, lung microbial flora, and mutation screening results, were also documented.

3.2.3 Data collection

In order to compare mutation screening results of SBAH CF patients to known frequencies of CF-causing variants in the South African CF population, thorough internet searches were performed via Google Scholar and PubMed. Key words employed in these searches included any combination of “cystic fibrosis”, “prevalence”, “incidence”, “mutation(s)”, “mutation frequency”, “Black”, “Coloured”, “Mixed ancestry”, “White”, and “South Africa”, “Republic of South Africa”. To ensure that the largest possible combination of journal publications were identified, “mutation(s)” and “mutation frequency” were also substituted with the terms “variant(s)” and “variant frequency”.

All the collected data were entered into one of two customised databases – data from patient files was entered into a SBAH CF patient database, while literature-extracted variant data was entered into a separate database constructed purely for reference purposes. In constructing the SBAH CF database, only the most recent weight, height and age values were recorded. However, in recording the presence or absence of clinical features, a binary matrix was formulated in which the presence or absence of a particular feature was respectively indicated by either a one or a zero. A dash was used to indicate instances where a particular

clinical feature, such as infertility, could not be determined due to a patient's age or death or was not indicated within the patient's medical record.

The literary database was constructed from published data pertaining to *CFTR* variants present within South African CF populations. The total number of chromosomes analysed per study in each ethnic group sampled, as well as the methods through which mutation analysis was performed was recorded. Each CF variant identified in each of the studies, as well as the number of times each of the respective variants were found in each ethnic group was also recorded. Both databases were constructed using Microsoft Excel® software with mean and standard deviation (SD) values being calculated where possible and appropriate.

3.2.4 Statistical analysis

In determining the statistical significance of clinical outputs relating to the age of diagnosis, weight, height, BMI, and sweat electrolyte concentrations of CF patients attending SBAH, p-values were determined through the use of the Mann-Whitney U test. The Fisher's exact test statistic was employed in determining statistical significance of differences observed when measuring the presence or absence of liver damage, pancreatic insufficiency, and lung microflora. Finally, Chi-Square analysis was performed when 1) comparing whether or not statistically significant differences exist between genetic screening results obtained from CF patients attending SBAH and the expected screening results as determined by research findings presented by Goldman *et al.* (17); 2) determining whether significant differences exist when comparing the genetic screening results obtained from CF patients attending SBAH to

the genetic screening results of published reports of CF cases within the RSA; and 3) determining the statistical significance of results obtained from comparing genetic screening results of various ethnic groups from collated published reports of CF cases in the RSA. All statistical analysis was performed using the R software package. All α-values were set to a significance level of 0.05.

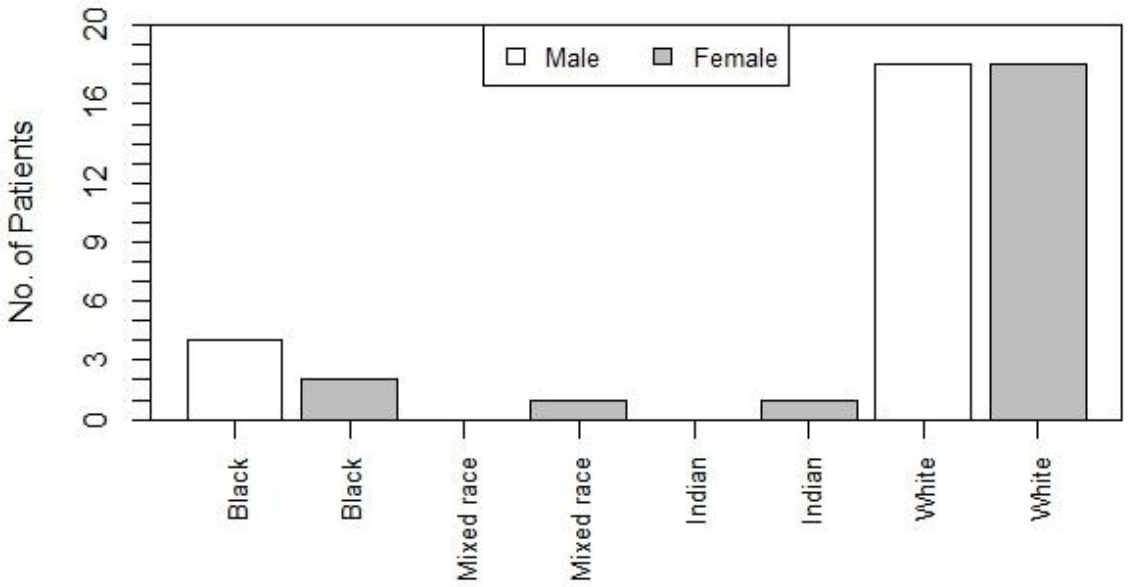
3.3 Results

3.3.1 Demographics

Although data were collected from 45 patients, it was found that one patient was not a true CF case. This patient's parents had been genetically screened with results indicating the presence of one known CF variant in each of the parents. The expected genotype of the patient could thus be determined. However, genetic screening could only identify one known variant in this patient thus rendering the patient a carrier of CF rather than a true CF patient. Data collected from this individual were therefore not included in any analysis. The average age at which the 44 CF patients attending SBAH were positively diagnosed was 12.85 years (SD±10.87). Notably, a significant proportion of the CF positive diagnoses (25%, $p<0.001$) were made in adult patients (age>18 years). The age at diagnosis in Black South African and White South African patients was 4.83 years (SD±2.27) and 14.62 years (SD±11.28), respectively ($p=0.022$). Black South African CF patients attending SBAH tended to be diagnosed as children/infants, while White South African CF patients were diagnosed at variable ages ranging between infancy and adulthood.

Of the 44 study participants, 22 were male and 22 female. The males consisted of four Black and 18 White South African CF patients, while the females consisted of two Black South African, one Mixed race, one Indian, and 18 White South African CF patients (Figure 5). Of the 44 CF patients attending the CF clinic at SBAH, 33 resided in Gauteng, nine lived in the North West province, one resided in Mpumalanga, and one resided in the Limpopo provinces of South Africa. Of the 44 CF patients attending SBAH, 12 adult cases of CF were diagnosed in four male and eight female patients.

Figure 5: Graphical representation of South African CF patients attending SBAH according to ethnicity and gender



3.3.2 Clinical presentation

As shown in Table 3, the average age, weight, height and BMI of the six Black South Africa CF patients attending the SBAH CF clinic was 5.8 years (SD±2.8), 16.5 kg (SD±2.8), 103.7 cm (SD±13.2), and 15.5 kg/m^2 (SD±2.03), respectively. Of the 36 White South African CF patients,

the average age, weight, height and BMI was 13.85 years (SD±11.03), 34.31 kg (SD±20.21), 133.64 cm (SD±33.93), and 17.32 kg/m^2 (SD±3.31), respectively. When compared to their Black South African counterparts, it was found that age (p=0.022), weight (p=0.046) and height (p=0.046) differed significantly. Although not indicated in Table 3, according to the standards outlined by the CDC (116), of the six Black South African CF patients attending the CF clinic at SBAH, one patient was clinically underweight, four were healthy, and one was overweight. The Mixed race South African CF patient had a BMI of 15.09 kg/m^2 and was considered healthy, while the Indian South African CF patient had a BMI of 12.48 kg/m^2 and was considered to be clinically underweight based on CDC guidelines (116). Of the 36 White South African CF patients, eight were clinically underweight, 22 were healthy, and four were clinically overweight. BMI measurements could not be determined for two White South African CF patients due to missing data.

Table 3: Average weight, height and BMI values in Black and White South African CF patients attending the SBAH CF clinic

Ethnicity	No. of Patients	Gender	Age (years)	Weight (kg)	Height (cm)	BMI
Black South African	4	Male	6.25 (SD±2.6)	17.5 (SD±2.59)	106 (SD±11.51)	15.61 (SD±1.08)
Black South African	2	Female	5.00 (SD±3.00)	14.5 (SD±1.4)	99 (SD±15)	15.4 (SD±3.17)
Black South African	6	**Total**	5.83 (SD±2.8)	16.5 (SD±2.8)	103.7 (SD±13.2)	15.5 (SD±2.03)
White South African	18	Male	10.88 (SD±9.96)	31.19 (SD±22.26)	126.21 (SD±34.14)	17.10 (SD±3.62)
White South African	18	Female	17.18 (SD±11.21)	37.8 (SD±16.96)	141.94 (SD±31.69)	17.56 (SD±2.91)
White South African	36	**Total**	13.85 (SD±11.03)	34.31 (SD±20.21)	133.64 (SD±35.93)	17.32 (SD±3.31)
p-value*			0.022	0.046	0.046	0.136

* p-value determined between Black and White South African CF patients attending SBAH using the Mann-Whitney U test

As shown in Table 4, it was found that 26 of the 44 CF patients (59.1%) attending SBAH had liver damage. Of these 26 patients, four were Black South African (two male and two female) and 20 were White South African (nine male and 11 female). Both the Mixed race and Indian patients were diagnosed with liver damage. Thus a total of four of the six (66.67%) Black South African CF and 20 of the 36 (55.56%) of White South African CF patients attending the CF clinic at SBAH presented with liver damage (p=0.571). The average Westaby score in all 26 patients was found to be 4. Additionally, five of the six (83.33%) Black South African CF patients and 31 of the 36 (86.11%) White South African CF patients attending SBAH were diagnosed with pancreatic insufficiency (p=0.48). Although the Mixed race patient was shown to be pancreatic insufficient, no mention of pancreatic insufficiency was found within the medical records of the Indian patient. Pancreatic function was therefore unknown in the Indian patient, but was identified in 84% of all remaining CF patients attending SBAH.

Table 4: Liver damage and pancreatic insufficiency associated with South African CF patients attending SBAH

Clinical finding	Black		Mixed race	Indian	White		Total	
	Male	Female	Female	Female	Male	Female	Male	Female
No. of patients with liver damage	2	2	1	1	9	11	11	15
Average Westaby score	4	4	-	-	4	4	≈ 4	≈ 4
No. of patients with pancreatic insufficiency	3	2	1	-	14	18	17	20

Of the 44 CF patients attending SBAH, and as shown in Table 5, 11 cases (24.44%) of intestinal obstruction (MIE and/or DIOS) was found, while GORD was observed 16 times (31.11%), and

CF-associated diabetes seen on seven occasions (15.56%). Osteoporosis was reported in two instances (4.44%), while infertility was reported eight times (17.78%). Further illustrated in Table 5, six CF patients (four male and two female) presented with only MIE/DIOS, while five patients (three male and two female) presented with only GORD. Three patients, all of whom were White South African females, presented with only CF-associated diabetes. One White South African male and two White South African female CF patients presented with only infertility.

Also shown in Table 5, three patients, one Indian female and two White South African females, presented with both MIE/DIOS and GORD, while one White South African female patient presented with MIE/DIOS, GORD and CF-associated diabetes. A Black South African male patient presented with MIE/DIOS, GORD and infertility, while two White South African female patients were diagnosed with MIE/DIOS, GORD and osteoporosis. Two White South African male patients experienced GORD and were infertile. It was established that a White South African female patient was both infertile and had CF-associated diabetes. Although not indicated in Table 5, two White South African patients (one male and one female) presented with rectal prolapse. Interestingly, it was also found that a familial history of CF was found in a total of 12 White South African patients (seven male and five female) from seven different unrelated families. Overall, MIE and/or DIOS was observed in only 24.4% of all CF patients attending SBAH CF clinic. GORD was observed in 31.1% of the CF population of SBAH.

Table 5: Illustration of South African patients attending the SBAH CF clinic who were diagnosed with any combination of MIE/DIOS, GORD, osteoporosis, infertility and/or CF-associated diabetes

Positive for:	Black		Mixed race	Indian	White		Total	
	Male	*Female*	*Female*	*Female*	*Male*	*Female*	*Male*	*Female*
Only MIE or DIOS	1	-	1	-	3	1	4	2
Only GORD	-	1	-	-	4	1	4	2
Only CF-associated Diabetes	-	-	-	-	2	1	2	1
Only Infertility	-	-	-	-	1	1	1	1
MIE/DIOS, GORD and infertility	1	-	-	-	-	-	1	-
MIE/DIOS, GORD and Osteoporosis	-	-	-	-	-	1	-	1
Both CF-associated Diabetes and Osteoporosis	-	-	-	-	-	1	-	1
Both MID/DIOS and GORD	-	-	-	1	-	1	-	2
Both GORD and CF-associated Diabetes	-	-	-	-	-	1	-	1
Both GORD and Infertility	-	-	-	-	2	2	2	2
CF-associated Diabetes and Infertility	-	-	-	-	-	1	-	1
Diagnosed with MID/DIOS, GORD and CF-associated Diabetes	-	-	-	-	-	1	-	1

3.3.3 Biochemical and Molecular test results

Sweat electrolyte concentrations had been determined in 21 of the 44 CF patients attending SBAH. The 21 patients included the one Mixed race and one Indian CF patient, but also included three Black South African and 16 White South African CF patients. As shown in Table 6, the average sweat chloride concentrations in Black South African CF patients attending SBAH was 73 mmol/L (SD±29.7), while it was found to be 101.3 mmol/L (SD±30.41) in White South African male and 96.83 mmol/L (SD±14.46) in White South African female SBAH CF patients. The average sweat electrolyte concentration in White South African CF patients attending SBAH was 99.5 mmol/L (SD±23.84). However, despite the fact that less

than 50% of patients attending the CF clinic at SBAH underwent sweat electrolyte testing and though not shown, only two patients (one Black South African and one White South African) had sweat chloride concentrations that were in the "grey" region of 30–60 mmol/L (20, 102). No significant difference was found between sweat electrolyte concentrations obtained from Black and White South African CF patients attending SBAH (p=0.751).

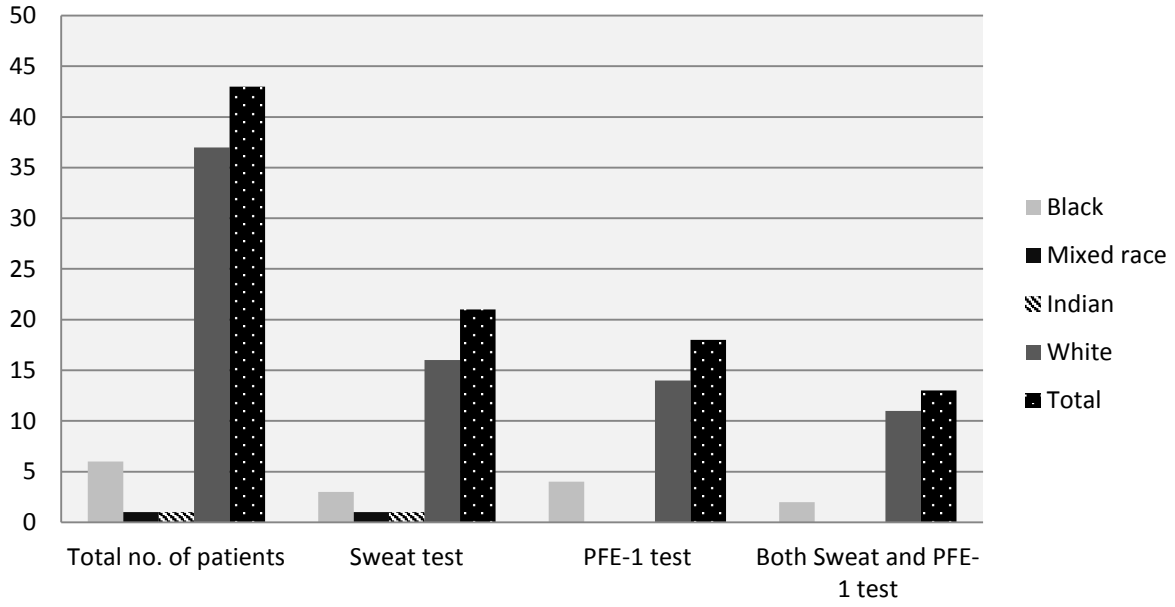
Table 6: Average sweat electrolyte of select South African CF patients attending SBAH

	Gender	Statistic	Sweat Electrolyte Concentration (mmol/L)	*P-value
Black	Male	Average	73 (SD±29.7)	-
White	Male	Average	101.3 (SD±30.41)	-
	Female	Average	96.83 (SD±14.46)	-
White	Male & Female	Average	99.46 (SD±23.84)	0.464
Black & White	Male & Female	Average	98.49 (SD±25.54)	0.751

* p-value determined between Black and White South African CF patients attending SBAH using the Mann-Whitney U test

A total of 17 CF patients attending SBAH underwent PFE-1 testing at least once. These included four Black South African male, 10 White South African male, and three White South African female patients. As both PFE-1 and sweat electrolyte testing are considered important clinical determinants of CF, it was found that only 13 of the 45 CF patients (two Black South African and 11 White South African) underwent testing for both (Figure 6). Although not shown in Figure 6, a second PFE-1 test was performed on two Black South African and four White South African CF patients, while three Black South African and 16 White South African patients' sweat electrolyte concentrations were determined on a second occasion.

Figure 6: Graphical representation of the number of South African patients attending the CF clinic at SBAH in relation to the number of patients who underwent sweat electrolyte and PFE-1 tests



Forty-two different bacterial species were found to have colonised the lungs of the 45 CF patients attending SBAH (Table 7). Of these 42 bacterial species, *Pseudomonas aeruginosa*, *Haemophilus parainfluenzae*, and *Staphylococcus aureus*, were respectively observed in 59%, 56%, and 44% of all CF patients attending SBAH. *P. aeruginosa* was identified in three of the six (50%) Black South African SBAH CF patients, while it was identified in 21 of the 36 (60%) White South African SBAH CF patients (p=0.679). Similarly, *H. parainfluenzae* was identified in four of the six (67%) Black South African SBAH CF patients, while it was identified in 19 of the 36 (54%) White South African SBAH CF patients (p=0.679). Finally, *S. aureus* was found in two of the six (33%) Black South African SBAH CF patients, while 16 of 36 White South African had been colonised with this micro-organism (p=0.679).

Table 7: Summary of the number of infections with various bacterial species identified in the lungs of South African CF patients attending SBAH CF clinic

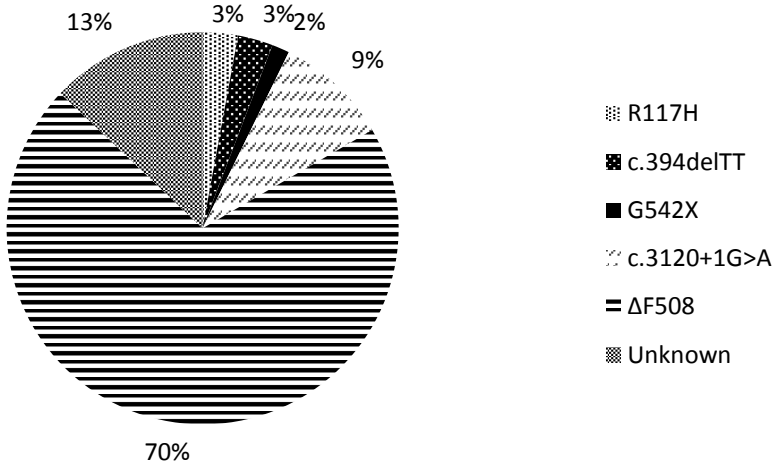
Bacterial species	Total observations/ethnic group				Bacterial species	Total observations/ethnic group			
	Black (N=6)	White (N=35)	Total (N=41)	% Values		Black (N=6)	White (N=35)	Total (N=41)	% Values
α-haemolytic *streptococcus*	0 (0.00%)	1 (2.86%)	1	2.44%	*Klebisella ozaenae*	0 (0.00%)	2 (5.71%)	2	4.88%
Acinetobacter baumannii	0 (0.00%)	1 (2.86%)	1	2.44%	*Klebsiella pneumoniae*	1 (16.67%)	4 (11.43%)	5	12.20%
Acinetobacter lwoffii	0 (0.00%)	1 (2.86%)	1	2.44%	*Moraxella spp.*	0 (0.00%)	5 (14.29%)	5	12.20%
Achromobacter spp.	0 (0.00%)	1 (2.86%)	1	2.44%	*Morganella morganii*	0 (0.00%)	1 (2.86%)	1	2.44%
ß-haemolytic Streptococcus group A	0 (0.00%)	3 (8.57%)	3	7.32%	*Mycobacterium tuberculosis* complex	0 (0.00%)	1 (2.86%)	1	2.44%
ß-haemolytic Streptococcus group F	0 (0.00%)	1 (2.86%)	1	2.44%	*Pseudallescheria boydii*	0 (0.00%)	1 (2.86%)	1	2.44%
Bordetella spp.	0 (0.00%)	1 (2.86%)	1	2.44%	*Pseudomonas spp.*	0 (0.00%)	8 (22.86%)	8	19.51%
Burkholderia cepacia	0 (0.00%)	1 (2.86%)	1	2.44%	*Pseudomonas aeruginosa**	3 (50.00%)	21 (60.00%)	24	58.54%
Candida spp.	0 (0.00%)	1 (2.86%)	1	2.44%	*Pseudomonas fluorescens*	0 (0.00%)	2 (5.71%)	2	4.88%
Candida albicans	2 (33.33%)	11 (31.43%)	13	31.71%	*Pseudomonas luteola*	0 (0.00%)	1 (2.86%)	1	2.44%
Candida dubliniensis	0 (0.00%)	2 (5.71%)	2	4.88%	*Rhodotorula spp.*	0 (0.00%)	1 (2.86%)	1	2.44%
Candida tropicalis	0 (0.00%)	1 (2.86%)	1	2.44%	*Serratia marcescens*	1 (16.67%)	0 (0.00%)	1	2.44%
Citrobacter freundii- Multi-resistant	1 16.67%	0 (0.00%)	1	2.44%	*Serratia marcescens* - Multi-resistant	1 (16.67%)	0 (0.00%)	1	2.44%
Enterobacter spp.	0 (0.00%)	1 (2.86%)	1	2.44%	*Staphylococcus* - coagulase negative	0 (0.00%)	7 (20.00%)	7	17.07%
Enterobacter cloacae	0 (0.00%)	2 (5.71%)	2	4.88%	*Staphylococcus aureus***	2 (33.33%)	16 (45.71%)	18	43.90%
Enterococcus faecalis	0 (0.00%)	1 (2.86%)	1	2.44%	MRSA	1 (16.67%)	2 (5.71%)	3	7.32%
Escherichia coli	1 (16.67%)	3 (8.57%)	4	9.76%	*Staphylococcus epidermidis*	0 (0.00%)	2 (5.71%)	2	4.88%
Haemophilus spp.	0 (0.00%)	1 (2.86%)	1	2.44%	*Stenotrophomonas maltopilia*	0 (0.00%)	3 (8.57%)	3	7.32%
Haemophilus influenzae	1 (16.67%)	11 (31.43%)	12	29.27%	*Streptococcus* spp.	0 (0.00%)	1 (2.86%)	1	2.44%
*Haemophilus parainfluenzae***	4 (66.67%)	19 (54.29%)	23	56.10%	*Streptococcus pneumoniae*	1 (16.67%)	8 (22.86%)	9	21.95%
Klebsiella oxytoca	0 (0.00%)	2 (5.71%)	2	4.88%	*Streptococcus salivarius*	0 (0.00%)	2 (5.71%)	2	4.88%

*p=0.679, **p=0.679, ***p=0.679. All p-values determined through use of the Fisher exact test.

3.3.4 CFTR molecular data

Of the 44 CF patients attending the CF clinic at SBAH, 42 had been screened genetically. Analysis of data obtained from the 42 SBAH CF patients who had genetic screening results showed that the ΔF508 variant was present at an allele frequency of 71%. The c.3120+1G>A variant had a frequency of 9.5%, while the R117H and c.394delTT variants were each represented by a frequency of 2.9%. The G542X variant was found at a frequency of 1.5%, while the frequency of unknown/unidentified variants was 13% (Figure 7). Mutation screening further revealed that three Black South African patients where homozygous for the c.3120+1G>A variant and 25 patients homozygous for the ΔF508 variant (one Mixed race and 24 White South African CF patients). Seven patients were compound heterozygous for the ΔF508 variant (one Black and six White South African CF patients). Three patients (the one Indian patient and two White South African patients) harboured the ΔF508 variant, but possessed an unknown second variant. Four patients (two Black and two White South African) had two unknown causative mutations each.

Figure 7: Frequency distribution (%) of the five positively identified mutational types present in 42 CF patients attending the SBAH CF clinic



Five pathogenic CF variants (c.3120+1G>A, c.394delTT, ΔF508, G542X, and R117H) were identified in the 42 CF patients attending the SBAH CF clinic. Two of these variants were identified in the Black South African CF group (c.3120+1G>A and ΔF508), one variant was observed in both the Mixed race and Indian CF patients (ΔF508), and all five variants identified in the White South African CF population. However, The c.3120+1G>A and ΔF508 mutations present in the Black South African CF patients accounted for 67% of the positively identified CF variants, while the five mutations present in the White South African CF patients (c.3120+1G>A, c.394delTT, ΔF508, G542X, and R117H) accounted for 91% of the positively identified CF variants. According to Goldman *et al.* (17), 21% and 83% of CF variants are expected to be positively identified in Black and White South African CF patients, respectively. The frequency of positively identified variants in the six Black South African CF patients attending SBAH was thus notably higher than that reported by Goldman *et al.* (17), which was also determined to by statistically significant ($p<0.001$). The number of positively identified variants in White South African CF patients attending SBAH did not differ significantly from the expected Goldman frequency ($p=0.08$).

In comparison to this, molecular data for nearly 400 South African CF patients have been reported previously (17, 99, 118-121). Within these 400 reported CF patients, 18 different CF-associated variants were identified. As shown in Table 8, 16 variants were identified in 258 White South African CF patients, while six were found in both the 39 Black and 76 Mixed race South African CF patients. However, in order to avoid potential overlaps in data, three published reports of CF patients from the South African population were excluded from analysis (110, 122, 123). Furthermore, in one publication, there was no stratification

according to race (99), and the results obtained from this particular publication are thus not reported. Notably, the 16 mutations reported in the White South African CF population represented 89.5% of all identified variants in this group, while the four mutations present in the Black and five mutations present in the Mixed race South African CF patients represented 29.5% and 51.3% of positively identified CF variants in these two groups, respectively.

Table 8: Summary of South African CF variants reported in literature

Ethnicity	Total number of patients tested	Percentage (%) value of causative CF variants identified	Total number of different variants found
Black	39	29.5	4
Mixed race	76	51.3	5
White	258	89.5	16

The following publications were used in the extraction of data for this table (17, 99, 118-121), while the following publications were excluded in order to avoid potential data duplications (110, 122, 123).

Of the 18 causative mutations reported in South African CF patients, 11 variants (c.1717-1G>A, c.2789+5G>A, c.3659delC, c.394delTT, c.621+1G>T, N1303K, Q493X, R117H, R553X, S549N, and W1282X) were found exclusively in the White South African CF population, while three (c.3196del54, c.54-1161_c.164+1603del2875, and G1249E) were identified in the Black South African population alone. Four variants (c.3272-26A>G, ΔF508, G542X, and G551D) were found in both the Mixed race and White South African CF population groups.

As shown in Table 9, the most commonly observed variant in Black South African CF patients, c.3120+1G>A, occurred at an overall frequency of nearly 25%. This was significantly lower (p=0.021) than the 58% observed in the Black South African CF patients attending SBAH. It was furthermore found that the c.3120+1G>A variant occurred in approximately 10%, and 0.4% of the Mixed race and White South African CF patients, respectively. This difference was statistically significant in each of the comparisons for this variant between the three ethnic groups (Black South African and Mixed race [p=0.006], Black South African and White South African [p<0.001], and Mixed race and White South African CF patients [p<0.001]).

The common Caucasoid variant, ΔF508, occurred in 38.2% of Mixed race and 77.5% of White South African CF patients, with the difference in frequency of this variant between the two ethnic groups being significantly different (p<0.001). No statistically significant difference (p=0.680) was however observed when comparing the frequency of this variant between the White South African CF patients at SBAH (82%) and those reported in literature (77.5%). Interestingly, no literature report has been found describing the presence of the ΔF508 variant in a Black South African CF patient. It has however been identified in a single Black South African CF patient attending SBAH in this study.

Table 9: Distribution of CF variants in the South African population according to literature

	Black	Mixed race	White
No. of patients	39	76	258
Number of Chromosomes	78	152	516
c.1717-1G >A	-	-	1 (0.2%)
c.2789+5G>A	-	-	1 (0.2%)
c.3120+1G>A	19 (24.4%)	15 (9.9%)	2 (0.4%)
c.3196del54	2 (2.6%)	-	-
c.3272-26A>G	-	1 (0.7%)	16 (3.1%)
c.3659delC	-	-	1 (0.2%)
c.394delTT	-	-	15 (2.9%)
c.621+1G>T	-	-	1 (0.2%)
c.54-1161_c.164+1603del2875	1 (1.3%)	-	-
deltaF508	-	58 (38.2%)	400 (77.5%)
G542X	-	2 (1.3%)	7 (1.4%)
G551D	-	2 (1.3%)	3 (0.6%)
N1303K	-	-	4 (0.8%)
Q493X	-	-	1 (0.2%)
R117H	-	-	1 (0.2%)
R553X	-	-	4 (0.8%)
S549N	-	-	1 (0.2%)
W1282X	-	-	4 (0.8%)

The following publications were used in the extraction of data for table 9 (17, 118-121), while the following publications were excluded in order to avoid potential data duplications (110, 122, 123). Dashed values indicate that literary support for the presence of the CF variant in question was not reported for that particular ethnic group.

3.4 Discussion and conclusion

3.4.1 Demographic and clinical presentation of patients attending the SBAH CF clinic

It has been widely reported and commonly accepted that CF is typically associated with the presentation of intestinal obstruction(s) (MIE and/or DIOS), pancreatic insufficiency, intestinal malabsorption and poor weight gain, as well as frequent and re-occurring chest

infections (20, 107, 108). CF-associated diabetes, infertility and osteoporosis were identified solely in White South African SBAH CF patients. However, since these patients tended to be older than other CF patients attending the SBAH CF clinic, and since these CF-associated complications tend to only manifest at a later age, this is not an entirely unsurprising finding (20).

What is interesting to note is that, of the seven reported cases of infertility, four occurred in White South African female patients and three in White South African male patients. Congenital bilateral absence of the vas deferens has been documented to occur at a low frequency within male CF patients rendering such patients infertile (20, 108, 124), but CF-associated infertility in females is however not well-documented, making this observation potentially noteworthy. Further investigation would be necessary in order to determine whether or not this trend is observed in other CF populations of South Africa.

It is known that CF is frequently underdiagnosed in Black South African patients, particularly as a result of confounding tuberculosis and HIV/AIDS infections, as well as protein energy malnutrition and chronic pulmonary infections (95, 99, 114, 115). Further compounding the problem of diagnosis of CF in the Black South African population is the incorrect notion that there are “very few reports of CF in Black Africans with minimal Caucasoid admixture” (19). Mutesa and Bours (95) state that “there is a general lack of awareness of CF in the African medical profession”. Wide-scale improvement in the knowledge available to healthcare professionals would thus be required in order to challenge and change these perceptions.

This study reports an initial and detailed collection of CF clinical data obtained from patients attending the CF clinic at SBAH. Further monitoring and maintenance of this information would be able to contribute to the improvement of available knowledge of CF in South Africa and to the establishment of a much needed national database.

3.4.2 Biochemical and Molecular test results

Although the South African Cystic Fibrosis Consensus Document (20) outlines the minimum CF diagnostic requirements when determining sweat conductivity, it was noted that of the 21 SBAH CF patients who had their sweat chloride concentrations determined, none underwent sweat conductivity analysis. It was furthermore observed that the Black CF patients had the lowest average sweat chloride concentrations (73 mmol/L, SD±29.7 mmol/L). However, no significant differences in sweat chloride concentrations were observed between Black South African and White South African CF patients attending SBAH.

It has been well documented that *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and *Haemophilus influenza* are the most commonly observed bacteria implicated in lung/pulmonary infections of South African CF patients (18, 20, 94, 99). The South African Cystic Fibrosis Foundation (20) additionally states that MRSA and multi-drug resistant *P. aeruginosa* are frequently observed in South African CF patients and are of significant interest within these patients. This is due not only to restricted treatment options in such instances, but also due to the decreased efficacy with which such infections can be treated in affected patients. Within the 44 patients attending the CF clinic at SBAH, the most commonly observed

bacterial species included *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and *Haemophilus parainfluenza* which were respectively present in approximately 59%, 44% and 56% of all patients. Literature reports suggest that *Haemophilus influenza* is a commonly-isolated bacterial species in CF patients (20). *Haemophilus influenza* infections were reported in only 29% of the patients attending the CF clinic at SBAH. Since the exact prevalence of *H. influenza* infections in the South African CF population is not known, significance cannot be tested between CF patients attending SBAH and CF patients in other regions of South Africa. Nevertheless, several environmental factors, such as air quality, risk of pathogen exposure, immune system and nutritional deficiencies, as well as lifestyle choices such as physical activity and supplement intake, may be responsible for the three commonly observed bacterial species in CF patients attending the SBAH CF clinic. A more thorough evaluation of patient's living conditions and location(s) would be required in order to test the significance of these observations.

3.4.3 CFTR molecular data

After screening six Black South African, one Mixed race, one Indian patient, and 34 White South African CF patients attending SBAH, it could be stated that unknown variants were present in only 13.1% of the CF chromosomes of those patients attending SBAH. The frequency of positively identified variants in Black and White South African SBAH CF patients was 67% and 91%, respectively. These values were well above the 21% and 73% previously reported for Black and White South African CF patients, respectively (17).

Goldman *et al.* (17, 110) have also shown that ΔF508 is present in 76% of all White South African CF patients, while the c.3120+1G>A variant was observed in 46% of all Black South African CF patients. The frequencies of these two variants in White and Black South African CF patients attending SBAH were similarly shown to be 85% and 58%, respectively. However, the study of collated literature revealed that the frequency of the ΔF508 in White South African CF patients is approximately 78%, while the frequency of the c.3120+1G>A variant in Black South African CF patients is as low as 24%.

When comparing the frequency occurrence of these results between the different ethnic groups, the statistically significant differences could be suggestive of the high levels of genetic diversity that one would expect to observe in admixture/Mixed race populations as well as populations of African origin (125, 126). However, it must be considered that the Black South African CF patients reported in literature and present at SBAH have small sample sizes (six Black South African CF patients attending SBAH, and 14 Black South African CF patients tested by Goldman et al. (17)). A larger cohort of Black South African CF patients should therefore be studied in order to add further confidence to this observation.

Regardless, the generally low percentage of identified variants within the Black and Mixed race South African CF populations implies that a higher percentage of identified variants in these populations could be achieved if whole gene, exome or whole genome sequencing is done, and population specific CF screening panels developed from this. However, the large percentage of identified variants within the White South African CF patients, as well as White

South African CF patients attending SBAH, indicates that existing genetic screening methods have been sufficiently developed for this particular population.

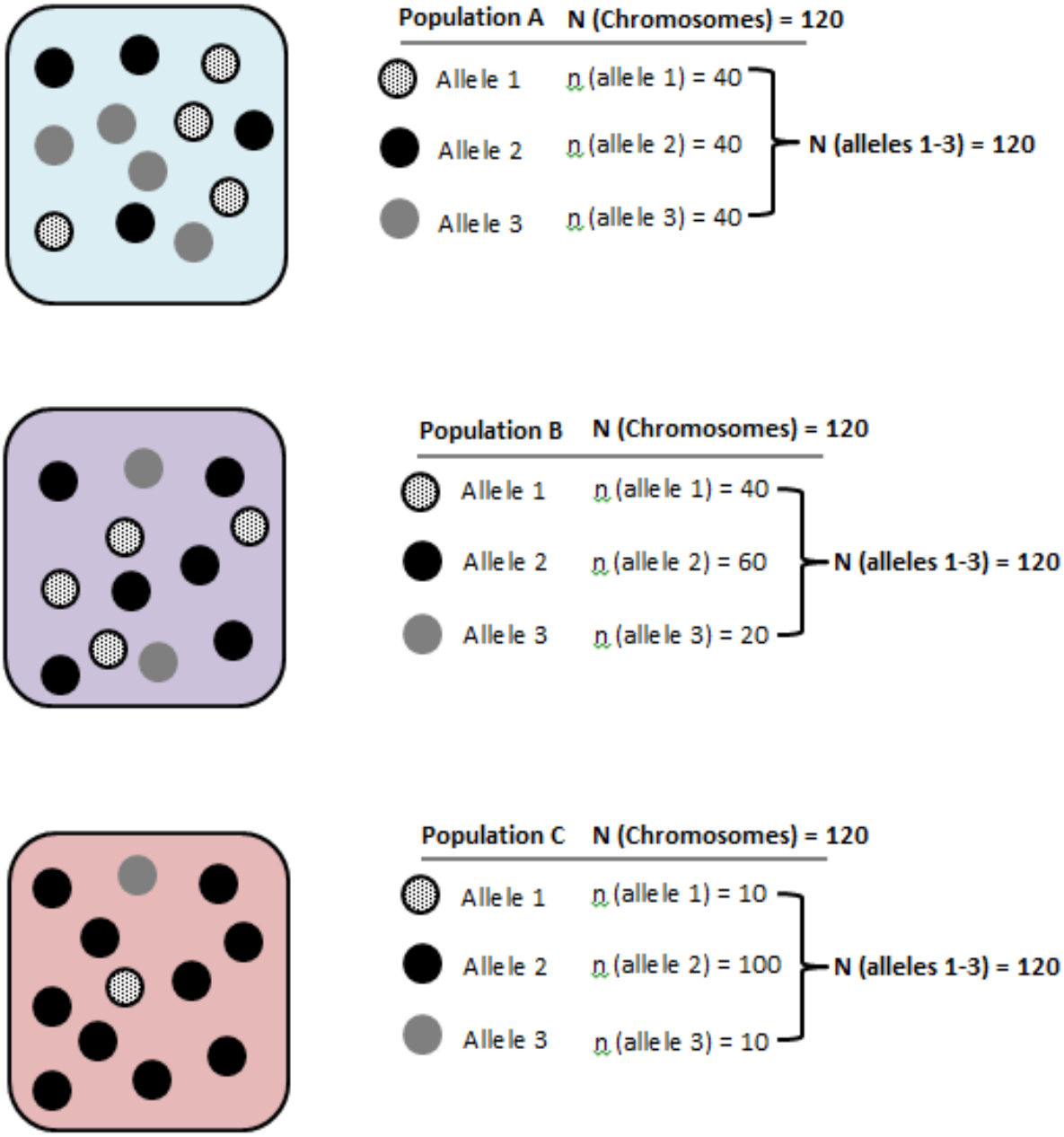
It has, in previous years, been known that there is an urgent need for a clear clinical and molecular characterisation of CF to be made for South African populations. This led to the construction of the CF consensus document of South Africa (20). Thus, in keeping with the outlines of this document, a thorough clinical and molecular investigation of CF patients attending SBAH has been undertaken and is presented in this study. However, in order to determine how the presentation of CF differs between SBAH patients and those attending any of the other CF clinics in South Africa, a collaborative effort will have to be undertaken by each of the CF clinics in South Africa in terms of collecting and storing CF patient data centrally. This is certainly not an impossible feat, but would require very careful thought and management in order to prevent data duplication and ensure patient confidentiality. The establishment of such a centralised database would also contribute to South Africa's agreement to the UN's CRC agreement in a positive light, thereby providing the framework from which to provide long-term assistance and support to CF patients with regards to healthcare services.

Chapter 4 – Cystic Fibrosis and diversity: Unification through variation

4.1 Introduction

Gene diversity is defined as the probability that two randomly chosen copies of a gene will be different alleles (30). Similarly, genetic diversity can refer not only to the variety of alleles that occur at a gene locus or population, but also how distinct such alleles are within and between studied populations - whether in micro-organisms, plants, humans or animals (9). To demonstrate this largely intuitive concept, we can consider an example of alleles associated with a particular disease in three different population groups. As shown in Figure 10, suppose that three equally sized populations consisting of 60 individuals, or 120 chromosomes, each hypothetically contain three different alleles (represented by the shaded circles). Further assume that each of the three different alleles are associated with the manifestation of elevated cholesterol levels. By applying Nei's definition of gene diversity (30), the probability of randomly selecting two different alleles for this hypothetical "high-cholesterol" gene would thus instinctively be greatest in population A and lowest in population C.

Figure 8: Hypothetical example of gene diversity for a high-cholesterol gene in three different populations



Although easy to determine when the number of alleles and/or populations studied is small and equal in size, genetic diversity becomes complicated to measure when uneven sample sizes are observed and large populations are studied. For this reason, several mathematical methods exist through which genetic/gene diversity can be determined - the Shannon-Weaver (1), Simpson, Simpson Diversity (2) and rarefaction (6, 7, 9, 13) diversity indices are just some of several examples of commonly-applied diversity methods.

Although slight differences exist in their mathematical composition, it is noteworthy to state that diversity indices determine diversity by using an assumption pertaining to sampling effort, and by secondarily considering two statistical parameters – sample evenness and species richness (1, 2, 6, 7, 65). By "equal sampling effort", it is implied that the same amount of time has been spent and effort exerted in generating sampling results from sampling region "A" as in sampling region "B". Thus, in a medical context, equal sampling effort would imply that an equal number of patients had been tested for the same disease-causing variants in each sampled population/region, as observed in our hypothetical example presented in Figure 10.

Sample evenness describes how the various alleles within a sampled region are distributed, while species richness gives an indication of how many distinct alleles are represented within a sampled region/population (6, 7). Once again, by using the hypothetical example in Figure 10, it can be shown that all three alleles were found a total of 40 times each in population "A". Thus, not only has the sample distribution (i.e. sample evenness) been described through

the evaluation of the total number of times each allele was identified in population "A", but also how many different alleles were studied in population "A" (i.e. species richness). Therefore, in population "A", it is observed that the three sampled alleles are evenly distributed. By comparison, in population "B", the three sampled alleles are not evenly distributed, but are still more evenly distributed than the three alleles that were sampled in population "C". Although described as individual entities, sample evenness and species richness are the core mathematical principles behind the use of the Shannon-Weaver, Simpson, Simpson Diversity and rarefaction methods, and never function independently of each other (8).

The mathematical concept of diversity, until recently, has predominantly been focused within ecological disciplines (6, 13, 14). Regardless, both Kalinowski (13) and Colwell *et al.* (14) have indicated the usefulness of considering genomic data to determine diversity in humans, and have developed methods through which diversity theory can be applied (64). However, the concept of diversity and extent to which diversity theory can be used to measure the genetic diversity of alleles associated with monogenic disorders, such as those causing cystic fibrosis (CF), has to our knowledge not yet been explored. We therefore endeavoured to 1) elucidate the most appropriate method of measuring genetic diversity of common disease-causing mutations in different CF populations world-wide, and 2) describe this diversity relative to the South African CF population.

4.2 Methods

4.2.1 Data assimilation

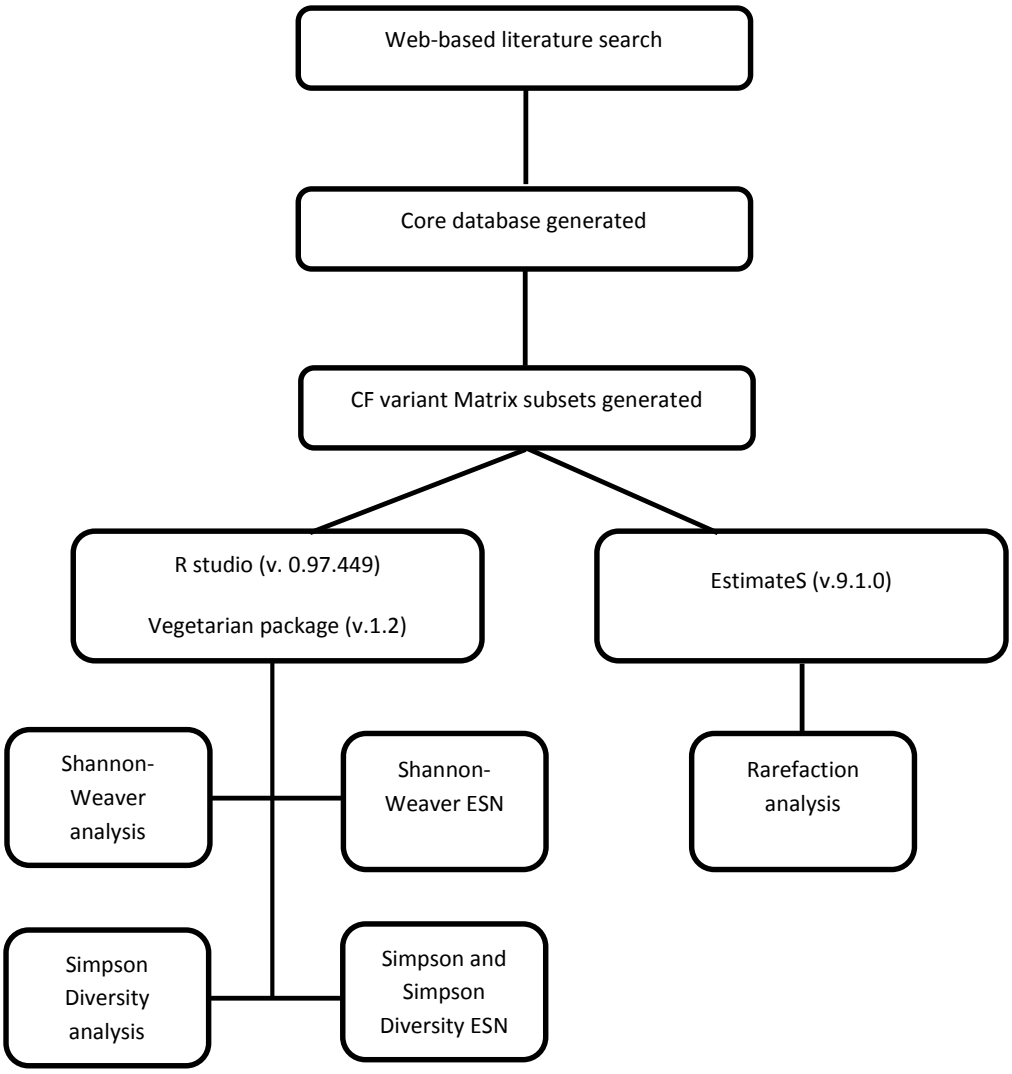
As outlined in Figure 9, a core database was generated from published information through the use of internet search terms. These search terms, which were investigated via Google Scholar and PubMed, included any combination of “Cystic Fibrosis”, “Cystic Fibrosis variants”, “Cystic Fibrosis variant frequency”, “Europe”, “North America”, “South America”, “USA”, “Australia”, “New Zealand”, “Asia”, and “Africa”. In order to ensure that the largest possible number of CF variant publications was obtained, the term “variant” was also substituted with “mutation”. Reports were thoroughly studied and databased according to continental region, the total number of chromosomes tested per population/region and the total number of times each published CF variant was identified.

In accordance with recommendations made by Gotelli and Colwell (65), data from countries, populations or regions in which less than 40 chromosomes had been sampled in total were excluded from the database. This was done in order to prevent any unnecessary skewing of results that could occur due to excessively large variation in sample sizes between studies countries, populations or regions. In order to allow for a comparison of genetic diversity across populations, variants associated uniquely with a single population/region were excluded.

4.2.2. Diversity analysis

Genetic diversity of select CF variants was measured using the Shannon-Weaver, Simpson Diversity, and rarefaction methods. Shannon-Weaver and Simpson Diversity analysis was performed utilising the Vegetarian package (v.1.2)(127) in R studio (v. 0.97.449) (128) in accordance with the methods proposed and described by Jost (6, 7). Following the methods described by Hill (9) and elaborated on by Jost (6, 7), Effective Species Numbers (ESNs)/Hill ratios were additionally calculated using the Vegetarian package (v.1.2.)(127) for each of these three diversity measures, while standard error (SE) values were reported through bootstrapping methods described by Chao *et al.* (129) as an in-built function of the Vegetarian package (v.1.2.)(127). Significance testing of Shannon and Simpson ESN results were performed through the use of the R Studio statistical program (128), and by using the Shannon t-test method as described by Hutcheson (130) and Simpson t-test method outlined by Gardener (131). Bonferroni correction for multiple comparison was performed using R Studio (128) through methods developed by Bonferroni (132) and outlined by Gardener (131). Individual-based rarefaction analysis was performed using EstimateS (v.9.1.0) (64), with 95% CIs being determined through methods described by Colwell *et al.* (14). In accordance with Colwell *et al.* (14, 64), sample order was not randomized and a single run was performed per dataset.

Figure 9: Algorithm used to determine world-wide CF variant diversity



4.3 Results

4.3.1 Data assimilation

Web-based internet searches identified several publications, of which 48 contained information that was relevant to the search criteria. However, a single, pivotal, publication presented CF variant data for 206 disease-associated variants in 52 different countries and was constructed from 115 different publications (133). Thus, for comparative purposes, this publication by Bobadilla *et al.* (133) was used as the main source of world-wide CF variant data. In order to include the most recent CF data from other countries, CF registry data replaced that presented by Bobadilla *et al.* (133) for five of the 52 countries (134-138). Additionally, country-wide data presented by Watson *et al.* (139) replaced the Bobadilla CF data for the USA (United States of America). However, when considering all the data presented when collating the nearly 165 publications that were collected in total, an approximated 67 000 published cases of CF were investigated on a global scale.

It was additionally found that Goldman *et al.* (17) presented the largest single report of CF variants in the South African population. This publication, highlighting 17 identified CF variants, was therefore used as the point of comparison between CF variants present in South African and other CF populations. However, in order to ensure that the most recent information was used within the South African CF population, CF variant data from an additional three publications (99, 119, 121) was collated with data presented by Goldman *et al*. (17). Since Masekela *et al.* (99) reported on the molecular status of several CF patients attending the SBAH CF clinic, data obtained from SBAH patient files with regard to genetic screening was not included into this database in order to avoid duplication of the data.

As reported by Bobadilla *et al.* (133), of the 52 different countries, 33 were European, seven were Middle Eastern and North African, nine were associated with North, Central or South America, two were Australasian and one was Asian. A total of 142 CF variants were reported in the 33 European countries, while 58 CF variants were described within the seven Middle Eastern and Northern African countries. Eight CF variants were described within the two Australasian countries, while two variants were reported within the Asian country. In accordance with recommendations made by Gotelli and Colwell (65), sample sizes of less than 40 CF-positive chromosomes were excluded from further analysis. Furthermore, variant data was limited to the 17 CF variants that were identified within the South African CF population. This resulted in inclusion of comparative variant data from a total of 32 European, five Middle Eastern or North African, seven North, Central or South American, and two Australasian countries (Table 10). Since each country does not necessarily screen for the same CF variants (using the same screening panels and screening methods), inclusion of variant results was not dependent on which method was used in each of the countries (108, 140, 141). All publications utilized in the construction of data presented in Table 10 were standardised according to the methods employed by Bobadilla (133) with percentage values indicating the total number of positively identified CF chromosomes in each country.

As shown in Table 10, the ΔF508 variant was found to be the most frequently occurring and was identified in each of the 46 countries. The G542X, N1303K, W1282X, R553X variants were observed the next most frequently, respectively being observed in 38, 33, 26 and 22 of the 46 countries. Conversely, a total of six of the 17 CF variants (c.394delTT, c.3659delC, S549N, c.3120+1G>A, c.3272-26A>G, and Q493X) were observed in less than 10 countries.

Table 10: World-wide frequency (%) occurrence of 17 CF variants

Country	N(Chromosomes)	deltaF508	G542X	N1303K	W1282X	R553X	G551D	c.1717-1G>A	R117H	c.621+1G>T	c.2789+5G>A	R1162X	c.394delTT	c.3659delC	S549N	c.3120+1G>A	c.3272-26A>G	Q493X
RSA	564	69.50	1.95	0.71	0.71	0.71	1.24	0.18	0.18	0.18	0.18	0.18	2.66	0.18	0.18	17.91	3.19	0.18
Belgium	1,830	83.06	3.60	3.55	0.93	0.33	0.27	1.80	1.42	-	1.20	0.82	0.38	0.44	-	0.22	1.75	0.22
UK	18,692	80.40	3.16	1.21	0.44	0.78	4.99	1.16	3.60	2.07	0.30	0.22	-	0.81	0.20	0.12	-	0.53
USA	34,173	82.93	3.30	1.59	2.75	1.51	2.41	0.55	0.68	1.61	0.47	0.37	-	0.35	0.18	1.08	-	0.21
Australia	4,266	87.20	1.88	1.15	0.63	0.47	4.81	0.73	2.20	-	0.12	0.19	0.12	0.16	0.12	-	-	0.14
Germany	10,436	88.18	2.24	2.24	0.69	2.34	1.98	0.98	0.61	0.30	-	0.39	-	-	0.05	-	-	-
Netherlands	2,463	91.46	2.24	1.35	0.63	1.48	-	1.79	2.07	-	0.69	1.23	-	-	-	-	1.18	-
Greece	2,097	52.90	4.10	3.30	0.70	-	0.50	-	1.20	5.00	-	-	-	-	-	0.60	0.80	-
Italy	2,608	83.44	5.87	3.95	1.46	1.69	0.73	2.60	-	-	0.19	-	-	-	-	-	0.04	-
Austria	1,516	62.90	3.30	0.60	-	1.70	1.20	-	0.50	-	-	1.90	-	-	-	-	-	-
France	17,854	67.70	2.94	1.83	0.91	0.86	0.74	1.35	-	-	-	-	-	-	-	-	-	-
Hungary	1,133	54.90	1.70	1.30	1.80	2.10	1.00	1.90	-	-	-	-	-	-	-	-	-	-
Poland	4,046	57.10	2.60	1.80	0.70	1.90	0.50	2.40	-	-	-	-	-	-	-	-	-	-
Spain	3,608	52.70	8.00	2.50	0.60	-	-	-	-	-	0.70	1.60	-	-	-	-	0.50	-
Turkey	1,067	24.50	2.60	2.90	0.70	-	-	-	-	2.60	3.90	0.60	-	-	-	-	-	-
Belarus	278	61.20	4.50	3.20	1.00	0.50	-	-	-	-	-	-	-	-	0.50	-	-	-
Bulgaria	948	63.60	3.90	5.60	1.00	-	-	0.80	-	-	0.80	-	-	-	-	-	-	-
Denmark	1,888	87.50	0.70	1.10	-	-	-	-	-	0.60	-	-	1.80	0.60	-	-	-	-
Ireland	801	70.40	1.00	-	-	-	5.70	0.60	2.40	0.80	-	-	-	-	-	-	-	-
Norway	410	60.20	0.60	0.60	-	-	1.20	-	3.00	-	-	-	4.20	-	-	-	-	-
Switzerland	1,268	57.20	2.60	1.20	1.10	14.00	-	2.70	-	-	-	-	-	-	-	-	-	-
Reunion Island	138	52.00	0.70	-	-	-	1.40	0.70	-	-	0.70	-	-	-	-	8.00	-	-
Brazil	820	47.70	7.20	2.40	1.30	0.70	-	-	-	-	-	2.50	-	-	-	-	-	-
Czech Republic	2,196	70.00	2.20	2.90	0.60	-	3.80	-	-	-	-	-	-	-	-	-	-	-
Romania	224	36.60	1.40	-	1.70	1.40	-	-	-	1.40	-	-	-	-	-	-	-	-
Russia	5,073	54.40	0.90	-	1.00	3.50	-	-	-	-	-	-	1.00	-	-	-	-	-
Slovakia	908	57.30	6.80	3.40	1.30	4.00	-	-	-	-	-	-	-	-	-	-	-	-

Country	N(Chromosomes)	deltaF508	G542X	N1303K	W1282X	R553X	G551D	c.1717-1G>A	R117H	c.621+1G>T	c.2789+5G>A	R1162X	c.394delTT	c.3659delC	S549N	c.3120+1G>A	c.3272-26A>G	Q493X
Slovenia	455	57.80	1.90	-	-	0.80	-	-	-	-	4.10	3.20	-	-	-	-	-	-
Sweden	1,357	66.60	0.60	-	-	-	-	-	0.60	-	-	-	7.30	5.40	-	-	-	-
Ukraine	1,055	65.20	-	2.40	0.50	3.60	1.80	-	-	-	-	-	-	-	-	-	-	-
Yugoslavia	709	68.90	4.00	0.80	-	-	0.50	-	-	0.50	-	-	-	-	-	-	-	-
Argentina	326	58.60	3.90	1.80	3.90	-	-	0.90	-	-	-	-	-	-	-	-	-	-
Mexico	374	41.60	-	-	-	0.50	-	-	0.50	1.20	-	-	-	-	1.90	-	-	-
New Zealand	636	78.00	2.00	1.90	-	-	4.40	-	-	1.10	-	-	-	-	-	-	-	-
Croatia	276	64.50	3.30	2.90	-	-	1.10	-	-	-	-	-	-	-	-	-	-	-
Macedonia	559	54.30	4.20	2.00	-	-	-	-	-	1.30	-	-	-	-	-	-	-	-
Lebanon	40	35.00	-	10.00	20.00	-	-	-	-	-	2.50	-	-	-	-	-	-	-
Tunisia	78	17.60	8.90	6.40	2.60	-	-	-	-	-	-	-	-	-	-	-	-	-
Columbia	48	35.40	6.30	2.10	2.10	-	-	-	-	-	-	-	-	-	-	-	-	-
Estonia	165	51.70	-	-	-	-	-	-	-	-	-	-	13.30	1.70	-	-	-	-
Finland	132	46.20	1.90	-	-	-	-	-	-	-	-	-	28.80	-	-	-	-	-
Lithuania	94	31.00	-	2.00	-	4.00	-	-	-	-	-	-	-	-	-	-	-	-
Portugal	739	44.70	1.60	0.70	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Chile	72	29.20	-	-	-	4.20	-	-	-	-	-	-	-	-	-	-	-	-
Venezuela	54	29.60	3.70	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Albania	270	72.40	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
UAE	86	26.90	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

All variant values are indicated as percentages, while the total number of positive CF chromosomes sampled per country is indicated by N(Chromosomes). A dash was used in cases where no data was present for a particular variant in a given country. A total of 11 publications were used in the construction of this particular matrix (17, 99, 119, 121, 133-139).

Since data is not available for every variant in each of the countries listed in Table 10, it was decided to derive three different matrices based on rational premise and specific criteria:

Matrix 1 - The "South African matrix", based on the inclusion of the three most prevalent alleles in South Africa (ΔF508, c.3120+1G>A and c.3272-26A>G). Additional alleles were included into the matrix if data existed for each of the countries already included based on the initial criteria.

Matrix 2 - Based on including the highest possible number of alleles for comparison

Matrix 3 - Based on including the highest possible number of countries for comparison

Matrix 1 was constructed based on the premise of the three most commonly observed CF variants in South Africa. Three countries (the Republic of South Africa [RSA], Belgium, and Greece) were found to harbour frequency data for each of the three variants in question (ΔF508, c.3120+1G>A and c.3272-26A>G), as well as for an additional five variants (G542X, G551D, N1303K, W1282X, and R117H). Thus, as shown in Table 11, eight variants were considered for diversity analysis in Matrix 1. When observing variant data for Matrix 1, it was found that in the three studied countries (Belgium, Greece and The RSA), two of the eight variants (ΔF508 and c.3120+1G>A) were present in more than 10% of the disease chromosomes in the RSA, while only ΔF508 was found to be the predominant allele in Belgium (83.1%) and Greece (52.9%). All remaining variants were identified at frequencies of less than 5% in these three countries.

Table 11: Variant subset data - Matrix 1

Country	Total Number Disease Chromosomes in Country	deltaF508	c.3120+1G>A	c.3272-26A>G	G542X	G551D	N1303K	W1282X	R117H
RSA	564	392 (69.5%)	101 (17.9%)	18 (3.2%)	11 (2.0%)	7 (1.2%)	4 (0.7%)	4 (0.7%)	1 (0.2%)
Belgium	1,830	1,520 (83.1%)	4 (0.2%)	32 (1.8%)	66 (3.6%)	5 (0.3%)	65 (3.6%)	17 (0.9%)	26 (1.4%)
Greece	2,097	1,109 (52.9%)	13 (0.6%)	17 (0.8%)	86 (4.1%)	11 (0.5%)	69 (3.3%)	15 (0.7%)	25 (1.2%)

Percentage values reflected in brackets are relative to the total number of disease chromosomes identified in each of the respective countries.

The G542X variant was the second most frequently observed variant in Belgium and Greece, occurring at frequencies of 3.6% and 4.1%, respectively. The c.3120+1G>A variant, reported at 17.9% of South African CF patients, was the second most frequently observed variant in this population. The R117H variant was the least frequently observed variant in The RSA (0.2%), while the c.3120+1G>A and G542X variants were the observed with the lowest frequencies in Belgium and Greece, respectively.

Construction of the second matrix (Matrix 2) was based on including as many variants as possible (at least 10), and ensuring that each variant had a reported frequency in as many countries possible. This included variant data from The RSA, Belgium, The UK, The USA, Australia, Germany and The Netherlands. To this end, a total of 12 variants (ΔF508, G542X, N1303K, W1282X, R553X, G551D, c.1717-1G>A, R117H, c.2789+5G>A, R1162X, c.3659delC, and Q493X), present in five countries (The RSA, Belgium, the United Kingdom [UK], the USA, and Australia), were utilised in the construction of Matrix 2 (Table 12).

The ΔF508 variant was the most frequently observed variant within each of the five countries occurring at a frequency ranging between 69.5% and 87.2%, while the G542X variant was the second most frequently observed variant in The RSA, Belgium and the USA – occurring at frequencies of 2.0%, 3.6%, and 3.3%, respectively. The G551D variant was the second most common variant in the UK and Australia and was identified in 5.0% and 4.8% of all CF patients, respectively.

Table 12: Variant subset data - Matrix 2

Country	Total Number Disease Chromosomes in Country	deltaF508	G542X	N1303K	W1282X	R553X	G551D	c.1717-1G>A	R117H	c.2789+5G>A	R1162X	c.3659delC	Q493X
RSA	564	392 (69.5%)	11 (2.0%)	4 (0.7%)	4 (0.7%)	4 (0.7%)	7 (1.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
Belgium	1,830	1,502 (83.1%)	66 (3.6%)	65 (3.6%)	17 (0.9%)	6 (0.3%)	5 (0.3%)	33 (1.8%)	26 (1.4%)	22 (1.2%)	15 (0.8%)	8 (0.4%)	4 (0.2%)
UK	18,692	15,032 (80.4%)	590 (3.2%)	226 (1.2%)	82 (0.4%)	146 (0.8%)	932 (5.0%)	216 (1.2%)	672 (3.6%)	14 (0.3%)	11 (0.2%)	38 (0.8%)	25 (0.5%)
USA	34,173	28,339 (82.9%)	1,128 (3.3%)	543 (1.6%)	940 (2.8%)	517 (1.5%)	825 (2.4%)	188 (0.6%)	231 (0.7%)	20 (0.5%)	16 (0.4%)	15 (0.4%)	9 (0.2%)
Australia	4,266	3,722 (87.2%)	80 (1.9%)	80 (1.2%)	27 (0.6%)	20 (0.5%)	205 (4.8%)	31 (0.7%)	94 (2.2%)	5 (0.1%)	8 (0.2%)	7 (0.2%)	6 (0.1%)

Percentage values reflected in brackets are relative to the total number of disease chromosomes identified in each of the respective countries.

Finally, Matrix 3 was constructed on the premise that as many comparable countries as possible was to be included, but with a secondary condition that at least five variants should be accounted for to allow for meaningful comparisons. As shown in Table 13, a total of five CF variants, namely, ΔF508, G542X, N1303K, W1282X, and R553X, were compared across a total of 15 countries (The RSA, Belgium, the UK, the USA, Australia, Germany, the Netherlands, Italy, France, Hungary, Poland, Belarus, Switzerland, Brazil and Slovakia). The ΔF508 variant was the most frequently observed variant in each of the 15 countries, occurring in a percentage range of between 47.7% and 91.5%. The G542X variant was the second most frequently occurring variant in 13 of the countries (The RSA, Belgium, the UK, the USA, Australia, the Netherlands, Italy, France, Poland, Belarus, Switzerland, Brazil and Slovakia), and was observed in between 1.9% and 7.2% in these countries. In Germany, the G542X and N1303K variants were both observed with a frequency of 2.2% and were jointly the second most frequently occurring variants if these patients. In Hungary, the second most frequently observed variant (R553X) was observed in 2.1% of the sampled population.

Table 13: Variant subset data - Matrix 3

Country	Total Number Disease Chromosomes in Country	deltaF508	G542X	N1303K	W1282X	R553X
RSA	564	392 (69.5%)	11 (2.0%)	4 (0.7%)	4 (0.7%)	4 (0.7%)
Belgium	1,830	1,520 (83.1%)	66 (3.6%)	65 (3.6%)	17 (0.9%)	6 (0.3%)
UK	18,692	15,032 (80.4%)	590 (3.2%)	226 (1.2%)	82 (0.4%)	146 (0.8%)
US	34,173	28,339 (82.9%)	1128 (3.3%)	543 (1.6%)	940 (2.8%)	517 (1.5%)
Australia	4,266	3,722 (87.2%)	80 (1.9%)	49 (1.2%)	27 (0.6%)	20 (0.5%)
Germany	10,436	9,202 (88.2%)	234 (2.2%)	234 (2.2%)	72 (0.7%)	244 (2.3%)
Netherlands	2,463	2,161 (91.5%)	52 (2.2%)	32 (1.4%)	15 (0.6%)	35 (1.5%)
Italy	2,608	2,176 (83.4%)	153 (5.9%)	103 (4.0%)	38 (1.5%)	44 (1.7%)
France	17,854	12,087 (67.7%)	525 (2.9%)	136 (1.8%)	163 (0.9%)	154 (0.9%)
Hungary	1,133	622 (54.9%)	19 (1.7%)	15 (1.3%)	20 (1.8%)	24 (2.1%)
Poland	4,046	2,310 (57.1%)	105 (2.6%)	73 (1.8%)	28 (0.7%)	77 (1.9%)
Belarus	278	170 (61.2%)	13 (4.5%)	9 (3.2%)	3 (1.0%)	1 (0.6%)
Switzerland	1,268	725 (57.2%)	33 (2.6%)	15 (1.2%)	14 (1.1%)	178 (14.0%)
Brazil	820	391 (47.7%)	59 (7.2%)	20 (2.4%)	11 (1.3%)	6 (0.7%)
Slovakia	908	520 (57.3%)	62 (6.8%)	31 (3.4%)	12 (1.3%)	36 (4.0%)

Percentage values reflected in brackets are relative to the total number of disease chromosomes identified in each of the respective countries.

4.3.2 Diversity analysis

4.3.2.1 Matrix 1

Since it is assumed that an equal sampling effort has been utilised in order for Shannon and Simpson Diversity analysis to be performed (1, 2), it was necessary to consider (as shown in Table 11) the variation that existed in the number of disease chromosomes studied in each of the countries included in Matrix 1. Therefore, in determining the diversity of the eight variants presented in Matrix 1 using Shannon, Simpson and Simpson Diversity indices, variant data was standardised according to the country in which the lowest number of disease chromosomes were sampled (the RSA, n=564). In order to achieve this, a simple random sampling approach was applied to the data presented in Table 10 and each of the CF matrices reconstructed. Table 14 thus indicates the randomly sampled variant data included in Matrix 1 for Shannon and Simpson Diversity analysis.

Table 14: Variant values used for Shannon and Simpson Diversity analysis in CF Matrix 1 through simple random sampling

Country	Total Number Disease Chromosomes in Country	deltaF508	c.3120+1G>A	c.3272-26A>G	G542X	G551D	N1303K	W1282X	R117H
RSA*	564	392	101	18	11	7	4	4	1
Belgium	564	458	1	1	24	2	23	6	11
Greece	564	427	4	4	36	5	29	6	10

The country marked with an asterix (*) indicates the country with the lowest number of disease chromosomes studied, and hence the reference population to which the number of chromosomes in other populations was standardised.

As shown in Table 15, the diversity between the eight variants in the three studied countries, when analysed through the Shannon index, was found to be 0.9, 0.6, and 0.8 in CF populations of the RSA, Belgium and Greece, respectively. By contrast, Simpson Diversity index results indicated diversity values of 0.4, 0.2, and 0.3 for each of the respective countries.

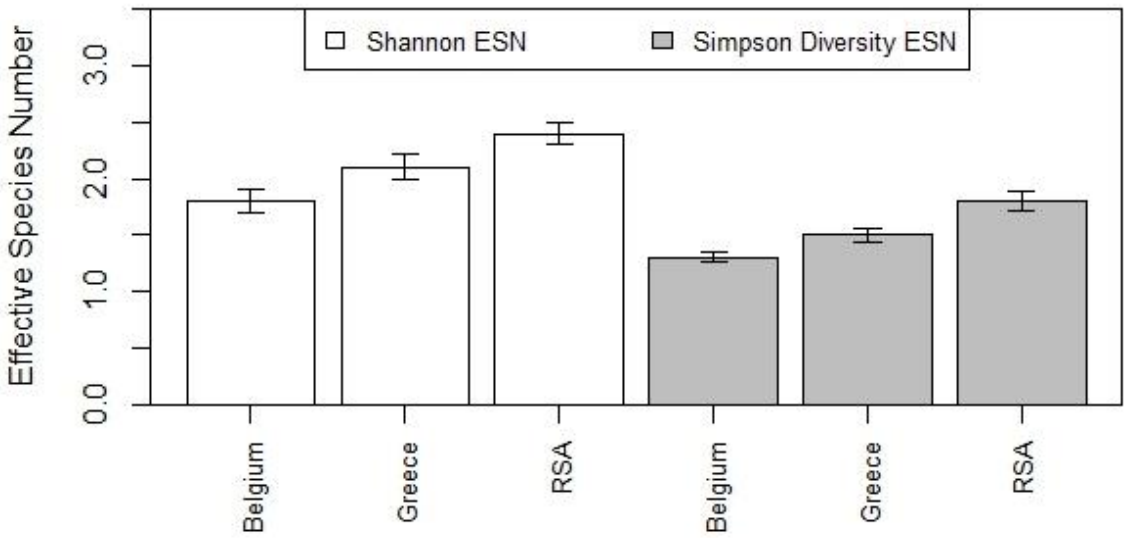
Table 15: Results from Shannon-Weaver and Simpson Diversity analysis of Matrix 1

Country	Shannon index		Simpson Diversity index	
	Index	SE	Index	SE
RSA	0.9	0.05	0.4	0.02
Belgium	0.6	0.05	0.2	0.02
Greece	0.8	0.06	0.3	0.03

SE = standard error.

However, in order to interpret Shannon, Simpson and Simpson Diversity index results, it was necessary to view the obtained results in the light of their respective effective species numbers (ESNs). As shown in Figure 10, it was found that Shannon ESNs indicated that the presence of approximately two equally distributed variants within each of the three studied countries would generate the observed diversity. Statistically significant differences, using p-values corrected by the Bonferroni method, were observed between Belgian CF populations and CF populations of the RSA ($p<0.001$). Comparatively, Simpson Diversity ESNs revealed that between one and two variants were likely to generate the observed diversity in each of the countries. P-values corrected through use of the Bonferroni method indicated that a statistically significant difference was once more observed in the diversity between the RSA and Belgian CF populations with regard to the variants analysed in Matrix 1 ($p=0.008$).

Figure 10: Shannon and Simpson Diversity Effective Species Number analysis of Matrix 1

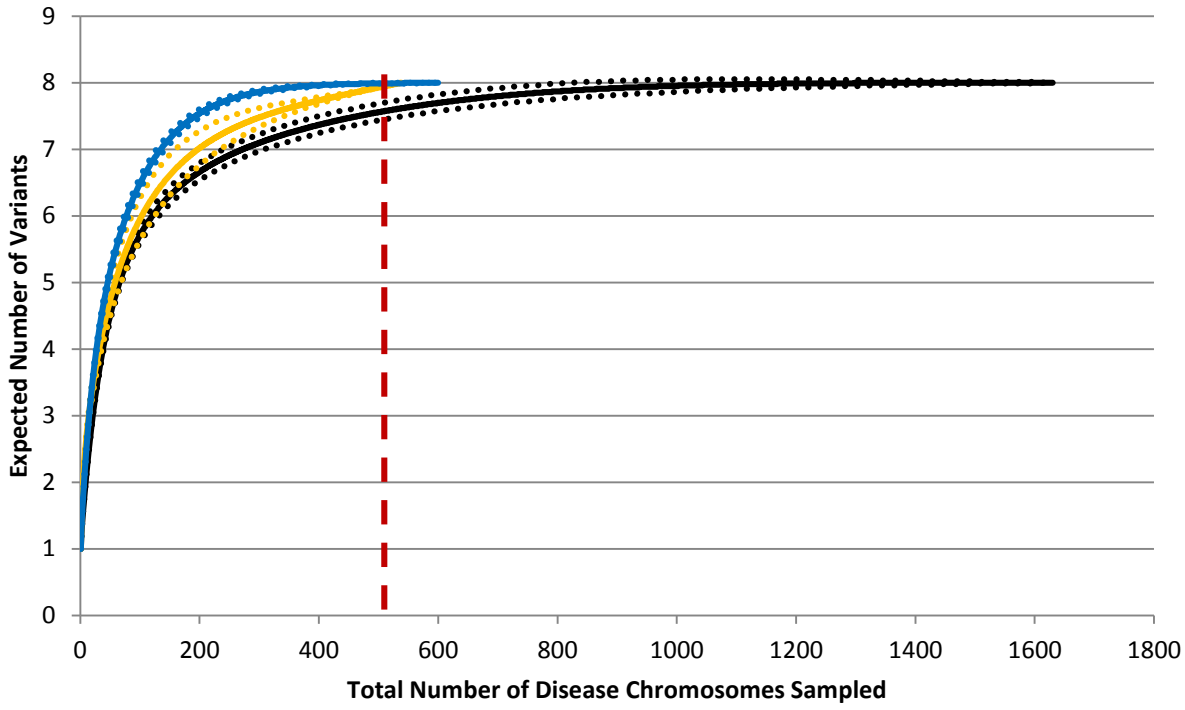


Standard error bars are indicated for each country, with ESN referring to Effective Species Number.

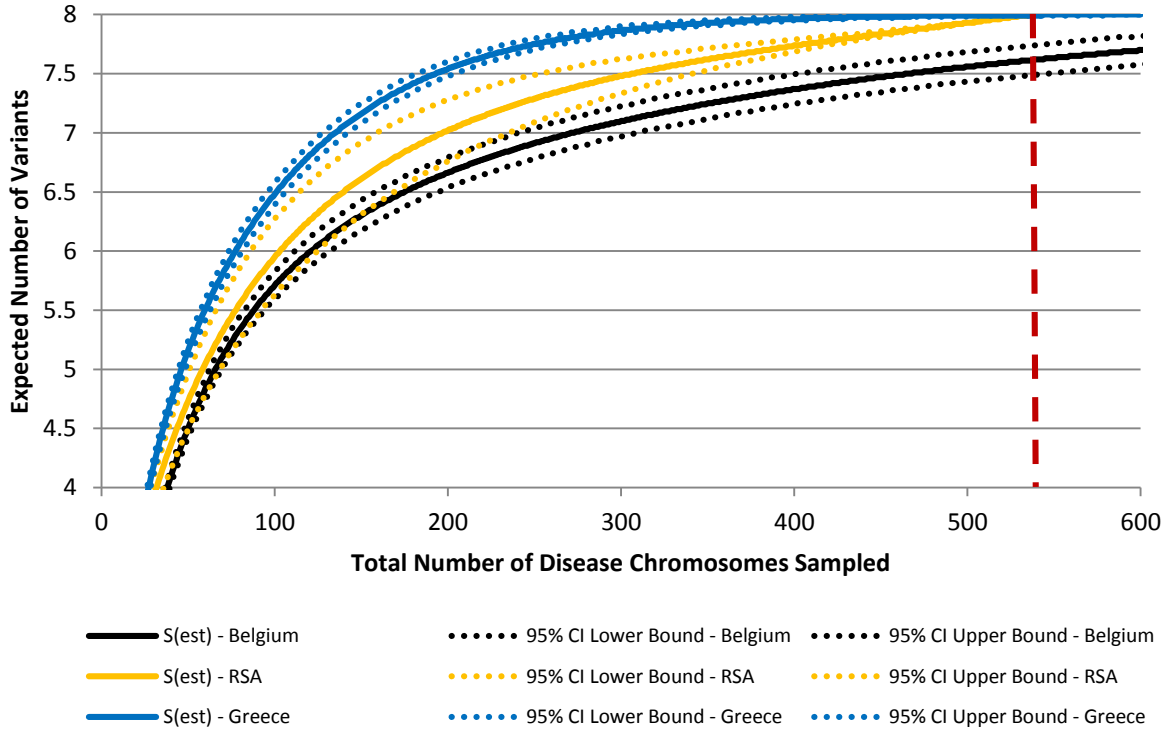
Rarefaction analysis was performed using the variant count values as presented in Table 11, and as shown in Figure 11 (A and B); the point of comparison was determined by the CF population of the RSA, where the total number of disease chromosomes that were sampled was equal to 564. Greece was shown to have the highest diversity at the point of comparison with regard to the eight variants studied, followed by the RSA and Belgium. Sample saturation was achieved for the eight selected variants within the Greek and Belgian CF populations, but was not achieved within the CF population of the RSA. Regardless, variant diversity was significantly different between the Belgian CF population and those of the RSA and Greece.

Figure 11: Rarefaction analysis of Matrix 1

A.



B.



4.3.2.2 Matrix 2

As presented in Table 12, variation was once more observed in the total number of disease chromosomes that were sampled in each of the countries included in Matrix 2. Therefore, in order to perform Shannon and Simpson Diversity analysis, it was necessary to employ a simple random sampling approach on the variant data. The RSA CF population was once more observed as having the lowest number of sampled disease chromosomes (n=564) and was thus used as the point of comparison for Matrix 2. Shannon and Simpson Diversity analysis was subsequently performed on variant data presented in Table 16.

Table 16: Relative values used for Shannon and Simpson Diversity analysis in CF Matrix 2

Country	Total Number Disease Chromosomes in Country	deltaF508	G542X	N1303K	W1282X	R553X	G551D	c.1717-1G>A	R117H	c.2789+5G>A	R1162X	c.3659delC	Q493X
RSA*	564	392	11	4	4	4	7	1	1	1	1	1	1
Australia	564	492	14	4	3	2	28	7	9	1	1	1	0
Belgium	564	458	24	23	6	1	2	12	11	7	4	1	1
UK	564	462	16	9	2	4	20	6	16	4	5	7	1
USA	564	450	24	12	24	12	14	2	6	0	1	4	0

The country marked with an asterix (*) indicates the country with the lowest number of disease chromosomes studied relative to the variants included in the matrix. This is the value used for variant standardisation across all countries included in the matrix.

While the Simpson Diversity index indicated that the diversity of the 12 selected variants ranged between 0.2 and 0.3 across the five sampled countries, the Shannon Diversity index indicated that diversity ranged between 0.5 and 0.8 for the same 12 variants and five countries (Table 17). However, in order to evaluate the full extent to which the diversity of the 12 selected variants differed between the five countries, it was necessary to once more compare results obtained from determining both Shannon and Simpson ESNs.

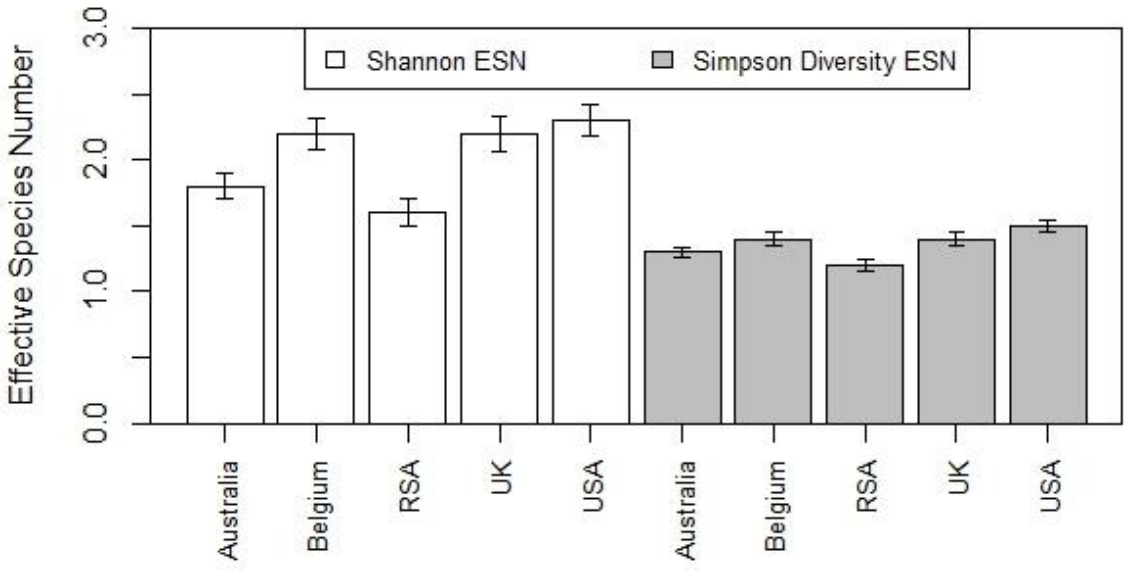
Table 17: Results from Shannon-Weaver and Simpson Diversity analysis of Matrix 2

Country	Shannon index		Simpson Diversity index	
	Index	SE	Index	SE
RSA	0.5	0.05	0.2	0.02
Australia	0.6	0.05	0.2	0.02
Belgium	0.8	0.06	0.3	0.03
UK	0.8	0.06	0.3	0.03
USA	0.8	0.06	0.3	0.03

SE = standard error

Shannon ESN values indicated that sampling approximately two equally distributed CF variants would generate the observed diversity in each of the countries, while Simpson and Simpson Diversity ESN values indicated that sampling between one and two evenly distributed variants in each of the countries was likely to generate the observed diversity results (Figure 12). Using p-values corrected through the Bonferroni method, Shannon ESN indicated significant differences in diversity exist between Australia and the USA ($p=0.014$), the RSA and Belgium ($p=0.001$), the RSA and the UK ($p=0.005$), the RSA and the USA ($p<0.001$), and the USA and the UK ($p<0.001$). Comparing Simpson ESN diversity between the various countries included in Matrix 2 indicated statistically significant differences between the RSA and the USA ($p=0.004$), and the USA and the UK ($p=0.004$) with regard to the 12 variants included in Matrix 2.

Figure 12: Shannon and Simpson Diversity Effective Species Number analysis of Matrix 2



Standard error bars are indicated for each country, with ESN referring to Effective Species Number.

For rarefaction analysis of Matrix 2, it was found that there was a computational limitation to EstimateS in that no more than 5,000 chromosomes, per country, could be analysed at a time. A simple random sample was therefore performed from variant data presented in Table 10 for the four countries in which more than 5,000 variants were analysed (France, Germany, the UK, and the USA). Rarefaction analysis was therefore performed on variant data presented in Table 18.

Table 18: Adjusted Matrix 2 variant data

Country	Total No. Disease Chromosomes in Country	deltaF508	G542X	N1303K	W1282X	R553X	G551D	c.1717-1G>A	R117H	c.2789+5G>A	R1162X	c.3659delC	Q493X
R.S.A.	564	392	11	4	4	4	7	1	1	1	1	1	1
Belgium	1,830	1,520	66	65	17	6	5	33	26	22	15	8	4
UK	5,000	4,068	158	57	25	24	234	51	162	17	10	42	25
USA	5,000	4,133	164	67	145	75	145	31	27	24	24	22	11
Australia	5,000	3,722	80	49	27	20	205	31	94	5	8	7	6

At the point of comparison where the total number of disease chromosomes sampled was equal to 428, the RSA CF population was found to have the highest diversity, followed by the CF populations of Belgium, the UK, the USA, and Australia, respectively (Figure 13). At the point of comparison, it was furthermore found that approximately 10 of the 12 variants would likely be identified in Australia, while approximately 11 of the 12 variants would be identified in Belgium, the UK and the USA. While sample saturation had been achieved for CF populations in Australia, Belgium, the UK and the USA, the variation in diversity was found to differ significantly between CF patients in Australia and those in Belgium, the UK and the USA (Table 19).

Figure 13: Rarefaction analysis of Matrix 2

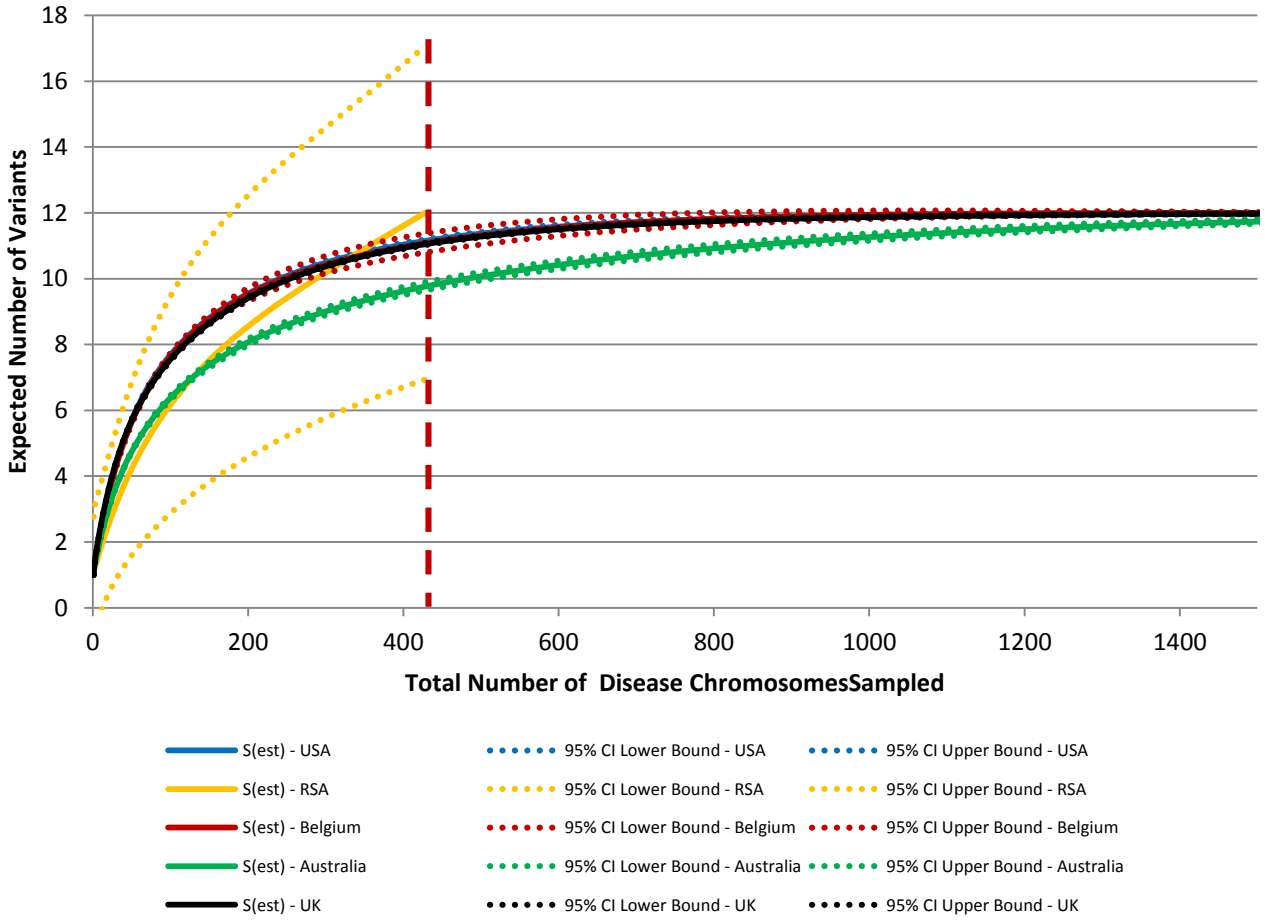


Table 19: Comparison of the significance of variant diversity between countries represented in Matrix 2

	USA	RSA	Belgium	Australia	UK
USA		0	0	1	0
RSA	0		0	0	0
Belgium	0	0		1	0
Australia	1	0	1		1
UK	0	0	0	1	

A significant difference in variant diversity is indicated by "1", while a "0" indicates that no significant difference exists between variant diversity results in the respective country/countries.

4.3.2.3 Matrix 3

As shown in Table 13, variation was observed in the total number of disease chromosomes studied per country included into Matrix 3. Therefore, as was performed for Matrix 1 and Matrix 2, a simple random sample was drawn from Table 10 in order to generate a standardised matrix for Shannon and Simpson Diversity analysis. Since Belarus had sampled the lowest number of disease chromosomes of the 15 countries included in Matrix 3 (n=278), this value was used as the point on which variant data was standardised. Table 20 thus indicates the randomly sampled variant data included in Matrix 3 for Shannon and Simpson Diversity analysis. However, since Simpson and Simpson Diversity index results are derivatives of each other, Simpson index results were not shown despite being determined.

Table 20: Relative values used for Shannon and Simpson Diversity analysis in CF Matrix 3

Country	Total Number Disease Chromosomes in Country	deltaF508	G542X	N1303K	W1282X	R553X
Belarus*	278	170	13	9	3	1
Australia	278	242	6	4	3	3
Belgium	278	240	10	6	3	2
Brazil	278	222	31	11	2	2
France	278	252	6	4	5	5
Germany	278	247	4	6	2	7
Hungary	278	242	6	6	7	7
Italy	278	234	19	8	4	6
The Netherlands	278	241	5	5	2	5
Poland	278	234	12	11	4	6
RSA	278	197	7	3	1	0
Slovakia	278	218	24	13	5	18
Switzerland	278	200	13	8	6	44
UK	278	227	10	3	1	1
USA	278	231	11	6	6	3

The country marked with an asterix (*) indicates the country with the lowest number of disease chromosomes studied relative to the variants included in the matrix. This is the value used for variant standardisation across all countries included in the matrix.

As shown in Table 21, Shannon index results obtained from analysis of Matrix 3 ranged between 0.9 in Swiss CF populations, and 0.3 in CF patients from Australia, Germany, the Netherlands, the RSA, and the UK. Simpson Diversity index results ranged between 0.4 in CF patients from Switzerland and Slovakia, and 0.1 in CF patients from Australia, France, Germany, the Netherlands, the RSA, and the UK.

Table 21: Results from Shannon-Weaver and Simpson Diversity analysis of Matrix 3

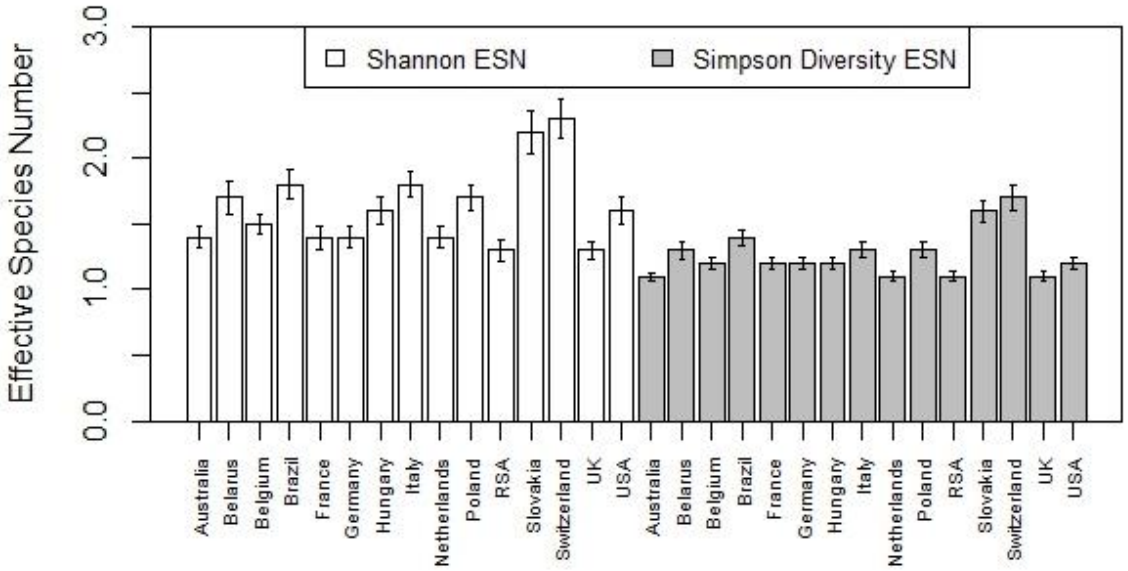
Country	Shannon index		Simpson Diversity index	
	Index	SE	Index	SE
Belarus	0.5	0.07	0.2	0.04
Australia	0.3	0.07	0.1	0.03
Belgium	0.4	0.06	0.2	0.03
Brazil	0.6	0.06	0.3	0.03
France	0.4	0.06	0.1	0.03
Germany	0.3	0.06	0.1	0.03
Hungary	0.5	0.06	0.2	0.03
Italy	0.6	0.06	0.2	0.03
The Netherlands	0.3	0.06	0.1	0.03
Poland	0.5	0.07	0.2	0.03
RSA	0.3	0.06	0.1	0.03
Slovakia	0.8	0.07	0.4	0.04
Switzerland	0.9	0.06	0.4	0.04
UK	0.3	0.06	0.1	0.02
USA	0.5	0.07	0.2	0.03

SE = standard error

However, in order to evaluate the full extent to which the diversity of the five selected variants differed between the 15 countries, it was once more necessary to view results in lieu of respective Shannon and Simpson ESN values. Thus, as shown in Figure 14, by sampling between one and two equally distributed variants, the Shannon and Simpson ESN calculated for each of the 15 countries would be sufficient to ascertain the diversity determined for the Shannon or the Simpson Diversity indices. Shannon ESN p-values, corrected for through the Bonferroni method, revealed significant differences between the five variants compared for CF patients in the following countries: Slovakia and Australia ($p=0.001$), Switzerland and Belgium ($p<0.001$), Slovakia and France ($p=0.002$), Switzerland and France ($p<0.001$), Slovakia and Germany ($p=0.001$), Switzerland and Germany ($p<0.001$), the UK and Italy ($p=0.005$), Slovakia and the Netherlands ($p=0.001$), Switzerland and the Netherlands ($p<0.001$), Slovakia

and the RSA (p<0.001), the UK and Slovakia (p<0.001), and the UK and Switzerland (p<0.001). In contrast, Simpson ESN p-values, corrected by the Bonferroni method, yielded statistically significant results between only Slovakia and Australia (p=0.04), and Switzerland and Australia (p=0.045).

Figure 14: Shannon, Simpson and Simpson Diversity ESN analysis of Matrix 3



Histograms in white indicate Shannon ESNs, while histograms in grey indicate Simpson Diversity ESNs. Standard error bars are indicated.

For the purposes of rarefaction analysis and due to computational limitations of EstimateS (v.9.1.0), it was once more required that variant data be re-evaluated for those countries in which more than 5,000 disease chromosomes were sampled within the studied country. Thus, as observed in Table 22, variant data was adjusted for four countries (the UK, the USA, Germany, and France) through the use of simple random sampling from Table 10. Random selection was governed to a maximum of 5,000 disease chromosomes.

Table 22: Adjusted Matrix 3 variant data

Country	Total No. Disease Chromosomes in each Country	deltaF508	G542X	N1303K	W1282X	R553X
Belarus	278	170	13	9	3	1
Australia	4,266	3,722	80	49	27	20
Belgium	1,830	1,520	66	65	17	6
Brazil	820	391	59	20	11	6
France	5,000	4,506	182	56	48	60
Germany	5,000	4,395	117	111	39	121
Hungary	1,133	622	19	15	20	24
Italy	2,608	2,176	153	103	38	44
The Netherlands	2,463	2,161	52	32	15	35
Poland	4,046	2,310	105	73	28	77
RSA	564	392	11	4	4	4
Slovakia	908	520	62	31	12	36
Switzerland	1,268	725	33	15	14	178
UK	5,000	4,068	158	57	25	24
USA	5,000	4,133	164	67	145	75

Rarefaction analysis was subsequently performed on variant data presented in Table 22, with the point of comparison being established by the CF population of Belarus (total number of disease chromosomes sampled = 278). It was found that, at the point of comparison and with regard to the five studied variants, CF patients in Slovakia had the highest diversity, followed by Hungary and Brazil (Figure 15). CF patients in Australia were shown to have the lowest diversity, followed by those patients present in the UK and Belgium. Sample saturation was achieved in all 15 countries and variant diversity was shown to differ significantly between several countries (Table 23).

Figure 15: Rarefaction analysis of adjusted Matrix 3 variant data

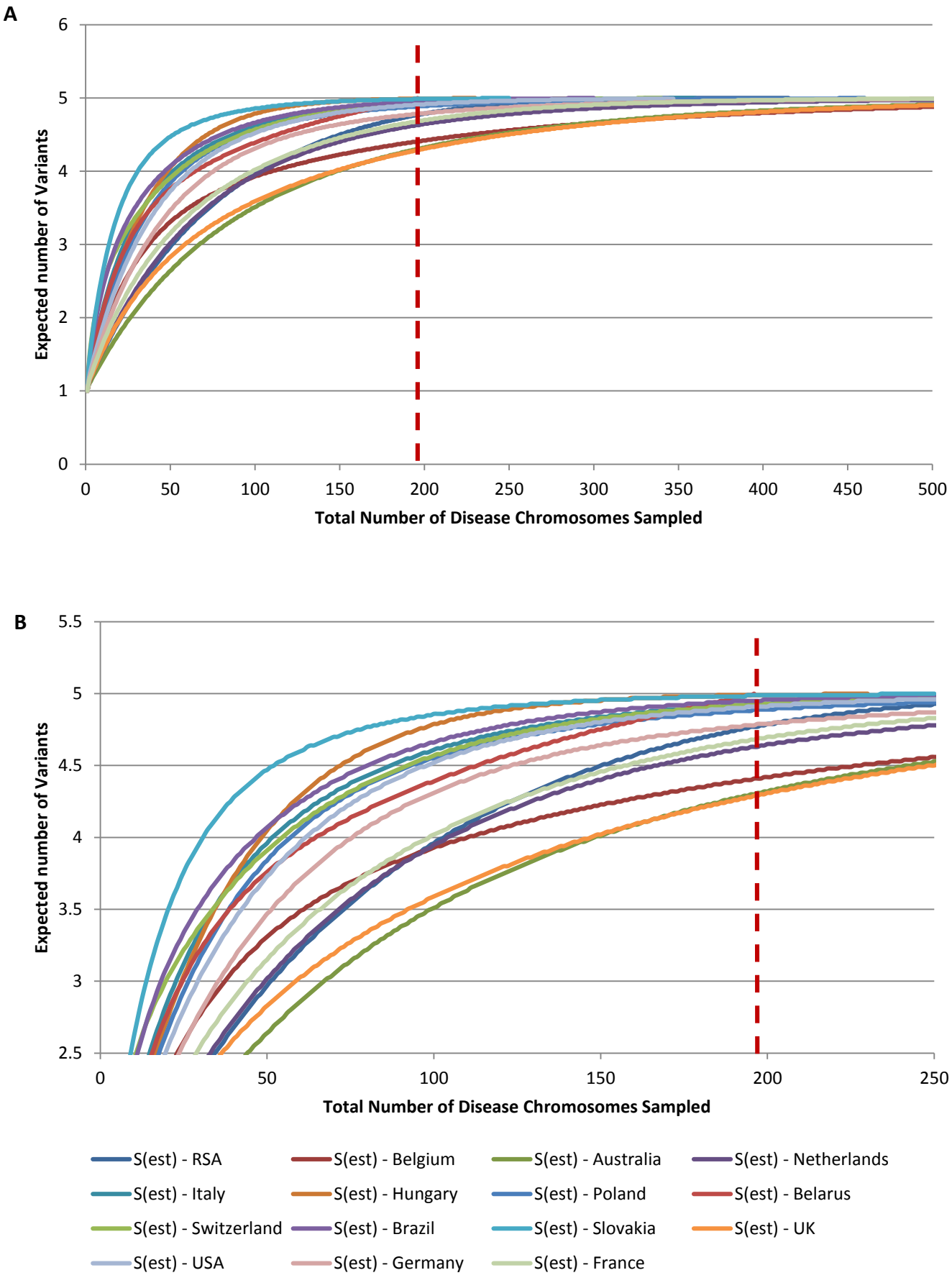


Table 23: Comparison of the significance of variant diversity between countries represented in Matrix 3

	RSA	Belgium	Australia	The Netherlands	Italy	Hungary	Poland	Belarus	Switzerland	Brazil	Slovakia	UK	USA	Germany	France
RSA		0	1	0	0	0	0	0	0	0	0	0	0	0	0
Belgium	0		0	0	1	1	1	0	1	1	1	0	1	1	0
Australia	1	0		1	1	1	1	0	1	1	1	1	1	1	1
The Netherlands	0	0	1		1	1	1	0	1	1	1	1	1	0	0
Italy	0	1	1	1		0	0	0	0	0	0	1	0	1	1
Hungary	0	1	1	1	0		0	0	0	0	0	1	0	1	1
Poland	0	1	1	1	0	0		0	0	0	0	1	0	1	1
Belarus	0	0	0	0	0	0	0		0	0	0	0	0	0	0
Switzerland	0	1	1	1	0	0	0	0		0	0	1	0	1	0
Brazil	0	1	1	1	0	0	0	0	0		0	1	0	1	1
Slovakia	0	1	1	1	0	0	0	0	0	0		1	0	1	1
UK	0	0	0	1	1	1	1	0	1	1	1		1	1	1
USA	0	1	1	1	0	0	0	0	0	0	0	1		1	1
Germany	0	1	1	0	1	1	0	0	1	1	1	1	1		0
France	0	0	1	0	1	1	1	0	0	1	1	1	1	0	

A significant difference in variant diversity is indicated by “1”, while a “0” indicates that no significant difference exists between variant diversity results in the respective country/countries.

The CF population in Australia was found to differ significantly from the most number of countries (12), followed by the UK (11) and Germany (10). Significant differences in variant diversity were found to occur nine times when comparing the Netherlands with other countries, and eight times when Belgium and France were compared with other countries. Six significant differences in variant diversity were identified when Italy, Hungary, Poland, Brazil, Slovakia, and the USA were compared to all of the other countries. Five significant differences in variant diversity were identified when comparing the Swiss CF population to others. The CF population of the RSA differed significantly from only the Australian CF population. No significant difference in CF diversity was found to exist when comparing to CF patients of Belarus.

Comparatively, Shannon ESN results yielded significance outcomes for those countries found to be significantly diverse from each other through rarefaction analysis, with the exception of three comparisons. These were Switzerland and France, Slovakia and the RSA, and Switzerland and the RSA. Shannon ESN comparison indicated that diversity between these three groups was statistically significant, while rarefaction did not yield statistically significant results for the same three groups.

4.4 Discussion

4.4.1 Matrix 1

Diversity results obtained from Matrix 1 indicates that Shannon and Simpson Diversity analysis is only useful if viewed in the light of their respective ESN values. This has been well

studied and described by Jost (6-8). However, when comparing rarefaction analysis to Shannon ESN and Simpson Diversity ESN analysis, it was shown that although all three methods indicated the level of diversity across each of the three countries included in this subset, the rarefaction method was the only method that described how well each region was sampled with regards to the eight variants included in this subset. Countries in which the rarefaction curve reached a plateau indicated that sampling within the corresponding CF populations had reached saturation with respect to the eight selected variants. This, as shown by Hurlbert (12), Kalinowski (13), and Colwell *et al.* (14, 15), is one of the benefits of rarefaction analysis. Unlike Shannon, Simpson and Simpson Diversity analysis, rarefaction was furthermore found to not be influenced by variations in sample size and accounted for such variations through the point of comparison (Figure 11).

Rarefaction analysis ultimately indicated that significant differences between the diversity of the variants in the studied countries was not dependent on the assumption of equal sampling effort. Thus, through the use of a point of comparison, the rarefaction method eliminates the necessity of using simple random sampling methods, as was required for Shannon and Simpson diversity measures. However, unlike rarefaction, methods exist through which correction can be made for multiple comparisons when using Shannon and Simpson ESN diversity methods. This is despite the fact that rarefaction methods incorporate unbiased estimators of variance in its determination of the 95% confidence intervals (14).

4.4.2 Matrix 2

As was observed in Matrix 1, Shannon and Simpson Diversity index results for Matrix 2 were informative only in the light of their corresponding ESNs. However, in order to compare the diversity of the 12 variants included in Matrix 2, it was necessary to employ a simple random sampling method to standardise variant information across the various countries so as to utilise the Shannon and Simpson diversity methods. This is, in essence, what rarefaction achieves through the point of comparison.

The rarefaction method indicated that the relatively high level of diversity observed for CF patients in the RSA was not significantly different from the levels of diversity observed in the remaining four countries. Conversely, Australia had 1) the lowest diversity values with regard to the 12 selected variants, and 2) diversity values that differed significantly from diversity data obtained from Belgium, the UK, as well as the USA. This implies that high levels of variant diversity does not guarantee that significant differences exist in diversity across studied countries, regions or populations; nor does low levels of diversity suggest that significant differences between studied countries, regions or populations do not exist. By contrast, Shannon ESN diversity comparisons, after correction, indicated that the greatest level of significant differences in diversity was observed between CF populations of the RSA. It must be considered though that both Shannon and Simpson ESN diversity comparisons are corrected for multiple comparisons, while that obtained from rarefaction is not. This may be argued to be a significant short-coming of the rarefaction method in comparison to the Shannon and Simpson diversity methods.

However, the lack of a plateau for the CF population of RSA indicates that sampling saturation has not been reached when investigating the 12 selected variants. Although a plateau was not reached, this once more indicates a distinct advantage that the rarefaction method has over the use of either the Shannon or the Simpson diversity methods when considering sampling effort. Nevertheless, in the context of Matrix 2, the lack of sample saturation could be indicative of South Africa's rich history and known genetic admixture. Regardless, further sampling would be required in order to determine the validity of the statement concerning admixture in the South African population.

4.4.3 Matrix 3

Studying rarefaction output once more indicated that variation in sample size did not influence the comparative capacity of the rarefaction method, although it is recognised that this method does not afford ways through which correction can be made for multiple comparison. Nevertheless, it was found that significance in variant diversity results for Matrix 3 followed the same general trend as that observed for Matrix 2 – high levels of variant diversity did not guarantee that significant differences in diversity would exist across studied countries, regions or populations. Conversely, low levels of variant diversity did not imply that significant differences between studied countries, regions or populations, did not exist.

Sample saturation was achieved for all countries included into Matrix 3 for analysis, and affirms this distinct advantage that the rarefaction method has over the Shannon and Simpson Diversity methods. However, the variability of significant differences observed

across the 15 countries with regard to the five variants could potentially be associated with the lack of correction for multiple comparison when using the rarefaction method as well as the strictness associated with Bonferroni correction in the instance of Shannon and Simpson ESN diversity comparisons.

4.5 Conclusions

Since diversity analysis was limited to published CF variant data, it must be considered that the diversity results obtained may not necessarily be a true reflection of the current state of diversity in any one country. However, in regions where CF registries are in place and make it compulsory for all individuals diagnosed with CF to be registered with the organisation(s) involved, as is observed in the UK (137), diversity results are likely to represent CF variant diversity more precisely than in regions/countries were such regulations do not exist or are not enforced.

Moreover, as indicated by Mutesa and Bours (95), Masekela *et al.* (99), and Westwood *et al.* (20), CF has historically not been associated with non-Caucasoid population groups in South Africa. This is particularly prominent in South African CF populations and is reflected through the routine use of different CF screening panels by the NHLS in South African CF patients of different ethnicities (141). Therefore, not only are literature reports for CF variants predominantly focused on those identified in Caucasoid patients, but literature reports of CF variants in non-Caucasoid populations are largely absent in South Africa (99). This could account for the lack of statistically significant differences observed when comparing variant

data generated from South African CF populations to those generated from non-South African CF populations. However, it must also be considered that South African populations generally have high levels of genetic diversity (59, 125, 126). This could be a contributing factor to the lack of sample saturation achieved for each of the three rarefaction graphs generated from CF data pertaining to South African CF patients. However, further sampling from non-Caucasoid CF patients would be required in order to increase confidence in this statement.

Although the three diversity methods determine diversity in their own unique way, only the rarefaction method is capable of illustrating the point of sampling saturation. This feature makes rarefaction analysis a useful and powerful tool in a medical context, as it provides a manner through which sampling efforts can be evaluated. Despite the fact that the rarefaction method determines variance and subsequent confidence intervals through use of unbiased estimators of variance (14, 15, 65), a potential pitfall of rarefaction is that methods correcting for multiple comparisons has not been described. Nevertheless, the restriction in access to CF variant data as a consequence of selective variant screening in different ethnic groups, as well as limited numbers of published reports of CF variants in CF patients truly reflective of the South African population, is largely the reason why comparisons have not been made with regard to the diversity of CF-variants in different population groups/different regions in South Africa. There is simply not enough information available at present in order to do so.

Chapter 5 – A global perspective of MLD: A review of incidence, prevalence and mutation data

5.1 Introduction

The arylsulfatase A (*ARSA*) gene was described in 1990 by Kreysing *et al*. and is implicated in the development of the panethnic, autosomal recessive disorder known as metachromatic leukodystrophy (MLD) (82). Found at position q13.33 on chromosome 22, the *ARSA* gene is 32kb long and encompasses eight exonic regions (142). The *ARSA* gene produces three mRNA fragments of 2.1kb, 3.7kb, and 4.8kb respectively. These three mRNA fragments, once translated, produce proteins that bind together to form the lysosomal enzyme known as arylsulfatase A (ASA) which is responsible for the hydrolysis of the polar glycolipid, cerebroside sulfate. As shown in Figure 16, cerebroside sulfate is a main constituent of myelin and binds to the lysosomal protein saposin B (SAPB), which is encoded for by the precursor prosaposin (*PSAP*) gene (143-146). The binding of SAPB to cerebroside sulfatase is the signal from which ASA hydrolyses the cerebroside sulfatase lipid.

MLD primarily results from mutations in the *ARSA* gene, but may also manifest due to mutations in the *PSAP* gene (147, 148). MLD variants in the *ARSA* gene prevent ASA synthesis, thus resulting in the accumulation of cerebroside sulfate. The accumulation of cerebroside sulfate ultimately causes the fatal process of central and peripheral nervous system demyelation, particularly affecting the select visceral organs and the white matter of the brain (81, 82, 149-151). In instances where deleterious mutations occur in the *PSAP* gene, the

concentration of SAPB present within cells is decreased and ASA is unable to effectively hydrolyse cerebroside sulfate. As a consequence, nervous system demyelation will occur. A third and final state of MLD exists and is known as the pseudodeficient state. Pseudodeficient MLD patients are considered phenotypically normal, but have decreased concentrations of ASA (147, 152).

Figure 16: Schematic representation of cerebroside sulfate hydrolysis

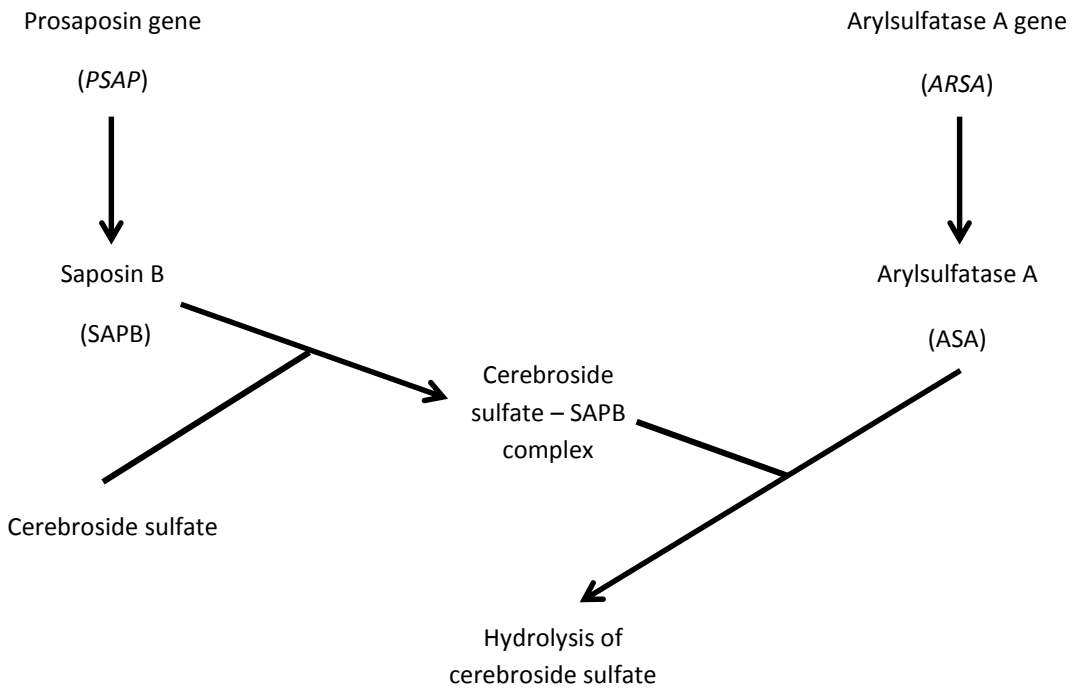


Figure 16 shows the process by which cerebroside sulfate is regulated by Saposin B (SAPB) and arylsulfatase A (ASA). SAPB is produced by the precursor prosaposin (*PSAP*) gene while arylsulfatase A (ASA) is produced by the arylsulfatase A (*ARSA*) gene. SAPB will bind to the polar glycolipid cerebroside sulfate. This signals ASA to bind to the cerebroside sulfate-SAPB complex. Attachment of ASA to the cerebroside sulfate-SAPB complex results in the down regulation and hydrolysis of cerebroside sulfate.

Four forms of MLD are known to exist and include the late-infantile, early juvenile, late juvenile and adult forms (147, 149, 152). Typically, the severity of MLD decreases as patient age increases. Thus, the late-infantile form of MLD is the most severe form, while the adult form is the least severe. MLD is however fatal, irrespective of which form patients have. Patients with the less severe adult form of MLD live for longer periods of time, while patients with infantile form MLD typically die before reaching ten years of age. A diagnosis of late-infantile form MLD is usually made between birth and two years of age, while early and late juvenile forms MLD are respectively diagnosed in children between four and six, and eight and 16 years of age. A diagnosis of adult form MLD is usually made in patients over 16 years of age (25-27, 81, 153).

Phenotypic expression of MLD across the four different forms is similar in many regards, with many of the same symptoms being expressed in each of the different forms of MLD. Nevertheless, differences in the physical manifestations of the four forms of MLD do exist. The following symptoms have been described to occur in late-infantile patients and usually begin manifesting at the age of one: deterioration of mental/cognitive and motor skills, dementia, impaired swallowing ability, speech abnormalities, muscle cramps/rigidity and hypertonia, convulsions/seizures, partial or complete paralysis, and blindness (25-27, 147, 149, 153, 154). Unfortunately, children who are diagnosed with the late-infantile form of MLD typically do not live past the age of 10 (153, 155, 156).

Early and late juvenile MLD forms share the same symptom manifestations as late-infantile form MLD. However, the clinical presentation of juvenile forms of MLD may manifest at a slower rate than that of infantile form MLD which are additionally associated with ataxia and impaired school performance. An additional difference between infantile and juvenile MLD symptom development is the fact that juvenile MLD patients will start showing deterioration in either motor skills or cognitive ability – simultaneous deterioration of both cognitive and motor skills is an exceptionally rare occurrence. Symptoms of either of the juvenile forms usually develop in children between the ages of four and 14 (25-27, 147, 153, 154).

By comparison, the physical manifestations of MLD in adult patients differ from the infantile form MLD in that the following symptoms are additionally observed in adult MLD patients: severe personality changes – including anxiety, bewilderment, decreased ability to make good judgements, disorganization and a decreased level of alertness, depression and addictive tendencies towards alcohol and drugs, the presence of an enlarged colon, complete paralysis/quadriplegia, and physical disturbances and tremors. Adult form MLD most commonly manifests in patients that are 16 years of age or older (25-27, 147, 149, 153, 154).

MLD can be diagnosed through five different methods which include prenatal testing, MRI and/or CT scans, electrophysiological studies, post-mortem testing, ASA concentration determination and genetic screening. Prenatal testing includes determining ASA concentrations present in the foetus and can commonly be achieved through amniocentesis

or chorionic villus sampling. Although an uncommon method to use, foetal blood foetoscopy is another way through which prenatal testing can be conducted.

As it is known that decreased concentrations/absence of ASA levels lead to the gradual demyelation of both the central and peripheral nervous systems. MRI and CT scans are therefore performed in order to determine the extent to which demyelation has occurred in patients. However, due to the large number of neurodegenerative disorders that will show demyelation of the central nervous system, it is advised that MRI and CT scan results for MLD patients be assessed in conjunction with either ASA concentration results and/or genetic screening results. This same statement holds true when using electrophysiological methods to diagnose MLD. The electrophysiological method indicates nerve conductivity velocity with slow nerve reactions potentially indicative of MLD. Post-mortem analysis can also be performed and includes pathological studies through white matter staining and liver tissue ASA determination (25, 81, 153, 157).

The two preferred methods through which MLD can be diagnosed include ASA concentration determination and mutation/genetic screening tests. ASA concentrations can be determined from various sources, most commonly from leukocyte and urine samples, but also from placental material, fibroblasts and serum samples. It is important to realise that although reference values exist for ASA concentrations, results indicative of MLD will be dependent on the method and tissue/sample type used to determine the ASA levels. Nevertheless, it is known that ASA concentrations will always be incredibly low or completely absent in MLD

patients when compared to reference values, irrespective of which method is applied to the samples (158-160).

It is also important to note that in instances where MLD exists as a result of SAPB deficiency, test results will indicate normal ASA levels, but increased concentration of SAPB. In patients who present with the pseudodeficiency state, test results will indicate decreased concentrations of ASA, but will possess enough residual activity in order to prevent nervous system degeneration. In order to differentiate between true and pseudodeficiency MLD states, genetic testing must be performed. As is frequently the case, genetic testing also decreases the probability that MLD patients will not be misdiagnosed with cerebral palsy, Batton's disease, attention deficit hyperactivity disorder, multiple sclerosis, or schizophrenia. Genetic testing is therefore the best tool currently available through which MLD can be accurately diagnosed (25, 81, 153, 157, 161).

Although only one known South African case of MLD has been reported to date (21), the incidence of this disorder ranges significantly in many other countries. The reported incidence of MLD in Caucasian patients of European descent ranges between 1/40 000 and 1/100 000, but an incidence as high as 1/75 live births has been reported for Habbanite Jews while an incidence of 1/2 500 is known to exist in Navajo Indian and Alaskan populations (81, 149, 153, 156, 162-165).

Interestingly, over 150 causative MLD variants had been described for the *ARSA* gene by 2010. Of these, three variants, namely c.459+1G>A, P426L and I179S, represent 40-65% of all causative mutations found in MLD patients of Caucasian origin, while an additional six variants contributed to a further 10% of all positively identified variants. (23, 24, 139, 150, 155, 166). Interestingly, the two polymorphisms associated with the pseudodeficient MLD state, c.*96A>G and c.1049A>G, reportedly occur at frequencies of between 10-20% within European populations of Caucasian descent (22, 147, 167).

Despite the significant amount of research conducted on this disorder, an exhaustive report on the significance of using mutation/variant diversity indices been explored. The research conducted was therefore aimed, firstly at updating the global view of mutation/variant frequency of MLD, and secondly at elucidating MLD variant diversity between various populations.

5.2 Materials and Methods

5.2.1 Data assimilation

Data assimilation was performed through the use of key-word, web-based searches utilising Google Scholar and Pubmed. Various combinations of the terms “metachromatic leukodystrophy”, “mutation(s)”, “mutation frequency”, “South Africa”, “Southern Africa”, “Africa”, “Europe”, “North America”, “South America”, “Australasia”, and “Asia” were used in

order to obtain relevant, journal publications. The terms “mutation” and “mutation frequency” were also substituted with the terms “variant” and “variant frequency”, in order to ensure that the largest possible number of MLD publications were obtained. References in these journal publications were also used to obtain any relevant documents that could supply additional information.

Relevant information pertaining to variants and variant frequencies was collected and entered into a customised database in Microsoft Excel®. The data captured included 1) the first author and year of the publication from which information was extracted, 2) the form of MLD the patients were diagnosed with, 3) the total number of patients/individuals studied according to each report, 4) the total number of patients/ individuals who underwent genetic screening, 5) the causative variants identified in the studied patients, and 6) the frequencies with which the causative variants were identified within the studied MLD population(s).

5.2.2 Diversity analysis

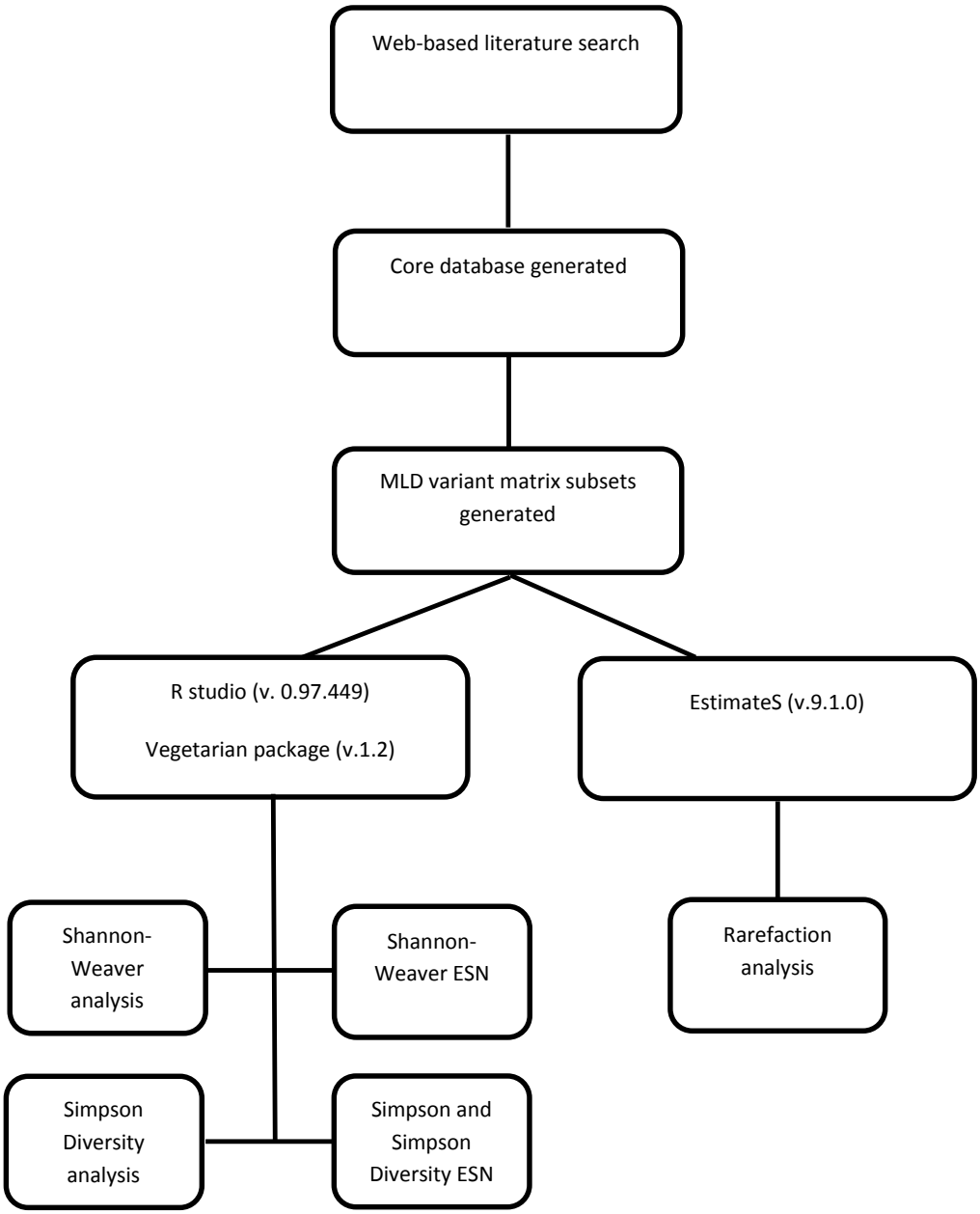
Variant data was obtained through key-word searches for journal publications and databased according to continental region and the total number of patients sampled. Distinctions were made between variants that were screened for but not identified in patients and variants that were not screened for at all, where the former were assigned zero values and the latter dashed values, respectively. Secondary data matrices, on which diversity analysis was

performed, were constructed according to synonymous variant data – i.e. only variant data common to each of the respective countries was included for diversity analysis.

Genetic diversity was determined through the use of three diversity methods, namely, the Shannon-Weaver, Simpson Diversity and rarefaction methods. As outlined in Figure 17, Shannon-Weaver and Simpson Diversity analysis was performed through use of the Vegan package (v.1.2)(127) operating from the R Studio platform (v.0.97.449)(128).

Effective species numbers (ESNs) for each of these indices were additionally determined using the Vegetarian package (v.1.2). Shannon-Weaver and Simpson Diversity indices were determined through methods described by Jost (6, 7), while effective species numbers (ESNs)/Hill ratios were determined through methods described by Hill (9) and Jost (6, 7). Standard error (SE) values were computed through the Vegan package (v.1.2) for both the Shannon-Weaver and Simpson Diversity indices as well as each of the respective ESNs through bootstrap methods developed by Chao *et al.* (129). Individual-based rarefaction analysis and 95% confidence intervals (CIs) were determined through methods described by Colwell *et al.* (14) using EstimateS (v.9.1.0) (64). As described by Colwell *et al.* (14, 64), sample order was not randomized and only one run was performed for each variant matrix.

Figure 17: Algorithm used to determine MLD variant diversity



5.3 Results

5.3.1 Variant comparison

Upon studying literature pertaining to MLD variants, it was found that over 150 variants are known to be associated with MLD (23). However, although a large number of causative variants have been described, only a few are mentioned relatively frequently in literature. In total, the literature survey elucidated 18 countries for which variant frequency data was available for commonly observed variants. Of the 18 countries, 14 were European (The Netherlands, Poland, Italy, Austria, Belgium, Czech and Slovak Republic, France, Germany, Great Britain, Greece, Portugal, Spain, Sweden, and Switzerland), while the remaining four countries were situated in either the Middle Eastern (Turkey, Ukraine and Israel) or Asian territories (India). Nevertheless, variant data of five variants (c.1204+1G>A, c.459+1G>A, D225H, I179S, P426L, c.763A>G) and two pseudodeficiency alleles (c.1049A>G, c.*96A>G) was compared across all 18 countries.

As shown in Table 24, at least 100 patients were examined in Germany and Great Britain; between 40 and 90 patients were studied in India, Poland and The Netherlands; while less than 10 patients were studied in Switzerland and Turkey. Furthermore, the most common variants in each country were c.459+1G>A and P426L. Each of these variants was reported in 13 of the 18 countries. D225H was the least commonly described variant and was reported only once (USA). The c.1204+1G>A variant was also only reported once and was described in Polish MLD patients. Portugal had the highest number of patients harbouring the c.459+1G>A

variant (58.8%). In total, these seven variants were observed in 35.7% of the 691 MLD patients from all 18 countries. Two variants, namely c.459+1G>A and P426L, were found in 16.34% and 11.56% of the 691 MLD patients, respectively. The five remaining variants were found in between 0.14% and 3.04% of all 691 MLD patients described in the literature. The pseudodeficiency variant, c.1049A>G, was found with the highest frequency in Great Britain, occurring in almost 8% of the tested patients.

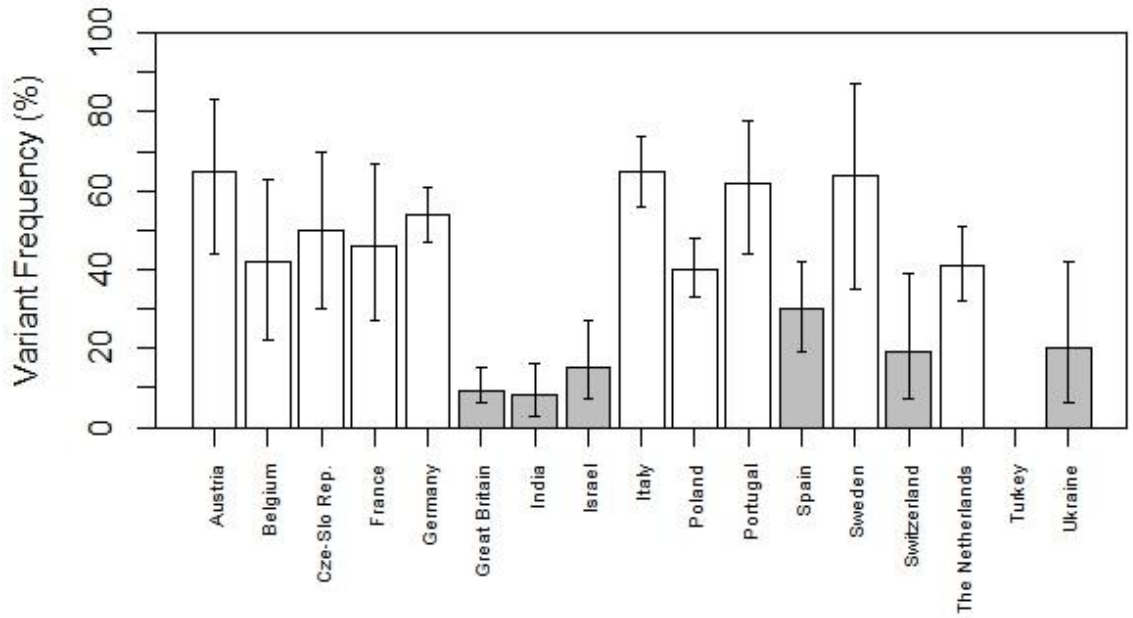
Table 24: Frequency of the most commonly reported allelic variants of MLD in 18 countries

Country	Number of patients	Number of chromosomes	c.459+1G>A	P426L	c.1049A>G	c.*96A>G	c.1204+1G>A	I179S	c.763+1G>A	Reference
Austria	13	26	6 (23.1%)	11 (42.3%)	-	-	-	-	-	(152)
Belgium	12	24	9 (37.5%)	1 (4.2%)	-	-	-	-	-	(152)
Czech & Slovak Republic	13	26	10 (38.5%)	3 (11.5%)	-	-	-	-	-	(152)
France	21	42	6 (14.3%)	6 (14.3%)	-	-	-	-	-	(152)
Germany	102	204	49 (24.0%)	61 (29.9%)	-	-	-	-	-	(152)
Great Britain	18	36	18 (50%)	3 (8.3%)	-	-	-	-	-	(152)
Great Britain	77	154	-	-	27 (17.5%)	20 (13.0%)	-	-	-	(153)
Greece	7	14	0	0	-	-	-	-	-	(152)
India	20	40	3 (7.5%)	0	6 (15.0%)	7 (17.5%)	0	0	3 (7.5%)	(164)
Israel	27	54	8 (14.8%)	-	-	-	-	-	-	(168)
Italy	23	46	12 (26.1%)	1 (2.2%)	1 (2.2%)	3 (6.5%)	0	0	3 (6.5%)	(169-171)
Italy	32	64	23 (35.9%)	1 (1.6%)	-	-	-	-	-	(152)
Poland	49	98	16 (16.3%)	16 (16.3%)	2 (2.0%)	0	5 (5.1%)	12 (12.2%)	0	(23)
Poland	38	76	16 (21.5%)	13 (17.1%)	-	-	-	-	-	(152)
Portugal	17	34	20 (58.8%)	1 (2.9%)	-	-	-	-	-	(152)
Spain	32	64	15 (23.4%)	4 (6.3%)	-	-	-	-	-	(152)
Sweden	13	26	7 (26.9%)	2 (7.7%)	-	-	-	-	-	(152)
Switzerland	7	14	5 (35.7%)	0	-	-	-	-	-	(152)
The Netherlands	3	6	0	0	1 (16.7%)	1 (16.7%)	0	0	1 (16.7%)	(172)
The Netherlands	49	98	8 (8.2%)	34 (34.7%)	-	-	-	-	-	(152)
Turkey	5	10	0	0	1 (10.0%)	0	0	0	3 (30.0%)	(155)
Ukraine	10	20	2 (10%)	2 (10%)	-	-	-	-	-	(152)

The total number of times each of the variants was identified in the MLD patients of each country is indicated, while the percentage frequency occurrence of each variant in each country is indicated in brackets. Variant information that was not provided is indicated by a dash, while a "0" indicates that the particular variant was screened for, but not identified within the studied country.

Three variants, c.459+1G>A, P426L, and I179S, are reported to occur at frequencies between 40-65% in MLD patients of European Caucasian descent. Thus, studying collated data presented in Table 24 and Figure 18, it was found that four European countries, namely Great Britain, Spain, Switzerland, and Ukraine, had frequencies that were outside of this expected range. A further two countries (India and Israel) did not fall within this range, while not one of these three variants were identified in a further country (Turkey). Interestingly, in Great Britain where the largest number of MLD patients had been tested (172 in total, Table 24), the c.459+1G>A, P426L and I179S variants collectively accounted for approximately 5% of all observed variants. Furthermore, only one country (Poland) harboured the I179S variant in its population.

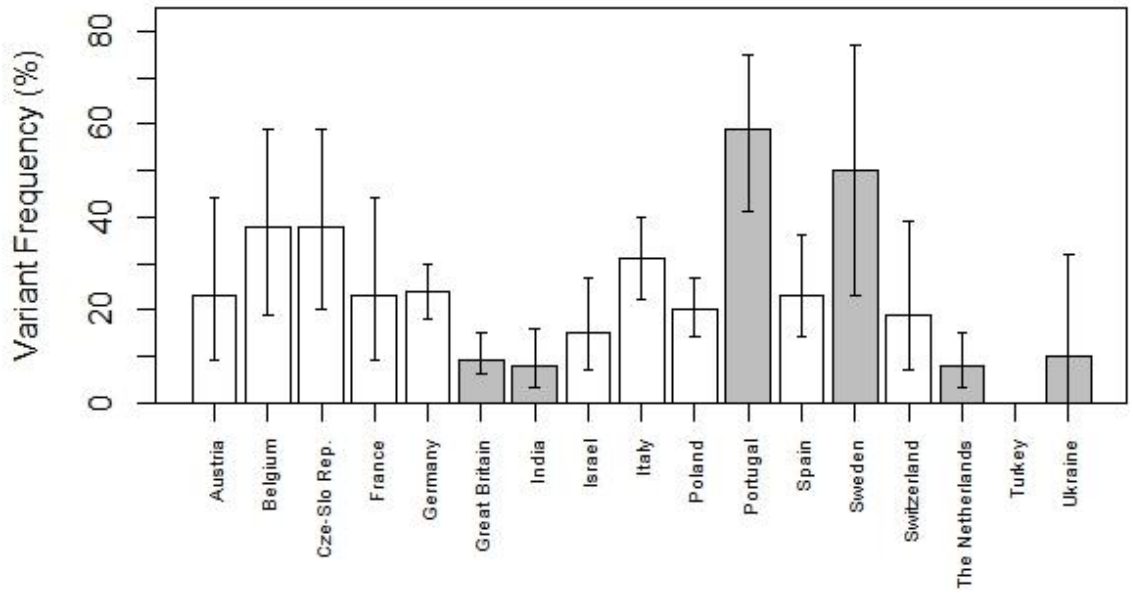
Figure 18: Combined frequency of the c.459+1G>A, P426L, and I179S variants



White-shaded bars indicate populations where the joint frequency of the common variants (c.459+1G>A, P426L and I179S) are within the expected range of 40-65%, while the grey-shaded bars indicate countries where the observed frequency of these variants were outside of the expected threshold of 40-65%. In all cases, 95% confidence intervals are indicated. The bar labelled Cze-Slo Rep indicates frequency value for the Czech and Slovak Republics.

When studying collated data presented in Table 24, and As shown in Figure 19, patients from only nine European countries harboured frequencies ranging between 15-43% for the c.459+1G>A variant. While this variant was not identified in Turkish patients diagnosed with MLD, the frequency of this variant was either less than 15% or greater than 43% in five European nations (Great Britain, Portugal, Sweden, the Netherlands, and the Ukraine) and one non-European nation (India).

Figure 19: Frequency of the c.459+1G>A variant (%) reported in various countries

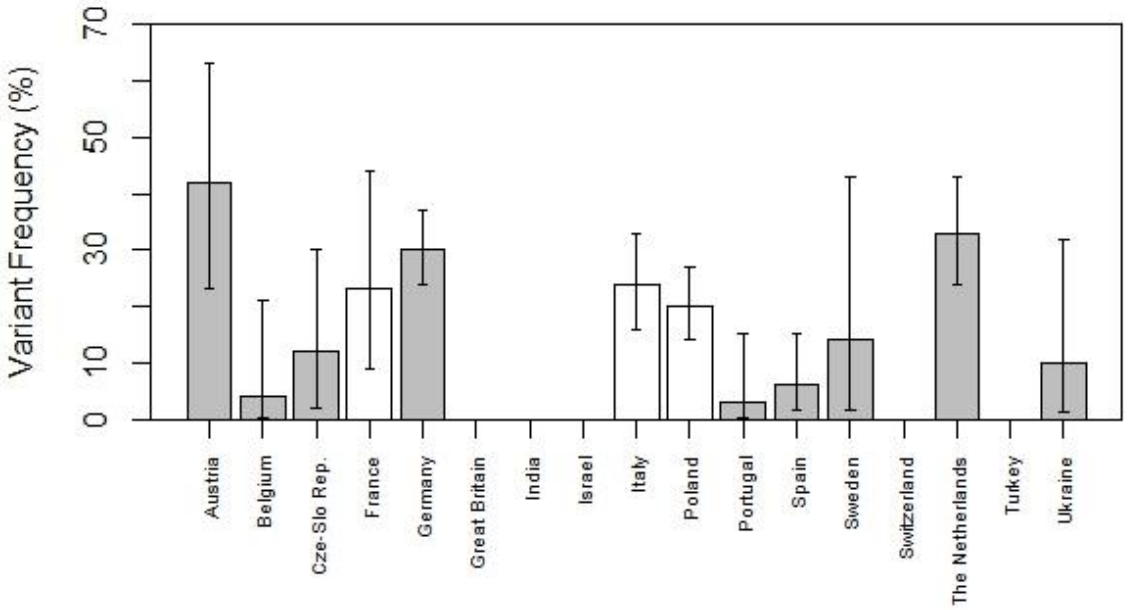


White bars indicate a frequency between 15% and 43%, while grey bars indicate frequencies that fall outside of the expected frequency range of 15-45%. In all cases, 95% confidence intervals are indicated. The bar labelled Cze-Slo Rep indicates frequency value for the Czech and Slovak Republics.

However, the occurrence of the P426L variant was far less consistent with the frequencies of 16-25% reported by Lugowska *et al.* (152) for European patients of Caucasian descent. As

illustrated in Figure 20 and from collated data presented in Table 24, only in patients from three European countries (France, Italy and Poland) did the frequency of this variant range between 16% and 25%. This variant was not observed in a single MLD patient in three European countries (Great Britain, Switzerland, and Turkey), and two non-European countries (India and Israel). The frequency of the P426L variant was either less than 16% or greater than 25% in nine countries, all of which were European (Austria, Belgium, the Czech and Slovak Republics, Germany, Portugal, Spain, Sweden, the Netherlands, and the Ukraine).

Figure 20: Frequency comparison of the P426L variant (%)



White bars indicate a frequency of 16-25%, bars indicate frequencies that fall outside of the expected frequency range of 16-25%. In all cases, 95% confidence intervals are indicated. The bar labelled Cze-Slo Rep indicates frequency value for the Czech and Slovak Republics.

Although the frequencies of the c.459+1G>A and P426L variants differ between all 18 countries, if data is aggregated according to global assemblages, it is found that the frequency of both of these causative variants is approximately 36% in the European, 14% in Arab and

7.5% in Indian MLD patients. No published data exists as yet to support the occurrence of this causative variant in American MLD patients. The P426L variant is present in approximately 15% and 2% of MLD patients present in European and Middle-Eastern populations, respectively. There is no indication that the P426L variant is present in the Indian population from published reports. When studying the presence of the c.459+1G>A variant in European, Arab and Indian populations, it is seen that this variant is respectively present in approximately 21%, 12%, and 8% of these MLD patients.

5.3.2 Variant Diversity

A total of 73 publications were identified as potentially good data sources, however only 10 of these were found to contain MLD variant data suitable for comparative analysis. Thus, a core database containing variant data pertaining to 18 populations as represented in 18 different countries was generated from these 10 publications. However, due to variability in study designs and testing methods, only seven MLD-associated *ARSA* variants could be included into this core database. Furthermore, if any of the seven *ARSA* variants were screened for within a population but not identified, a value of zero was assigned to the variant(s) in question. If selective screening was performed within the studied population and any of the seven variants was not screened for, a dash was assigned to the variant(s) in question (Table 24).

In order to determine the diversity of MLD variants across different countries, a matrix subset was constructed containing variant information for those variants found synonymously across

the various countries (Table 25). The subset included variants which were screened for but not found, and excluded variants that were not screened for. Thus variant information from five countries was included, namely India, Italy, Poland, the Netherlands, and Turkey. Shannon and Simpson analysis was conducted using variant counts, as presented in Table 26, for India, Italy and Poland. Variant sample sizes from Turkey or The Netherlands were not large enough to use population data or to perform simple random sample to standardise according to the size of the MLD population in India. Thus, although Shannon and Simpson diversity could be calculated, diversity results could not be compared across the different countries.

Table 25: MLD variant subset used in diversity analysis

Country	Total No. Disease Chromosomes in each Country	c.1049A>G	c.763+1G>A	c.*96A>G	c.459+1G>A	P426L	c.1204+1G>A	I179S
India	40	6 (31.6%)	3 (15.8%)	7 (36.8%)	3 (15.8%)	0	0	0
Italy	46	1 (5%)	3 (15%)	3 (15%)	12 (60%)	1 (5%)	0	0
Poland	98	2 (3.9%)	0	0	16 (31.4%)	16 (31.4%)	5 (9.8%)	12 (23.5%)
The Netherlands	6	1 (33.3%)	1 (33.3%)	1 (33.3%)	0	0	0	0
Turkey	10	1 (25%)	3 (75%)	0	0	0	0	0

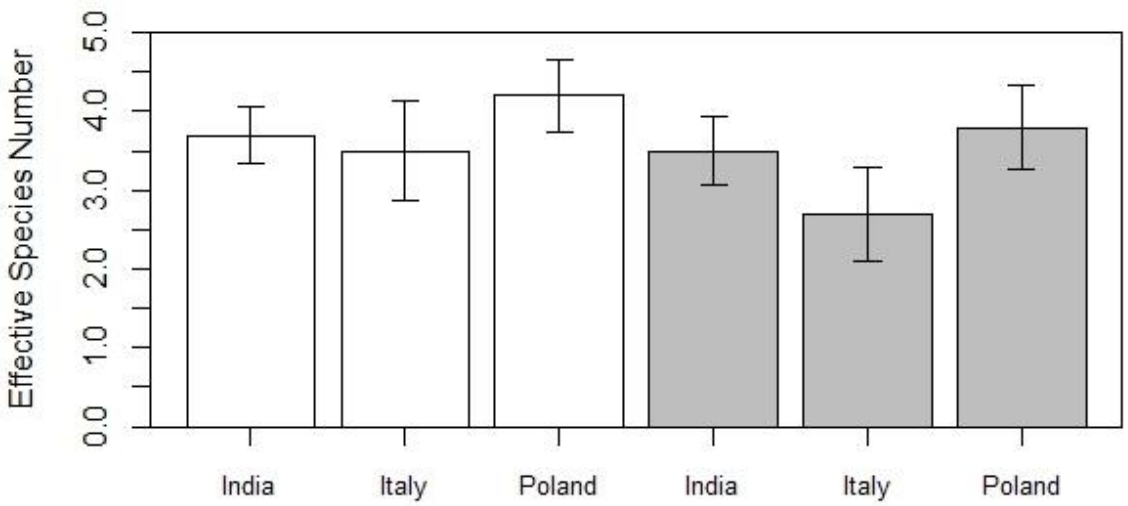
The total number of times each of the variants was identified in each of the countries is indicated as whole numbers, while percentage values are indicated in brackets.

Table 26: Shannon, Simpson and Simpson Diversity analysis of MLD variant subset data

	Shannon index		Simpson Diversity index	
Country	Index	SE	Index	SE
India	1.31	0.36	0.71	0.44
Italy	1.24	0.63	0.63	0.60
Poland	1.42	0.45	0.74	0.53

In order to further interpret Shannon and Simpson Diversity results it was necessary to investigate their corresponding effective species numbers (ESNs) (6-9, 11). Thus, as shown in Figure 21, Shannon and Simpson ESNs revealed that Poland required approximately four equally frequent variants to be present in order to generate the observed diversity. Evaluation of ESNs of MLD patients in both India and Italy revealed that between three and four equally frequent variants would be required to generate the observed diversity values calculated for each of these countries. However, since sample sizes could not be standardized, comparison of ESN values between Poland, India and Italy could not be performed using either the Shannon or Simpson diversity methods.

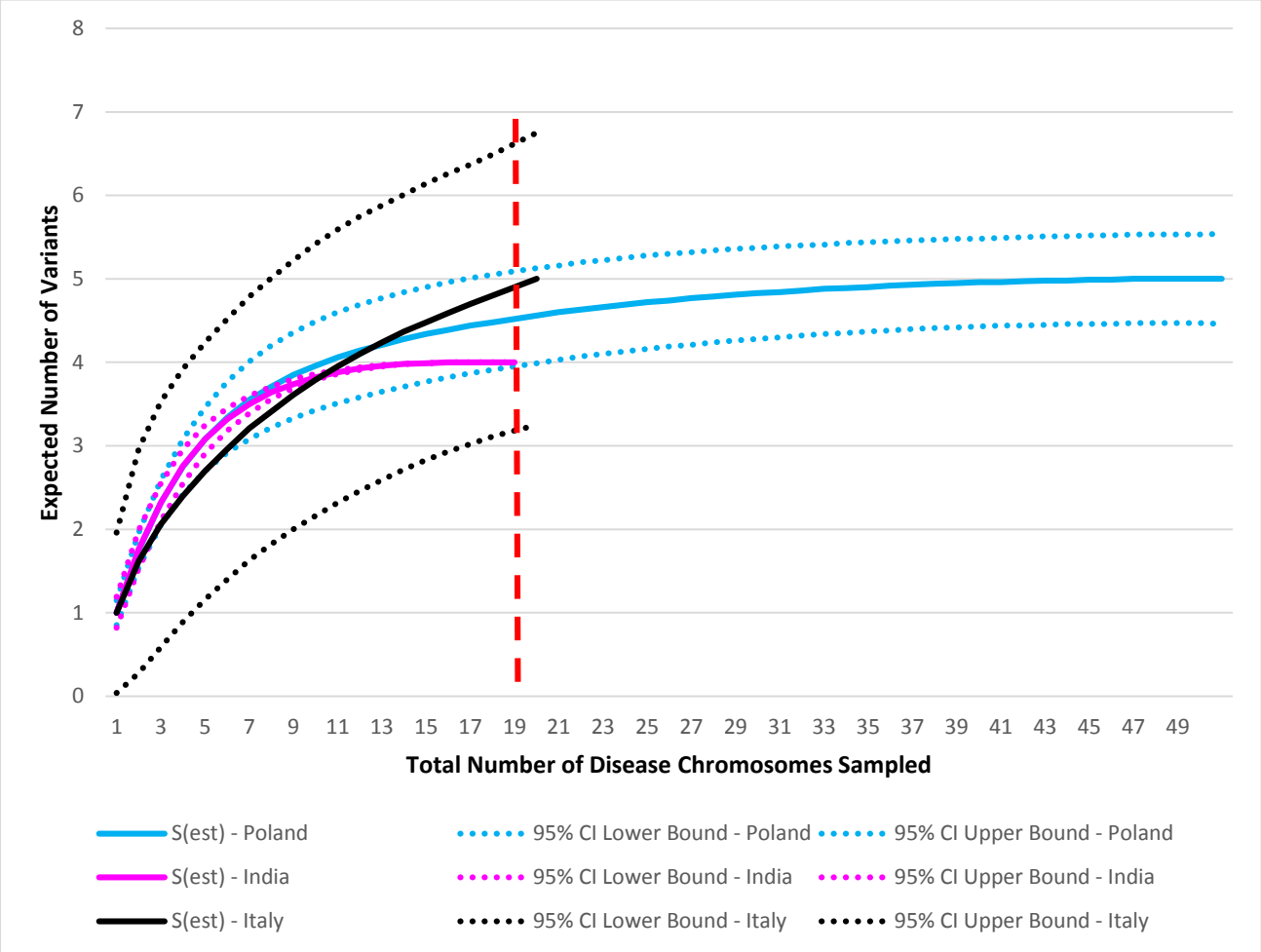
Figure 21: Shannon and Simpson ESNs for MLD variant subset data



Histograms in white indicate Shannon ESNs, while histograms in grey indicate Simpson Diversity ESNs. Standard error bars are indicated.

By comparison, and as shown in Figure 22, rarefaction analysis revealed that the point of comparison was established by the Indian MLD population, where the total number of disease chromosomes sampled was 19. Thus, at the point of comparison it was found that Italy had the highest diversity, followed by Poland and India. No significant difference in variant diversity was shown to exist between these three countries, and sample saturation was achieved by the point of comparison in India, but after this point in Polish MLD patients. Sample saturation was not achieved for Italian MLD patients, indicating the need to study larger sample sizes within this population. Rarefaction analysis could unfortunately not be performed on Dutch and Turkish MLD patients due to insufficient sample numbers.

Figure 22: Rarefaction analysis of MLD variant subset data



5.4 Discussion and Conclusion

5.4.1 Variant comparison

It is known that the c.459+1G>A, P426L and I179S variants account for 40-65% of all identified variants in MLD patients of European Caucasian descent (23, 24, 152, 155, 166). However, this was found to not be the case in four of the 14 European countries (Great Britain, Spain,

Switzerland and Ukraine). All four of these countries had frequencies less than 40%. Not one of these variants was observed in Turkish MLD patients. At least one of these variants was identified in Indian and Israeli patients. Similarly, the c.459+1G>A variant is reportedly observed in 15–43% of all European MLD patients of Caucasian descent (152). Of the 14 European countries that were studied, five countries had frequencies that fell outside of this range. A frequency higher than 43% was observed in MLD patients from Portugal and Sweden, while a frequency lower than 15% was observed in British, Dutch and Ukrainian MLD patients. This variant was not observed in Turkish MLD patients, but was respectively observed in 15% and 7.5% of Israeli and Indian patients. P426L is also reported to occur commonly and is said to be found in approximately 16-25% of all Caucasian patients of European ancestry (152).

Based on all the assimilated reports, it was found that the frequency of c.459+1G>A as well as P426L varied considerably from country to country (Figure 19 and Figure 20). However, it must be considered that MLD is a comparatively rare monogenic disorder with an approximated birth prevalence ranging between one in every 24,000 live births in Poland (24) to one in every 170,000 live births in Germany (163). Additionally complicating the accurate diagnosis of MLD is the fact that it is often misdiagnosed due to phenotypic similarities shared with other neurodegenerative disorders (25, 173-175). Nevertheless, variant data shown here suggests the potential existence of 1) rare or private population-specific variants, and 2) novel variants. Due to its rarity, the complete scope of mutational burden in such countries may thus only be fully elucidated upon large-scale population screening of those suffering from leukodystrophies in general, and not just MLD (174, 176).

5.4.3 Variant Diversity comparison

It was observed that, although the Shannon and Simpson Diversity indices could be used to determine the diversity of select variants in different MLD populations, it was not possible to compare diversity across the different MLD populations. This challenge was due mostly to small sample sizes which prohibited the standardisation of MLD variant data through the application of simple random sampling methods. However, since the rarefaction method does not require sample standardisation to be performed prior to application, this method was not only able to determine diversity of the select MLD variants in the respective populations, but also allowed comparison of diversity outputs across the respective populations. This indicates a distinct benefit over the Shannon and Simpson Diversity methods.

Unlike Shannon and Simpson methods, the rarefaction method was additionally able to indicate sample saturation – i.e. able to indicate the minimum number of disease chromosomes that would need to be studied in order to find at least one copy of each of the representative variants. Since it was shown that sampling saturation was achieved only for Indian MLD patients with respect to the seven MLD variants, larger sample sizes would be required in all other MLD populations under investigation. It must however be remembered that MLD is a comparatively rare monogenic disorder with birth prevalence values having been described as ranging between approximately 1/24,400 and 1/167,000 live births (24, 163). Despite the fact that known founder MLD populations, such as the Navajo Indians (162, 177), Alaskan Eskimos (178), as well as Arab and Habbanite Jew populations in Israel (81, 150) have been described, sample saturation is likely to remain a problem within MLD populations and variant diversity determination.

Chapter 6 – Concluding Discussion

This dissertation was founded on the hypothesis that diversity theories which are commonly applied in ecological studies can successfully be applied within the medical field in order to investigate the molecular diversity of disease-associated variants. Although this was found to be true, it was also found that distinct differences exist between the various methods. Firstly, where Shannon, Simpson and rarefaction are all able to determine the diversity of variants within and across different populations/countries, methods correcting for multiple comparison are available for Shannon and Simpson analysis, whereas such methods have not yet been described for rarefaction.

Secondly, rarefaction considers that uneven sample sizes may be present across different populations/regions and makes allowance for this in its comparisons. In contrast, the Shannon and Simpson methods assume an equal sampling effort whether this is truly the case or not. In instances where this is not the case, simple random selection should first be performed on population/sampled data before applying Shannon or Simpson methods. This is not only unnecessarily laborious in the light of rarefaction, but makes reproducing results impossible when first applying simple random selection before performing Shannon or Simpson analysis.

Thirdly, within the context in which these diversity measures have been applied, rarefaction analysis describes diversity as a function of variant richness allowing results to be compared across sampled regions/populations without need for further processing/manipulation. Shannon and Simpson diversity methods, by contrast, must be accompanied by their

respective effective species numbers in order to be truly comparable with reference to variant diversity.

The final advantage that use of the rarefaction method can provide in the context of genetic disorders is that of sample saturation. Rarefaction curves are able to indicate whether or not sample saturation has been achieved in a selected population/region, while neither Shannon nor Simpson analysis is able to indicate sample saturation. Ultimately, the rarefaction method is able to indicate the minimum number of disease chromosomes that would need to be studied in order to identify each of the selected variants at least once. This is a distinguishing feature of the rarefaction method over other diversity methods and holds immense potential in the light of sequence data – this method has the power to significantly contribute to the development and/or refinement of population-specific screening tools in a low-cost manner if used as a means through which to optimise genetic screening panel development.

Therefore; despite the fact that Shannon, Simpson and rarefaction diversity methods were originally developed in the context of ecological studies, the use of rarefaction in the context of genetic disorders is a potentially powerful tool in order to study variant diversity across different populations/regions. Despite offering correction for multiple comparisons, the use of Shannon or Simpson indices is not advised in the context of genetic disorders due to the several disadvantages that these methods present. It was however found that, for this study, although the rarefaction method was the most valuable method to use in determining variant

diversity, it was additionally associated with somewhat less-desirable outcomes associated with 1) the source of the data and 2) computational limits of EstimateS.

Given the fact that CF is considered to be one of the most common monogenic disorders in South Africa, a clear clinical picture of CF across the various population groups of South Africa is not well-described and not always recent, if available. In an attempt to generate such information through studying the patient files of 44 CF patients attending SBAH, it was found that variations in clinical presentation exist between CF patients of different ethnicities attending the SBAH CF clinic. Although a clear clinical picture of CF is still not available for the South African CF population, a definite clinical profile of CF in the various ethnic groups of patients attending SBAH has been determined with regard to clinical, biochemical and molecular presentation. However, if a collaborative effort is established amongst all South African CF clinics, improved epidemiological data could be collected, which in turn may result in improved overall healthcare for CF patients of all ethnicities within South Africa, and not just those attending SBAH.

Additionally, since different variants are typically screened for on a selective basis in the various ethnic groups of South African CF patients (141), diversity analysis across the different CF populations using any of the diversity methods presented in this dissertation was unfortunately not possible when variant data was stratified according to ethnicity. However, South Africa experiences an exceptionally high disease burden due to both communicable and non-communicable diseases and the cost to the South African Government, affected

South African individuals and their families is considerable (84). Several groups (20, 89) have highlighted the lack of specialised and dedicated staff pertaining to genetic disorders within the health sector of South Africa. Large-scale collaborative efforts with regard to both common and rare monogenic disorders (such as CF and MLD, respectively) would therefore not only aid in the positive contribution towards the human genetics policy guidelines for the management and prevention of genetic disorders, birth defects and disabilities (84) through the potential establishment of a national registry/registries, but also afford the opportunity to be truly comparable on a global scale. Collaborative efforts would also provide the means through which sequencing data could be generated in large enough quantities in different regions or populations to be analysed and realistically compared through the rarefaction method.

References

1. Shannon CE, Weaver W. The mathematical concept of communication. 1325 South Oak Street, Champaign, IL: University of Illinois Press; 1949.
2. Simpson EH. Measurement of diversity. Nature. 1949;163:688.
3. Gini C. Variabilitae mutabilita. University of Cagliari: StudiE conomico-Giuridici; 1912.
4. Greenberg JH. The measurement of linguistic diversity. Language. 1956;32:109-15.
5. Pielou EC. An introduction to mathematical ecology. New York: Wiley-Interscience; 1969.
6. Jost L. Entropy and diversity. Oikos. 2006;113(2):363-75.
7. Jost L. Partitioning Diversity into Independent Alpha and Beta Components. Ecology. 2007;88(10):2427-39.
8. Jost L. The Relation between Evenness and Diversity. Diversity. 2010;2:207-32.
9. Hill MO. Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology. 1973;54(2):427-32.
10. Hill MO, H. G. Gauch J. Detrended Correspondence Analysis: An Improved Ordination Technique. Vegetatio. 1980;42(1/3):47-58.
11. Macarthur RH. Patterns of species diversity. Biological Reviews. 1965;40(4):510-33.
12. Hurlbert SH. The Nonconcept of Species Diversity: A Critique and Alternative Parameters. Ecology. 1971;52(4):577-86.
13. Kalinowski ST. Counting alleles with rarefaction: Private alleles and hierarchical sampling designs. Conservation Genetics. 2004;5:539-43.
14. Colwell RK, Chao A, J.Gotelli N, Lin S-Y, Mao CX, Chazdon RL, et al. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. Journal of Plant Ecology. 2012;5(1):3–21.
15. Colwell RK, Mao CX, Chang J. Interpolating, Extrapolating, and Comparing Incidence-Based Species Accumulation Curves. Ecology. 2004;85(10):2717-27.
16. Chao A, Jost L. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. Ecology. 2012;93(12):2533–47.
17. Goldman A, Graf C, Ramsay M. Molecular diagnosis of cystic fibrosis in South African populations. South African Medical Journal. 2003;93(7):518-9.
18. Westwood T, Henderson B, Ramsay M. Diagnosing cystic fibrosis in South Africa. The South African Medical Journal. 2006;96:304-6.
19. Padoa C, Goldman A, Jenkins T, Ramsay M. Cystic fibrosis carrier frequencies in populations of African origin. Journal of Medical Genetics. 1999;36:41-4.
20. Westwood T, Masakela R, Mer M, Ramsay M, Willcox P, Baird C, et al. The South African cystic fibrosis consensus document. 2012.
21. Cole G. Autopsy findings in mental patients. South African Medical Journal. 1977;52:534-6.
22. Cesani M, Capotondo A, Plati T, Sergi LS, Fumagalli F, Roncarolo MG, et al. Characterization of New Arylsulfatase A Gene Mutations Reinforces Genotype-Phenotype Correlation in Metachromatic Leukodystrophy. Human Mutation. 2009;30:E936-E45.
23. Ługowska A, Płoski R, Włodarski P, Tylki-Szymańska A. Molecular bases of metachromatic leukodystrophy in Polish patients. Journal of Human Genetics. 2010;55 394–6.

24. Ługowska A, Ponińska J, Krajewski P, Broda G, Płoski R. Population Carrier Rates of Pathogenic *ARSA* Gene Mutations: Is Metachromatic Leukodystrophy Underdiagnosed? PLoS ONE. 2011;6(6):1-5.
25. mldfoundation.org. [cited 2014 16 February]. Available from: http://mldfoundation.org/MLD-101-testing.html.
26. ulf.org. [cited 2014 16 February]. Available from: http://ulf.org/metachromatic-leukodystrophy-mld.
27. www.rightdiagnosis.com. [cited 2014 16 February]. Available from: http://www.rightdiagnosis.com/m/metachromatic_leukodystrophy/intro.html.
28. Halliburton R. Introduction to population genetics. First ed. Upper Saddle River, NJ, 07458: Pearson Prentice Hall; 2004.
29. Herrick JB. Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. Archive of Internation Medicine 1910(6):517-21.
30. Nei M. Molecular Evolutionary Genetics. New York: Columbia University Press; 1987.
31. www.nature.com. [cited 2014 10 February]. Available from: http://www.nature.com/scitable/topicpage/rare-genetoc-disorders-learning-about-genetic-disease.979.
32. www.ncbi.nlm.nih.gov. [cited 2014 10 February]. Available from: http://www.ncbi.nlm.nih.gov/Omim/mimstats.html.
33. Dawkins R. Selfish genes and selfish memes. Excerpts from Richard Dawkings, The Selfish Gene. Second Edition ed. Oxford: Oxford University Press; 1989.
34. Muller HJ. Our load of mutations. The American Journal of Human Genetics. 1950;2:111-76.
35. Muller HJ. The relation of recombination to mutational advance. Mutation Research. 1964;1:2-9.
36. Darwin C. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. : Harvard University Press; 1859.
37. Wright S. Breeding Structure of Populations in Relation to Speciation. The American Naturalist. 1940;74(752):232-48.
38. Wright S. Evolution in mendelian populations. Genetics. 1931;16(2):97-159.
39. Wright S. Isolation by distance. Genetics. 1943;28:114-38.
40. Mayr E. Animal Species and Evolution. Cambridge, MA: Harvard University Press; 1963.
41. Peires JB. The central beliefs of the Xhosa cattle-killing. The Journal of African History. 1987;28:43-63.
42. www.sahistoryonline.org.za. [cited 2014 20 January]. Available from: http://www.sahistoryonline.org.za/people/king-shaka-zulu.
43. www.sahistoryonline.org.za. [cited 2014 21 January]. Available from: http://www.sahistoryonline.org.za/people/king-mzilikazi.
44. Peires JB. The late great plot: The official delusion concerning the Xhosa cattle killing. African Studies Association 1985;12:253-79.
45. Walter EV. Rise and fall of the Zulu power World politics. 1966;18:546-63.
46. Golan D. The Life Story of King Shaka and Gender Tensions in the Zulu State. History in Africa. 1990;17:95-111.
47. Debenham F. The Reader's Digest Great World Atlas. First Edition ed. Regis House, Adderley Street, Cape Town: The Reader's Digest Association Limited; 1961.

48. www.glogster.com. [cited 2014 27 January]. Available from: http://www.glogster.com/evayoko/the-great-migration/g6-618uulbp5c5t6krktc8aea0.
49. www.bigpictureofthebible.com. [cited 2014 27 January]. Available from: http://www.bigpictureofthebible.com/south-africa-mission.
50. www.lasalle.edu. [cited 2014 27 January]. Available from: http://www.lasalle.edu/~mcinneshin/344/week10.htm.
51. Nei M, Maruyama T, Chakraborty R. The Bottleneck Effect and Genetic Variability in Populations. Evolution. 1975;29(1):1-10.
52. Beighton P, Botha MC. Inherited disorders in the Black population of southern Africa. Part 1. Historical and demographic background; genetic haematological conditions. The South African Medical Journal. 1986;69:247-9.
53. Botha MC, Beighton P. Inherited disorders in the Afrikaner population of southern Africa. Part 1. Historical and demographic background, cardiovascular, neurological, metabolic and intestinal conditions. The South African Medical Journal. 1983;64:609-12.
54. Foster MW, Sharp RR. Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. Genome Research. 2002;12:844-50.
55. Leland KN, Odling-Smee J, Myles S. How culture shaped the human genome:bringing genetics and the human sciences together. Nature Reviews. 2010;11:137-48.
56. www.voortrekker-history.co.za. [cited 2015 18 August]. Available from: www.voortrekker-history.co.za/early_history.php#.VdNDbvnvO00.
57. Vink M. "The World's Oldest Trade": Dutch Slavery and Slave Trade in the Indian Ocean in the Seventeenth Century. Journal of World History. 2003;14(2):131-77.
58. Guelke L, Shell R. An early colonial landed gentry: land and wealth in the Cape Colony 1682-1731. Journal of Historical Geography. 1983;9(3):265-86.
59. de Wit E, Delport W, Rugamika CE, Meintjes A, Möller M, Helden PDv, et al. Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. Hum Genet 2010;128:145–53.
60. Wright S. Evolution and the genetics of populations. II The theory of gene frequencies. Chicago: University of Chicago Press; 1969.
61. Crow JF, Kimura M. An introduction to populatioin genetics theory. New York: Harper and Row; 1970.
62. Smith F, Grassle JF. Sampling properties of a family of diversity measures. Biometrics. 1977;33:283-92.
63. Burnham KP, Overton WS. Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika. 1978;65:625-33.
64. Colwell RK. *Estimate S*: Statistical estimation of species richness and shared species from samples. 9.1.0 ed. University of Connecticut, USA2013.
65. Gotelli NJ, Colwell RK. Estimating species richness. In: Magurran AE MBe, editor. Frontiers in Measuring Biodiversity. New York: Oxford University Press; 2011. p. 39–54.
66. Payton ME, Greenstone MH, Schenker N. Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? Journal of Insect Science. 2003;3:34-9.
67. Peet RK. The measurements of species diversity. Reviews of Ecology and Systematics. 1974;5:285-307.
68. Margalef R. La teoria de la informacion en ecologia. Mem Real Acad Cience Artes Barcelona 1957;32:373-449.

69. Gotelli NJ, Chao A. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In: S.A. L, editor. Encyclopedia of Biodiversity. 5. 2nd ed. Waltham, MA: Academic Press; 2013. p. 195-211.
70. Williams CB. Patterns in the balance of nature. London 1964.
71. Leberg PL. Estimating allelic richness: Effects of sample size and bottlenecks Molecular Ecology. 2002;11:2445-9.
72. Grünwald NJ, Hoheisel G-A. Hierarchical analysis of diversity, selfing, and genetic differentiation in populations of the Oomycete *Aphanomyces euteiches* Phytopathology. 2006;96:1134-41.
73. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, et al. Topographical and temporal diversity of the human skin microbiome Science. 2009;324:1190-2.
74. Tilman D, Reich PB, Knops JMH. Biodiversity and ecosystem stability in a decade-long grassland experiment. Nature. 2006;44:629-32.
75. Davis Parker EJ. Ecological implications of clonal diversity in pathenogenetic morphospecies. American Zoologist. 1979;19:753-62.
76. Berger WH, Parker FL. Diversity of planktonic *Forminifera* in deep-sea sediments. Science. 1970;168:1345-7.
77. van Nood E, Vrieze A, Nieuwdorp M, Fuentes S, Zoetendal EG, Vos WMd, et al. Duodenal infusion of donor feces for recurrent *clostridium difficile* The New England Journal of Medicine. 2013;368:407-15.
78. www.genecards.org. [cited 2014 10 February]. Available from: www.genecards.org.
79. www.genet.sickkids.on.ca. [cited 2014 10 February]. Available from: http://www.genet.sickkids.on.ca/StatisticsPage.html.
80. www.cftr2.org. [cited 2014 10 February]. Available from: http://www.cftr2.org/mutations/history.php.
81. Zlotogora J, Bach G, Barak Y, Elian E. Metachromatic Leukodystrophy in the Habbanite Jews: High Frequency in a Genetic Isolate and Screening for Heterozygotes. The American Journal of Human Genetics. 1980;32:663 -9.
82. Kreysing J, Figura Kv, Gieselmann V. Structure of the arylsulfatase A gene. The European Journal of Biochemistry. 1990;191:627-31.
83. Zeegers MPA, Poppel Fv, Vlietinck R, Spruijt L, Ostrer H. Founder mutations among the Dutch. European Journal of Human Genetics. 2004;12:591-600.
84. Government TSA. Human genetics policy document for the management and prevention of genetic disorders, birth defects and disability. In: Health Do, editor.: The South African Government, Department of Health; 2001. p. 1-85.
85. Maake N. In: Rensburg Jv, editor.: Stats SA; 2014.
86. Africa SS. Mid-year population estimates. 2013.
87. Africa SS. Households by monthly income category per municipality. Statistics South Africa, 2011.
88. Association SAC. [cited 2014 24 February]. Available from: www.sacfa.org.
89. Kromberg JGR, Sizer EB, Christianson AL. Genetic services and testing in South Africa. The Journal of Community Genetics. 2012;4(3):413-23.
90. Kerem B-s, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, et al. Identification of the Cystic Fibrosis Gene: Genetic Analysis. Science. 1989;245(4922):1073-80.

91. Riordan JR, Rommens JM, Kerem B-s, Alon N, Rozmahel R, Grzelczak Z, et al. Identification of the Cystic Fibrosis Gene: Cloning and Characterization of Complementary DNA. Science. 1989;245(4922):1066-73.
92. Rommens JM, Zengerling-Lentes S, Kerem B-s, Melmer G, Buchwald M, Tsui L-C. Physical Localization of Two DNA Markers Closely Linked to the Cystic Fibrosis Locus by Pulsed-Field Gel Electrophoresis. The American Journal of Human Genetics. 1989;45:932-41.
93. OMIM. #602421 [cited 2014 10 February]. Available from: http://www.omim.org/entry/602421.
94. Schwarz M, Gardner A, Jenkins L, Norbury G, Renwick P, Robinson D. Testing Guidelines for molecular diagnosis of Cystic Fibrosis. Clinical Molecular Genetics Society, 2008.
95. Mutesa L, Bours V. Diagnostic Challenges of Cystic Fibrosis in Patients of African Origin. Journal of Tropical Pediatrics. 2009;55(5):281-6.
96. Tsui L-C. The spectrum of cystic fibrosis mutations. Trends in Genetics. 1992;8:392-8.
97. Welsch MJ, Smith AE. Molecular mechanism of cystic fibrosis transmembrane conductance receoptor chloride channel dysfunction in cystic fibrosis. The American Journal of Human Genetics. 1992;50:1178-84.
98. Wilschanski M, Zielenski J, Markiewicz D, Tsui L-C, Corey M, Levison H, et al. Correlation of sweat chloride concentration with classes of the cystic fibrosis transmembrane conductance regulator gene mutations. The Journal of Pediatrics. 1995;127:705-10.
99. Masekela R, Zampoli M, Westwood AT, White DA, Green RJ, Olorunju S, et al. Phenotypic expression of the 3120+1GNA mutation in non-Caucasian children with cystic fibrosis in South Africa. Journal of Cystic Fibrosis. 2013;12:363-6.
100. Dominici R, Franzini C. Fecal Elastase-1 as a Test for Pancreatic Function: a Review. Clinical Chemistry and Laboratory Medicine. 2002;40:325-32.
101. Williams SGJ, Evanson JE, Barrett N, Hodson ME, Boultbee JE, Westaby D. An ultrasound scoring system for the diagnosis of liver disease in cystic fibrosis Journal of Hepatology 1995;22:513-21.
102. Shwachman H, Mahmoodian A. Pilocarpine iontopheresis sweat testing: results of seven years' experience. . Basel: Karger; 1967.
103. Gaskin K, Gurwitz D, Durie P, Corey M, Levison H, Forstner G. Improved respiratory prognosis in patients with cystic fibrosis with normal fat absorption. The Journal of Pediatrics. 1982;100:857-62.
104. Terbrack HG, Gürtler KH, Klör HU, Lindemann H. Human pancreatic elastase 1 concentration in faeces of healthy children and children with cysitic fibrosis. Gut. 1995;37:A253.
105. Löser C, Möllgaard A, Fölsch UR. Faecal elastase 1: a novel, highly sensitive, and specific tubeless pancreatic function test. Gut. 1996;39:580-6.
106. Hitzeroth HW, Petersen EM, Herbert J, Denter M. Preventing cystic fibrosis in the RSA. The South African Medical Journal. 1991;80:92-8.
107. Rosenstein BJ, Cutting GR. The diagnosis of cystic fibrosis: A consensus statement. The Journal of Pediatrics. 1998;132:589-95.
108. Farrell PM, Rosenstein BJ, White TB, Accurso FJ, Castellani C, Cutting GR, et al. Guidelines for Diagnosis of Cystic Fibrosis in Newborns through Older Adults: Cystic Fibrosis Foundation Consensus Report. The Journal of Pediatrics. 2008;153:S4-S14.

109. www.nhlbi.nih.gov. [cited 2014 10 February]. Available from: https://www.nhlbi.nih.gov/health/health-topics/topics/cf/diagnosis.html.
110. Goldman A, Claustres M, Guittard C, Labrum R, Desgeorges M, Wallace A, et al. The molecular basis of cystic fibrosis in South Africa. Clinical Genetics. 2001;59:37-41.
111. LeGrys VA. Sweat testing for the diagnosis of cystic fibrosis: Practical considerations. The Journal of Pediatrics. 1996;129:892-7.
112. Coakley J, Scott S, Doery J, Greaves R, Talsma P, Whitham E, et al. Australian Guidelines for the Performance of the Sweat Test for the Diagnosis of Cystic Fibrosis. Clinical and Biochemistry Review. 2006;27:S1-S7.
113. Jenkins T. Medical genetics in South Africa. The Journal of Medical Genetics 1990;27:760-79.
114. Henley LD, Hill ID. Errors, Gaps, and Misconceptions in the Disease-Related Knowledge of Cystic Fibrosis Patients and Their Families. Pediatrics 1990;85:1008-14.
115. Duff AJA. Cultural issues in cystic fibrosis. Journal of Cystic Fibrosis. 2003;2:38–41.
116. www.cdc.gov. [cited 2014 20 March]. Available from: http://www.cdc.gov/growthcharts.
117. cdc.gov. [cited 2014 20 March]. Available from: http://mcdc.gov/en/HealthSafetyTpoics/HealthyWeight/AssessingYourWeight/BodyMassIndex/BMIChildrenTeens.
118. Carles S, Desgeorges M, Goldman A, Thiart R, Guittard C, Kitazos CA, et al. First report of CFTR mutations in black cystic fibrosis patients of southern African origin. Journal of Medical Genetics. 1996;33:802-4.
119. De Carvalho CL, Ramsay M. CFTR structural rearrangements are not a major mutational mechanism in black and coloured southern African patients with cystic fibrosis. South African Medical Journal. 2009;99(10):724.
120. Herbert JS, Retief AE. The frequency of the delta F508 mutation in the cystic fibrosis genes of71 unrelated South African cystic fibrosis patients. South African Medical Journal. 1992;82:13 -5.
121. Westwood T, Brown R. Cystic fibrosis in black patients: Western Cape experiences. South African Medical Journal. 2006;96(4):288-9.
122. des Georges M, Guittard C, Templin C, Altiéri J-P, Carvalho Cd, Ramsay M, et al. WGA Allows the Molecular Characterization of a Novel Large CFTR Rearrangement in a Black South African Cystic Fibrosis Patient. Journal of Molecular Diagnostics. 2008;10(6):544–8.
123. Osborne L, Santis G, Schwarz M, Klinger K, Dörk T, McIntosh I, et al. Incidence and expression of the N1303K mutation of the cystic fibrosis (CFTR) gene. Human Genetics. 1992;89:653-8.
124. O'Sullivan BP, Freedman SD. Cystic fibrosis. Lancet. 2009;373:1891-904.
125. Sirugo G, Hennig BJ, Adeyemo AA, Matimba A, Newport MJ, Ibrahim ME, et al. Genetic studies of African populations: An overview on disease susceptibility and response to vaccines and therapeutics. Hum Genet. 2008;123:557-98.
126. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The genetic structure and history of Africans and African Americans. Science. 2009;324:1035-44.
127. Charney N, Record S. Vegetarian: Jost Diversity Measures for Community Data. 1.2 ed2012.
128. Team RC. R: A language and environment for statistical computing. 0.97.449 ed. Vienna, Austria: R Foundation for Statistical Computing; 2014.

129. Chao A, Jost L, Chiang SC, Jiang Y-H, Chazdon RL. A two-stage probabilistic approach to multiple-community similarity indices. Biometrics. 2008;64:1178-86.
130. Hutcheson K. A Test for Comparing Diversities Based on the Shannon Formula. Journal of Theoretical Biology. 1970;29:151-4.
131. Gardener M. Community Ecology: Analytical methods using R and Excel®. First edition ed: Pelagic Publishing; 2014. p. 200-9.
132. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilita: Libreria internazionale Seeber; 1936.
133. Bobadilla JL, Jr MM, Fine JP, Farrell PM. Cystic Fibrosis: A worldwide analysis of CFTR mutations-correlation with incidence data and application to screening. Human Mutation. 2002;19:575-606.
134. Gesundheitswesen ZfQuMi. Berichtsband Qualitätssicherung mukoviszidose 2012. B Werbeagentur, Henning Bock, Ermekeilstraße 48, 53113 Bonn: 2013.
135. Register BM. Summary report. Registre Belge de la Mucoviscidose, Belgium, Scientific Institute of Public Health, (WIV-ISP): 2011.
136. Registry DC. Dutch Cystic Fibrosis Registry - Report on the year 2012. Nederlandse Cystic Fibrosis Stichting, 2013.
137. Registry UC. Cystic Fibrosis Trust Annual data report 2011. 11 London Road, Bromley, Kent: 2013.
138. Australia CF. Cystic fibrosis in Australia. Unit 26, 5 Inglewood Place, Norwest Business Park, Baulkham Hills NSW 2153, Australia: Cystic Fibrosis Australia, 2012.
139. Watson MS, Cutting GR, Desnick RJ, Driscoll DA, Klinger K, Mennuti M, et al. Cystic fibrosis population carrier screening: 2004 revision of American College of Medical Genetics mutation panel. Genetics in Medicine. 2004;6(5):387-91.
140. Dequeker E, Stuhrmann M, Morris MA, Casals T, Castellani C, Claustres M, et al. Best Practice guidelines for molecular genetic diagnosis of cystic fibrosis and CFTR-related disorders - updated European recommendations. European Journal of Human Genetics. 2009;17:51-65.
141. http://www.sacftr.co.za. [cited 2015 25 August]. Available from: http://www.sacftr.co.za/whatisCF.asp.
142. OMIM. #607574 [cited 2014 16 February]. Available from: http://www.omim.org/entry/607574.
143. Li S-C, Kihara H, Serizawa S, Li Y-T, Fluharty AL, Mayes JS, et al. Activator Protein Required for the Enzymatic Hydrolysis of Cerebroside Sulfate. Deficiency in urine of patients affected with cerebroside sulfatase activator deficiency and identity with activators for the enzymatic hydrolysis of GM1 ganglioside and glocotriosynceramide. The Journal of Biological Chemistry. 1985;260:1867-71.
144. Li S-C, Sonnino S, Tettamanti G, Li Y-T. Characterization of a Nonspecific Activator Protein for the Enzymatic Hydrolysis of Glycolipids. The Journal of Biological Chemistry. 1988;263:6588-91.
145. Ciaffoni F, Tatti M, Boe A, Salvioli R, Fluharty A, Sonnino S, et al. Saposin B binds and transfers phospholipids. Journal of Lipid Research. 2006;47:1045-53.
146. OMIM. #249900 [cited 2014 16 February]. Available from: http://www.omim.org/entry/249900.
147. Biffi A, Cesani M, Fumagalli F, Carro UD, Baldoli C, Canale S, et al. Metachromatic leukodystrophy – mutation analysis provides further evidence of genotype–phenotype correlation. Clinical Genetics. 2008 74:349–57.

148. OMIM. #176801 [cited 2014 16 February]. Available from: http://www.omim.org/entry/176801.
149. Polten A, Fluharty A, Fluharty CB, Kaapler J, Figura Kv, Gieselmann V. Molecular basis of different forms of metachromatic leukodystrophy. The New England Journal of Medicine. 1991;324:18-22.
150. Heinisch U, Zlotogora J, Kafert S, Gieselmann V. Multiple Mutations Are Responsible for the High Frequency of Metachromatic Leukodystrophy in a Small Geographic Area. The American Journal of Human Genetics. 1995;56:51-7.
151. Long J, Covington C, Delaney-Black V, Nordstrom B. Allelic Variation and Environmental Lead Exposure in Urban Children. AACN Clinical Issues. 2002;13(4):550-6.
152. Lugowska A, Amaral O, Berger J, Berna L, Bosshard NU, Chabas A, et al. Mutations c.459+1G>A and p.P426L in the ARSA gene: Prevalence in metachromatic leukodystrophy patients from European countries. Molecular Genetics and Metabolism. 2005;86:353-9.
153. Barth ML, Fensom A, Harris A. The arylsulphatase A gene and molecular genetics of metachromatic leucodystrophy. Journal of Medical Genetics. 1994;31:663-6.
154. Gieselmann V, Franken S, Klein D, Mansson JE, Sandhoff R, Lüllmann R, et al. Metachromatic leukodystrophy: consequences of sulphatide accumulation. Acta Paediatrics Supplement. 2003;443:74-9.
155. Önder E, Sinici I, Sönmez FM, Topçu M, Özkara HA. Identification of two novel arylsulfatase A mutations with a polymorphism as a cause of metachromatic leukodystrophy. Neurological Research. 2009;31:60-6.
156. Gomez-Lira M, Perusi C, Mottes M, Pignatti PF, Manfredi M, Rizzuto N, et al. Molecular genetic characterization of two metachromatic leukodystrophy patients who carry the T799G mutation and show different phenotypes; description of a novel null-type mutation. Human Genetics. 1998;102:459–63.
157. Poenaru L, Castelnau L, Besançon A-M, Nicolesco H, Akli S, Theophil D. First Trimester Prenatal Diagnosis of Metachromatic Leukodystrophy on Chorionic Villi by 'Immunoprecipitation-electrophoresis'. Journal of Inherited Metabolic Disease. 1988;11:123-30.
158. Education B. [cited 2014 18 February]. Available from: https://www.bcm.edu/cancergeneticslab/test_details.cmf?testcode=4538&show=1.
159. www.mushealth.com. [cited 2014 18 February]. Available from: http://www.mushealth.com/lab/content.aspx?id=150322.
160. www.sbmf.org. [cited 2014 18 February]. Available from: http://www.sbmf.org/index.php/clinicaltests/page/43537.
161. Natowicz MR, Prence EM, Chatcurvedi P, Newburg DS. Urine sulfatides and the diagnosis of metachromatic leukodystrophy. Clinical Chemistry. 1996;42(2):232-8

162. Holve S, Hu D, McCandless SE. Metachromatic leukodystrophy in the Navajo: Fallout of the American-Indian wars of the Nineteenth century. American Journal of Medical Genetics. 2001;101:203-8.
163. Heim P, Claussen M, Hoffmann B, Conzelmann E, Gärtner J, Harzer K, et al. Leukodystrophy Incidence in Germany. American Journal of Medical Genetics. 1997;71:475–8.

164. Shukla P, Vasisht S, Srivastava R, Gupta N, Ghosh M, Kumar M, et al. Molecular and structural analysis of metachromatic leukodystrophy patients in Indian population. Journal of the Neurological Sciences. 2011;301:38-45.
165. Holmes L, Cornes MJ, Foldi B, Miller F, Dabney K. Clinical Epidemiologic Characterization of Orthopaedic and Neurological Manifestations in Children With Leukodystrophies. Journal of Pediatrics and Orthopaedics. 2011;31:587–93.
166. Berná L, Gieselmann V, Poupětouvá H, Hřebíček M, Elleder M, Ledvinová J. Novel Mutations Associated With Metachromatic Leukodystrophy: Phenotype and Expression Studies in Nine Czech and Slovak Patients. American Journal of Medical Genetics. 2004;129A:277–81.
167. Coulter-Mackie MB, Gagnier L, Beis MJ, Applegarth DA, Cole DEC, Gordon K, et al. Metachromatic leucodystrophy in three families from Nova Scotia, Canada: a recurring mutation in the arylsulphatase A gene. Journal of Medical Genetics. 1997;34:493-8.
168. Zlotogora J, Furman-Shaharabani Y, Harris A, Barth ML, Figura Kv, Gieselmann V. A single origin for the most frequent mutation causing late infantile metachromatic leucodystrophy. Journal of Medical Genetics. 1994;31:672-4.
169. Regis S, Corsolini F, Ricci V, Duca MD, Filocamo M. An unusual arylsulfatase A pseudodeficiency allele carrying a splice site mutation in a metachromatic leukodystrophy patient. European Journal of Human Genetics. 2004;12:150-4.
170. Regis S, Corsolini F, Stroppiano M, Cusano R, Filocamo M. Contribution of arylsulfatase A mutations located on the same allele to enzyme activity reduction and metachromatic leukodystrophy severity. Human Genetics. 2002;110:351-5.
171. Grossi S, Regis S, Rosano C, Corsolini F, Uziel G, Sessa M, et al. Molecular Analysis of ARSA and PSAP Genes in Twenty-one Italian Patients with Metachromatic Leukodystrophy: Identification and Functional Characterization of 11 Novel ARSA Alleles. Human Mutation. 2008;29:E220-E30.
172. Luyten JAFM, Wenink PW, Steenbergen-Spanjers GCH, Wevers RA, Amstel HKPv, Jong JGNd, et al. Metachromatic leukodystrophy: A 12-bp deletion in exon 2 of the arylsulphatase A gene in a late infantile variant. Human Genetics. 1995;96:357-60.
173. Bonkowsky JL, Nelson C, Kingston JL, Filloux FM, Mundorff MB, Srivastava R. The burden of inherited leukodystrophies in children. Neurology. 2010;75:718–25.
174. Coelho JC, Wajner M, Burin MG, Vargas CR, Giugliani R. Selective screening of 10,000 high-risk Brazilian patients for the detection of inborn errors of metabolism. European Journal of Pediatrics. 1997;156:650-4.
175. Kohlschütter A, Bley A, Brockmann K, Gärtner J, Krägeloh-Mann I, Rolfs A, et al. Leukodystrophies and other genetic metabolic leukoencephalopathies in children and adults. Brain & Development. 2010;32:82–9.
176. Eto Y, Kawame H, Hasegawa Y, Ohashi T, Ida H, Tokoro T. Molecular characteristics in Japanese patients with lipidosis: Novel mutations in metachromatic leukodystrophy and Gaucher disease. Molecular and Cellular Biochemistry. 1993;119:179-84.
177. Pastor-Soler NM. Metachromatic leukodystrophy in the Navajo Indian population: A splice site mutation in intron 4 of the arylsulfatase A gene. Human Mutation. 1994;4:199-207.
178. Pastor-Soler NM, Schertz EM, Rafi MA, Gala Gd, Wenger DA. Metachromatic leukodystrophy among southern Alaskan Eskimos: molecular and genetic studies. Journal of Inherited Metabolic Disease. 1995;18 326-32.

Appendix A: Informed consent for SBAH patients

PATIENT / PARTICIPANT'S INFORMATION LEAFLET & INFORMED CONSENT FORM FOR CLINICAL TRIAL / NON-INTERVENTION STUDY

TITLE OF STUDY:
RAREFACTION AS A TOOL TO DETERMINE VARIANT DIVERSITY IN MONOGENIC DISORDERS.

Consent and assent:
If there are children younger than 7 years in your study, the parents give consent on their behalf and you will need to adapt the information leaflet by substituting "you" with "your child".
For children between 7 and 18 years, parents give consent for their child to participate in the study and the child gives assent. Adapt the form below for that purpose too. Both information leaflets and the consent /assent form have to be included with your application.

Dear Mr. / Mrs. ..... Date …...../…...../…....

1) INTRODUCTION

You are invited to volunteer for a research study. This information leaflet is to help you to decide whether if you would like to participate. Before you agree to take part in this study you should fully understand what is involved. If you have any questions, which are not fully explained in this leaflet, do not hesitate to ask the investigator. You should not agree to take part unless you are completely happy about all the procedures involved. In the best interest of your health, it is strongly recommended that you discuss with or inform your personal doctor of your possible participation in this study, wherever possible.

2) THE NATURE AND PURPOSE OF THIS STUDY

You are invited to take part in a research study. The aim of this study is to evaluate

the presence of monogenic/inherited disorders occurring in South Africa. In doing so, we wish to focus our attention on Cystic Fibrosis (CF) in South Africa by determining which mutations are found the most in South African CF patients. We also hope to be able to calculate more accurate statistics for CF mutations as well as the prevalence of CF in South Africa. This will be done so as to better inform the clinician/doctor about CF, which would allow a CF diagnosis to be made as early as possible. Results from this study will also be applied to the development of a genetic testing tool. This tool would assist physicians and clinicians alike to make a more accurate and quicker diagnosis as to the patient's condition. The sooner a diagnosis can be made, the sooner the patient may start treatment for the condition.

3) EXPLANATION OF PROCEDURES TO BE FOLLOWED

This study will only extract information from patient files. Only information on positively diagnosed Cystic Fibrosis patients will be examined. Information that will be extracted from patient files will include the following where available: Hospital number, date of birth, age, race, region from which the patient originates, which symptoms were present during the time of diagnosis, as well as which symptoms are currently experienced by the patient, family history of cystic fibrosis, and the test results of any molecular or biochemical analysis performed on the patient. Each patient's hospital number will be given a randomly designated code. Only the patient's physician(s) will know which code responds to which patient. This will be done in order to protect the identity of each patient and at no time will any of the patients' names be made public to anyone for any reason.

Once all of the data has been collected, it will be entered into a database using the randomly designated codes and stringent statistics will be performed on the data. The statistics that will be used in this study will help identify which mutations are seen most frequently in the CF patients attending Steve Biko Academic Hospital. It will also aid in the eventual development of a diagnostic tool capable of making the diagnosis of CF easier for the patient's physician/clinician. This could also allow the patient to receive treatment for the symptoms/condition sooner and would help alleviate some of the complications associated with this disorder quicker than is currently possible for many South African CF patients.

4) RISK AND DISCOMFORT INVOLVED.

There is no risk or discomfort that will be caused/associated with this study. No form of communication or physical contact will be made with the patients and/or guardians unless made by the patient's physician/clinician (to gain informed consent, for questions that would like to be answered etc.). The work performed in this project does not require anything other than permission to study and anonymously record the contents of the patient's file(s).

5) POSSIBLE BENEFITS OF THIS STUDY.

One of the long term goals of this study is to aid in the development and production of a diagnostic tool able to identify disease-causing mutations with greater accuracy, in a shorter period of time and at a lower cost than currently available. The information that is collected in this study will not only bring us closer to achieving these goals, but will also help us identify which mutations occur most commonly in the various ethnic groups living in South Africa. This will help us to bring better diagnostic services to those affected by Cystic Fibrosis in South Africa in the future. However, we cannot, in any way, guarantee that gathering and studying the information generated from the patient files will be of any immediate help or use to the patient in question.

6) I understand that if I do not want to participate in this study, I will still receive standard treatment for my illness.

7) I may, at any time, withdraw from this study.

8) HAS THE STUDY RECEIVED ETHICAL APPROVAL?

This Protocol was submitted to the Faculty of Health Sciences Research Ethics Committee, University of Pretoria and written approval has been granted by that committee. The study has been structured in accordance with the Declaration of Helsinki (last update: October 2008), which deals with the recommendations guiding doctors in biomedical research involving humans/subjects. A copy of the Declaration may be obtained from the investigator should you wish to review it.

This study forms part of a larger study for which ethical approval has already been granted (no. 4-2013)

9) INFORMATION If I have any questions concerning this study, I should either contact:

Dr Refiloe Masekela tel : (012) 354 5272 or cell: +2779 489 0936

Or
The University of Pretoria's Research Ethics committee
Faculty of Health Sciences
HW Snyman South Building
Rooms 2.33 – 2.35
31 Bophelo Road
Gezina
Pretoria
Tel: (012) 354 1677

10) CONFIDENTIALITY

All records obtained whilst in this study will be regarded as confidential. Results will be published or presented in such a fashion that patients remain unidentifiable.

11) CONSENT TO PARTICIPATE IN THIS STUDY.

I have read or had read to me in a language that I understand the above information before signing this consent form. The content and meaning of this information has been explained to me. I have been given the opportunity to ask questions and am satisfied that they have been answered satisfactorily. I understand that I will not receive any kind of payment if I volunteer to participate in this study

I understand that if I do not participate in this study, that it will not alter my management in any way. I hereby volunteer to take part in this study.

I have received a signed copy of this informed consent agreement.

.....	.....
Patient / Guardian signature	Date
.....	.....
Person obtaining informed consent	Date
.....	.....
Witness	Date

VERBAL PATIENT INFORMED CONSENT (applicable when patients cannot read or write)

I, the undersigned, Dr ……..…, have read and have explained fully to the patient, named …….. and/or his/her relative, the patient information leaflet, which has indicated the nature and purpose of the study in which I have asked the patient to participate. The explanation I have given has mentioned both the possible risks and benefits of the study and the alternative

treatments available for his/her illness. The patient indicated that he/she understands that he/she will be free to withdraw from the study at any time for any reason and without jeopardizing his/her treatment. I hereby certify that the patient has agreed to participate in this study.

Patient's Name _____
(Please print)

Investigator's Name _____
(Please print)

Investigator's Signature _____Date _____

Witness's Name _____Witness's Signature _____Date _____
(Please print)

Witness - sign that he/she has witnessed the process of informed consent)

Appendix B: Ethical Approval Certificate – 40-2014

The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.
- FWA 00002567, Approved dd 22 May 2002 and Expires 20 Oct 2016.
- IRB 0000 2235 IORG0001762 Approved dd 22/04/2014 and Expires 22/04/2017.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences Research Ethics Committee

08/05/2014

Approval Certificate
New Application

Ethics Reference No.: 40/2014

Title: Rarefaction as a tool to determine variant diversity in monogenic disorders

Dear Ms Jeanne van Rensburg

The **New Application** as supported by documents specified in your cover letter for your research received on the 28/01/2014, was approved by the Faculty of Health Sciences Research Ethics Committee on the 08/05/2014.

Please note the following about your ethics approval:
- Ethics Approval is valid for 2 years
- Please remember to use your protocol number (40/2014) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, or monitor the conduct of your research.

Ethics approval is subject to the following:
- The ethics approval is conditional on the receipt of 6 monthly written Progress Reports, and
- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

Dr R Summers, MBChB, MMed (Int), MPharMed
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes 2004 (Department of Health).

♦ Tel:012-3541330 ♦ Fax:012-3541367 Fax2Email: 0866515924 ♦ E-Mail: fhsethics@up.ac.za
♦ Web: //www.healthethics-up.co.za ♦ H W Snyman Bld (South) Level 2-34 ♦ Private Bag x 323, Arcadia, Pta, S.A., 0007

Appendix C: Ethical Approval Certificate – 4-2013

The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.

- **FWA** 00002567, Approved dd 22 May 2002 and Expires 20 Oct 2016.
- **IRB** 0000 2235 IORG0001762 Approved dd 13/04/2011 and Expires 13/04/2014.

Universiteit van Pretoria
University of Pretoria

Faculty of Health Sciences Research Ethics Committee
Fakulteit Gesondheidswetenskappe Navorsingsetiekkomitee

DATE: 30/01/2013

NUMBER	**4/2013**
TITLE OF THE PROTOCOL	A Molecular Investigation of Cystic Fibrosis in South Africa
PRINCIPAL INVESTIGATOR	Prof. Micheal Pepper **Dept:** Immunology, University of Pretoria. **Cell:** 072 209 6324 **E-Mail:** michael.pepper@up.ac.za
SUB INVESTIGATOR/S	Dr. Cheryl Stewart **E-Mail:** cherylsstewart@gmail.com u13297440@tuks.co.za
STUDY COORDINATOR	Prof. Pepper
SUPERVISOR	Prof. Pepper **E-Mail:** michael.pepper@up.ac.za
STUDY DEGREE	MSc student (Ms. Jeanne van Rensburg)
SPONSOR COMPANY	None
CONTACT DEATAILS OF SPONSOR	Not applicable
SPONSORS POSTAL ADDRESS	Not applicable
MEETING DATE	30/01/2013

The Protocol and Informed Consent Document were approved on 30/01/2013 by a properly constituted meeting of the Ethics Committee subject to the following conditions:

1. The approval is valid for 4 years period [till the end of December 2016], and
2. The approval is conditional on the receipt of 6 monthly written Progress Reports, and
3. The approval is conditional on the research being conducted as stipulated by the details of the documents submitted to and approved by the Committee. In the event that a need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

Members of the Research Ethics Committee:

Prof RSK Apatu	MBChB (Legon,UG); PhD (Cantab); PGDip International Research Ethics (UCT)
Prof M J Bester	(female)BSc (Chemistry and Biochemistry); BSc (Hons)(Biochemistry); MSc(Biochemistry); PhD (Medical Biochemistry)
Mrs N Briers	(female) BSc (Stell); BSc Hons (Pretoria); MSc (Pretoria); DHETP (Pretoria)
Dr IK Dada	BSc.HB; MB ChB (UNZA); MA Appl. Pop. Research (EXON) ; MPH (UP)
Prof R Delport	(female)BA et Scien, B Curationis (Hons) (Intensive care Nursing), M Sc (Physiology), PhD (Medicine), M Ed Computer Assisted Education
Prof MM Ehlers	(female) BSc (Agric) Microbiology (Pret); BSc (Agric) Hons Microbiology (Pret); MSc (Agric) Microbiology (Pret); PhD Microbiology (Pret); Post Doctoral Fellow (Pret)
Dr R Leech	(female) B.Art et Scien; BA Cur; BA (Hons); M (ECI); PhD Nursing Science
Mr SB Masombuka	BA (Communication Science) UNISA; Certificate in Health Research Ethics Course (B compliant cc)
Dr MP Mathebula	(female)Deputy CEO: Steve Biko Academic Hospital; MBCHB, PDM, HM
Prof A Nienaber	(female) BA(Hons)(Wits); LLB; LLM; LLD(UP); PhD; Dipl.Datametrics(UNISA) – Legal advisor
Mrs MC Nzeku	(female) BSc(NUL); MSc(Biochem)(UCL, UK) – Community representative

2013/02/04MS: dd 2013/02/04: C:\Documents and Settings\user\My Documents\Protokolle\Grade briewe\Letters 2013\4.doc