



UNIVERSITEIT VAN PRETORIA  
UNIVERSITY OF PRETORIA  
YUNIBESITHI YA PRETORIA

# Investigation of Cystic Fibrosis transmembrane conductance regulator variants in South African patients with Cystic Fibrosis

By  
Odette le Grange

Submitted in fulfilment of the requirements for the degree  
Masters of Science Bioinformatics

In the Department of Biochemistry, Genetics and Microbiology  
Faculty of Natural and Agricultural Sciences  
University of Pretoria  
28 May 2023

Primary supervisor:  
Prof. Michael Pepper  
Institute for Cellular and Molecular Medicine  
Department of Immunology  
Faculty of Health Science

Co-supervisor:  
Prof. Fourie Joubert  
Centre for Bioinformatics and Computational Biology  
Department of Biochemistry, Genetics and Microbiology  
Faculty of Natural and Agricultural Sciences

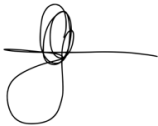
Collaborator:  
Dr. Cheryl Stewart

Funding bodies:  
SAMRC  
NRF  
University of Pretoria

*Declaration of Authorship*

I, Odette le Grange, declare that the dissertation, which I hereby submit for the degree of Masters in Science in Bioinformatics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signed by Odette le Grange



At the University of Pretoria on 28 May 2023

*Dedicated to my parents, Emile and Elsie.*

# Table of Contents

Acknowledgements .....	6
Preface .....	7
List of Figures .....	10
List of Abbreviations .....	11
<b>Chapter 1: Literature review .....</b>	<b>12</b>
<b>1.1. Introduction.....</b>	<b>13</b>
<b>1.2. Incidence.....</b>	<b>14</b>
1.2.1. Misdiagnosis .....	15
1.2.2 Decline in incidence.....	15
<b>1.3. Population allele frequencies .....</b>	<b>16</b>
1.3.1. Evolution of the high CF carrier frequency in the European population.....	19
1.3.2. Evolution of the CF carrier frequency in African populations .....	20
<b>1.4. CFTR variants in Africa .....</b>	<b>21</b>
1.4.1. A timeline of CFTR variants in Africa and surrounding geography .....	21
1.4.2. A timeline of CFTR variants in South Africa .....	23
<b>1.5. Registry capture .....</b>	<b>24</b>
<b>1.6. Early diagnosis: Improved outcome, life expectancy and quality of life .....</b>	<b>26</b>
<b>1.7. Definition of disease and genotype-phenotype correlation.....</b>	<b>28</b>
<b>1.8. Challenges to implementation of NBS .....</b>	<b>29</b>
<b>1.9. Evaluation of current CFTR variant screening panels .....</b>	<b>31</b>
<b>1.10. Conclusion .....</b>	<b>33</b>
<b>1.11. References:.....</b>	<b>35</b>
<b>Chapter 2: Methodology .....</b>	<b>39</b>
<b>2.1. Introduction .....</b>	<b>40</b>
<b>2.2. Ethics Approval .....</b>	<b>41</b>
<b>2.3. Data Collection .....</b>	<b>41</b>
<b>2.4. Molecular Biology – sample collection and NGS .....</b>	<b>41</b>
<b>2.5. QC and Trimming .....</b>	<b>42</b>
<b>2.6. Mapping and Variant Detection.....</b>	<b>42</b>
<b>2.7. In Silico Validation and Identification of Potential Variants .....</b>	<b>44</b>
<b>2.8. Variant Effect Prediction .....</b>	<b>45</b>
<b>2.9. Assessment of genotype information .....</b>	<b>48</b>
<b>2.10. Validation of Variants.....</b>	<b>48</b>
2.10.1. PCR Protocol:.....	49

2.10.1.1. Primer design.....	49
2.10.1.2. Optimisation of PCR reactions: .....	54
<b>2.11. Amendment of genotype information after confirmation with Sanger.....</b>	<b>55</b>
<b>2.12. Conclusion.....</b>	<b>55</b>
<b>2.13. References: .....</b>	<b>56</b>
<b>Chapter 3: Results.....</b>	<b>57</b>
<b>3.1. Introduction .....</b>	<b>58</b>
<b>3.2. QC, Mapping and Variant detection .....</b>	<b>58</b>
3.2.1. FASTQC:.....	58
3.2.2. Mapping: .....	59
<b>3.3. Variant Detection and in silico validation:.....</b>	<b>59</b>
<b>3.4. Variant Annotation and Effect Prediction.....</b>	<b>60</b>
<b>3.5. Compilation of a Master Variant list .....</b>	<b>63</b>
3.5.1. CFTR2 annotations:.....	63
3.5.2. Variant pathogenicity scores .....	64
3.5.2.1. VVP percentile scores:.....	64
3.5.2.2. Candidates for validation with Sanger .....	64
<b>3.6. Amendment of genotype information following most recent NGS .....</b>	<b>67</b>
<b>3.7. Sanger Sequencing confirmation results .....</b>	<b>71</b>
3.7.1. PCR Results.....	71
3.7.2. Sanger sequencing results.....	71
<b>3.8. Amendment of genotype information following confirmation with Sanger sequencing ....</b>	<b>73</b>
<b>3.9. Concluding remarks and summary of results.....</b>	<b>76</b>
<b>3.10. References: .....</b>	<b>77</b>
<b>Chapter 4 &amp; 5: Discussion and Conclusion .....</b>	<b>78</b>
<b>4.1. Introduction .....</b>	<b>79</b>
<b>4.2. QC, Mapping and Variant detection .....</b>	<b>79</b>
<b>4.3. In silico validation .....</b>	<b>80</b>
<b>4.4. Variant annotation and effect prediction.....</b>	<b>80</b>
<b>4.5. CFTR variants in a diverse population.....</b>	<b>82</b>
<b>5.1. Conclusion and future work.....</b>	<b>84</b>
<b>References: .....</b>	<b>86</b>

## *Acknowledgements*

Enormous gratitude goes to Prof. Pepper. Thank you so much for giving me a project and a purpose for the last few years. Thank you for creating a safe space where there is room for imperfection, honesty, and growth. Thank you for encouraging me to try harder and believing that I can reach the high standard that you set. Thank you for caring about my mental and physical health, and for continually understanding that we are all human. I think it is a reminder that we all need sometimes, and you are so great at recognizing the humanity in all of us. Thank you for making it possible for me to pursue my passion for science and for being such a kind-hearted person underneath all the serious responsibility that you carry.

To Prof. Fourie Joubert, my most heartfelt gratitude is due. Fourie, your extreme kindness and care is something I have not encountered often. You give your all to every single student, myself included, without a second thought or hesitation. You are there for it all, and your love for people as well as science has showed me how to approach everything in life with attention, compassion, heart, and soul. I am still not sure how you manage to be so involved with all of our lives and still have the time to do anything else. Thank you for your encouragement, trust and guidance. This dissertation would absolutely not exist without you, much less would my sanity. I hope to need your help in many of my future projects, as I know I will have it if I ask.

To my mom, thank you for being there to listen even when you don't have any idea what I am talking about. You have always had so much faith in my ability to do it all, and I think that kept me going more than anything else. Thank you for raising me to have a curious and critical mind, and to always have conviction in my beliefs. Thank you for giving me the room to find myself.

To my dad, the ever-so-recluse cheerleader, thank you for being in my corner and keeping me from giving up. You speak of me with such pride to everyone who will listen, and I have only ever wanted to live up to that amazing person you think I am. Your love keeps me pushing to be better and work harder every day.

To my best friend, Siya. Thank you for loving me the way you do. All of the tears, frustration, heartbreak, anger, sadness and dismay were caught on your shoulders so that I could keep moving forward. You are my stability, my joy, my sunshine in every moment. Your honesty, clarity and directness kept my head clear enough to "do the things". Your love and gentle nudges kept my heart in it. These last few years have been so hard, but you have anchored me through it all.

Finally, to Bella. I am pretty sure I would have gone insane multiple times without our coffee breaks and rant sessions. We have been side-by-side with every step of our academic lives thus far, and it is a bit surreal to me that we won't be living our lives at the same pace and in the same place anymore. I admit, I am a bit lost without you. Thank you for your friendship over the years, it's wild that this chapter is over but I am so excited for what's next for us.

## *Preface*

Cystic fibrosis is a genetic disorder resulting from variants in the CFTR gene causing greatly reduced life expectancy. It is characterised by the accumulation of thick, viscous mucus secretions in various organ systems. This sticky mucus is the result of dysregulation of the transmembrane movement of chloride and sodium ions across the epithelium (Kuller, Baughman et al. 1999) which commonly affects the gastrointestinal, pulmonary and genitourinary systems (Elborn 2016). The thick mucus and dysfunctional ion transport decreases mucociliary action particularly within the respiratory tract, allowing for bacterial colonisation by *Pseudomonas aeruginosa*, *Haemophilus influenzae* and *Staphylococcus aureus*. Chronic bacterial infection and the corresponding prolonged inflammatory response ultimately lead to severe illness and compromise the airway (Bell, De Boeck et al. 2015).

The means with which to care for CF patients worldwide has seen drastic improvement over the last few years. Diagnostic protocols have gained in specificity and sensitivity, with high variant detection rates being achieved using new-born screening in populations with higher frequencies of known, common variants. Additionally, advances in treatment have enabled patients with these common variants to have greater life expectancy and overall slower progression of disease. When it comes to diagnosis of CF in South Africa, the variant detection rates are markedly lower when using standard gene panels that were designed for European populations. The premise of this project, when initiated in 2013, was that there may be variants common to South African populations that are not of European origin in a similar way that F508del is common to some European populations, and that a more specific gene panel should be designed to have a higher variant detection rate if this is the case.

The F508del variant is the most common variant in CF patients of European ancestry, likely because of evolutionary pressure and population bottlenecks. It is speculated that carriers of CF have experienced some degree of protection from several prevalent pathogens, a phenomenon known as *heterozygote advantage*. Variants of the CFTR gene have been speculated to provide a protective effect against pathogens that cause typhoid, tuberculosis, and cholera. Examples of heterozygote advantage have been demonstrated in African populations, such as the evolution of high population frequencies of variants that are causative of sickle cell anaemia in malaria-ridden regions because of their protective effect. A lack of significant genetic bottlenecks does not negate heterozygous advantage of mutated genes. Thus, it is not unlikely that African ethnolinguistic groups have evolved to have higher frequencies of CFTR variants in response to

endemic typhoid fever, TB infection, and recurrent cholera outbreaks. However, the lack of genetic bottlenecks does contribute to genetic diversity and the lower probability of a few common variants having high frequency in the population.

This project originally sought to identify variants that are shared/common to the South African population, and 65 South African patients with CF on whom a molecular diagnosis could not be confirmed were sampled using next generation sequencing (NGS) of the exonic regions of their CFTR gene. Some pathogenic variants were identified and confirmed by the original researcher and collaborator, Dr. Cheryl Stewart. However, many advancements in variant discovery, annotation and effect prediction have been made since 2013. As a result, the analysis of the NGS data for these patients was repeated to ensure that the current best practices were used to identify as many of the pathogenic variants as possible. The results of this investigation show that the cohort is genetically diverse, as expected with African ethnolinguistic groups. Thus, it is unlikely that a gene panel can be designed with great enough sensitivity that would make it an effective diagnostic tool. Alternative approaches need to be investigated to provide thorough molecular diagnosis in South African patients who are suspected of having CF.

In this dissertation, the first chapter seeks to review the literature pertaining to CF in South Africa as well as globally. First, it evaluates the incidence of CF in South Africa and the factors leading to likely underestimation of incidence, such as misdiagnosis. The literature pertaining to CFTR variants is explored: variants that have been identified in the population and the evolution of high carrier frequencies of CF, as well as the challenges to achieving a high variant detection rate for African ethnolinguistic groups. Furthermore, the challenges and benefits of a national CF registry - and some of the insights gained from its implementation - are discussed. The importance of early diagnosis is highlighted by the impact that it has on CF patients and their families. This is discussed with reference to disease progression, life expectancy and quality of life. The definition of disease is also reviewed in this chapter since phenotype may vary with genotype and the “classic” presentation of symptoms needs to take diverse populations into consideration. The importance of NBS and the challenges related to its implementation are also explored, with the hope that screening can alleviate some of the issues associated with misdiagnosis. Finally, the currently available variant screening panels are reviewed and solutions for diverse populations are discussed. This chapter provides valuable context for the necessity of establishing the spectrum of CFTR variants in South African patients to improve diagnosis and, ultimately, the quality of the lives of South African people affected by CF.



The second chapter establishes the methodology used to investigate the CFTR variants in this cohort of South African patients with CF. Next generation sequencing data from a cohort of South African patients with CF was subjected to quality control and pre-processing, and variant discovery across the exonic regions of the CFTR gene was performed. Thereafter, variant effect prediction was performed, and potentially pathogenic variants were identified. Lastly, these potential variants were validated experimentally using traditional Sanger sequencing and a final list of candidates was compiled for addition to the CFTR2 database and future functional studies to evaluate.

The third chapter provides the results of the investigation. The output of the pre-processing stages is provided, followed by mapping statistics. The results are provided for the concordance between the four variant call sets as well as an *in silico* tool for corroboration of likely true positive variant calls. The next step was to perform variant annotation and effect prediction, the summary and heatmaps of which have been provided for all four variant call sets. Pathogenicity scores were then evaluated, together with the CFTR2 functional annotations, to determine a Master List of variants that are potentially pathogenic. Finally, the potential variants that have been confirmed with Sanger sequencing are provided, and the genotyping of each of the patients is also presented.

The last two chapters form the discussion and conclusion, respectively. Here the results are fully discussed and thoroughly investigated in the context of the literature. The gaps in research are discussed, as well as the challenges that might be faced going forward. Lastly, the impact of this research and its commercial application is presented.

In South Africa, greater knowledge of the CFTR variant spectrum and clinical presentation is needed for early, effective diagnosis and treatment of patients with CF. This is not only important for successful intervention and improvement of the quality of life of the patients and their families, but also for the alleviation of the burden on our healthcare services that are ultimately tasked with treating patients that are diagnosed too late and require more resources to manage.

*List of Figures*

<b>Figure 1:</b> Concept of gradient thermocycler with each row programmed to a different temperature and each column being used to optimize a different pair of primers.	54
<b>Figure 3.1:</b> a) Sequence quality for the forward and reverse reads before trimming. b) Adapter content for the forward and reverse reads before adapter removal.	58
<b>Figure 3.2:</b> a) Sequence quality for the forward and reverse reads after trimming. b) Adapter content for the forward and reverse reads after adapter removal.	58
<b>Figure 3.3:</b> An example of the mapping using Bowtie2 (a) and BWA (b), visualized with IGV, for sample CA0144930. c) Mapping with Bowtie2 for sample CA014493, zoomed out in IGV.	59
<b>Figure 3.4:</b> a) Venn diagram of the concordance between the different variant calling algorithms. b) Graph comparing the size of each variant list. c) Graph showing elements specific or shared.	60
<b>Figure 3.5:</b> a) Venn diagram of the concordance between the different variant calling algorithms and BAYSIC. b) Graph comparing the size of each variant list. c) Graph showing elements specific or shared.	60
<b>Figure 3.6:</b> VEP results for the CASAVA variant call set. a) Summary statistics of the VEP output for the CASAVA variant call set. b) Heatmap for the variants in each of the samples.	61
<b>Figure 3.7:</b> VEP results for the CLC Genomics variant call set. a) Summary statistics of the VEP output for the CLC Genomics variant call set. b) Heatmap for the variants in each of the samples.	62
<b>Figure 3.8:</b> VEP results for the Freebayes variant call set. a) Summary statistics of the VEP output for the Freebayes variant call set. b) Heatmap for the variants in each of the samples.	62
<b>Figure 3.9:</b> VEP results for the GATK variant call set. a) Summary statistics of the VEP output for the GATK variant call set. b) Heatmap for the variants in each of the samples.	63
<b>Figure 3.10:</b> eCDF plot used to normalise the background distribution of percentile scores for CFTR from gnomAD.	64
<b>Figure 3.11:</b> Visual representation of the genotyping results for the cohort. a) Distribution of samples that could or could not be fully genotyped using the NHLS panel. b) Distribution of samples that had variants discovered using NGS (excluding samples that were completely genotyped using the NHLS panel).	70
<b>Figure 3.12:</b> Ethnicity distribution of potentially pathogenic variants found in samples using NGS or gene panel screening before confirmation with Sanger sequencing.	71
<b>Figure 3.13:</b> Visual representation of the genotyping results for the cohort. a) Distribution of samples that could or could not be fully genotyped using the NHLS panel. b) Distribution of samples that had variants discovered using NGS (excluding samples that were completely genotyped using the NHLS panel).	75
<b>Figure 3.14:</b> Ethnicity distribution of potentially pathogenic variants found in samples using NGS or gene panel screening after confirmation with Sanger sequencing.	76

## *List of Abbreviations*

ACMG: American College of Medical Genetics	LMIC: Low to Middle Income Country
ACOG: American College of Obstetricians and Gynaecologists	LRT: Likelihood Ratio Test
AIDS: Acquired Immunodeficiency Syndrome	NBS: New-born Screening
BAM: Binary Alignment Map	NGS: Next Generation Sequencing
Bp: Base Pairs	NHLS: National Health Laboratory Service
BQSR: Base Quality Score Recalibration	PCR: Polymerase Chain Reaction
BWA-MEM: Burrows-Wheeler Alignment Algorithm - Maximal Exact Match	QC: Quality Control
CF: Cystic Fibrosis	RCWMCH: Red Cross War Memorial Children's Hospital
CFTR: Cystic Fibrosis Transmembrane Conductance Regulator	RM: Ready Mix
CLRT: Corrected Likelihood Ratio Test	SA: South Africa
CMJAH: Charlotte Maxeke Johannesburg Academic Hospital	SACFR: South African Cystic Fibrosis Registry
CNV: Copy Number Variation	SAM: Sequence Alignment Map
CRD: Cumulative Rank Distribution	SBAH: Steve Biko Academic Hospital
DNA: Deoxyribonucleic Acid	SDC: Secretary Diarrhoea including Cholera
ECFS: European Cystic Fibrosis Society	SNP: Single Nucleotide Polymorphism
GATK: Genome Analysis Toolkit	SV: Structural Variant
GiaB: Genome in a Bottle	TB: Tuberculosis
HIC: High income country	TH: Tygerberg Hospital
HIV: Human Immunodeficiency Virus	UCT: University of Cape Town
IGV: Integrated Genomics Viewer	UP: University of Pretoria
IRT: Immunoreactive Trypsinogen	US: United States
Kbp: Kilo Base Pairs	UTR: Untranslated Region
	VCF: Variant Call Format
	VEP: Variant Effect Predictor
	VUS: Variants of Unknown Significance
	WES: Whole Exome Sequencing

## *Chapter 1: Literature review*

Cystic Fibrosis in South Africa

### 1.1. Introduction

From diagnosis to treatment using modulator therapies, CF research has made extraordinary progress in the last few decades. The methodology underpinning the sweat chloride test, the diagnostic gold standard for CF, has been improved (Bell, Mall et al. 2020). New-born screening (NBS) programmes have been implemented and updated. In addition, molecular studies are becoming more prevalent. Patient registries have also seen improvement, enabling more thorough evaluation of incidence and outcomes (Elborn, Bell et al. 2016). Though a few studies have sought to elucidate the molecular nature of CF in genetically diverse populations, CF patients from African ethnolinguistic groups are yet to be extensively studied. CF patients from these groups have been left behind due to the classic belief that CF is a genetic disease predominant in patients of European ancestry and a rare disease in African groups (Mutesa and Bours 2009). This false assumption has created challenges regarding misdiagnosis and early intervention, estimations of incidence, and treatment availability.

The considerable difference between high and lower-to-middle income countries (HICs and LMICs) can be observed in almost all areas of CF research and care including diagnosis, nutrition and growth, lung function and quality of life (Mehta, Macek et al. 2010, Bell, Mall et al. 2020). There is also a notable difference within and between countries, as a function of inherent social inequality. Currently, alternative measures of diagnosis in LMICs is limited to CFTR panel testing designed for populations of European origin (Bell, Mall et al. 2020). However, rapid molecular diagnosis is becoming more accessible in LMICs and provides an alternative method of diagnosis when traditional sweat testing is unavailable, with the caveat of needing the appropriate spectrum of CFTR variants specific to each population. This remains a challenge as the distribution of CFTR variant frequencies depends on the genetic admixture of the population. Thus, next generation sequencing provides an avenue for improving commercial testing kits by sequencing all CFTR exons, UTRs, and CNVs in each population (Bell, Mall et al. 2020).

The *Lancet Respiratory Medicine Commission on the Future of Cystic Fibrosis Care* has released a report on the current state of care of CF worldwide and addressed the future implications for research (Bell, Mall et al. 2020). The report focused on various key areas in CF diagnosis and treatment, as well as the challenges faced. It reviewed several advances made in the field to date, such as the improvement of early diagnosis using NBS and improvements in therapy and care. When reviewing the epidemiology of CF, it was found that CFTR variants causing severe disease often lead to a spectrum of clinical manifestations and as such it is likely that environmental factors and

modifier genes may play a role in the severity of the disease. The spectrum of disease ranges from CF with pancreatic insufficiency to no disease, depending on the CFTR variant and the associated dysfunctional CFTR protein (Cutting 2010, Gallati 2014, Bell, Mall et al. 2020). Additionally, with increasing numbers of patients surviving until adulthood, several complications have arisen. These include cystic fibrosis-related diabetes, metabolic bone disease, gastrointestinal malignancy, and comorbidities including the increase of mental health conditions (Plant, Goss et al. 2013).

Lower-to-middle income countries may not benefit from current diagnostic algorithms if their basic diagnostic techniques lag behind the protocols used in HICs (Bell, Mall et al. 2020). It is undeniable that great strides have been made to improve the diagnosis and care of CF patients. However, this progress is largely limited to HICs. Not only are the reviews limited to these regions, but their recommendations for NBS tend to overlook African countries. The reasons for this seem to range from the available statistics for CF incidence in these countries to the belief that NBS is not worthwhile on the continent due to lack of resources to deal with the consequences of a positive finding. Unfortunately, this carries implications for the CF populations in these countries who will not benefit from the progress made in the rest of the world. A detailed health economic analysis is required to assess the true return on investment of NBS versus managing patients who are diagnosed late.

### *1.2. Incidence*

Cystic fibrosis was originally thought to be a genetic disorder reserved for patients of “European ancestry” (Mutesa and Bours 2009). However, this has been disproven as a growing number of patients are diagnosed with CF in countries across Africa, Asia, the Middle East and South America (Bell, Mall et al. 2020). The incidence of CF in Europe is around 1/3000 to 1/6000, which subsequently confers a high carrier rate in this population (Scotet, Gutierrez et al. 2020, Scotet, L'Hostis et al. 2020). The incidence of CF in South Africa remains elusive; however, the estimation of incidence is improving with the implementation of a national registry (Zampoli On Behalf Of The Msac 2019). In 2020, the incidence of CF in South Africa was estimated to be 1:3000 in people of European origin, 1:10300 in mixed race individuals and 1:14000 in black South Africans (Padoa, Goldman et al. 1999, Westwood, Henderson et al. 2006, da Silva Filho, Zampoli et al. 2020).

Modelling of incidence and prevalence of CF is challenging. Improved treatment increases the survival of patients with CF and thus increases the prevalence in the population (Bell, Mall et al. 2020). Additionally, immigration from countries with a lower incidence or undetected variants complicates detection since these variants are not necessarily included in screening protocols (Bell,

Mall et al. 2020). Furthermore, ethnic-specific birth rates, availability of preconception carrier screening and variable registry recording also impact the accurate modelling of CF incidence and prevalence (Bell, Mall et al. 2020). The prevalence of CF is also difficult to determine due to differential quality of literature and patient registration in different countries (Mirtajani, Farnia et al. 2017), as well as lack of thorough NBS data (Bell, Mall et al. 2020, Scotet, L'Hostis et al. 2020). Furthermore, there is a higher frequency of registered CF patients based in Europe than elsewhere and it is speculated that this is due to awareness and access to healthcare (Mirtajani, Farnia et al. 2017), rather than a higher incidence of CF cases. The actual frequency of CF in African ethnolinguistic groups, including those in South Africa, is difficult to determine accurately due to lack of access to diagnosis, misdiagnosis of diseases with similar symptoms that plague Africa, and the high infant mortality rate (Padoa, Goldman et al. 1999).

### *1.2.1. Misdiagnosis*

Recently published registry data suggests that many infants in South Africa are dying of CF that is incorrectly diagnosed as malnutrition (due to undernutrition) or infectious disease, and that NBS could prevent this (Zampoli, Verstraete et al. 2021). CF is likely remaining undiagnosed and misdiagnosed in African countries as the early symptoms can be confused with the symptoms associated with severe acute malnutrition (previously known as protein-energy malnutrition), tuberculosis, chronic pulmonary infections and HIV/AIDS (Mutesa and Bours 2009, Bell, Mall et al. 2020). Many infants in LMICs are born into poverty, and are faced with malnutrition that also presents with impaired pancreatic exocrine function and impaired development (Bhutta, Berkley et al. 2017). Malnutrition is a large contributor to child death in Africa, with about 45% of all child deaths having been associated with malnutrition (W.H.O. 2022). Additionally, CF commonly presents with malnutrition stemming from pancreatic insufficiency (McCarthy, O'Carroll et al. 2015). In South Africa, up to 37% of children present with severe malnutrition as a result of likely undernutrition when they are diagnosed (Zampoli, Verstraete et al. 2021). Thus, it is vital to provide an early, definitive CF diagnosis so that the caregivers of these infants can intervene with appropriate and rigorous nutritional intervention needed to combat the malnutrition from which their infant is suffering, and thereby improve long term nutritional status and overall outcome (Farrell, Kosorok et al. 2001).

### *1.2.2 Decline in incidence*

It is likely that preconception carrier screening and prenatal diagnosis would lead to a decrease in the incidence of CF; however, this requires large population studies to be confirmed (Bell, Mall et

al. 2020). A long-term study in Brittany, France was conducted to monitor the incidence of CF over 35 years (Scotet, Dugueperoux et al. 2012). The area is known to have a particularly high incidence of CF, but the incidence has declined by 40% since the late 1970's. The study sheds light on the influence of health policy on the incidence of CF as the observed trend coincides with the implementation of prenatal diagnosis and NBS. Prenatal diagnosis in this region is recommended for known carriers during pregnancy. CF is diagnosed prenatally through the presence of an echogenic bowel. The authors quantified that 35.8% of the total decline in incidence could be accounted for when including pregnancy termination following prenatal diagnosis. Although this study focuses on a European population, it is notable that the health policies available as well as the characteristics of the population contribute to the incidence of CF in a population. The effect of health policy and access to healthcare on the incidence of CF in a population may extend to countries where healthcare is less accessible and cultural stigmas often prevent effective care (Nyblade, Stockton et al. 2019).

The reasons for the declining trend in incidence over time in countries with established CF populations include demographic factors, implementation of health policy for prevention, and cultural influence (Scotet, L'Hostis et al. 2020). The demographic changes include population admixture, decreasing consanguinity and decreasing fertility. The implementation of genetic-based health policies is also said to influence the incidence of CF through prevention using prenatal diagnosis, pre-implantation diagnosis, family testing, prenatal screening, and population carrier testing (Scotet, L'Hostis et al. 2020). It should be noted here that many of these strategies have been difficult to implement in LMICs. Cultural influence is also a factor as it impacts on access to care and on attitudes to health-related practices including genetic testing, prenatal diagnosis and pregnancy termination. The above factors are thus said to impact on the trend in incidence and vary by region and population. Furthermore, many areas which have observed a decline in incidence are those that have implemented prenatal or population carrier screening (Scotet, L'Hostis et al. 2020).

### *1.3. Population allele frequencies*

Knowledge of CFTR variant distribution in Africa remains incomplete, despite a growing number of studies across the continent. It was found through retrospective review that most studies are from Northern and Southern Africa, with F508del being the most common variant with several other variants overlapping with studies on European, Middle Eastern, Arabian African and American populations (Stewart and Pepper 2016). Some of the challenges faced when investigating



variant distribution in Africa include the significant diversity brought by intra- and inter-country migration (Campbell and Tishkoff 2008). The implications for molecular diagnostic testing of CFTR variants in CF patients from diverse populations have also been described (Schrijver, Pique et al. 2016). Among these implications are the racial-ethnic disparities in the CFTR variant spectrum and the subsequent low sensitivity of screening and molecular diagnostic tests that leaves these patients at risk for later identification of CF (Schrijver, Pique et al. 2016, Pique, Graham et al. 2017). There was also a recent appeal for researchers to study the full CFTR gene in South Africans of various ancestries and worldwide so that the molecular diagnostic tests could be improved and allow for an appropriate gene panel to be implemented (Wonkam 2016). It is hoped that this will improve carrier detection in South Africans in African ethnolinguistic groups.

To this end, the frequencies of CFTR variants in populations in African ethnolinguistic groups are starting to be elucidated and it is likely that there are variants that originated in Africa rather than originating through admixture with populations of European origin (from which F508del is believed to have been introduced), such as 3120+1G>A (Padoa, Goldman et al. 1999). It is further speculated that when more CF patients are diagnosed, the frequency of the 3120+1G>A variant will be accurately determined and that other CFTR variants will be identified, thereby improving estimates of carrier frequency and prevalence of CF in Africa (Padoa, Goldman et al. 1999). Carrier frequencies for specific variants are also variable in different populations, with unexpected variant frequencies being found in an African American cohort (Monaghan, Bluhm et al. 2004). Further studies are needed to determine the incidence of variants that are not currently included in current panels and that a review of the allele frequencies in different populations is necessary before revising the current panels. The authors also mentioned the findings by Bobadilla *et al.* that speculated that extending the panel by six common African American variants would improve the detection rate by 1.2% for this population (Bobadilla, Macek et al. 2002). Furthermore, there were significant ethnic differences in allele frequency when screening 364 890 individuals from the US, and that identifying the variants that are limited to specific ethnicities is necessary for thorough screening of CF (Rohlf, Zhou et al. 2011). A substantial proportion (22.7%) of the alleles identified in African Americans are not part of the standard ACMG/ACOG panel and four variants in this group have not been found to be present in any other ethnicities (Rohlf, Zhou et al. 2011). Furthermore, the ACMG/ACOG threshold of >1% overall allele frequency results in population-specific variants being excluded from the panel, despite being more common in certain populations. The inclusion of these variants would improve the detection rate significantly (Rohlf, Zhou et al. 2011).

Accurate screening of CFTR variants is a challenge due to differences in the prevalence of these variants across populations as well as ethnic heterogeneity and increasing population admixture (Grody, Cutting et al. 2001). Even within European populations, a heterogeneous distribution of CF allele frequencies can be observed across geographical locations, and the allele diversity has been shown to vary significantly between populations with the same disease (Lao, Andres et al. 2003). The sensitivity of neonatal screening has also been criticised when the variant screening is limited to the F508del variant, as certain populations have higher frequencies of other variants (Bobadilla, Macek et al. 2002). Though F508del is said to be the most common variant associated with CF, there are other variants which are more common in different populations. This may be due to a founder effect as the variants have had a long time period to spread across different groups (Bobadilla, Macek et al. 2002). Examples of these variants include G542X (Loirat, Hazout et al. 1997), N1303K, and G551D (Cashman, Patino et al. 1995). In addition, some variants that are less common in the European cohort are more prevalent in other countries/regions. As such, they may be more significant in screening programs and a thorough knowledge of regional variants is necessary to achieve high sensitivity (Bobadilla, Macek et al. 2002). Furthermore, the targeted CFTR screening panels and variant databases are biased towards European variants, and fail to incorporate a great proportion of likely deleterious variants that have been found in other geographical populations (Lim, Silver et al. 2016).

Investigation of the molecular nature of CF in African patients has been conducted in 12 African countries, revealing 79 variants (Stewart and Pepper 2016). The “common” F508del variant, typically found at a frequency of 70-90%, was not found in patients from four countries (Sudan, Rwanda, Cameroon and Zimbabwe), but some studies did not include this variant in their screening protocols. F508del variant frequency was 48% in the remaining eight countries (Stewart and Pepper 2016). This suggests that there may be other causative variants that are present at a higher frequency in these populations. From the 12 countries reporting CFTR variants on 2344 chromosomes, the most frequently reported variants are F508del, 3120+1G>A, G542X, N1303K, W1282X, E1104X, 711+1G>T, 3272-26A>G and 394delTT (Stewart and Pepper 2016).

The South African Cystic Fibrosis Registry (SACFR) has furthermore confirmed that genotype is correlated with ancestry in the South African CF cohort (Zampoli, Verstraete et al. 2021). This can be seen by the fact that white South Africans and patients of mixed ancestry most commonly present with F508del, as expected. The second most common variant in patients of mixed ancestry

is the 3120+1G>A variant, which is also the most common variant in the black South Africans recorded in the registry. There are still many patients lacking a full molecular diagnosis, with 11% of the patients having an incomplete or unknown genotype after screening, and majority being of mixed ancestry or black South Africans (Zampoli, Verstraete et al. 2021). This not only provides evidence for CF presence on the continent, but also for the need to investigate the variants specific to each population (Stewart and Pepper 2016).

Though many countries have been able to achieve variant detection rates of over 95%, only two African countries have enough data to even report detection rates: Algeria with a detection rate of 60-69% and South Africa with a rate of 70-79% (Consortium 1994, Ikpa, Bijvelds et al. 2014, Stewart and Pepper 2016). It is proposed that a more thorough molecular investigation of CFTR variants in individuals from African ethnolinguistic groups be conducted so that the variant detection rate can be improved and thereby improve patient care (Stewart and Pepper 2016, Zampoli, Verstraete et al. 2021). The challenge of identifying South African-specific variants can be approached using next generation sequencing (NGS), whole exome sequencing (WES), and/or targeted sequencing of the CFTR gene (Van Rensburg, Alessandrini et al. 2018). Furthermore, it has been suggested that differences in CFTR variant frequencies between populations might be determined in the future using population-based genomic variant frequencies from international genome projects (Bell, Mall et al. 2020). As NGS becomes cheaper, the challenge of low variant detection frequency might be overcome and population differences may be accounted for, especially when looking in multicultural cities (Davis, D'Odorico et al. 2013, Loukas, Thodi et al. 2015, Bell, Mall et al. 2020).

### *1.3.1. Evolution of the high CF carrier frequency in the European population*

There remains a high frequency of deleterious CFTR alleles in the European population despite the recent start of a decline in incidence (Scotet, L'Hostis et al. 2020). This high allele frequency is proposed to have been propagated via a high variant rate, founder effect, genetic drift and/or balancing selection/heterozygote advantage (Bertranpetit and Calafell 1996). Heterozygote advantage seems to have gained some evidence, as it is supported by the high incidence of the disease in Europe which coincides with its population history and migration patterns (Bertranpetit and Calafell 1996). Furthermore, some evidence supports heterozygote advantage against infectious diseases such as typhoid fever, secretory diarrhoea including cholera (SDC), and tuberculosis (Anderson, Allan et al. 1967, Hansson 1988, Chao, de Sauvage et al. 1994, Pier, Grout et al. 1998, van de Vosse, Ali et al. 2005). However, it may be that only resistance of CF carriers

to tuberculosis had the necessary selective pressure capable of generating the observed allele persistence and high incidence of CF in Europe (Poolman and Galvani 2007). This is further supported by a negative association between CF carriers and TB incidence rate (Bosch, Bosch et al. 2017).

### *1.3.2. Evolution of the CF carrier frequency in African populations*

The p.Phe508del variant is the most common in the European CF population. This is suspected to be the result of a population bottleneck combined with selective advantage (Bertranpetit and Calafell 1996). African populations did not have the same migratory patterns at the time but may have undergone similar selective pressures by the same pathogens. Thus, it may be that different CFTR variants provided a heterozygous advantage for both European and African populations, but only some underwent a bottleneck in which the diversity of variants was reduced and a few variants remained and increased in frequency over time (Bobadilla, Macek et al. 2002). This is supported by the known extent of genetic diversity on the African continent. Variants that are found more commonly in smaller populations (including some African ethnolinguistic groups) are thought to have arisen in a similar manner to p.Phe508del through a founder-effect, and spread in the populations over time (Bobadilla, Macek et al. 2002). However, these seem to be more prevalent in northern parts of Africa and Mediterranean populations as a result of historic migration patterns (Bobadilla, Macek et al. 2002).

The heterozygote frequency of some deleterious CFTR alleles in African groups is surprisingly higher than in the rest of the world, including Europe (Lim, Silver et al. 2016). Though these variants are proposed to be non-disease causing as homozygotes, it is speculated that they contribute to disease in compound heterozygotes (Lim, Silver et al. 2016). Lim, *et al.* have speculated that these deleterious alleles are only causative of CF in compound heterozygotes, which would explain the lower incidence of CF in African countries (Lim, Silver et al. 2016). However, this assumes that the estimates of CF incidence in Africa are correct and that it is not highly prevalent on the continent. If carriers are protected from typhoid fever, cholera and/or tuberculosis, endemic to many African countries, then it follows that a high frequency of deleterious alleles will have evolved in the population and the incidence of the disease is higher than is currently estimated. To address this question, further studies are needed to identify the deleterious allele frequency in the “normal” population, and diagnostic protocols will need to be improved.

#### 1.4. *CFTR variants in Africa*

##### 1.4.1. *A timeline of CFTR variants in Africa and surrounding geography*

In 1991, Lucotte *et al.* described the frequency of the common F508del variant in Algeria by evaluating 24 Algerian patients using allele-specific polymerase chain reaction. The frequency of this variant was 0.43. This is lower than the estimation for Northern Europe at the time and consistent with a North-South decreasing gradient of this variant across geographical distributions observed (Lucotte, Barre *et al.* 1991). The next year, Loumi *et al.* described a homozygous variant of the M470V polymorphism and deletion in exon 10 of the *CFTR* gene in an Algerian child displaying severe cystic fibrosis symptoms (Loumi, Cuppens *et al.* 1992). In 1995, Kerem *et al.* published a study on the incidence of CF and variant distribution in different Jewish ethnic groups across Israel (Kerem, Kalman *et al.* 1995). The frequency of CF in Ashkenazi Jews was found to be like that of most European populations at the time. However, the non-Ashkenazi Jewish population displayed considerable variability depending on the country. It was also found that the disease was caused by different variants in each ethnic group.

In 1997, Tzetis *et al.* described the landscape of *CFTR* variants in the Greek population and discovered five novel variants (Tzetis, Kanavakis *et al.* 1997). The same year, Macek *et al.* described an increase in the detection rate after identifying common CF variants in African American patients (Macek, Mackova *et al.* 1997). The coding and intronic sequences of African American patients were evaluated for variants in *CFTR*. The 3120+1G->A variant was found at a relatively high frequency and was also found in a native African patient. It was said that African American patients have a characteristic variant profile of variants common in the population originating from Africa and that including these variants in screening drastically improved the detection rate of CF in African American patients (Macek, Mackova *et al.* 1997). Also in 1997, *CFTR* variants in Saudi Arabian children with severe CF were studied (el-Harith, Dork *et al.* 1997). Again, the coding and intronic regions of the *CFTR* gene were evaluated. Two novel variants as well as several prominent variants were found, namely 1548delG, 406-2A->G, 3120 + 1G->A, N1303K and 1548delG.

Two years later, the frequencies of variants in populations in African ethnolinguistic groups were studied (Padoa, Goldman *et al.* 1999). Carrier screening was performed for the variants found in African CF patients: 3120+1G-->A, D1270N, A559T, S1255X and 444delA. The study predicted the incidence of CF in South African black patients to be between 1 in 784 and 1 in 13924 births. It was furthermore suggested that reasons for the low detection of CF in this group included malnutrition, tuberculosis, and the additional diseases commonly found in African patients. In

1999, Lissens *et al.* published a study on the frequency of a splice variant in Egyptian males with CBAVD (Lissens, Mahmoud *et al.* 1999). Five years later, Feuillet-Fieux *et al.* evaluated CFTR variants in black cystic fibrosis patients (Feuillet-Fieux, Ferrec *et al.* 2004). Four novel variants were reported: IVS2+28, 459T>A, EX17a\_EX18del, and IVS22+IG>A.

Naguib *et al.* described CF detection and CFTR variant analysis in Egyptian children in 2007 (Naguib, Schrijver *et al.* 2007). They concluded that the incidence of CF was underestimated and that large studies needed to be conducted in Egypt to determine more accurately the incidence and the molecular and clinical patterns of CF in the population. Messaoud *et al.* then speculated that additional variants might be found in the promoter or intronic regions of the CFTR gene of the Tunisian Mediterranean population (Messaoud, Verlingue *et al.* 1996). In 2009, the CFTR variant spectrum in 68 Tunisian CF patients was studied (Fredj, Messaoud *et al.* 2009). Almost all patients had at least one variant and several were noted: F508del, E1104X, N1303K, 711+1T>G, W1282X, G542X, R1158X, 4016insT and R785X. Three novel variants were also identified: I1203V, 1811+5A>G, and 4268+2T>G. The study also highlighted the need for complete scanning of CFTR to ensure efficient screening of patients in North Africa. In 2008, Lakeman *et al.* evaluated Turkish and North African immigrants with CF living in Europe and found the sensitivity of the common CFTR variant panels to be too low for appropriate screening in multi-ethnic societies (Lakeman, Gille *et al.* 2008).

In 2009, Mutesa *et al.* evaluated the CFTR and ENaC variants of 60 Rwandan patients with CF-like symptoms (Mutesa, Azad *et al.* 2009). Three CFTR variants, one novel missense variant (p.A204I), and one 5T/7T variant were identified in five patients. Two years later, the first study of CFTR variants in Libyan CF patients was published (Hadj Fredj, Fattoum *et al.* 2011). The study evaluated the coding and intronic regions of the CFTR gene in 10 Libyan CF patients and found four variants, namely F508del, c.1670delC, N1303K and E1104X. The following year, another study on CFTR variants in 37 Egyptian patients was published (El-Seedy, Pasquet *et al.* 2013). Four variants were found: c.1418delG, c.2620-15C>G, c.3877G>A as well as the novel variant c.3718-24G>A. Six polymorphisms were also described: M470V, P1290P, c.2562T>G, c.1584G>A, c.4389G>A, c.869+11C>T. Only the M470V polymorphism had previously been described in this population. In 2013, the novel frameshift variant 3729delAinsTCT was discovered in a Tunisian cystic fibrosis patient (Hadj Fredj, Boudaya *et al.* 2013). Later, Ibrahim *et al.* described CF in Sudanese children (Ibrahim, Fadl Elmola *et al.* 2014). Only three of the patients underwent variant analysis and were confirmed to have CFTR gene variants.

Recently, Phillips and colleagues evaluated known pathogenic variants in patients of European and African ancestry with chronic pancreatitis, including pathogenic variants in CFTR (Phillips, LaRusch et al. 2018). They found that the variants were less common in African ethnolinguistic groups than in European patients, and concluded that the complex risk factors for pancreatitis in African groups requires more evaluation.

#### *1.4.2. A timeline of CFTR variants in South Africa*

The frequency of the common F508del variant in South African CF patients was first studied in 1992 (Denter, Ramsay et al. 1992). The frequency in patients of European descent was found to be 0.81; however, this variant was not found in the one black and one Indian patient studied. Another study found similar results and noted a 0.53 frequency of the variant in patients of mixed ancestry (Herbert and Retief 1992). Later, known CF variants in 140 white South African families were investigated (Goldman, Jenkins et al. 1994). Again, F508del was the most frequent variant, followed by G542X. Four additional variants were found at low frequency: R553X, S549N, 621+1G>T and N1303K.

In 1996, Carles and colleagues published the results of their investigation of CFTR variants in three black South African patients with CF (Carles, Desgeorges et al. 1996). The 3120 + 1G->A variant, known to be prevalent in black populations, was found in all three patients with one patient being homozygous for the variant. The G1249E variant was also found, as well as a novel in-frame deletion. Later, variants evaluated across three South African populations revealed that F508del occurred with the highest frequency in the white CF patients, followed by 3272-26A>G, 394delTT and G542X (Goldman, Labrum et al. 2001). In the admixed population, lower frequencies of the common variants were observed, namely F508del (0.43) and 3120+1G>A (0.29). In the black population, the 3120+1G>A variant occurred at an estimated frequency of 0.46. It was confirmed that although screening can detect variants in the white population with relatively high sensitivity, the variants in black and admixed populations require further evaluation.

In 2003, the diagnosis of CF in South Africa was evaluated using gene panels designed for the study (Goldman, Graf et al. 2003). The aim was “to improve the sensitivity and efficiency of diagnostic testing for CF in South Africa” by designing panels of variants specific to the different population groups of South Africa. It was found that a large proportion of patients could have their diagnosis confirmed by the detection of CFTR variants (Goldman, Graf et al. 2003). This is

important, as sweat tests are not always accessible or reliable in South Africa. The distribution of variants was also variable for the different populations (Goldman, Graf et al. 2003). This confirmed that gene panels need to be designed according to these differences. Although a large proportion of the white South African patients could be confirmed, only 21% of the black CF patients were confirmed using the gene panels, highlighting a need to investigate the variants in this group (Goldman, Graf et al. 2003). Westwood *et al.* referenced the panel designed by Goldman and colleagues in an editorial (Westwood, Henderson et al. 2006). It was said that although this panel had so far shown the best results in South African patients, the variant detection rate still required significant improvements and warranted further study.

In 2008, des Georges and colleagues investigated the molecular nature of unidentified CFTR alleles in six samples from a previous study (des Georges, Guittard et al. 2008). They tested for large rearrangements and found a novel deletion in a black South African patient who was heterozygous for the 3120+1G>A variant, and subsequently designed a test to detect it. However, it was found that exon CNVs of CFTR are not likely to have a large contribution to the variant mechanism in CF in black and coloured South African patients and thus should not be included in the gene panels for these patients (De Carvalho and Ramsay 2009). In 2013, the phenotypic expression of the 3120+1G>A variant in 30 black and mixed race children in South Africa was evaluated (Masekela, Zampoli et al. 2013). 47% of the participants were homozygous for the variant, and a further 53% were found to be heterozygous. It was found that malnutrition and failure to thrive were the most common clinical features of CF in the participants.

The recent advances in molecular diagnosis and recognition of the wide spectrum in CF clinical manifestations has led to the revision of the diagnostic nomenclature and criteria used in South Africa (Zampoli On Behalf Of The Msac 2019). Additionally, the characterisation of CFTR variants in CF patients is vital for treatment (Zampoli On Behalf Of The Msac 2019). Furthermore, the introduction of a critically important local CF registry will continue to improve the knowledge of CF in South Africa and help to identify and improve underperforming aspects of CF care in the country (Zampoli On Behalf Of The Msac 2019).

### 1.5. Registry capture

The lack of CF patient registries in addition to misdiagnosis of CF on the African continent continues to obscure the true incidence and prevalence of this disease in African populations. It also means that the true population variation in carrier frequency cannot be calculated and that the



distribution and penetrance of CFTR variants among Africans remains unknown (Bell, Mall et al. 2020). The use of CF registries in African and Asian countries is severely lacking and it is likely that more than 50% of these countries have no registry whatsoever. This negatively impacts investigation of the population-associated risk estimates, and serious efforts are thus needed to improve CF registry capture globally (Mirtajani, Farnia et al. 2017). Additionally, the quality of national registries will impact CF research and care across the board. Registry data can also allow targeted therapies to be implemented for patients who are identified to be at higher risk of mortality (Bell, Mall et al. 2020). Registry data will also shed light on the overall health of CF populations and the complications found within populations of older patients (George, Banya et al. 2011, Goss, Sykes et al. 2018). Furthermore, the lack of well-established registries in LMICs may lead to an underestimation of the number of CF patients in these countries and worldwide (Bell, Mall et al. 2020). Lastly, most registries tend to record CF according to the specific country instead of ethnicity (Bell, Mall et al. 2020). Thus, many patients are unable to benefit from information relating to population-specific differences in diagnosis, consanguinity-derived comorbidities, diet, environment, socio-economic factors and access to specialised care (Bell, Mall et al. 2020).

In the case of South Africa, the implementation of a national registry has drastically improved the information available. It has helped to gauge the state of CF diagnosis, disease, and treatment in the country. Though South Africa is considered a middle-income country, there are various resources available for patients in the country. The first annual report of the SACFR (South African Cystic Fibrosis Registry) has provided a vast source of valuable data for 447 patients recorded in the registry (Zampoli, Verstraete et al. 2021). There are 16 CF care centres that participate in the national registry and that provide specialised care for CF patients. The evaluation of CF disease in this population demonstrates deviations in clinical presentation from the “classic” phenotypic characteristics described in European cohorts. It has been observed that more than 80% of patients did not present with meconium ileus, a classic signature of CF, and 88.4% of CF patients present with pancreatic insufficiency. Most notably, South African patients seem to present with worse overall lung function and nutrition than patients in Europe and North America. It has also been observed that although allele frequencies of the F508del variant and the 3120+1G>A variant were relatively high (63.1% and 9.5%, respectively), up to 7% of the allele variants were unknown, and that 11% of the patients were left with an incomplete molecular diagnosis. Furthermore, the report has provided valuable information regarding treatment. Though many traditional treatments are widely used (such as antibiotics, inhaled and oral steroids,

inhaled hypertonic saline, etc.), the limited data available points to virtually no patients receiving modulator therapies (Zampoli, Verstraete et al. 2021). This is especially relevant to the patients with homozygous or heterozygous F508del variants, as these patients stand to benefit greatly from the treatments that are becoming well established in European CF populations where this variant is common (Zaher, ElSaygh et al. 2021).

#### *1.6. Early diagnosis: Improved outcome, life expectancy and quality of life*

In high-income countries with established CF registries and care centres, great improvements in survival and life expectancy have been observed. The life expectancy of CF patients in Europe has risen to more than 40 years (Kerem, Viviani et al. 2014, MacKenzie, Gifford et al. 2014). In contrast, the median survival age of South African patients was only 20.8 years in 2008 (Westwood 2008), with the median age of the SACFR cohort being 14.7 years in 2018 (Zampoli, Verstraete et al. 2021). Furthermore, the median age of diagnosis in some European countries that do not utilise NBS is 5.0 months (de Monestrol, Klint et al. 2011), whereas the median age at diagnosis in South Africa, which does not have an established NBS programme, is 7.6 months (Zampoli, Verstraete et al. 2021). In contrast, the median age at diagnosis in European countries that utilise NBS is 1 month (Tridello, Castellani et al. 2018).

Various factors contribute to the prognosis and outcome of CF patients. Poor clinical outcomes are attributed to factors that include early and severe infection, insufficient adherence to treatment, and low socioeconomic status (Bell, Mall et al. 2020). Furthermore, CF patients who are malnourished due to CF-related pancreatic insufficiency are typically more likely to present with failure to thrive, steatorrhea and fat soluble vitamin deficiency (McCarthy, O'Carroll et al. 2015). Patients who are diagnosed later will not benefit from early nutritional intervention and subsequently suffer from malnourishment, failure to thrive and steatorrhea (McCarthy, O'Carroll et al. 2015). Malnourished and underweight CF patients are also at higher risk of lung disease progression and death (Kerem, Viviani et al. 2014, McCarthy, O'Carroll et al. 2015). Furthermore, late diagnosis also places enormous psychological and financial strain on the parents of CF patients. The families of late-diagnosed CF patients experience anxiety, trauma and self-doubt because of a lack of adequate health care and many have retrospectively reflected that they feel NBS would have significantly improved the psychological impact on their family and improved their feelings towards the medical sector, as well as prevented the pain they experience (Kharrazi and Kharrazi 2005). In contrast, earlier detection and diagnosis of CF before the onset of

symptoms improves long-term outcome (Dankert-Roelse and te Meerman 1995, Waters, Wilcken et al. 1999).

To this end, NBS enables earlier diagnosis of CF and allows nutritional intervention to be performed early on, which contributes to significant improvements in patient outcome and is correlated with lung function (Steinkamp, Rodeck et al. 1990, Farrell, Kosorok et al. 2001, Martinez-Costa, Escribano et al. 2005, Stephenson, Mannik et al. 2013). However, access to healthcare has been posed as a significant challenge to CF diagnosis and treatment in lower income countries. South Africa shows great disparity between its private and public healthcare systems, and ancestry as well as socioeconomic status have been shown to significantly affect outcome (Zampoli, Verstraete et al. 2021). In addition, earlier diagnosis and specialised treatment has led to the recent increase in average survival of CF patients in LMICs, posing an additional challenge of providing care to adults by physicians who mainly specialise in paediatrics (Bell, Mall et al. 2020, da Silva Filho, Zampoli et al. 2020). The issue of financial constraints as well as the overwhelming burden of other diseases has also created a unique challenge for CF care in lower income countries (da Silva Filho, Zampoli et al. 2020). It has been said that the outcome and care of South African patients can be greatly improved if research is driven in the direction of identifying CF variants relevant to all SA population groups, maintaining a country specific CF database/registry, establishing a solid foundation for a new-born screening programme and exploring novel means through which a positive clinical diagnosis could be made (Van Rensburg, Alessandrini et al. 2018).

Furthermore, patients from LMICs are often still faced with several challenges despite earlier recognition (Bell, Mall et al. 2020). The sweat chloride test is still considered to be the best tool for CF diagnosis, confirmation of the relevance of CFTR variants, and validation of CF-negative patients (Bell, Mall et al. 2020, Zampoli, Verstraete et al. 2021). However, high-quality sweat testing is difficult in LMICs and availability of tests is subject to availability of resources, with access to reliable sweat testing being limited to a few main cities in South Africa. New-born screening can be done in South Africa using an IRT/DNA protocol akin to that which is currently used in Europe and HICs. However, there are still some challenges including significant cost, reliance on effective sweat chloride testing, and the use of gene panels with limited variant detection rate in diverse populations.

### *1.7. Definition of disease and genotype-phenotype correlation*

Another challenge to diagnosis is that the definition of disease is currently biased towards European-typical CF phenotypes (Lim, Silver et al. 2016). This needs to be re-evaluated to be more considerate of variable phenotypes and CFTR alleles that are present in diverse populations around the world. This will facilitate the improvement of both CF diagnosis as well as the development of appropriate screening panels in these populations. While screening for variants specific to the respective ethnicities will improve the detection rates of variants in these populations (Monaghan, Bluhm et al. 2004, Rohlf, Zhou et al. 2011), sufficient data on the genotype-phenotype correlations in rare CFTR variants is needed for effective application of CFTR modulators (Gentzsch and Mall 2018).

Correlation between genotype and phenotype has also been said to differ between organ systems (Mickle and Cutting 2000). Furthermore, patients show disparity in genotype-phenotype correlation in different populations and a distinct phenotype presentation may occur with its own correlated CFTR subtype (Hamosh, FitzSimmons et al. 1998, Padoa, Goldman et al. 1999). Thus, novel variants detected by NGS may end up being classified as variants of unknown clinical significance (Steward, Parker et al. 2017). Most recently, the phenotype, genotype, nutrition and pulmonary function of black South African children with CF was compared to those with the F508del genotype (Owusu, Morrow et al. 2020). The 3120+1G>A variant was found most frequently, and the patients were more malnourished than the controls. The patients presented with neonatal bowel obstruction less frequently while the nutrition, while pulmonary function and mortality were similar between both groups.

The CFTR2 project has been established in part to alleviate the difficulty in predicting phenotypic outcome associated with rare variants by thoroughly annotating variants with their associated clinical features. However, the database is currently over-representative of variants from individuals of European origin, with 95% of the database comprising variants from this group (Lim, Silver et al. 2016). The under-representation of the genetic variation in global populations still poses challenges for NBS and carrier testing, and carries implications for the accuracy of incidence and prevalence estimates of CF (Kabra, Kabra et al. 2006, Bell, Mall et al. 2020). Compounding the problem, the core screening panel recommended by ACMG is biased towards European disease presentation. This is through including ‘classic disease presentation’ as well as a variant frequency cut-off of 0.1 or more in its criteria for selection of variants (Grody, Cutting et al. 2001, Watson, Cutting et al. 2004, Lim, Silver et al. 2016). Thus, these recommendations are

inappropriate for populations with rare variants as well as populations that may have non-classical disease presentation. This compounds the difficulty in obtaining effective carrier detection in global populations and may contribute to under-diagnosis in these populations. As more rare variants are discovered, a need for personalized treatment using CFTR modulators emerges. This approach has been demonstrated using a case study of unconventional CF presentation in an African woman (McCravy, Quinney et al. 2020). The mutational profile (c.1373delG and c.571T>G) and its clinical manifestation were investigated, followed by testing of appropriate CFTR modulators. Extensive definition of the effect of genotype on CFTR function and late presentation of CF in an African American woman was provided and provides an example of how the field will need to evolve to effectively treat CF patients with ethnolinguistically diverse ancestries.

### *1.8. Challenges to implementation of NBS*

In HICs, adding DNA analysis to the screening protocol has removed the need for a second blood sample when coupled with the immunoreactive trypsinogen assay which lowers anxiety in the family and allows for earlier diagnosis. Though it is too early to determine the long-term impact of NBS for CF on survival in most countries, it is likely to maximise survival through early diagnosis and management (Scotet, L'Hostis et al. 2020). African countries stand to benefit greatly from NBS for CF by helping to overcome the problem of misdiagnosis (Mutesa and Bours 2009). Furthermore, targeted treatment of the underlying cause, instead of symptomatic treatment, has grown in popularity as the genetic mechanisms are being discovered and precise CFTR modulators are being introduced (Gentzsch and Mall 2018). These rapidly evolving treatment options have been speculated to likely change the way that screening is approached (Bell, Mall et al. 2020).

Recent reviews focus primarily on European advancements with the authors speaking to the situation worldwide without including a thorough assessment of the state of CF care in Africa (Bell, Mall et al. 2020, Scotet, L'Hostis et al. 2020). Despite this oversight, they provide insight into the evolution of CF in HICs, which may one day have applications in Africa. The implementation of NBS has improved the estimation of the incidence of CF which was previously biased by under-diagnosis and under-reporting (Scotet, L'Hostis et al. 2020). However, this is still a challenge in LMICs such as those in Africa (Bell, Mall et al. 2020). Furthermore, many reviews have evaluated whether NBS for CF might be worthwhile (Scotet, Gutierrez et al. 2020). However, many of them fail to mention African countries despite the growing wealth of knowledge that is emerging here. African countries are usually omitted because the incidence of CF and the available resources are

considered too low to warrant CF care centres in these regions. Additionally, most African countries lack the capacity to diagnose CF as sweat testing is a highly technical procedure. However, there were several challenges that had to be overcome by European countries when implementing and improving NBS programmes (Scotet, Gutierrez et al. 2020), that will likely be applicable when considering NBS in Africa. Ensuring that resources such as laboratory sufficiency and follow-up care were available were some of the initial obstacles. Additionally, NBS is not considered worthwhile without the resources for sustained and efficient upkeep of the necessary elements with high delivery quality (Scotet, Gutierrez et al. 2020). Furthermore, a collaborative effort to address each of the issues including equitable access should be ensured in the region for which NBS is proposed, as well as sufficient genetic counselling to balance the psychosocial risks (Scotet, Gutierrez et al. 2020).

The European Cystic Fibrosis Society (ECFS) has provided guidelines for the implementation of NBS and has suggested that a minimum incidence of 1:7000 in a country indicates that NBS might be worthwhile (Castellani, Duff et al. 2018). This has, however, been criticized and it has been speculated that an incidence of 1:25000 might be more appropriate since there is a wide range of incidence data in various countries (Scotet, Gutierrez et al. 2020). Like Africa, the Latin American populations have great genetic diversity and therefore difficulty in determining the incidence of CF using molecular criteria. However, despite this complication and the lack of reliable data and diagnostic sensitivity, many still deem NBS to be worthwhile for CF in these countries (Scotet, Gutierrez et al. 2020). This is based on an “expected high number of cases and the late age in diagnosis” suggesting that NBS may enable earlier diagnosis and improved survival (Scotet, Gutierrez et al. 2020). This is likely also applicable to South Africa. Based on many of the recommendations provided by various reviews (Scotet, Gutierrez et al. 2020), African countries may find NBS for CF to be worthwhile if they are able to utilise a collaborative approach with dedicated staff and resources as well as sufficient access to diagnosis and follow-up care (Scotet, Gutierrez et al. 2020).

To make NBS for CF worthwhile in many African countries, especially South Africa, there are a few obstacles that need to be overcome, following the recognized recommendations (Scotet, Gutierrez et al. 2020). First, incidence needs to be thoroughly recorded in national registries, as has now been done in SA. Second, the sensitivity of screening panels needs to be markedly improved by including population-specific variants to achieve the recommended sensitivity of 95%. This may not be possible without NGS. Third, protocols for early diagnosis in these countries will need to be optimized which will require improvement of current testing methods in

the available facilities (such as sweat chloride tests, IRT/DNA protocols, and gene panels), or the development of new, innovative diagnostic methods. Lastly, infrastructure will need to be altered to accommodate monitoring of tests, diagnosed patients, treatment implementation and the involvement of a dedicated specialist CF team (Scotet, Gutierrez et al. 2020).

### *1.9. Evaluation of current CFTR variant screening panels*

Many reviews are available evaluating the clinical sensitivity of available variant panels for CF in diverse populations (Hughes, Stevens et al. 2016). For many of the variant panels investigated, the panels have the lowest sensitivity in the black population. For example, the *Illumina* MiSeqDx CF 139-Variant Assay improves the sensitivity in this population by 20%, as it has a more comprehensive list of variants that are included in the screening protocol (Hughes, Stevens et al. 2016). This panel's increased sensitivity, reliability, and tolerance towards impurities may make it suitable for NBS in a diverse population (Hughes, Stevens et al. 2016). Furthermore, approaches to screening for CFTR variants in an ethnically diverse population have been compared (Currier, Sciortino et al. 2017). The first approach is characterised by screening with the standard ACMG gene panel after an elevated IRT, known as a “second-tier” test. The second approach is also characterised by standard panel screening after elevated IRT; however, it includes additional “third-tier” screening using a population-specific panel (Currier, Sciortino et al. 2017). The latter is best in the case of a relatively diverse population and even the broadest of CFTR panels would miss 21% of cases if used alone. This is especially important in ethnically diverse populations where the panels are often missing variants that were previously unreported or novel (Currier, Sciortino et al. 2017). Thus, adding CFTR sequencing to the protocol or expanding the panel will likely be beneficial in under-represented populations (Currier, Sciortino et al. 2017). Lastly, research into the variants found in these under-represented populations is essential and should be accompanied by thorough registry capture (Currier, Sciortino et al. 2017).

The differential detection rates across geographical regions when using variant panels developed by ACMG (American College of Obstetricians and Gynecologists and American College of Medical Genetics 2011) were also evaluated with the goal of improving the detection rate in Italy, a country which has a relatively high level of genetic heterogeneity between the northern and southern regions (Lucarelli, Porcaro et al. 2017). The authors developed, validated and tested an NGS-based assay, and found that their panel of 188 variants had a detection rate of up to 95.6%. This relatively new approach is specifically suited to diverse populations as a larger number of variants can be screened. By using NGS, the first step is to check for variants with the population-

customized, validated 188-variant panel. If no variants are found (or only one), the whole CFTR region can be unmasked, and additional variants can be identified without any additional laboratory tests (Lucarelli, Porcaro et al. 2017). This technology is expected to improve genotyping of CFTR as it is fast, simple, and provides predictable identification of complex alleles (which are difficult to screen using conventional assays). However, though this technology is suitable for use in diagnosis, there is still a need for conventional panels in order to limit cost and time during the initial search for variants (Lucarelli, Porcaro et al. 2017).

Equitable diagnosis of CF has been evaluated, suggesting the use of NGS as an alternative to conventional screening (Shum, Bennett et al. 2021). The development of panels that are more inclusive of variants across different ethnicities provides a solution to the current disparity seen across populations (Bobadilla, Macek et al. 2002). However, the unavailability of data for the variants present in understudied populations is likely to make this difficult (Shum, Bennett et al. 2021). One solution for equitable diagnosis regardless of ethnic background is to implement complete gene sequencing of CFTR, which is becoming more cost-effective and is easily implemented through automation pipelines (Shum, Bennett et al. 2021). The authors argue that this approach is favourable for minority populations whose variants are underrepresented in panels, but then go on to say that most improvement will likely be seen in countries with sufficient infrastructure and accessible healthcare to facilitate effective disease management (Shum, Bennett et al. 2021). Furthermore, continued use of insensitive panels in ethnically diverse countries will further accentuate the inequality in healthcare, and that this needs to be addressed. The authors provide a solution to this by arguing that if a country includes these panels in their diagnostic protocol and no variants are found in a patient with high IRT, that the patient then be referred for further testing. Furthermore, the use of full gene sequencing will need to be evaluated regarding each individual healthcare system, as access to healthcare may be a higher priority than faster diagnosis (Shum, Bennett et al. 2021). Lastly, variants of unknown significance (VUS) continue to contribute hesitancy towards full gene sequencing, as it introduces uncertainty and complicates functional interpretation and genetic counselling (Sosnay, Siklosi et al. 2013).

The methods and feasibility of exome sequencing with *a priori* analysis restriction as a universal second-tier test in NBS has been evaluated (Ruiz-Schultz, Sant et al. 2021). This was to provide an alternative procedure for second-tier or confirmatory testing that is more affordable and scalable to new conditions. Variants from multiple datasets were used to improve interpretation within the pipeline. The pipeline was validated using NBS specimens representing four genetic disorders,



including CF. It was found that the pipeline achieved a 100% detection rate when validated with *in silico* data sets and 11% of the variants required manual curation. It was concluded that the pipeline is effective and allows for restriction of analysis to variants in single genes or full exome analysis if necessary. The sequence data quality and performance of the *Swift* accel-amplicon CFTR Panel have also been evaluated. This is an amplicon/library preparation kit that amplifies the CFTR gene using 87 amplicons and allows fast and affordable sequencing when combined with the *Illumina* MiSeq Nano kit v2 (Leung, Watson et al. 2020). This is an effective population screening method with a high detection rate, and the sequencing data has shown appropriately high coverage correlated with the GC content of each exon and almost 100% on-target reads. Furthermore, the data generated by this investigation should be considered when laboratories consider this method for carrier screening (Leung, Watson et al. 2020).

### 1.10. Conclusion

The incidence of CF in many African countries remains understudied and under-estimated. However, as health care systems improve in hospitals and clinics across the continent, more cases will emerge. It is likely that there is a high carrier frequency of deleterious CFTR alleles in African ethnolinguistic groups as a result of heterozygote advantage and the evolutionary pressures present. Thus, there is likely to be a much higher incidence of CF in the population than is currently diagnosed and recorded. Achieving adequate care for all CF patients in African countries is challenged by misdiagnosis, socio-economic status and access to healthcare, lack of registry data, genetic heterogeneity and diversity of CFTR variants, atypical disease presentation and progression, as well as available treatment options. Early diagnosis of CF is crucial as it has been shown that life expectancy, quality of life and treatment efficacy can be dramatically improved if CF is diagnosed early in life. Since sweat testing requires a high level of technical skill and is only available in major cities and tertiary hospitals in South Africa, molecular diagnostic protocols are favoured but rely on gene panels that are unsuitable for diverse populations. Diagnostic protocols are biased towards the clinical presentation of CF patients of European descent and variant databases and are under-representative of global genetic diversity. To achieve equitable diagnosis of CF globally, efforts must be made to address the over-representation of European CF population data and research.

Africa is addressing the under-representation of variant data found in patients with CF and with time this will serve to reinforce the need for CFTR molecular diagnostic protocols that are appropriate for use in diverse populations. Gene panel testing with a few, “common” variants

and/or panels inclusive of population-specific variants may serve as a first-line approach to molecular confirmation of CF diagnosis as they are more affordable and accessible. However, in order to achieve greater variant detection rates using population-specific panels, the spectrum of variants present in the South African population still requires investigation. Finally, evaluation of the full spectrum of CFTR variants will need to be available in the cases where these panels do not provide a full molecular diagnosis, following an elevated IRT assay. NGS will soon become a more affordable and accessible solution for diagnosis in diverse populations and will pave the way for standardised NBS in South Africa and other African countries to ensure that misdiagnosis is prevented. Despite the consistent oversight of African CF patients by international research groups, there remains an argument to be made for implementing efficient NBS programs for CF. However, despite the clear benefits of implementing NBS for CF in Africa, each country will need to evaluate whether it is worthwhile to allocate resources to NBS and whether the infrastructure can accommodate the implementation of dedicated follow-up care centres. They will also need to determine the extent that low socioeconomic status will have on clinical outcome, as well as the impact of early and severe infection and adherence to treatment. Finally, research will need to be driven towards modulator therapies that can effectively treat CF patients with diverse variants.

### 1.11. References:

- American College of Obstetricians and Gynecologists, A. and A. American College of Medical Genetics (2011). "ACOG Committee Opinion No. 486: Update on carrier screening for cystic fibrosis." *Obstet Gynecol* **117**(4): 1028-1031.
- Anderson, C. M., J. Allan and P. G. Johansen (1967). "Comments on the possible existence and nature of a heterozygote advantage in cystic fibrosis." *Bibl Paediatr* **86**: 381-387.
- Bell, S. C., K. De Boeck and M. D. Amaral (2015). "New pharmacological approaches for cystic fibrosis: promises, progress, pitfalls." *Pharmacol Ther* **145**: 19-34.
- Bell, S. C., M. A. Mall, H. Gutierrez, M. Macek, S. Madge, J. C. Davies, P. R. Burgel, E. Tullis, C. Castanos, C. Castellani, C. A. Byrnes, F. Cathcart, S. H. Chotirmall, R. Cosgriff, I. Eichler, I. Fajac, C. H. Goss, P. Drevinek, P. M. Farrell, A. M. Gravelle, T. Havermans, N. Mayer-Hamblett, N. Kashirskaya, E. Kerem, J. L. Mathew, E. F. McKone, L. Naehrlich, S. Z. Nasr, G. R. Oates, C. O'Neill, U. Pypops, K. S. Raraigh, S. M. Rowe, K. W. Southern, S. Sivam, A. L. Stephenson, M. Zampoli and F. Ratjen (2020). "The future of cystic fibrosis care: a global perspective." *Lancet Respir Med* **8**(1): 65-124.
- Bertranpetit, J. and F. Calafell (1996). "Genetic and geographical variability in cystic fibrosis: evolutionary considerations." *Ciba Found Symp* **197**: 97-114; discussion 114-118.
- Bhutta, Z. A., J. A. Berkley, R. H. J. Bandsma, M. Kerac, I. Trehan and A. Briend (2017). "Severe childhood malnutrition." *Nat Rev Dis Primers* **3**: 17067.
- Bobadilla, J. L., M. Macek, Jr., J. P. Fine and P. M. Farrell (2002). "Cystic fibrosis: a worldwide analysis of CFTR mutations--correlation with incidence data and application to screening." *Hum Mutat* **19**(6): 575-606.
- Bosch, L., B. Bosch, K. De Boeck, T. Nawrot, I. Meyts, D. Vanneste, C. A. Le Bourlegat, J. Croda and L. da Silva Filho (2017). "Cystic fibrosis carriership and tuberculosis: hints toward an evolutionary selective advantage based on data from the Brazilian territory." *BMC Infect Dis* **17**(1): 340.
- Campbell, M. C. and S. A. Tishkoff (2008). "African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping." *Annu Rev Genomics Hum Genet* **9**: 403-433.
- Carles, S., M. Desgeorges, A. Goldman, R. Thiart, C. Guittard, C. A. Kitazos, T. J. de Ravel, A. T. Westwood, M. Claustres and M. Ramsay (1996). "First report of CFTR mutations in black cystic fibrosis patients of southern African origin." *J Med Genet* **33**(9): 802-804.
- Cashman, S. M., A. Patino, A. Martinez, M. Garcia-Delgado, Z. Miedzzybrodzka, M. Schwarz, A. Shrimpton, C. Ferec, O. Raguenes, M. Macek, Jr. and et al. (1995). "Identical intragenic microsatellite haplotype found in cystic fibrosis chromosomes bearing mutation G551D in Irish, English, Scottish, Breton and Czech patients." *Hum Hered* **45**(1): 6-12.
- Castellani, C., A. J. A. Duff, S. C. Bell, H. G. M. Heijerman, A. Munck, F. Ratjen, I. Sermet-Gaudelus, K. W. Southern, J. Barben, P. A. Flume, P. Hodkova, N. Kashirskaya, M. N. Kirszenbaum, S. Madge, H. Oxley, B. Plant, S. J. Schwarzenberg, A. R. Smyth, G. Taccetti, T. O. F. Wagner, S. P. Wolfe and P. Drevinek (2018). "ECFS best practice guidelines: the 2018 revision." *J Cyst Fibros* **17**(2): 153-178.
- Chao, A. C., F. J. de Sauvage, Y. J. Dong, J. A. Wagner, D. V. Goeddel and P. Gardner (1994). "Activation of intestinal CFTR Cl<sup>-</sup> channel by heat-stable enterotoxin and guanylin via cAMP-dependent protein kinase." *EMBO J* **13**(5): 1065-1072.
- Consortium, C. F. G. A. (1994). "Population variation of common cystic fibrosis mutations. The Cystic Fibrosis Genetic Analysis Consortium." *Hum Mutat* **4**(3): 167-177.
- Currier, R. J., S. Sciortino, R. Liu, T. Bishop, R. Alikhani Koupaei and L. Feuchtbaum (2017). "Genomic sequencing in cystic fibrosis newborn screening: what works best, two-tier predefined CFTR mutation panels or second-tier CFTR panel followed by third-tier sequencing?" *Genet Med* **19**(10): 1159-1163.
- Cutting, G. R. (2010). "Modifier genes in Mendelian disorders: the example of cystic fibrosis." *Ann N Y Acad Sci* **1214**: 57-69.
- da Silva Filho, L., M. Zampoli, M. Cohen-Cymerknoh and S. K. Kabra (2020). "Cystic fibrosis in low and middle-income countries (LMIC): A view from four different regions of the world." *Paediatr Respir Rev*.
- Dankert-Roelse, J. E. and G. J. te Meerman (1995). "Long term prognosis of patients with cystic fibrosis in relation to early detection by neonatal screening and treatment in a cystic fibrosis centre." *Thorax* **50**(7): 712-718.
- Davis, K. F., P. D'Odorico, F. Laio and L. Ridolfi (2013). "Global spatio-temporal patterns in human migration: a complex network perspective." *PLoS One* **8**(1): e53723.
- De Carvalho, C. L. and M. Ramsay (2009). "CFTR structural rearrangements are not a major mutational mechanism in black and coloured southern African patients with cystic fibrosis." *S Afr Med J* **99**(10): 724.
- de Monestrol, I., A. Klint, P. Sparen and L. Hjelte (2011). "Age at diagnosis and disease progression of cystic fibrosis in an area without newborn screening." *Paediatr Perinat Epidemiol* **25**(3): 298-305.
- Denter, M., M. Ramsay and T. Jenkins (1992). "Cystic fibrosis. Part I. Frequency of the delta F508 mutation in South African families with cystic fibrosis." *S Afr Med J* **82**(1): 7-10.
- des Georges, M., C. Guittard, C. Templin, J. P. Altieri, C. de Carvalho, M. Ramsay and M. Claustres (2008). "WGA allows the molecular characterization of a novel large CFTR rearrangement in a black South African cystic fibrosis patient." *J Mol Diagn* **10**(6): 544-548.
- el-Harith, E. A., T. Dork, M. Stuhmann, H. Abu-Srair, A. al-Shahri, K. M. Keller, M. J. Lentze and J. Schmidtke (1997). "Novel and characteristic CFTR mutations in Saudi Arab children with severe cystic fibrosis." *J Med Genet* **34**(12): 996-999.
- El-Seedy, A., M. C. Pasquet, H. Shafiek, M. El-Komy, A. Kitzis and V. Ladeveze (2013). "20 Cystic fibrosis in Egypt: New mutational detection of the CFTR gene in patients from Alexandria, northern Egypt." *Journal of Cystic Fibrosis* **12**: S53.

- Elborn, J. S., S. C. Bell, S. L. Madge, P. R. Burgel, C. Castellani, S. Conway, K. De Rijcke, B. Dembski, P. Drevinek, H. G. Heijerman, J. A. Innes, A. Lindblad, B. Marshall, H. V. Olesen, A. L. Reimann, A. Sole, L. Viviani, T. O. Wagner, T. Welte and F. Blasi (2016). "Report of the European Respiratory Society/European Cystic Fibrosis Society task force on the care of adults with cystic fibrosis." *Eur Respir J* **47**(2): 420-428.
- Farrell, P. M., M. R. Kosorok, M. J. Rock, A. Laxova, L. Zeng, H. C. Lai, G. Hoffman, R. H. Laessig and M. L. Splaingard (2001). "Early diagnosis of cystic fibrosis through neonatal screening prevents severe malnutrition and improves long-term growth. Wisconsin Cystic Fibrosis Neonatal Screening Study Group." *Pediatrics* **107**(1): 1-13.
- Feuillet-Fieux, M. N., M. Ferrec, N. Gigarel, L. Thuillier, I. Sermet, J. Steffann, G. Lenoir and J. P. Bonnefont (2004). "Novel CFTR mutations in black cystic fibrosis patients." *Clin Genet* **65**(4): 284-287.
- Fredj, S. H., T. Messaoud, C. Templin, M. des Georges, S. Fattoum and M. Claustres (2009). "Cystic fibrosis transmembrane conductance regulator mutation spectrum in patients with cystic fibrosis in Tunisia." *Genet Test Mol Biomarkers* **13**(5): 577-581.
- Gallati, S. (2014). "Disease-modifying genes and monogenic disorders: experience in cystic fibrosis." *Appl Clin Genet* **7**: 133-146.
- Genzsch, M. and M. A. Mall (2018). "Ion Channel Modulators in Cystic Fibrosis." *Chest* **154**(2): 383-393.
- George, P. M., W. Banya, N. Pareek, D. Bilton, P. Cullinan, M. E. Hodson and N. J. Simmonds (2011). "Improved survival at low lung function in cystic fibrosis: cohort study from 1990 to 2007." *BMJ* **342**: d1008.
- Goldman, A., C. Graf, M. Ramsay, F. Leisegang and A. T. Westwood (2003). "Molecular diagnosis of cystic fibrosis in South African populations." *S Afr Med J* **93**(7): 518-519.
- Goldman, A., T. Jenkins and M. Ramsay (1994). "Analysis of 40 known cystic fibrosis mutations in South African patients." *Clin Genet* **46**(6): 398-400.
- Goldman, A., R. Labrum, M. Claustres, M. Desgeorges, C. Guittard, A. Wallace and M. Ramsay (2001). "The molecular basis of cystic fibrosis in South Africa." *Clin Genet* **59**(1): 37-41.
- Goss, C. H., J. Sykes, S. Stanojevic, B. Marshall, K. Petren, J. Ostrenga, A. Fink, A. Elbert, B. S. Quon and A. L. Stephenson (2018). "Comparison of Nutrition and Lung Function Outcomes in Patients with Cystic Fibrosis Living in Canada and the United States." *Am J Respir Crit Care Med* **197**(6): 768-775.
- Grody, W. W., G. R. Cutting, K. W. Klinger, C. S. Richards, M. S. Watson, R. J. Desnick and A. o. G. S. C. A. A. C. o. M. G. Subcommittee on Cystic Fibrosis Screening (2001). "Laboratory standards and guidelines for population-based cystic fibrosis carrier screening." *Genet Med* **3**(2): 149-154.
- Hadj Fredj, S., M. Boudaya, S. Oueslati, S. Sahnoun, C. Sahli, H. Siala, K. Boussetta, A. Bibi and T. Messaoud (2013). "New frameshift CF mutation 3729delAinsTCT in a Tunisian cystic fibrosis patient." *J Genet* **92**(1): 81-83.
- Hadj Fredj, S., S. Fattoum, A. Chabchoub and T. Messaoud (2011). "First report of cystic fibrosis mutations in Libyan cystic fibrosis patients." *Ann Hum Biol* **38**(5): 561-563.
- Hamosh, A., S. C. FitzSimmons, M. Macek, Jr., M. R. Knowles, B. J. Rosenstein and G. R. Cutting (1998). "Comparison of the clinical manifestations of cystic fibrosis in black and white patients." *J Pediatr* **132**(2): 255-259.
- Hansson, G. C. (1988). "Cystic fibrosis and chloride-secreting diarrhoea." *Nature* **333**(6175): 711.
- Herbert, J. S. and A. E. Retief (1992). "The frequency of the delta F508 mutation in the cystic fibrosis genes of 71 unrelated South African cystic fibrosis patients." *S Afr Med J* **82**(1): 13-15.
- Hughes, E. E., C. F. Stevens, C. A. Saavedra-Matiz, N. P. Tavakoli, L. M. Krein, A. Parker, Z. Zhang, B. Maloney, B. Vogel, J. DeCeli-Germana, C. Kier, R. D. Anbar, M. N. Berdella, P. G. Comber, A. J. Dozor, D. M. Goetz, L. Guida, Jr., M. Kattan, A. Ting, K. Z. Voter, C. New York State Cystic Fibrosis Newborn Screening, P. van Roey, M. Caggana and D. M. Kay (2016). "Clinical Sensitivity of Cystic Fibrosis Mutation Panels in a Diverse Population." *Hum Mutat* **37**(2): 201-208.
- Ibrahim, S. A., M. A. Fadl Elmola, Z. A. Karrar, A. M. Arabi, M. A. Abdullah, S. K. Ali, F. Elawad, T. E. Ali, M. B. Abdulrahman, S. O. Ahmed and A. S. Gundi (2014). "Cystic fibrosis in Sudanese children: First report of 35 cases." *Sudan J Paediatr* **14**(1): 39-44.
- Ikpa, P. T., M. J. Bijvelds and H. R. de Jonge (2014). "Cystic fibrosis: toward personalized therapies." *Int J Biochem Cell Biol* **52**: 192-200.
- Kabra, S. K., M. Kabra, S. Shastri and R. Lodha (2006). "Diagnosing and managing cystic fibrosis in the developing world." *Paediatr Respir Rev* **7 Suppl 1**: S147-150.
- Kerem, E., Y. M. Kalman, Y. Yahav, T. Shoshani, D. Abeliovich, A. Szeinberg, J. Rivlin, H. Blau, A. Tal, L. Ben-Tur and et al. (1995). "Highly variable incidence of cystic fibrosis and different mutation distribution among different Jewish ethnic groups in Israel." *Hum Genet* **96**(2): 193-197.
- Kerem, E., L. Viviani, A. Zolin, S. MacNeill, E. Hatzigorou, H. Ellemunter, P. Drevinek, V. Gulmans, U. Krivec, H. Olesen and E. P. R. S. Group (2014). "Factors associated with FEV1 decline in cystic fibrosis: analysis of the ECFS patient registry." *Eur Respir J* **43**(1): 125-133.
- Kharrazi, M. and L. D. Kharrazi (2005). "Delayed diagnosis of cystic fibrosis and the family perspective." *J Pediatr* **147**(3 Suppl): S21-25.
- Lakeman, P., J. J. Gille, J. E. Dankert-Roelse, H. G. Heijerman, A. Munck, A. Iron, H. Grasemann, A. Schuster, M. C. Cornel and L. P. Ten Kate (2008). "CFTR mutations in Turkish and North African cystic fibrosis patients in Europe: implications for screening." *Genet Test* **12**(1): 25-35.
- Lao, O., A. M. Andres, E. Mateu, J. Bertranpetit and F. Calafell (2003). "Spatial patterns of cystic fibrosis mutation spectra in European populations." *Eur J Hum Genet* **11**(5): 385-394.
- Leung, M. L., D. J. Watson, C. N. Vaccaro, F. Mafra, A. Wenocur, T. Wang, H. Hakonarson and A. Santani (2020). "Evaluating sequence data quality from the Swift Accel-Amplicon CFTR Panel." *Sci Data* **7**(1): 8.

- Lim, R. M., A. J. Silver, M. J. Silver, C. Borroto, B. Spurrier, T. C. Petrossian, J. L. Larson and L. M. Silver (2016). "Targeted mutation screening panels expose systematic population bias in detection of cystic fibrosis risk." *Genet Med* **18**(2): 174-179.
- Lissens, W., K. Z. Mahmoud, E. El-Gindi, A. Abdel-Sattar, S. Seneca, A. Van Steirteghem and I. Liebaers (1999). "Molecular analysis of the cystic fibrosis gene reveals a high frequency of the intron 8 splice variant 5T in Egyptian males with congenital bilateral absence of the vas deferens." *Mol Hum Reprod* **5**(1): 10-13.
- Loirat, F., S. Hazout and G. Lucotte (1997). "G542X as a probable Phoenician cystic fibrosis mutation." *Hum Biol* **69**(3): 419-425.
- Loukas, Y. L., G. Thodi, E. Molou, V. Georgiou, Y. Dotsikas and K. H. Schulpis (2015). "Clinical diagnostic Next-Generation sequencing: the case of CFTR carrier screening." *Scand J Clin Lab Invest* **75**(5): 374-381.
- Loumi, O., H. Cuppens, R. Bakour, M. Benabadji, M. Baghriche, P. Marynen and J. J. Cassiman (1992). "An Algerian child homozygous for the M470V polymorphism and for a deletion of two nucleotides in exon 10 of the CFTR gene, shows severe cystic fibrosis symptoms." *Genet Couns* **3**(4): 205-207.
- Lucarelli, M., L. Porcaro, A. Biffignandi, L. Costantino, V. Giannone, L. Alberti, S. M. Bruno, C. Corbetta, E. Torresani, C. Colombo and M. Seia (2017). "A New Targeted CFTR Mutation Panel Based on Next-Generation Sequencing Technology." *J Mol Diagn* **19**(5): 788-800.
- Lucotte, G., E. Barre and S. Berriche (1991). "Frequency of the cystic fibrosis mutation delta F508 in Algeria." *Hum Genet* **87**(6): 759.
- Macek, M., Jr., A. Mackova, A. Hamosh, B. C. Hilman, R. F. Selden, G. Lucotte, K. J. Friedman, M. R. Knowles, B. J. Rosenstein and G. R. Cutting (1997). "Identification of common cystic fibrosis mutations in African-Americans with cystic fibrosis increases the detection rate to 75%." *Am J Hum Genet* **60**(5): 1122-1127.
- MacKenzie, T., A. H. Gifford, K. A. Sabadosa, H. B. Quinton, E. A. Knapp, C. H. Goss and B. C. Marshall (2014). "Longevity of patients with cystic fibrosis in 2000 to 2010 and beyond: survival analysis of the Cystic Fibrosis Foundation patient registry." *Ann Intern Med* **161**(4): 233-241.
- Martinez-Costa, C., A. Escribano, F. Nunez Gomez, L. Garcia-Maset, J. Lujan and L. Martinez-Rodriguez (2005). "[Nutritional intervention in children and adolescents with cystic fibrosis. Relationship with pulmonary function]." *Nutr Hosp* **20**(3): 182-188.
- Masekela, R., M. Zampoli, A. T. Westwood, D. A. White, R. J. Green, S. Olorunju and M. Kwofie-Mensah (2013). "Phenotypic expression of the 3120+1G>A mutation in non-Caucasian children with cystic fibrosis in South Africa." *J Cyst Fibros* **12**(4): 363-366.
- McCarthy, C., O. O'Carroll, A. Franciosi and N. McElvaney (2015). Factors Affecting Prognosis and Predicting Outcome in Cystic Fibrosis Lung Disease.
- McCravy, M. S., N. L. Quinney, D. M. Cholon, S. E. Boyles, T. J. Jensen, A. A. Aleksandrov, S. H. Donaldson, P. G. Noone and M. Gentsch (2020). "Personalised medicine for non-classic cystic fibrosis resulting from rare CFTR mutations." *Eur Respir J* **56**(1).
- Mehta, G., M. Macek, Jr., A. Mehta and G. European Registry Working (2010). "Cystic fibrosis across Europe: EuroCareCF analysis of demographic data from 35 countries." *J Cyst Fibros* **9** Suppl 2: S5-S21.
- Messaoud, T., C. Verlingue, E. Denamur, O. Pascaud, I. Quere, S. Fattoum, J. Elion and C. Ferec (1996). "Distribution of CFTR mutations in cystic fibrosis patients of Tunisian origin: identification of two novel mutations." *Eur J Hum Genet* **4**(1): 20-24.
- Mickle, J. E. and G. R. Cutting (2000). "Genotype-phenotype relationships in cystic fibrosis." *Med Clin North Am* **84**(3): 597-607.
- Mirtajani, S., P. Farnia, M. Hassanzad, J. Ghanavi and A. Velayati (2017). "Geographical distribution of cystic fibrosis; The past 70 years of data analysis." *Biomedical and Biotechnology Research Journal (BBRJ)* **1**: 105 - 112.
- Mirtajani, S., P. Farnia, M. Hassanzad, J. Ghanavi and A. Velayati (2017). "Geographical distribution of cystic fibrosis; The past 70 years of data analysis." *Biomedical and Biotechnology Research Journal (BBRJ)* **1**: 105 - 112.
- Monaghan, K. G., D. Bluhm, M. Phillips and G. L. Feldman (2004). "Preconception and prenatal cystic fibrosis carrier screening of African Americans reveals unanticipated frequencies for specific mutations." *Genet Med* **6**(3): 141-144.
- Mutesa, L., A. K. Azad, C. Verhaeghe, K. Segers, J. F. Vanbellinghen, L. Ngendahayo, E. K. Rusingiza, P. R. Mutwa, S. Rulisa, L. Koulischer, J. J. Cassiman, H. Cuppens and V. Bours (2009). "Genetic analysis of Rwandan patients with cystic fibrosis-like symptoms: identification of novel cystic fibrosis transmembrane conductance regulator and epithelial sodium channel gene variants." *Chest* **135**(5): 1233-1242.
- Mutesa, L. and V. Bours (2009). "Diagnostic challenges of cystic fibrosis in patients of African origin." *J Trop Pediatr* **55**(5): 281-286.
- Naguib, M. L., I. Schrijver, P. Gardner, L. M. Pique, S. S. Doss, M. A. Abu Zekry, M. Aziz and S. Z. Nasr (2007). "Cystic fibrosis detection in high-risk Egyptian children and CFTR mutation analysis." *J Cyst Fibros* **6**(2): 111-116.
- Nyblade, L., M. A. Stockton, K. Giger, V. Bond, M. L. Ekstrand, R. M. Lean, E. M. H. Mitchell, R. E. Nelson, J. C. Sapag, T. Siraprasiri, J. Turan and E. Wouters (2019). "Stigma in health facilities: why it matters and how we can change it." *BMC Med* **17**(1): 25.
- Owusu, S. K., B. M. Morrow, D. White, S. Klugman, A. Vanker, D. Gray and M. Zampoli (2020). "Cystic fibrosis in black African children in South Africa: a case control study." *J Cyst Fibros* **19**(4): 540-545.
- Padoa, C., A. Goldman, T. Jenkins and M. Ramsay (1999). "Cystic fibrosis carrier frequencies in populations of African origin." *J Med Genet* **36**(1): 41-44.
- Phillips, A. E., J. LaRusch, P. Greer, J. Abberbock, S. Alkaade, S. T. Amann, M. A. Anderson, J. Baillie, P. A. Banks, R. E. Brand, D. Conwell, G. A. Cote, C. E. Forsmark, T. B. Gardner, A. Gelrud, N. Guda, M. Lewis, M. E. Money, T. Muniraj, B. S. Sandhu, S. Sherman, V. K. Singh, A. Slivka, G. Tang, C. M. Wilcox, D. C. Whitcomb and D. Yadav (2018). "Known genetic susceptibility factors for chronic pancreatitis in patients of European ancestry are rare in patients of African ancestry." *Pancreatolgy* **18**(5): 528-535.
- Pier, G. B., M. Grout, T. Zaidi, G. Meluleni, S. S. Mueschenborn, G. Banting, R. Ratcliff, M. J. Evans and W. H. Colledge (1998). "Salmonella typhi uses CFTR to enter intestinal epithelial cells." *Nature* **393**(6680): 79-82.
- Pique, L., S. Graham, M. Pearl, M. Kharrazi and I. Schrijver (2017). "Cystic fibrosis newborn screening programs: implications of the CFTR variant spectrum in nonwhite patients." *Genet Med* **19**(1): 36-44.

- Plant, B. J., C. H. Goss, W. D. Plant and S. C. Bell (2013). "Management of comorbidities in older patients with cystic fibrosis." *Lancet Respir Med* **1**(2): 164-174.
- Poolman, E. M. and A. P. Galvani (2007). "Evaluating candidate agents of selective pressure for cystic fibrosis." *J R Soc Interface* **4**(12): 91-98.
- Rohlf, E. M., Z. Zhou, R. A. Heim, N. Nagan, L. S. Rosenblum, K. Flynn, T. Scholl, V. R. Akmaev, D. A. Sirko-Osadsa, B. A. Allitto and E. A. Sugarman (2011). "Cystic fibrosis carrier testing in an ethnically diverse US population." *Clin Chem* **57**(6): 841-848.
- Ruiz-Schultz, N., D. Sant, S. Norcross, W. Dansithong, K. Hart, B. Asay, J. Little, K. Chung, K. F. Oakeson, E. L. Young, K. Eilbeck and A. Rohrwasser (2021). "Methods and feasibility study for exome sequencing as a universal second-tier test in newborn screening." *Genet Med* **23**(4): 767-776.
- Schrijver, I., L. Pique, S. Graham, M. Pearl, A. Cherry and M. Kharrazi (2016). "The Spectrum of CFTR Variants in Nonwhite Cystic Fibrosis Patients: Implications for Molecular Diagnostic Testing." *J Mol Diagn* **18**(1): 39-50.
- Scotet, V., I. Dugueperoux, P. Saliou, G. Rault, M. Roussey, M. P. Audrezet and C. Ferec (2012). "Evidence for decline in the incidence of cystic fibrosis: a 35-year observational study in Brittany, France." *Orphanet J Rare Dis* **7**: 14.
- Scotet, V., H. Gutierrez and P. M. Farrell (2020). "Newborn Screening for CF across the Globe-Where Is It Worthwhile?" *Int J Neonatal Screen* **6**(1): 18.
- Scotet, V., C. L'Hostis and C. Ferec (2020). "The Changing Epidemiology of Cystic Fibrosis: Incidence, Survival and Impact of the CFTR Gene Discovery." *Genes (Basel)* **11**(6).
- Shum, B. O. V., G. Bennett, A. Navilebasappa and R. K. Kumar (2021). "Racially equitable diagnosis of cystic fibrosis using next-generation DNA sequencing: a case report." *BMC Pediatr* **21**(1): 154.
- Sosnay, P. R., K. R. Siklosi, F. Van Goor, K. Kaniecki, H. Yu, N. Sharma, A. S. Ramalho, M. D. Amaral, R. Dorfman, J. Zielinski, D. L. Masica, R. Karchin, L. Millen, P. J. Thomas, G. P. Patrinos, M. Corey, M. H. Lewis, J. M. Rommens, C. Castellani, C. M. Penland and G. R. Cutting (2013). "Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene." *Nat Genet* **45**(10): 1160-1167.
- Steinkamp, G., B. Rodeck, J. Seidenberg, I. Ruhl and H. von der Hardt (1990). "[Stabilization of lung function in cystic fibrosis during long-term tube feeding via a percutaneous endoscopic gastrostomy]." *Pneumologie* **44**(10): 1151-1153.
- Stephenson, A. L., L. A. Mannik, S. Walsh, M. Brotherwood, R. Robert, P. B. Darling, R. Nisenbaum, J. Moerman and S. Stanojevic (2013). "Longitudinal trends in nutritional status and the relation between lung function and BMI in cystic fibrosis: a population-based cohort study." *Am J Clin Nutr* **97**(4): 872-877.
- Steward, C. A., A. P. J. Parker, B. A. Minassian, S. M. Sisodiya, A. Frankish and J. Harrow (2017). "Genome annotation for clinical genomic diagnostics: strengths and weaknesses." *Genome Med* **9**(1): 49.
- Stewart, C. and M. S. Pepper (2016). "Cystic fibrosis on the African continent." *Genet Med* **18**(7): 653-662.
- Tridello, G., C. Castellani, I. Meneghelli, A. Tamanini and B. M. Assael (2018). "Early diagnosis from newborn screening maximises survival in severe cystic fibrosis." *ERJ Open Res* **4**(2).
- Tzetzis, M., E. Kanavakis, T. Antoniadis, S. Doudounakis, G. Adam and C. Kattamis (1997). "Characterization of more than 85% of cystic fibrosis alleles in the Greek population, including five novel mutations." *Hum Genet* **99**(1): 121-125.
- van de Vosse, E., S. Ali, A. W. de Visser, C. Surjadi, S. Widjaja, A. M. Vollaard and J. T. van Dissel (2005). "Susceptibility to typhoid fever is associated with a polymorphism in the cystic fibrosis transmembrane conductance regulator (CFTR)." *Hum Genet* **118**(1): 138-140.
- Van Rensburg, J., M. Alessandrini, C. Stewart and M. S. Pepper (2018). "Cystic fibrosis in South Africa: A changing diagnostic paradigm." *S Afr Med J* **108**(8): 624-628.
- W.H.O. (2022). "Factsheet: Children." from <https://www.afro.who.int/health-topics/child-health#:~:text=About%2045%25%20of%20all%20child,than%20children%20in%20developed%20regions>.
- Waters, D. L., B. Wilcken, L. Irwing, P. Van Asperen, C. Mellis, J. M. Simpson, J. Brown and K. J. Gaskin (1999). "Clinical outcomes of newborn screening for cystic fibrosis." *Arch Dis Child Fetal Neonatal Ed* **80**(1): F1-7.
- Watson, M. S., G. R. Cutting, R. J. Desnick, D. A. Driscoll, K. Klinger, M. Mennuti, G. E. Palomaki, B. W. Popovich, V. M. Pratt, E. M. Rohlf, C. M. Strom, C. S. Richards, D. R. Witt and W. W. Grody (2004). "Cystic fibrosis population carrier screening: 2004 revision of American College of Medical Genetics mutation panel." *Genet Med* **6**(5): 387-391.
- Westwood, A. (2008). "The prognosis of cystic fibrosis in South Africa: a 33 year study. ." *J Cyst Fibros* **7**:458.
- Westwood, T., B. Henderson, M. Ramsay, Medical and A. Scientific Advisory Committee of the South African Cystic Fibrosis (2006). "Diagnosing cystic fibrosis in South Africa." *S Afr Med J* **96**(4): 304, 306.
- Wonkam, A. (2016). "Cystic fibrosis: the urgent need to report on mutations among patients of African descent." *South African Respiratory Journal* **22**: 34.
- Zaher, A., J. ElSaygh, D. ElSori, H. ElSaygh and A. Sanni (2021). "A Review of Trikafta: Triple Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Modulator Therapy." *Cureus* **13**(7): e16144.
- Zampoli, M., J. Verstraete, M. Frauendorf, R. Kassanje, L. Workman, B. M. Morrow and H. J. Zar (2021). "Cystic fibrosis in South Africa: spectrum of disease and determinants of outcome." *ERJ Open Res* **7**(3).
- Zampoli On Behalf Of The Msac, M. (2019). "Cystic fibrosis: What's new in South Africa in 2019." *S Afr Med J* **109**(1): 16-19.

## *Chapter 2: Methodology*

Investigating CFTR variants in South African patients with Cystic Fibrosis.

## 2.1. Introduction

Currently, the diagnostic protocol for suspected CF in South Africa involves a sweat test and confirmatory gene panel testing. Lower variant detection rates for molecular diagnosis of CF using gene panels have been observed in ethnically diverse populations (Currier, Sciortino et al. 2017). South Africa has a variant detection rate of 70-79% (Stewart and Pepper 2016), using a gene panel developed by the American College of Medical Genetics and slightly adapted for a few “population-specific” variants using limited genetic studies of the population (Goldman, Graf et al. 2003). However, the variant detection rate is >95% in many European populations (Bell, Mall et al. 2020). Thus, the original hypothesis for this project was that thorough investigation of CFTR variants in South African patients with CF would yield variants that are “common” to the population and help to develop a better population-specific gene panel. To address this hypothesis, Next-Generation Sequencing (NGS) of the CFTR gene of 65 CF patients and five parents was used to investigate the variants that may be causative of disease in a South African population. Patients were selected after gene panel sequencing yielded an incomplete genotype.

The aims of this study were:

- Perform variant detection in the CFTR gene from the raw NGS sequencing data of a cohort of South African patients with CF.
- Identify a list of variants from the NGS data that have the potential to be pathogenic.
- Validate the list of potential variants experimentally using Sanger sequencing.

The objectives of this study were:

- Perform QC and trimming on the raw NGS sequencing data for the cohort of South African patients with CF.
- Map to the reference genome and identify variants in the CFTR gene on chromosome 7.
- Compare the variant calls from the different methods of variant detection applied to the data.
- Perform in-silico validation of the variants using pibase and BAYSIC.
- Compare the variant calls before and after the validation.
- Compile a master variant list and format for input into VEP.
- Filter according to predicted consequence and evaluate to determine if variants are potentially pathogenic.
- Design primers and prepare amplicons for Sanger sequencing.
- Validate the primers experimentally using Sanger sequencing and subsequent data analysis.
- Compile a final list of experimentally validated CFTR variants in a South African cohort to inform gene panel testing.



## 2.2. Ethics Approval

Ethics approval for the overarching study was granted in 2013 by the Faculty of Health Sciences Research Ethics Committee at the University of Pretoria (UP; approval number 4/2013) and the Faculty of Health Sciences Human Research Ethics Committee at the University of Cape Town (UCT; approval number 433/2013). Most recent Annual Renewal by the University of Pretoria Faculty of Health Sciences Research Ethics Committee was approved on 2022-08-10 and is valid until 2023-08-12. Ethics approval for this MSc project (NAS039/2021) was granted by the Faculty of Health Sciences Research Ethics Committee on 2021-05-12 and has been renewed annually.

## 2.3. Data Collection

There are 70 participants in this study – 65 patients and five parents of patients – from various hospitals and clinics including Steve Biko Academic Hospital (SBAH) in Pretoria, Tygerberg Hospital in Stellenbosch, Charlotte Maxeke Johannesburg Academic Hospital (CMJAH) and the CF and asthma clinics at the Red Cross War Memorial Children’s Hospital (RCWMCH) in Cape Town. For these patients, the patient files are available in the form of various electronic databases and include some patient demographic data, medical history, symptoms upon presentation, symptoms at last visit, annual lung function test results and lung microbial flora. The raw sequencing data was obtained from the storage server of the UP Centre for Bioinformatics and Computational Biology after being granted access following ethics approval. An earlier component of this study performed variant calling and annotation using the CLC Genomics Workbench software on the raw sequencing data, producing \*.vcf files for each patient (Dr. C. Stewart, post-doctoral student). These files are securely stored on the storage server of the UP Centre for Bioinformatics and Computational Biology, with controlled access. Variant calling was also previously performed by Dr. Stewart on the raw data using *Illumina*’s in-house CASAVA software, and these \*.vcf files were also obtained and securely stored on the cluster.

## 2.4. Molecular Biology – sample collection and NGS

The following has been provided by the previous researcher (Dr. C. Stewart):

Informed consent was obtained from all participants from whom our collaborating clinicians collected blood samples. At SBAH, 19 individuals associated with the CF clinic agreed to participate in this study, as well as five parents of patients. Of these, eight are patients who had already been genotyped and thus served as positive controls and six lacked a molecular diagnosis of CF. At the RCWMCH, the necessary consent was obtained from 31 patients, nine of whom were members of the asthma clinic with equivocal sweat test results. We also included eight

CMJAH patients and seven patients from Tygerberg Hospital in our study. The QIAamp DNA Blood Midi Kit was used to prepare the DNA from the blood samples. A NanoDrop spectrophotometer was then used for DNA quantification. Each eluate was then visualised using agarose gel electrophoresis. The DNA was prepared for courier to Ambry Genetics in California, USA or GeneWiz in New Jersey, USA where NGS was performed using the *Illumina* MiSeq platform (Ravi, Walton et al. 2018). The sequencing methodology enriched for all 27 exons, at least 20bp into the 5' and 3' ends of each intron, the 5' and 3' UTR, intron 10's poly-T tract and the deep intronic variants c.1679+1634A>G and c.3717+12191C>T. Some initial analysis (including base calling and extracting cluster intensities) was conducted by Ambry and a sequence quality filtering script was executed using *Illumina* CASAVA version 1.8.2.

### 2.5. *QC and Trimming*

Quality control, trimming, mapping and variant detection were performed on the data using Galaxy and the CLC Genomics Workbench 7 (Dr. C. Stewart). However, as there was minimal consensus between the variants detected as well as the confidence in the base calls, this step was updated, repeated and includes packages from newer versions of some software such as GATK 4. Thus, only one combined .vcf was evaluated from the previous variant calling.

For the purposes of this dissertation, the raw NGS data first underwent quality control using FASTQC (Andrews 2010), and MultiQC (Ewels, Magnusson et al. 2016) was utilized for visualisation of the FASTQ files for all samples simultaneously. The over-represented sequences identified were used as search terms in a standard BLASTN search to determine their origin. Trimming was performed on the reads using Trimmomatic-0.36 (Bolger, Lohse et al. 2014) with the paired end function activated. The *Illumina* adapter sequences were removed using the TruSeq3-PE-2.fa file as input and the first 15bp at the beginning of the reads were trimmed using the HEADCROP:15 parameter. This produced four output files per sample, one paired and one unpaired for both forward and reverse reads. These results were again put through FASTQC and visualized with MultiQC to ensure that the reads received appropriate trimming and were of adequate to be used in subsequent analysis.

### 2.6. *Mapping and Variant Detection*

Mapping sequencing data to a reference genome and producing high-quality variant calls that can be used in further analyses has been made possible through BWA and the GATK best practices pipeline, implemented using the BCBIO python pipeline (<https://doi.org/10.5281/zenodo.3564938>). The associated protocols used are discussed in the

public access manuscript published by developers at the Broad Institute (Van der Auwera, Carneiro et al. 2013). The first protocol describes the steps involved in preparing the sequence data by converting FASTQ files to analysis-ready BAM files that can be used to call variants. The first step is to map the FASTQ sequences to the GRCh37 version of the human genome using BWA-MEM. GRCh37 was again used as a reference to enable comparison against the results of Dr. C. Stewart. This involves preparing the reference sequence, mapping the data to the reference, converting to BAM, sorting and marking duplicates, local realignment around indels, and base quality score recalibration (BQSR). The second protocol described in the manuscript is that of calling variants with HaplotypeCaller to identify sites of variation relative to the reference genome, focused on SNPs and Indels. This creates a .vcf file from the .bam file, containing raw calls that need filtering before any further analyses. This involves determining the basic parameters to be used in the analysis, such as genotyping mode, output mode, emission confidence threshold and calling confidence threshold. This is followed by calling the variants in the sequence data, which needs to be followed with application of the appropriate filters. Variant filtering involves flagging false-positive artifacts of the sequencing method from the original VCF file based on sequence alignment and variant calling metadata (Roy, Coldren et al. 2018).

The trimmed reads, as well as the appropriate bed file (containing the regions that were sequenced), were used as input into the BCBIO pipeline. This pipeline allows for streamlined mapping and variant calling on multiple samples using parameters and tools as specified in a configuration .yaml file for each sample. The reads were mapped onto the hg19 version of the human reference genome (GRCh37) using BWA (Li 2013) and Bowtie2 (Langmead and Salzberg 2012). BQSR was activated as a parameter using GATK CountCovariates and TableRecalibration (Poplin, Ruano-Rubio et al. 2018), and variant detection was performed. The pipeline was run four times, once for BWA and once for Bowtie2, using the GATK HaplotypeCaller (Poplin, Ruano-Rubio et al. 2018) and FreeBayes (Garrison and Marth 2012) variant callers respectively. Mapping statistics were obtained using QualiMap (Okonechnikov, Conesa et al. 2015) and compared between BWA and Bowtie2. The average mapping quality for the samples mapped using BWA was higher than that of Bowtie2 and so the samples mapped with BWA were used in subsequent analysis. The two algorithms were also compared visually using IGV (Robinson, Thorvaldsdottir et al. 2017), and the difference in mapping quality was confirmed.

The .vcf files for each sample produced by mapping with BWA and variant calling with the GATK HaplotypeCaller and Freebayes algorithms were merged using bcftools merge (BCFtools 2011)

into single .vcf files for each algorithm respectively. The .vcf files produced by CLC Genomics workbench (C. Stewart) were also merged, as well as the *Illumina* CASAVA data. The merged .vcf files were compared using bcftools isec (BCFtools 2011). These files were intersected to determine the concordance between the variants called by the different algorithms and whether one algorithm might be missing more variants than another. This step produced lists of variants called by each variant caller, variants unique to each variant caller, as well as a list of the variants shared between all four and the respective complements. This comparison was visualised using jvenn (Bardou, Mariette et al. 2014).

### 2.7. *In Silico* Validation and Identification of Potential Variants

This step was initiated by Dr. C. Stewart on the CLC Genomics Workbench and *Illumina* CASAVA data sets, but was repeated on the raw data following the variant detection described above. The variant calls from the different approaches were planned to be validated with Pibase (Forster, Forster et al. 2013) as well as BAYSIC (Cantarel, Weaver et al. 2014). This was also done by Dr. C. Stewart) on her initial variant calls. Pibase is used to validate the best genotype at the positions of interest but has not been established as a best practice tool for validation of variant calls since the algorithms for variant filtering (especially those used by GATK-4) have been drastically improved and the genotyping rules applied by Pibase remain ambiguous and arbitrary. As a result, Pibase was not used for validation of the variant calls as it was suspected it would remove valuable variants and not provide a satisfactory increase in information. In contrast, BAYSIC determines a posterior probability for each variant called and may be a valuable tool for combining variant calls from different tools (Cantarel, Weaver et al. 2014). While pibase uses a consensus-based approach, BAYSIC performs a Bayesian latent class analysis to estimate false positive and false negative error rates and determine a more accurate set of variant calls. For this reason, BAYSIC was used to further validate the variant call sets and combine the variant calls from GATK, Freebayes, CLC Genomics and *Illumina* CASAVA into a “more accurate” set of variant calls. The cutoff threshold was set to a posterior probability of 0.9.

This analysis enabled the comparison of variant calls before and after the validation step, which was started by Dr. C. Stewart but needed to be repeated and subsequently completed. For the Pibase validation step, four analyses were performed for each sample, one for each variant detection pipeline. Pibase uses a list of positions, the reference genome and a sorted, MD-tagged and indexed .bam file as input (Samtools), and then determines the best genotype at each position and assigns a quality classification. The pibase\_bamref initial step was performed four times for

each sample (one for each calling pipeline) with the relevant .vcf and .bam files used as input, as well as the reference genome. The resulting output was input into the pibase\_consensus module, which assigns a “Best Quality” tag based on 10 rule-based genotype decisions. The next step was to flag SNPs using the pibase\_flagsnps module, which flags a genotype if it is different to the reference genotype. This was followed with pibase\_c\_to\_contig to convert the pibase chromosome numbers into the conventional contig names (in this case, all contig names appear as “chr7”). Lastly, the pibase\_to\_vcf module would have been used to convert the pibase output to .vcf files. The .vcf files for each sample would then have been merged for each of the four calling pipelines, producing four merged .vcf files that could be compared using bcftools isec as was done previously. However, it was determined that this tool would not provide valuable results and the output was discarded at this stage. BAYSIC provided a combined.vcf file as the output of its analyses. This was intersected with the shared.vcf file created by intersecting the four variant call sets, using BCFtools isec (BCFtools 2011) in both instances, providing a list of shared variants between the two lists and the complements. This was also visualized using jvenn (Bardou, Mariette et al. 2014).

### 2.8. Variant Effect Prediction

The main Variant Effect Prediction tools that were used include LRT\_pred, MutationTaster, Provean, CADD and FATHMM. The Ensembl Variant Effect Predictor determines the effect of variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions (McLaren, Gil et al. 2016). VEP can be used in analysis, annotation, and prioritization of genomic variants in coding and non-coding regions (McLaren, Gil et al. 2016).

Two broad categories of genomic variants can be annotated with VEP. The first is sequence variants with well-defined changes (including SNVs, insertions, deletions, multiple base pair substitutions, microsatellites, and tandem repeats). The second is larger structural variants (>50 nucleotides) including CNVs or insertions and deletions of DNA. For protein annotation, VEP uses a variety of methods to predict the effect of the amino acid change and can predict how deleterious a variant may be (McLaren, Gil et al. 2016). Plug-ins such as FATHMM and MutationTaster were used for pathogenicity prediction scores. MutationTaster shows the pathogenic potential of DNA sequence variants and predicts the consequences of amino acid substitutions including intronic and synonymous changes, short insertion and/or deletion variants and variants spanning intron-exon borders (Mutation Taster 2020). FATHMM predicts the

functional consequences of coding variants (non-synonymous single nucleotide variants) and non-coding variants and can be used to distinguish between disease-causing variants and neutral polymorphisms in inherited disease (FATHMM 2020). The CADD plug-in was also used to score and prioritize the deleteriousness of SNVs and insertion/deletions variants in the human genome (OmicX\_CADD 2020). LRT\_pred is a tool used for the likelihood ratio test (LRT). This is a statistical test between two models and how well they compare, i.e. whether a more complex model is better than a simpler model. PROVEAN is a tool that will be used to predict whether an amino acid substitution or indel has an impact on the biological function of a protein and if they are predicted to be functionally important (OmicX\_PROVEAN 2020). Annotations from SIFT, PolyPhen, Condel and ClinVar were also evaluated for pathogenicity of variants using VEP.

The four variant call sets were input separately into Ensembl's VEP with the relevant plug-ins activated, and the resulting output annotations were downloaded (McLaren, Gil et al. 2016). VEP also provided some summary statistics for each set of variants.

Heatmaps were then constructed that represent each of the variants and the samples in which these variants have been detected. This was done in python using the Seaborn module. These heatmaps were then visualised using the Matplotlib module in python and the figures were downloaded.

The four datasets provided a mass of variants that needed to be combined before comparing to known databases and further filtering according to type and consequence. Thus, it was deemed necessary to decide on criteria to use when determining the final Master variant list. These criteria were:

- Variant was present in GATK call set.
- Variant was part of the BAYSIC set of variants with a posterior probability of 0.9 *and* variant was present in more than one variant call set.
- Variant had been validated previously (by Dr. C. Stewart) – these were not be re-validated and were included in the final, validated list of variants.

The Master Variant list was manually annotated with CFTR2 database hits to identify any potential CF-causing variants that had previously been identified and to provide functional information for the variants in the database (CFTR2 2011). Each variant was used as a search term in the CFTR2

database as multiple queries cannot be generated simultaneously. Additionally, variants can be recorded using different naming conventions thus necessitating multiple queries per variant.

The list of variants was then prepared for prioritisation with VVP (Flygare, Hernandez et al. 2018). Multi-allelic variants were decomposed using vt (Tan, Abecasis et al. 2015), re-annotated with VEP, and then VVP was run on the resulting .vcf file, as suggested by the VVP documentation. This was done using the gnomAD CFTR variants as background. VVP assigns a raw score for each variant using a CLRT method that incorporates allele frequency, sequence conservation, type of sequence change, zygosity and gene-specific burden. These raw scores can then be normalized to percentile scores using CRD curves per gene, so that the scores can be compared and prioritized (Flygare, Hernandez et al. 2018). The background distribution of percentile scores for CFTR (ENST00000003084) was queried from the background (\*.dist) file generated from gnomAD when VVP was initially run on the master list of variants. This was saved to a .csv file that was edited and inputted into a pandas data-frame in python. This data-frame was used to create an eCDF plot using the seaborn module and this was visualised using matplotlib.pyplot. This normalised the data so that the raw scores could be looked up to the corresponding percentile scores on the y-axis. Each VVP raw score from the master variant list was looked up on the x-axis of the plot and the corresponding y-value was obtained from the plot for the coding and non-coding variants, respectively. These percentile scores were then added back to the spreadsheet containing the various annotations for the master variant list. Finally, these scores were used in addition to the different annotations (described below) for pathogenicity prioritization.

The coding regions were not filtered by allele frequency as originally proposed. This is because the deleterious variant allele frequency in the CFTR gene has been found to violate the assumption that pathogenic variants are typically found at a low frequency in the general/healthy population and as such there may be some potential variants lost if they are filtered accordingly (Lim, Silver et al. 2016, Flygare, Hernandez et al. 2018). However, none of the variants identified as pathogenic exceeded 1% MAF, so filtering would not have altered the results. Predictions from ClinSig, Condel, LRT\_pred, CADD, SIFT, Polyphen2, FATHMM, MutationTaster, GERP++\_RS, and PROVEAN (McLaren, Gil et al. 2016) were used for analysis of the predicted effect of the variants in the coding region.

The information as described above was used in the decision criteria for compiling an initial list of likely pathogenic variant candidates. These were split according to variant consequence, and this

is presented as six tables. The existing databases, RedCap and Microsoft Access, and reports (Dr. C. Stewart) provided information regarding the genotypes identified using the NHLS CFTR panels, as well as patient information (where available). This provided an indication of the known variants and which patients lacked a complete molecular diagnosis.

### *2.9. Assessment of genotype information*

The existing genotype information and confirmation results (received from Dr. C. Stewart) were amended with the potentially pathogenic variant results from the NGS analysis before confirmation with Sanger sequencing. The genotype information was evaluated to determine which patients would have been able to be completely genotyped using the gene panel or NGS before validation, and which patients had variants identified that are not present on the panel. Furthermore, ethnicity was recorded according to the ability of gene panel screening or NGS to effectively genotype individuals before validation.

### *2.10. Validation of Variants*

This step was performed by Dr. C. Stewart on her list of potential variants; however, the experimental validation of variants needed to be repeated with the updated variant list. The list of likely pathogenic variants was validated using traditional Sanger sequencing. Primers were designed that flank the locations of the variants identified by NGS. These were used to amplify (with PCR) and sequence (with Sanger) the regions of interest. Variants identified by the NHLS screen were excluded, except for  $\Delta F508$  which was used as a positive control for the Sanger protocol, as well as two others that were identified by NGS but lacked NHLS panel information. The positions of interest were determined from the output of the VEP analysis and visualised in IGV (Robinson, Thorvaldsdottir et al. 2017). A list of possible primer pairs was then determined using PrimerQuest (IDT), as this includes a secondary structure/dimer formation checking function. OligoAnalyzer was used to check for the formation of hairpin loops, homodimers and heterodimers. Subsequently, the primers were checked with BLASTN to assure specificity to the appropriate region of the human genome (PrimerQuest also provides the functionality for this within the tool).

The appropriate primers were then ordered from Whitehead Scientific/IDT and used for amplicon preparation. The PCR products were sent to Inqaba Biotech for purification and subsequent Sanger sequencing using the standard sequencing parameters for the ABI 3500XL Genetic Analyzer (POP7™ and BrilliantDye™ Terminator v3.1). SnakVar version 2.4.3 (Kim, Kim et al. 2021) was then used to analyse the Sanger .ab1 output files. It performs QC, trimming, alignment



to the CFTR gene of the human genome and variant identification (including heterozygous indel identification) in a single step, with simple reports provided as output. This enabled confirmation or invalidation of the variants. Three samples (those for the c.\*1043A>C variant) could not be analysed with SnackVar, as it falsely identified the forward trace files as being reverse and *vice versa*. Thus, the ThermoFisher Cloud platform was used to analyse these samples. All Sanger figures (trace files, alignments, and variant positions) are available in Supplementary 3.

### 2.10.1. PCR Protocol:

#### 2.10.1.1. Primer design

**Step One:** Visualise in IGV and get sequence at the locations of interest.

- 1.1. Get the .vcf files for each patient.
- 1.2. Open each .vcf file in IGV and go to location of interest.
- 1.3. Copy sequence (extra 100bp before and after the variant; about 1kbp for searching in PrimerQuest).
- 1.4. Search for a primer in PrimerQuest (default search parameters, except amplicon length set between 300bp and 750bp; no amplicon was designed to be longer than 750bp).

**Step Two:** Check for uniqueness with NCBI BLASTN

**Step Three:** Use OligoAnalyzer to check each primer for formation of hairpins, self-dimers or cross-dimers and confirm conditions (T<sub>m</sub>, etc.).

Table 2.1: Variants to be validated with Sanger (primary list; likely pathogenic/CFTR2 confirmed pathogenic):

Variant	Location	Allele	Sample to confirm	Primer pair	Amplicon length	Optimal annealing temperature (°C)	Optimal no. of cycles
p.Met1Thr, p.Phe17SerfsX8 ***	117120150-117120192		CF8754900; CF1782680; CF3019852, CF3594271, CF7930867, CF9295572	FWD: GCG TAG TGG GTG GAG AAA G REV: GTG CCA AGA AGA CAA TCA AGT G	702bp	55	30
c.2T>C, p.Met1Thr	7:117120150-117120150	C	CF8754900; CF1782680	***			
R75X, c.223C>T, p.Arg75Ter	7:117149146-117149146	T	CF4602380	**			
L218X, c.653T>A, p.Leu218Ter	7:117175375-117175375	A	CF4062212	FWD: GCT CAG AAC CAC GAA GTG TT REV: CGG TAG CTC ATG CCT GTA ATA TC	702bp	55	30
p.Arg303AlafsTer16, c.906_907insGCCACTTTGCAATGT GAAAATGTTTACTCAGCAAGATGTT TTCITTGATCTACAGTTGTTATTA ATTGTGATTTGGAGCTATAGCAGTT GTCGCAGTTTACATCGGAAAGCA GCCTATGTG	7:117180171-117180171	TCGGAAGGCAG CCTATGTGGCCA CTTTGCAATGTG AAAATGTTTACT CACCAACATGTT TTCITTGATCTT ACAGTTGTTAT AAITGTGATTG GAGCTATAGCA GTTGTGCGAGTT TTACA	CF4062212	FWD: TCA ATG TTC CTC AAA GCC A REV: CAG AAT GAG ATG GTG GTG AAT A	674bp	55	30

Variant	Location	Allele	Sample to confirm	Primer pair	Amplicon length	Optimal annealing temperature (°C)	Optimal no. of cycles
c.1148T>A, p.Leu383Ter	7:117182101-117182101	A	CF2173052	FWD: ACC TTC ACA TGC TTC CTT AAC C REV: ACC TGG CCA TTC CTC TAC TT	711bp	55	30
p.Gly458Val, p.Gly451Ter *****	117188836-117188858		CA4932026, CF1697504, CF4544212, CF4833948, CF5158167, CF4283433	FWD: ACA GCT TTG AAA GAG GAG GAT TA REV: CCT TCC AGC ACT ACA AAC TAG AA	665bp	55	30
c.1351G>T, p.Gly451Ter	7:117188836-117188836	T	CA4932026, CF1697504, CF4544212, CF4833948, CF5158167	*****			
R709X, p.Arg709X, c.2125C>T	7:117232346-117232346	T	CF2173052	FWD: AAA CTC ATG GGA TGT GAT TCT TTC REV:TGA GTG TGT CAT CAG GTT CAG	402bp	55	35
c.3373G>T, p.Gly1125Ter	7:117254672-117254672	T	CA4932026, CF1697504, CF2349244, CF5158167, CF5181003, CF5830853, CF6803591, CF9830825	FWD: GCT CAT CTG GAT ACA GGA TCT C REV: CCT GAA TAA GGA AAC AGG TGA AAG	733bp	55	35
R1158X, p.Arg1158X, c.3472C>T	7:117267579-117267579	T	CF2843425	FWD:GGC TTA CAT AAC TGA GAA TTA GGT G REV: GCC AGG ACT TAT TGA GAA GGA	506bp	55	35
p.Gln1382X, , p.Gln1411Pro, c.4242+1G>T' ***** (ordered as 1382, 1411, c.4242)	117305520-117305619		CF3239825; CF2349244; CF6757915	FWD: TGA TTG TGG CTA ACG CTA TAT CA REV: GAA ATG TGC CTC TCA ACT TTG TC	693bp	50.6	35
Q1382X, p.Gln1382X, c.4144C>T	7:117305520-117305520	T	CF3239825	*****			
182delT, p.Phe17SerfsX8, c.50delT	7:117120191-117120192	-	CF3019852, CF3594271, CF7930867, CF9295572	***			
p.Arg75Ter; p.Leu88IlefsTer22, p.Trp57Leu **	117149092-117149183		CF1697504, CF3803349, CF7527369, CF4602380, CA1615190, CA4932026, CF0018616, CF1697504, CF2349244, CF3115703, CF4544212, CF4833948, CF5830853, CF9830825	FWD: CTG CCA CAG TTC TAA ACC AAT AAA REV: GTA AAT TGC CAC CCG TGT TC	731bp	54.5	30
c.262_263del, p.Leu88IlefsTer22	7:117149181-117149183	-	CF1697504, CF3803349, CF7527369	**			
p.Arg104_Ala107del; p.Ser158IlefsTer2 *	117170990-117171152		CF6268769; CF1323468	FWD: CTT GTC TCC CAC TGT TGC TAT AA REV: AGG CTG TGT GAG TCA TCT TAA C	732bp	55	30
c.473del, p.Ser158IlefsTer2	7:117171151-117171152	-	CF6268769	*			
p.Tyr627MetfsTer36; p.Lys684AsnfsX38/p.Gln685ThrfsX4; **** (ordered as: 627: 684/685T)	117232095-117232267		CF0235490; CF1534048; CF3803349; CF6175627, CF6724226, CF7527369	FWD: GTC TGT AAA CTG ATG GCT AAC AAA REV: CTG CTC AGA ATC TGG TAC TAA GG	462	50.6	35
c.1879del, p.Tyr627MetfsTer36	7:117232095-117232096	-	CF0235490; CF1534048; CF3803349	****			
2184delA, p.Lys684AsnfsX38, c.2052delA; 2184insA, p.Gln685ThrfsX4, c.2052dupA	7:117232266-117232267	-	CF1534048, CF6175627, CF6724226, CF7527369	****			
p.His856SerfsTer5, c.2561_2562insGG	7:117235054-117235054	GG	CF4062212; CF3239825	FWD: CAC AAT GGT GGC ATG AAA CTG	445bp	50.6	35

Variant	Location	Allele	Sample to confirm	Primer pair	Amplicon length	Optimal annealing temperature (°C)	Optimal no. of cycles
				REV: TCA GTA GTG GTT CTA CTT GTT GAT T			
p.Ser877PhefsTer29, c.2630del	7:117242889-117242890	-	CF4495056	FWD: CCC AGG AAC ACA AAG CAA AG REV: TGT CAC CTC ACC CAA CTA ATG	364	55	35
c.3963+9G>C; p.Ser1297LeufsTer31 *****	117292905-117292994		CF6188367; CA0144930, CA1615190, CF0018616, CF1697504, CF3512286, CF3594271, CF4471587, CF4833948, CF5158167, CF5181003, CF6188367, CF7527369, CF8213552, CF9111494, CF9830825	FWD: TGT TCA CAA GGG ACT CCA AAT A REV: TAC CAG TGA GGA GAG AAG TAG G	665bp	55	30
c.3889del, p.Ser1297LeufsTer31	7:117292905-117292906	-	CF6188367	*****			
p.Ala1465AspfsTer91; p.Gln1463_1le1464ins *9	117307107-117307108		CA1615190, CA4932026, CA8443975, CF0014912, CF0018616, CF0235490, CF1323468, CF1534048, CF1782680, CF2349244, CF2433640, CF3019852, CF3239825, CF3512286, CF3719491, CF3796568, CF3803349, CF4062212, CF4223536, CF4283433, CF4379523, CF4471587, CF4495056, CF4832869, CF4833948, CF4869626, CF5107567, CF5158167, CF5181003, CF5384911, CF5830853, CF5865254, CF6175627, CF6188367, CF6268769, CF6746590, CF6757915, CF6803591, CF7600423, CF778750, CF7930867, CF8754900, CF9111494, CF9442098, CF9830825	FWD: TTT GAG CCT GTG CCA GTT REV: GAT TGA CAT TTA GAG CTG CCT TTC	609bp	60	35
p.Ala1465AspfsTer91	7:117307107-117307107	AATT	CF4062212	*9			
3905insT, p.Leu1258PhefsX7, c.3773dupT, or c.3773_3774insT	7:117282541-117282541	T	CF0014912	FWD: CTT CCA CTG GTG ACA GGA TAA A REV: CCA AGG CTC CCA CTG TAA AT	531bp	55	30
p.Gly458Val, c.1373G>T	7:117188858-117188858	T	CF4283433	*****			
S549N, p.Ser549Asn, c.1646G>A,	7:117227854-117227854	A	CF1133987	FWD: GGA CCT ATG GAT GAT CTA CAC ATA TT REV: CCA AGA TAC GGG CAC AGA TT	640bp	55	30
c.170G>T, p.Trp57Leu	7:117149093-117149093	T	CA1615190, CA4932026, CF0018616,	**			

Variant	Location	Allele	Sample to confirm	Primer pair	Amplicon length	Optimal annealing temperature (°C)	Optimal no. of cycles
			CF1697504, CF2349244, CF3115703, CF4544212, CF4833948, CF5830853, CF9830825				
S1118F, p.Ser1118Phe, c.3353C>T,	7:117251848-117251848	T	CF3796568	FWD: AGA ATG GCA CCA GTG TGA A REV: CCC TTC AAT CAC AGA ATT GCT ATC	710	55	30
c.4232A>C, p.Gln1411Pro	7:117305608-117305608	C	CF2349244	*****			
1525-1G->A/ c.1393-1G>A; p.Phe508del *8	117199517-117199647		CF4062212; CF1133987, CF1478689, CF1697504, CF1782680, CF2843425, CF3115703, CF3512286, CF3796568, CF3803349, CF4495056, CF4544212, CF4602380, CF5107567, CF5181003, CF5384911, CF5980227, CF6268769, CF7527369, CF7760687, CF7930867, CF8213552, CF9295572, CF9830825, CF9862557	FWD: CCC TTC TCT GTG AAC CTC TAT C REV: TGA GGA CGT TTG TCT CAC TAA T	736bp	55	30
1525-1G->A, or c.1393-1G>A	7:117199517-117199517	A	CF4062212	*8			
c.1680-1G>T	7:117230406-117230406	T	CA4932026, CF0018616, CF1697504, CF2349244, CF3115703, CF4544212, CF4833948, CF5365245, CF5830853	FWD: CTT CAA GGG CAG GAA CTG TAT AA REV: GCA TGA GGC GGT GAG AAA	743bp	54.5	30
4374+1G->T, c.4242+1G>T	7:117305619-117305619	T	CF6757915	*****			
c.312_323del, p.Arg104_Ala107del	7:117170990-117171002	-	CF1323468	*			
F508del, c.1521_1523del, p.Phe508del	7:117199644-117199647	-	CF1133987, CF1478689, CF1697504, CF1782680, CF2843425, CF3115703, CF3512286, CF3796568, CF3803349, CF4495056, CF4544212, CF4602380, CF5107567, CF5181003, CF5384911, CF5980227, CF6268769, CF7527369, CF7760687, CF7930867, CF8213552, CF9295572, CF9830825, CF9862557	*8			
p.Gln1463_Leu1464ins LeuLeuCysProLeuCysAsnValLysMetPheThrHisGlnHisValPhePheAspLeuThrValValLeuAsnCysAspTrpSerTyrSerSerCysArgSerPheThrSerLysProGln	7:117307108-117307108	CTGCTCTGCCCA CTTTGCAATGTG AAAATGTTTACT CACCAACATGTT TTC TTGATCTT ACAGTTGTTATT AATTGTGATTG GAGCTATAGCA GTTGTCCGAGTT	CA1615190, CA4932026, CA8443975, CF0014912, CF0018616, CF0235490, CF1323468, CF1534048, CF1782680,	*9			

Variant	Location	Allele	Sample to confirm	Primer pair	Amplicon length	Optimal annealing temperature (°C)	Optimal no. of cycles
		TTACATCTAAGC CCCAA	CF2349244, CF2433640, CF3019852, CF3239825, CF3512286, CF3719491, CF3796568, CF3803349, CF4062212, CF4223536, CF4283433, CF4379523, CF4471587, CF4495056, CF4832869, CF4833948, CF4869626, CF5107567, CF5158167, CF5181003, CF5384911, CF5830853, CF5865254, CF6175627, CF6188367, CF6268769, CF6746590, CF6757915, CF6803591, CF7600423, CF7778750, CF7930867, CF8754900, CF9111494, CF9442098, CF9830825,				
p.Gln1463_Ile1464insLeuLeuTrpPro LeuCysAsnValLysMetPheThrHisGlnHisV alPhePheAspLeuThrValValIleAsnCysAspT rpSerTyrSerSerCysArgSerPheThrSerLysPr oGln	7:117307108- 117307108	GCTGCTCTGGCC ACITTTGCAATGT GAAAATGTTTAC TCACCAACATGT TTTCTTTGATCT TACAGTTGTTAT TAAITGTGATTG GAGCTATAGCA GTTGTGCGCAGTT TTACATCTAAGC CCCAA					
c..3963+9G>C	7:117292994- 117292994	C	CA0144930, CA1615190, CF0018616, CF1697504, CF3512286, CF3594271, CF4471587, CF4833948, CF5158167, CF5181003, CF6188367, CF7527369, CF8213552, CF9111494, CF9830825	*****			
c.*1043A>C	7:117308205- 117308205	C	CF2954129; CF6268769; CF4495056	FWD: GCT CAC AGA CCT TTG AAC TAG A REV: GCT GGC TGG GAA TCA TAC A	584bp	54.5	30

\*\*,\*\*\*, ... : these variants fall within close proximity, so one primer pair was designed to confirm all variants in the respective region.

Equipment required to complete this protocol:

- a. Thin-walled PCR tubes OR PCR plate
- b. PCR system/thermocycler
- c. Kappa Taq polymerase Ready Mix
  - i. Taq Polymerase
  - ii. MgCl<sub>2</sub>
  - iii. Buffer (for Taq)
- d. Autoclaved dH<sub>2</sub>O OR nuclease-free water
- e. PCR primers
  - i. Primer stock solutions: 100µM concentration

- ii. Primer working solutions: 10 $\mu$ M concentration (100 $\mu$ L prepared)
- f. DNA (concentration determined via nanodrop)
  - i. Stock DNA solutions of 50ng/ $\mu$ L were then prepared so that 1 $\mu$ L of DNA could be easily added to the reaction mixture.
    - i. In the case of nanodrop values <50ng/ $\mu$ L, the DNA for that sample was directly aliquoted and the volume used in the reaction was adjusted.

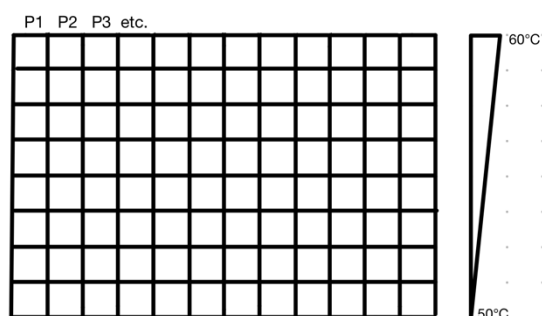
Table 2.2: Components of PCR reaction mixes (final volume: 25 $\mu$ L)

	Volume used in each reaction ( $\mu$ L)
KAPA Taq RM	12.5
dH <sub>2</sub> O	10.5
Forward primer	0.5
Reverse primer	0.5
DNA	1.0 *

\* In the case of DNA concentrations <50ng/ $\mu$ L (using the NanoDrop spectrometer), the DNA for that sample was directly aliquoted and the volume used in the reaction was adjusted in conjunction with the dH<sub>2</sub>O volume to maintain a final reaction volume of 25 $\mu$ L.

#### 2.10.1.2. Optimisation of PCR reactions:

Since ideal annealing temperatures and number of cycles varies between primers, the optimal conditions needed to be established before generating amplicons for each sample. This was done using a gradient thermocycler, which can be programmed to have different annealing temperatures per row. Thus, one run can be used to run PCR reactions for 12 different primers, each at 8 different temperatures (see **Figure 1** below). The temperatures ranged between 50.6°C and 60°C.



**Figure 1:** Concept of gradient thermocycler with each row programmed to a different temperature and each column being used to optimize a different pair of primers.

### *2.11. Amendment of genotype information after confirmation with Sanger*

After confirmation of true-positive variants using Sanger sequencing, the genotype information was again amended. This enabled evaluation of the clinical utility of NGS in providing a full molecular diagnosis. The genotype information was re-evaluated to determine which patients would have been able to be completely genotyped using the gene panel or NGS, and which patients had variants identified that are not present on the panel. Furthermore, ethnicity was also recorded according to the ability of gene panel screening or NGS to effectively genotype individuals.

### *2.12. Conclusion*

The investigation of the spectrum of CFTR variants in South African patients required evaluation of raw NGS data and benefitted from the incorporation of information from Dr. C. Stewart. This methodology enabled thorough evaluation of the molecular nature of the CFTR gene for the South African patients with CF. The methods used are fully reproducible, follow the current best practices of variant discovery, and include “gold-standard” confirmatory Sanger sequencing to ensure valid results are provided for the patients and their families.

### 2.13. References:

- Andrews, S. (2010). "FASTQC. A quality control tool for high throughput sequence data."
- Bardou, P., J. Mariette, F. Escudie, C. Djemiel and C. Klopp (2014). "jvenn: an interactive Venn diagram viewer." *BMC Bioinformatics* **15**: 293.
- BCFtools (2011). *bcftools*. github. H. Li, B. Handsaker, P. Danecek, S. McCarthy and J. Marshall. <https://samtools.github.io/bcftools/>.
- Bell, S. C., M. A. Mall, H. Gutierrez, M. Macek, S. Madge, J. C. Davies, P. R. Burgel, E. Tullis, C. Castanos, C. Castellani, C. A. Byrnes, F. Cathcart, S. H. Chotirmall, R. Cosgriff, I. Eichler, I. Fajac, C. H. Goss, P. Drevinek, P. M. Farrell, A. M. Gravelle, T. Havermans, N. Mayer-Hamblett, N. Kashirskaya, E. Kerem, J. L. Mathew, E. F. McKone, L. Naehrlich, S. Z. Nasr, G. R. Oates, C. O'Neill, U. Pypops, K. S. Raraigh, S. M. Rowe, K. W. Southern, S. Sivam, A. L. Stephenson, M. Zampoli and F. Ratjen (2020). "The future of cystic fibrosis care: a global perspective." *Lancet Respir Med* **8**(1): 65-124.
- Bolger, A. M., M. Lohse and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* **30**(15): 2114-2120.
- Cantarel, B. L., D. Weaver, N. McNeill, J. Zhang, A. J. Mackey and J. Reese (2014). "BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity." *BMC Bioinformatics* **15**(1): 104.
- CFTR2 (2011). "The Clinical and Functional TRanslation of CFTR (CFTR2)."
- Currier, R. J., S. Sciortino, R. Liu, T. Bishop, R. Alikhani Koupaei and L. Feuchtbaum (2017). "Genomic sequencing in cystic fibrosis newborn screening: what works best, two-tier predefined CFTR mutation panels or second-tier CFTR panel followed by third-tier sequencing?" *Genet Med* **19**(10): 1159-1163.
- Ewels, P., M. Magnusson, S. Lundin and M. Käller (2016). "MultiQC: summarize analysis results for multiple tools and samples in a single report." *Bioinformatics* **32**(19): 3047-3048.
- Flygare, S., E. J. Hernandez, L. Phan, B. Moore, M. Li, A. Fejes, H. Hu, K. Eilbeck, C. Huff, L. Jorde, G. R. M and M. Yandell (2018). "The VAAST Variant Prioritizer (VVP): ultrafast, easy to use whole genome variant prioritization tool." *BMC Bioinformatics* **19**(1): 57.
- Forster, M., P. Forster, A. Elsharawy, G. Hemmrich, B. Kreck, M. Wittig, I. Thomsen, B. Stade, M. Barann, D. Ellinghaus, B. S. Petersen, S. May, E. Melum, M. B. Schilhabel, A. Keller, S. Schreiber, P. Rosenstiel and A. Franke (2013). "From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software." *Nucleic Acids Res* **41**(1): e16.
- Garrison, E. and G. Marth (2012). "Haplotype-based variant detection from short-read sequencing." *arXiv* **1207**.
- Goldman, A., C. Graf, M. Ramsay, F. Leisegang and A. T. Westwood (2003). "Molecular diagnosis of cystic fibrosis in South African populations." *S Afr Med J* **93**(7): 518-519.
- Kim, Y. G., M. J. Kim, J. S. Lee, J. A. Lee, J. Y. Song, S. I. Cho, S. S. Park and M. W. Seong (2021). "SnackVar: An Open-Source Software for Sanger Sequencing Analysis Optimized for Clinical Use." *J Mol Diagn* **23**(2): 140-148.
- Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." *Nat Methods* **9**(4): 357-359.
- Li, H. (2013). "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM."
- Lim, R. M., A. J. Silver, M. J. Silver, C. Borroto, B. Spurrier, T. C. Petrossian, J. L. Larson and L. M. Silver (2016). "Targeted mutation screening panels expose systematic population bias in detection of cystic fibrosis risk." *Genet Med* **18**(2): 174-179.
- McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek and F. Cunningham (2016). "The Ensembl Variant Effect Predictor." *Genome Biol* **17**(1): 122.
- Okonechnikov, K., A. Conesa and F. García-Alcalde (2015). "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data." *Bioinformatics* **32**(2): 292-294.
- Poplin, R., V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. Van der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M. J. Daly, B. Neale, D. G. MacArthur and E. Banks (2018). "Scaling accurate genetic variant discovery to tens of thousands of samples." *bioRxiv*: 201178.
- Ravi, R. K., K. Walton and M. Khosroheidari (2018). "MiSeq: A Next Generation Sequencing Platform for Genomic Analysis." *Methods Mol Biol* **1706**: 223-232.
- Robinson, J. T., H. Thorvaldsdottir, A. M. Wenger, A. Zehir and J. P. Mesirov (2017). "Variant Review with the Integrative Genomics Viewer." *Cancer Res* **77**(21): e31-e34.
- Stewart, C. and M. S. Pepper (2016). "Cystic fibrosis on the African continent." *Genet Med* **18**(7): 653-662.
- Tan, A., G. R. Abecasis and H. M. Kang (2015). "Unified representation of genetic variants." *Bioinformatics* **31**(13): 2202-2204.



## *Chapter 3: Results*

The results for the investigation of CFTR variants in South African patients with Cystic Fibrosis.

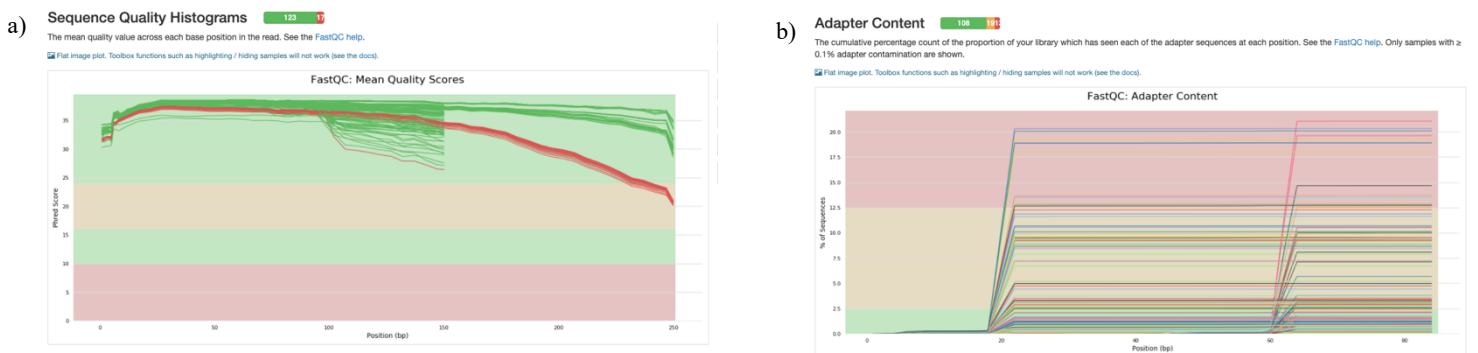
### 3.1. Introduction

This section seeks to present the results obtained through following the methodology outlined in the previous section, with the goal of thoroughly evaluating the variants discovered in this cohort. The analysis is focused on the *Illumina* NGS results and integrates all previous gene panel screening test results. All contributions from Dr. C. Stewart have been indicated where applicable.

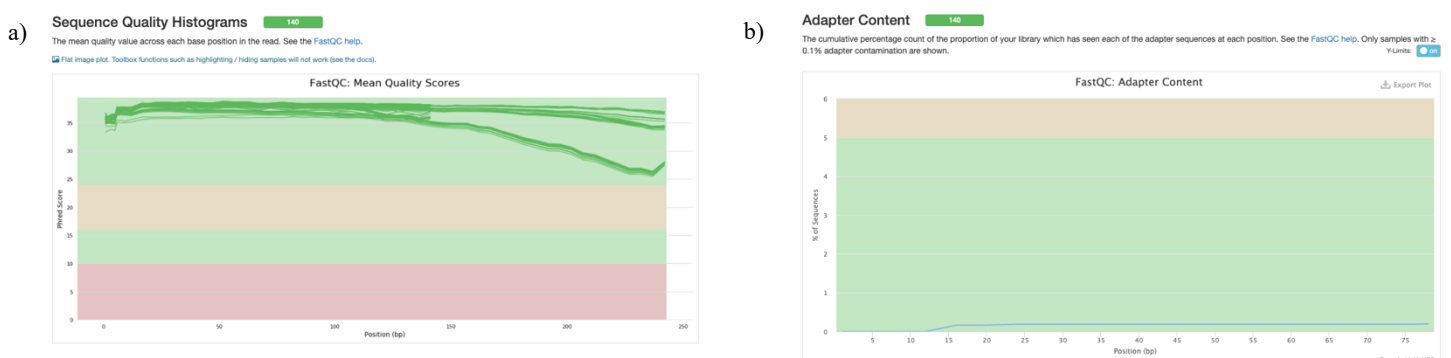
### 3.2. QC, Mapping and Variant detection

#### 3.2.1. FASTQC:

Before trimming, many of the reads were contaminated with adapter sequences (**Figure 3.1**) and the quality of the reads dropped towards the end of the reads. It was also observed that FASTQC failed many of the samples for having “over-represented sequences”. This was not deemed to be an issue that needed resolving, as the sequencing was enriched for the CFTR gene, and the over-represented sequences matched to CFTR using a conventional BLASTN search. After trimming and adapter removal (**Figure 3.2**), the remaining reads for all samples met the minimum quality thresholds and the adapter sequences had been removed and so these forward and reverse paired-end reads were used in the downstream analysis.



**Figure 3.1:** a) Sequence quality for the forward and reverse reads before trimming, b) Adapter content for the forward and reverse reads before adapter removal. Graphs generated with FASTQC and visualised with MultiQC (Andrews 2010, Ewels, Magnusson et al. 2016).



**Figure 3.2:** a) Sequence quality for the forward and reverse reads after trimming, b) Adapter content for the forward and reverse reads after adapter removal. Graphs were generated with FASTQC and visualised with MultiQC (Andrews 2010, Ewels, Magnusson et al. 2016).

### 3.2.2. Mapping:

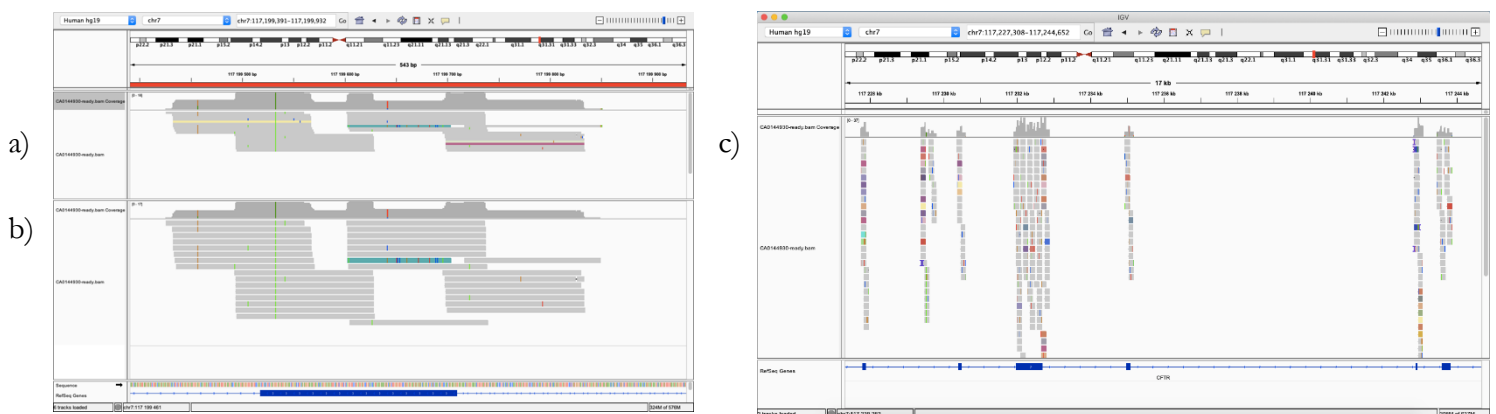
The BCBIO pipeline ran slightly slower when mapping with BWA; however, this reduction in speed was compensated for by higher mapping quality (see Tables 3.1 and 3.2). The average mapping quality for the samples mapped with BWA was 59.3, whereas the average mapping quality for the samples mapped with Bowtie2 was 40.67 (Okonechnikov, Conesa et al. 2015). The mapping to the CFTR region of hg19 is visualised with IGV for one of the samples (**Figure 3.3**). Though the two mapping algorithms do not contradict each other, BWA provided higher quality mapping and so these .bam files were used in the downstream analysis.

**Table 3.1: Summary statistics provided by Qualimap for the samples mapped with Bowtie2.**

Number of samples	70
Total number of mapped reads	86,770,885
Mean samples coverage	23,888.87
Mean samples GC-content	41.18
Mean samples mapping quality	40.67
Mean samples insert size	245.29

**Table 3.2: Summary statistics provided by Qualimap for the samples mapped with BWA.**

Number of samples	70
Total number of mapped reads	87,017,064
Mean samples coverage	23,910.08
Mean samples GC-content	41.18
Mean samples mapping quality	59.3
Mean samples insert size	245.26

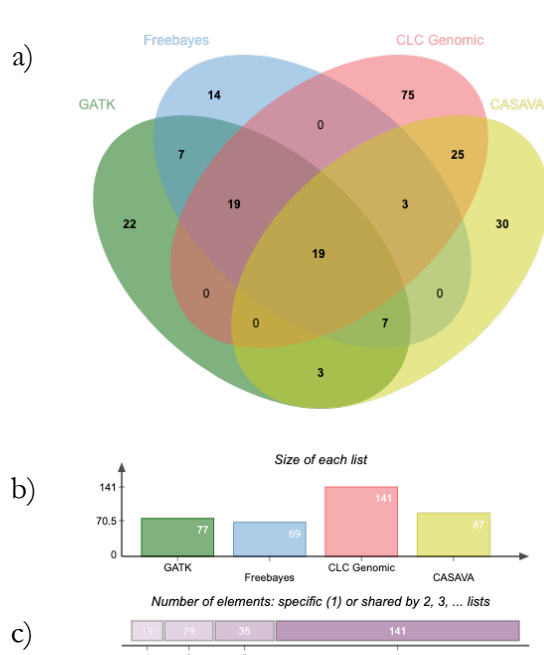


**Figure 3.3:** An example of the mapping using Bowtie2 (a) and BWA (b), visualized with IGV, for sample CA0144930. c) Mapping with Bowtie2 for sample CA014493, zoomed in in IGV.

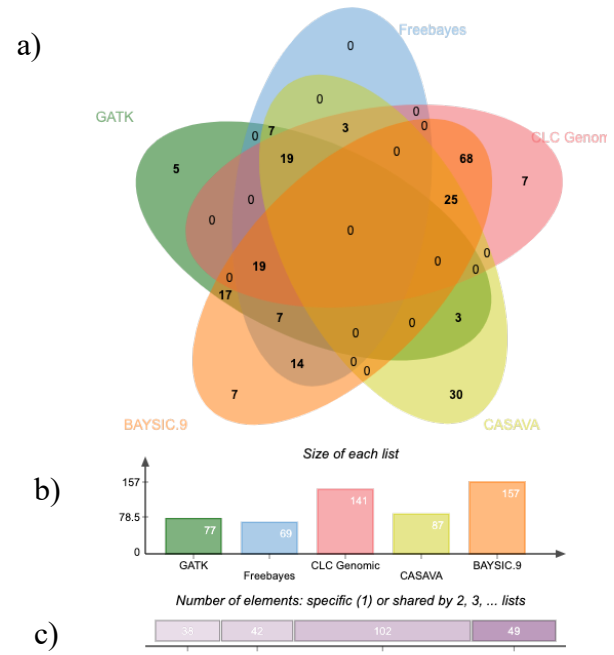
### 3.3. Variant Detection and *in silico* validation:

The concordance of variant calls can be seen in the Venn diagram below (**Figure 3.4a**). The overall concordance between the four variant calling algorithms was low. Only 19 positions were common to all four variant sets, 159 positions were unique to one variant caller, 35 were called by two callers and 29 were called by 3 callers (**Figure 3.4c**). When these results were combined with the BAYSIC call set, there were no variants that were present in all sets (**Figure 3.5a**). This also helped to confirm that 30 of the *Illumina* CASAVA variants had a low probability of being true-positive calls and were unsuitable for use in downstream validation steps. The 19 variants that were concordant between the four calling methods were not validated with BAYSIC, but a different set of 19

variants was concordant between GATK, Freebayes, CLC Genomics and BAYSIC. This may point to a set of high-quality variants that are suitable for experimental validation.



**Figure 3.4:** a) Venn diagram of the concordance between the different variant calling algorithms. b) Graph comparing the size of each variant list. c) Graph showing elements specific or shared. Produced using jvenn (Bardou, Mariette et al. 2014).



**Figure 3.5:** a) Venn diagram of the concordance between the different variant calling algorithms and BAYSIC. b) Graph comparing the size of each variant list. c) Graph showing elements specific or shared. Produced using jvenn (Bardou, Mariette et al. 2014).

### 3.4. Variant Annotation and Effect Prediction

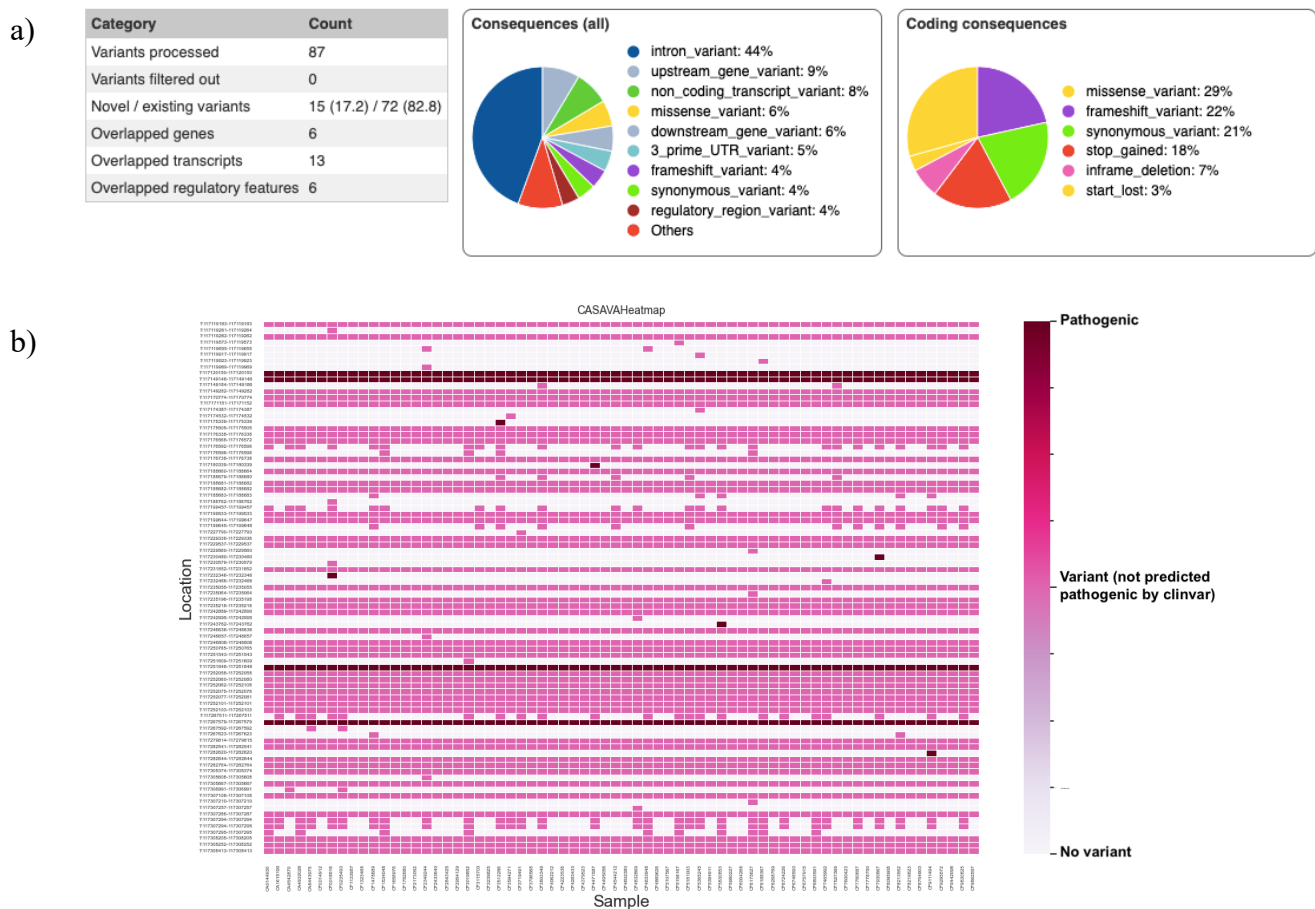
The summary statistics of the variant effect prediction for the *Illumina* CASAVA set of variant calls (**Figure 3.6a**) show that all 87 variants were processed, with 15 novel variants identified. 44% of the variants were found to be intron variants, followed by upstream gene variants and non-coding transcript variants being the second and third most common. For the coding variants, 29% were identified as missense variants, followed by 22% being frameshift variants. It can be seen from the heatmap (**Figure 3.6b**) that many of the variants were detected in all (or almost all) of the samples, which may be due to a lack of stringent quality filtering by the *Illumina* CASAVA algorithm. Furthermore, four variants, identified as pathogenic by ClinVar, were detected in all samples.

The VEP results for the CLC Genomics dataset (.vcf files prepared by Dr. C. Stewart) are summarised in **Figure 3.7a**. All 159 variants were processed, with 77 novel variants and 49% of the variants being intronic. Again, the next most frequent variant types are upstream gene variant and non-coding transcript variant. When looking at the heatmap for this dataset (**Figure 3.7b**), common variants also seem to be evident. The upstream gene, PDE1C was not filtered out and

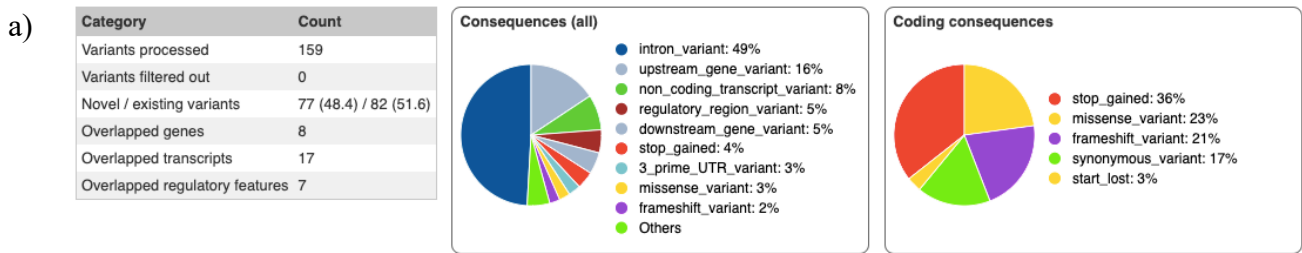
remains in the call set due to mapping to the entire chromosome instead of only mapping to the CFTR gene region. Furthermore, a few variants were again detected in all samples, which may be indicative of a common variant or sequencing artefact.

The summary statistics for the VEP results of the Freebayes variant call set (**Figure 3.8a**) show that only 69 variants were processed, with 14 of these being novel. The variant types are more evenly distributed in this dataset; the most common variant type is missense making up 18% of the variants. The heatmap also displays fewer variants overall as well as fewer variants being detected in all samples (**Figure 3.8b**).

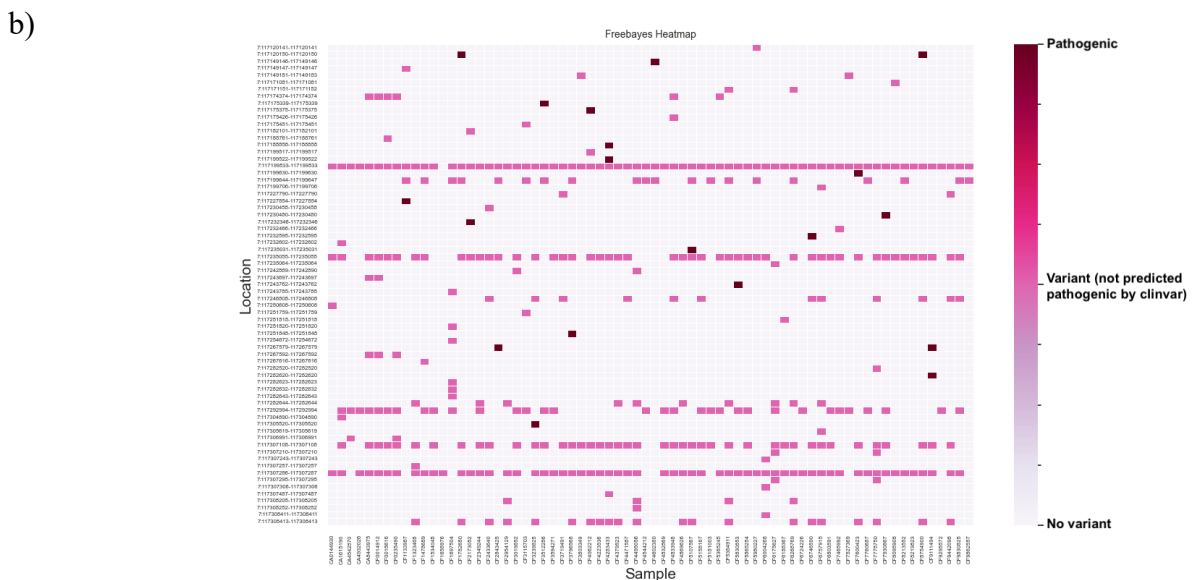
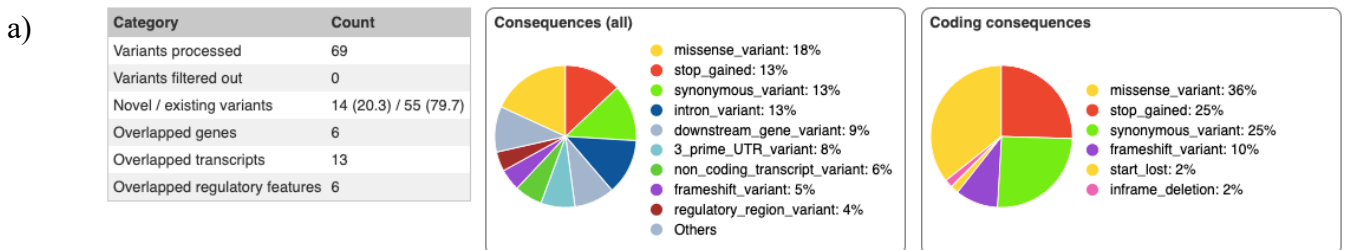
Similar results are observed for the summary statistics of the VEP results for the GATK variant set (**Figure 3.9a**); 77 variants were processed by VEP and 20 of these were novel. The most common variant type is missense, representing 14% of the variants. The heatmap (**Figure 3.9b**) again identifies a few variants that are common to a subset of samples, without any variants that are present in all samples.



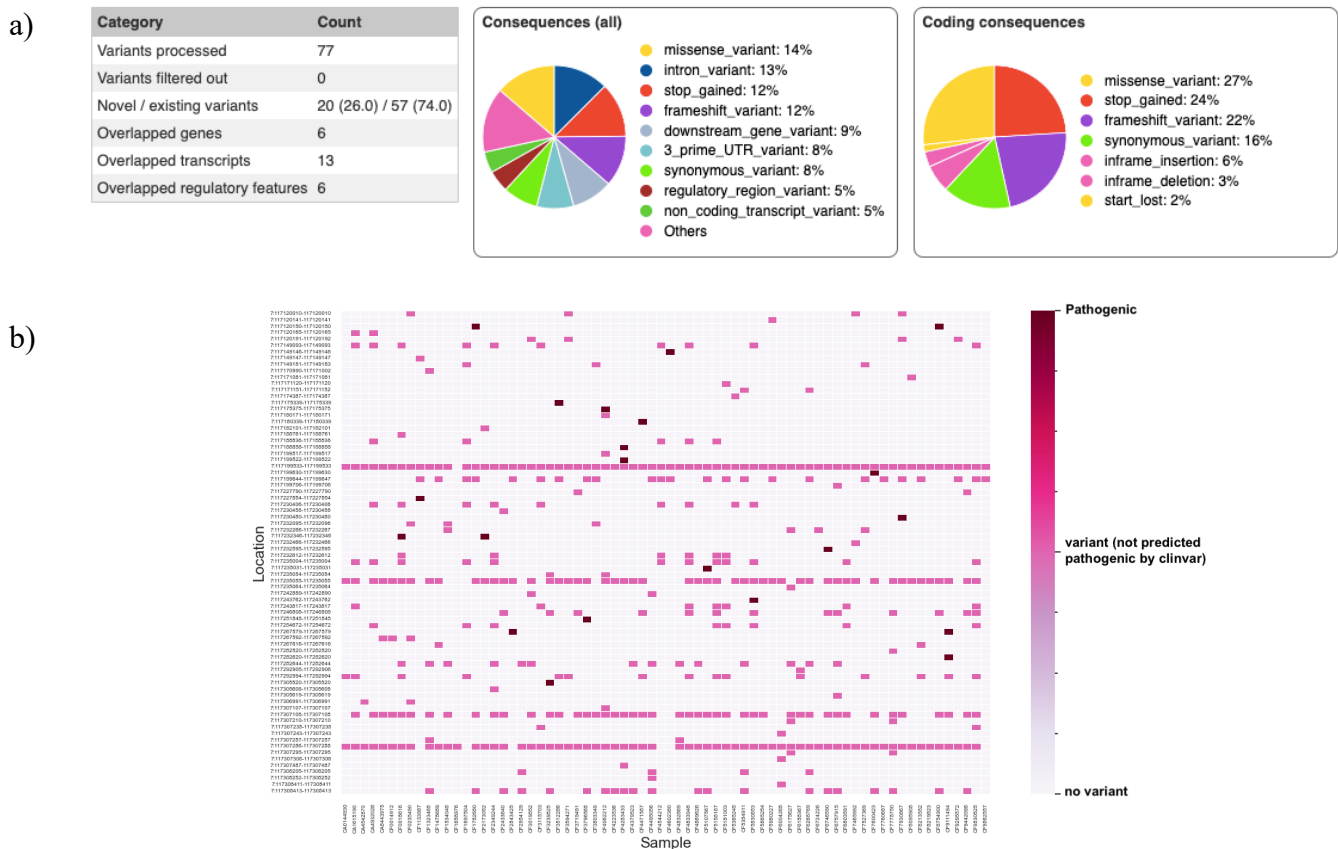
**Figure 3.6:** VEP results for the CASAVA variant call set. a) Summary statistics of the VEP output for the CASAVA variant call set. b) Heatmap for the variants in each of the samples.



**Figure 3.7:** VEP results for the CLC Genomics variant call set. a) Summary statistics of the VEP output for the CLC Genomics variant call set. b) Heatmap for the variants in each of the samples.



**Figure 3.8:** VEP results for the Freebayes variant call set. a) Summary statistics of the VEP output for the Freebayes variant call set. b) Heatmap for the variants in each of the samples.



**Figure 3.9:** VEP results for the GATK variant call set. a.) Summary statistics of the VEP output for the GATK variant call set. b.) Heatmap for the variants in each of the samples.

### 3.5. Compilation of a Master Variant list

The initial Master Variant List consists of 102 variants that fulfilled the criteria. 19 of these variants had already been confirmed using Sanger sequencing in some patients and one had been invalidated (Dr. C. Stewart). Additionally, five of the variants are present on the NHLs panel which was used to do initial genotyping of the patient cohort. Three of the variants have also been suggested for inclusion on a South African population-specific gene panel (Goldman, Graf et al. 2003).

#### 3.5.1. CFTR2 annotations:

The CFTR2 database provides valuable functional information, particularly regarding the variants that have been proven to be functionally causative of CF and those that are not pathogenic/disease-causing. Three of the variants were recorded in the CFTR2 database as “not causative of CF”: p.Arg75Gln (missense variant), p.Arg1162Leu (missense variant) and 125G/C (5'-UTR variant). Two missense variants were found at positions where other variants have been recorded in CFTR2 as “not causative of CF”: p.Val470Met (p.Met470Val was recorded as “not causative of CF”) and p.Met807Ile (p.Ile807Met was recorded as “not causative of CF”).

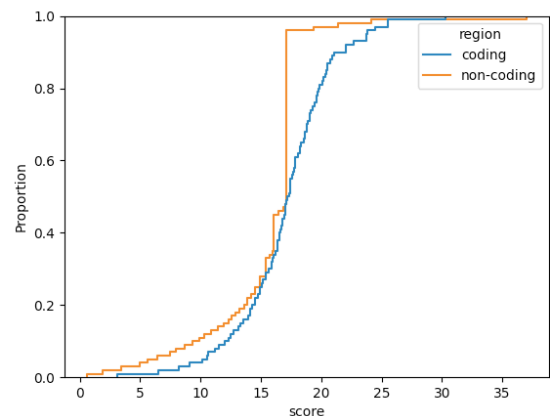
23 variants were annotated in the CFTR2 database as having been confirmed to functionally cause CF. Nine of these variants were not included as potentially pathogenic variants by Dr. C. Stewart. Finally, eight variants were found at the same positions as variants which have been annotated in CFTR2 as “causative of CF”. 70 variants did not have any functional annotations in the CFTR2 database.

### 3.5.2. Variant pathogenicity scores

Variant pathogenicity scores are particularly useful for variants that have not yet been functionally annotated in the CFTR2 database, as well as variants that have been found in positions where different nucleotide changes have been found to be functionally significant. After selecting variants identified as causative of CF (using CFTR2) and variants with severe amino acid consequences (i.e., frame-shift variants, stop-gain variants, etc.), a few missense variants were identified as being *likely pathogenic* using the various pathogenicity predictors. The full spreadsheet with all annotations is available in Supplementary 1. The final candidates for confirmation with Sanger sequencing are provided in Table 3.3-3.9 below.

#### 3.5.2.1. VVP percentile scores:

The results of the VVP analysis for further pathogenicity scoring are presented here. Each score for each variant, looked up from the eCDF plot (Figure 3.10) using Python, is available in the Supplementary 1. The VVP scores corroborated the candidate variants and provided another level of confidence in the predicted pathogenic variants.



**Figure 3.10:** eCDF plot used to normalise the background distribution of percentile scores for CFTR from gnomAD.

#### 3.5.2.2. Candidates for validation with Sanger

The potentially pathogenic variants which needed to be confirmed with Sanger are provided in Table 3.3-3.9 below. The variants have been split according to amino acid consequence.

**Table 3.3: Start-lost variants**

Variant	#Uploaded_variation	NGS SAMPLE	Allele	Consequence	IMPACT	HGVSc	HGVSp	CLIN_SIG	clinvar_clnsig	Gene panel	CFTR2 annotation	Variant caller
p.Met1Thr	rs397508476	CF1782680, CF8754900	C	start_lost	HIGH	ENST0000000308 4.6:c.21>C	ENSP0000000308 4.6:p.Met1?	pathogenic/likely pathogenic	Pathogenic/Likely _pathogenic			GATK Freebayes CLC CASAVA



Table 3.4: Stop-gain variants

Variant	#Uploaded_variation	NGS SAMPLE	Allele	Consequence	IMPACT	HGVSc	HGVSp	CLIN_SIG	clinvar_clnsig	Gene panel	CFTR2 annotation	Variant caller
p.Arg75Ter	rs121908749	CF4602380	T	stop_gained	HIGH	ENST0000000308 4.6:c.223C>T	ENSP0000000308 4.6:p.Arg75Ter	pathogenic	Pathogenic		yes	GATK Freebayes CLC CASAVA
p.Leu218Ter	rs397508777	CF4062212	A	stop_gained	HIGH	ENST0000000308 4.6:c.653T>A	ENSP0000000308 4.6:p.Leu218Ter	not_provided,pathogenic	Pathogenic		yes	GATK Freebayes CLC BAYSIC.9
p.Arg303AlafsTer16		CF4062212	TCGGAAGGCA GCCTATGTGG CCACTTGGCAA TGTGAAAATG TTTACTACCA ACATGTTTCT TTGATCTTACA GTTGTTATTAA TTGTGATTGG AGCTATAGCA GTTGTCGAG TTTACATCGG AAGGAGCCT ATGTG	stop_gained,frame shift_variant	HIGH	ENST0000000308 4.6:c.906_907msG CCACTTGGCAA TGTGAAAATG TTTACTACCA ACATGTTTCT TTGATCTTACA GTTGTTATTAA TTGTGATTGG AGCTATAGCA GTTGTCGAG TTTACATCGG AAGGAGCCT ATGTG	ENSP0000000308 4.6:p.Arg303AlafsTer16	-	-			GATK BAYSIC.9
p.Leu383Ter		CF2173052	A	stop_gained	HIGH	ENST0000000308 4.6:c.1148T>A	ENSP0000000308 4.6:p.Leu383Ter		-			GATK Freebayes CLC BAYSIC.9
p.Gly451Ter		CA4932026, CF1697504, CF4544212, CF4833948, CF5158167	T	stop_gained	HIGH	ENST0000000308 4.6:c.1351G>T	ENSP0000000308 4.6:p.Gly451Ter		-			GATK BAYSIC.9
p.Arg709X	rs121908760	CF0018616, CF2173052	T	stop_gained	HIGH	ENST0000000308 4.6:c.2125C>T	ENSP0000000308 4.6:p.Arg709Ter	pathogenic	Pathogenic		yes	GATK Freebayes CLC CASAVA
p.Gly1125Ter		CA4932026, CF1697504, CF2349244, CF5158167, CF5181003, CF5830853, CF6803591, CF9830825	T	stop_gained	HIGH	ENST0000000308 4.6:c.3373G>T	ENSP0000000308 4.6:p.Gly1125Ter		-			GATK Freebayes BAYSIC.9
p.Arg1158X	rs79850223	CF2843425, CF9111494	T	stop_gained	HIGH	ENST0000000308 4.6:c.3472C>T	ENSP0000000308 4.6:p.Arg1158Ter	pathogenic	Pathogenic		yes	GATK Freebayes CLC CASAVA
p.Gln1382X	rs397508684	CF3239825	T	stop_gained	HIGH	ENST0000000308 4.6:c.4144C>T	ENSP0000000308 4.6:p.Gln1382Ter	pathogenic	Pathogenic		yes	GATK Freebayes CLC BAYSIC.9

Table 3.5: Frame-shift variants

Variant	#Uploaded_variation	NGS SAMPLE	Allele	Consequence	IMPACT	HGVSc	HGVSp	CLIN_SIG	clinvar_clnsig	Gene panel	CFTR2 annotation	Variant caller
p.Phe17SerfsX8, c.50delT	rs397508742, rs397508714	CF3019852, CF3594271, CF7930867, CF9295572	-	frameshift_variant	HIGH	ENST0000000308 4.6:c.50del	ENSP0000000308 4.6:p.Phe17SerfsTer8	pathogenic	-		yes	GATK BAYSIC.9
p.Leu881lefsTer22	rs75414777;rs121908769	CF1697504, CF3803349, CF7527369	-	frameshift_variant	HIGH	ENST0000000308 4.6:c.262_263del	ENSP0000000308 4.6:p.Leu881lefsTer22	pathogenic	-	NHLS	yes	GATK, Freebayes, BAYSIC.9
p.Ser1581lefsTer2		CF5384911, CF6268769	-	frameshift_variant	HIGH	ENST0000000308 4.6:c.473del	ENSP0000000308 4.6:p.Ser1581lefsTer2		-			GATK Freebayes CLC CASAVA
p.Tyr627MetfsTer36		CF0235490, CF1534048, CF3803349	-	frameshift_variant	HIGH	ENST0000000308 4.6:c.1879del	ENSP0000000308 4.6:p.Tyr627MetfsTer36		-			GATK BAYSIC.9
p.Lys684AsnfsX38	rs116497484;rs121908746;rs113169227;rs121908786	CF1534048, CF6175627, CF6724226, CF7527369	-	frameshift_variant	HIGH	ENST0000000308 4.6:c.2046del	ENSP0000000308 4.6:p.Lys684AsnfsTer38	pathogenic	-		yes	GATK only
p.His856SerfsTer5		CF3239825, CF4062212	GG	frameshift_variant	HIGH	ENST0000000308 4.6:c.2561_2562msGG	ENSP0000000308 4.6:p.His856SerfsTer5		-			GATK BAYSIC.9
p.Ser877PhefsTer29		CF3019852, CF4495056	-	frameshift_variant	HIGH	ENST0000000308 4.6:c.2630del	ENSP0000000308 4.6:p.Ser877PhefsTer29		-			GATK Freebayes CLC CASAVA
p.Ser1297LeufsTer31		CF6188367	-	frameshift_variant	HIGH	ENST0000000308 4.6:c.3889del	ENSP0000000308 4.6:p.Ser1297LeufsTer31		-		yes (at this position)	GATK BAYSIC.9
p.Ala1465AspfsTer91		CF4062212	AATT	frameshift_variant	HIGH	ENST0000000308 4.6:c.4388_4389msAATT	ENSP0000000308 4.6:p.Ala1465AspfsTer91		-			GATK BAYSIC.9
p.Leu1258PhefsX7		CF0014912	T	frameshift_variant	HIGH	ENST0000000308 4.6:c.3773dup	ENSP0000000308 4.6:p.Leu1258PhefsTer7		-		yes	CLC CASAVA BAYSIC.9

Table 3.6: Indels

Variant	#Uploaded_variation	NGS SAMPLE	Allele	Consequence	IMPACT	HGVSc	HGVSp	CLIN_SIG	clinvar_clnsig	Gene panel	CFTR2 annotation	Variant caller
p.Arg104_Ala107del		CF1323468	-	inframe_deletion	MODERATE	ENST0000000308 4.6:c.312_323del	ENSP0000000308 4.6:p.Arg104_Ala107del		-			GATK BAYSIC.9
p.Phe508del	rs120706083;rs113993960	CF1133987, CF1478680, CF1697504, CF1782680, CF2843425, CF3115703, CF3512286, CF3796568, CF3803349, CF4495056, CF4544212, CF4602380, CF5107367, CF5181003, CF5384911, CF5980227, CF6268769, CF7527369, CF7760687, CF7930867, CF8213552, CF9295572, CF9830825, CF9862557	-	inframe_deletion	MODERATE	ENST0000000308 4.6:c.1521_1523del	ENSP0000000308 4.6:p.Phe508del	likely_pathogenic, pathogenic, drug_response, risk_factor	-	NHLS	yes	GATK Freebayes CASAVA
p.Gln1463_Ile1464insLeu.Leu.CysProLeu.Cys.Asn.Val.LysHisValPheHisGlnLeuThrValValleA		CA1615190, CA4932026, CA8445975, CF0014912, CF0018616, CF0235490, CF1323468, CF1534048, CF1782680, CF2349244, CF2433540, CF3019852, CF3239825, CF3512286, CF3719491, CF3796568	CTGCTCTGCC ACITTGCAATG TGA AAAATGTTT ACTCACCAACA TGT TTTCTTGG ATCTTACAGTT	inframe_insertion	MODERATE	ENST0000000308 4.6:c.4389_4390ms CTGCTCTGCC ACTTGTGCAATG TGA AAAATGTTT ACTCACCAACA	ENSP0000000308 4.6:p.Gln1463_Ile1464insLeu.Leu.CysProLeu.Cys.Asn.Val.LysHisValPheHisGlnLeuThrValValleA		-			GATK only

Variant	#Uploaded_variation	NGS SAMPLE	Allele	Consequence	IMPACT	HGVSc	HGVSp	CLIN_SIG	clinvar_clnsig	Gene panel	CFTR2 annotation	Variant caller
snCysAsp1TrpSerTyrSerSerCysArgSerPheThrSerLysProGln		CF380339, CF4062212, CF422336, CF4283433, CF4379523, CF4471387, CF4491056, CF4832869, CF4833948, CF4869626, CF5107567, CF5158167, CF5181003, CF5384911, CF5830853, CF5865254, CF6175627, CF6188367, CF6268769, CF6746590, CF6757915, CF6803591, CF7600423, CF7778750, CF7930867, CF8754900, CF9111494, CF9442098, CF9830825	GTTATTAAATGGTATTTGGAGCATTATAGCAGTTGTCGCAGTTTATCATCTAAGCCCAA			TGTTTTCTTTGATCTTACAGTTTATTAATTTGATTTGGAGCATTATAGCAGTTTATAGCAGTTTTCGCAGTTTACATCTAAGCCCAA	AspLeuThrValValIleAsnCysAspTrpSerTyrSerSerCysArgSerPheThrSerLysProGln					
p.Gln1463_Ile1464insLeuLeuTrpProLeuCysAsnValLysMetPheThrHisGlnHisValPheAspLeuThrValValLeuAsnCysAsp1TrpSerTyrSerSerCysArgSerPheThrSerLysProGln	rs1800136	CA1615190, CA4932026, EA8443975, CF0014912, CF0018616, CF0235490, CF1323468, CF1534048, CF1762580, CF2349244, CF2433640, CF3019852, CF3239825, CF3512286, CF3719491, CF3796568, CF3803349, CF4062212, CF422336, CF4283433, CF4379523, CF4471387, CF4491056, CF4832869, CF4833948, CF4869626, CF5107567, CF5158167, CF5181003, CF5384911, CF5830853, CF5865254, CF6175627, CF6188367, CF6268769, CF6746590, CF6757915, CF6803591, CF7600423, CF7778750, CF7930867, CF8754900, CF9111494, CF9442098, CF9830825	GCTGCTCTGGCCACTTTGCCAATGCAAAATGTTACTACCAACATGTTTCTTTGATCTTACATTTGATCTTACAGTTTGTGATCTTACAGTTTGTGATTTGGAGTTTACATCTAAGCCCAA	inframe_insertion	MODERATE	NM_000492.4:c.4389_4390insCTGCTCTGCCACTTTGCCAATGCAAAATGTTACTACCAACATGTTTCTTTGATCTTACAGTTTGTGATCTTACAGTTTGTGATTTGGAGTTTACATCTAAGCCCAA	NP_000483.3:p.Gln1463_Ile1464insLeuLeuTrpProLeuCysAsnValLysMetPheThrHisGlnHisValPheAspLeuThrValValLeuAsnCysAsp1TrpSerTyrSerSerCysArgSerPheThrSerLysProGln					GATK only

Table 3.7: Splice variants

Variant	#Uploaded_variation	NGS SAMPLE	Allele	Consequence	IMPACT	HGVSc	HGVSp	CLIN_SIG	clinvar_clnsig	Gene panel	CFTR2 annotation	Variant caller
c.1393-1G>A	rs397508200	CF4062212	A	splice_acceptor_variant	HIGH	ENST00000003084.6:c.1393-1G>A		pathogenic	-		yes	GATK Freebayes CLC BAYSIC.9
c.1680-1G>T		CA4932026, CF0018616, CF1697504, CF2349244, CF3115703, CF4544212, CF4833948, CF5365245, CF5830853	T	splice_acceptor_variant	HIGH	ENST00000003084.6:c.1680-1G>T					yes (variants at this position)	GATK BAYSIC.9
c.4242+1G->T	rs372227120	CF6757915	T	splice_donor_variant	HIGH	ENST00000003084.6:c.4242+1G>T		pathogenic	-		yes	GATK Freebayes CLC BAYSIC.9

Table 3.8: Missense variants

Variant	#Uploaded_variation	NGS SAMPLE	Allele	Consequence	IMPACT	HGVSc	HGVSp	CLIN_SIG	clinvar_clnsig	Gene panel	CFTR2 annotation	Variant caller
p.Gly458Val	rs121909009	CF4283433	T	missense_variant	MODERATE	ENST00000003084.6:c.1373G>T	ENSP00000003084.6:p.Gly458Val	pathogenic	Pathogenic		yes	GATK Freebayes BAYSIC.9
p.Ser549Asn	rs121908755	CF1133987	A	missense_variant	MODERATE	ENST00000003084.6:c.1646G>A	ENSP00000003084.6:p.Ser549Asn	pathogenic,drug_response	Pathogenic,drug_response	NHLS	yes	GATK Freebayes CLC BAYSIC.9
p.Trp57Leu		CA1615190, CA4932026, CF0018616, CF1697504, CF2349244, CF3115703, CF4544212, CF4833948, CF5830853, CF9830825	T	missense_variant	MODERATE	ENST00000003084.6:c.170G>T	ENSP00000003084.6:p.Trp57Leu				yes (at this position)	GATK BAYSIC.9
p.Ser1118Phe	rs146521846	CF3796568	T	missense_variant	MODERATE	ENST00000003084.6:c.3353C>T	ENSP00000003084.6:p.Ser1118Phe	pathogenic	Pathogenic		yes	GATK Freebayes CASAVA
p.Gln1411Pro	rs150177304	CF2349244	C	missense_variant	MODERATE	ENST00000003084.6:c.4232A>C	ENSP00000003084.6:p.Gln1411Pro	uncertain_significance	Uncertain_significance		yes (at this position)	GATK, CASAVA

Table 3.9: Non-coding variants

Variant	#Uploaded_variation	NGS SAMPLE	Allele	Consequence	IMPACT	HGVSc	HGVSp	CLIN_SIG	clinvar_clnsig	Gene panel	CFTR2 annotation	Variant caller
c.3963+9G>C		CA0144930, CA1615190, CF0018616, CF1697504, CF3512286, CF3594271, CF4471587, CF4833948, CF5158167, CF5181003, CF6188367, CF7527369, CF8213552, CF9111494, CF9830825	C	intron_variant	MODIFIER	ENST00000003084.6:c.3963+9G>C					yes (at this position)	GATK Freebayes CLC BAYSIC.9
c.*1043A>C	rs10234329	CF2954129, CF4495056, CF5384911, CF6268769	C	3_prime_UTR_variant	MODIFIER	ENST00000003084.6:c.*1043A>C		benign,likely_benign				GATK Freebayes CLC CASAVA

### 3.6. Amendment of genotype information following most recent NGS

The genotype information and previous confirmation results from Dr. C. Stewart, amended with potentially pathogenic variant results from the NGS analysis before validation, are available in Table 3.10 below. All heterozygous variants which were identified as being potentially pathogenic were found in cis configuration, so these results are not specifically reported in the tables.

**Table 3.10: Existing genotype information and variants identified using NGS**

Sample	Ethnicity	NHLS genotype	NGS genotype (C. Stewart)	Confirmed By C. Stewart	NGS Variant 1	NGS Variant 2	NGS Variant 3	NGS Variant 4	NGS Variant 5	NGS Variant 6
CA0144930	No Data	No Data	No Data		c.3963+9G>C					
CA1615190	No Data	No Data	No Data		p.Trp57Leu	c.3963+9G>C	p.Gln1463_IIe1464ins41			
CA4542870	No Data	No Data	No Data							
CA4932026	No Data	No Data	No Data		p.Trp57Leu	p.Gly451Ter	c.1680-1G>T	p.Gly1125Ter	p.Gln1463_IIe1464ins41	
CA8443975	No Data	No Data	No Data		p.Gln1463_IIe1464ins41					
CF0014912	No Data	No Data	No Data		p.Gln1463_IIe1464ins41	p.Leu1258PhefsX7				
CF0018616	Black	U/U	R709X/p.Ser427ThrsX16	Confirmed	p.Trp57Leu	p.Ser427ThrsTer16	c.1680-1G>T	p.Arg709X	c.3963+9G>C	p.Gln1463_IIe1464ins41
CF0235490	White	U/U	N/N	N/A	p.Tyr627MetfsTer36	p.Gln1463_IIe1464ins41				
CF1133987	White	DF508/Ser549Asn*	No Data		p.Ser549Asn	DF508				
CF1323468	Black	U/U	U/U	N/A	p.Arg104_Ala107del	p.Gln1463_IIe1464ins41				
CF1478689	Mixed	DF508/U	DF508/Gly173GlnfsX21	Confirmed	DF508	p.Gly1173GlnfsTer21				
CF1534048	Black	U/U	U/U	N/A	p.Tyr627MetfsTer36	p.Lys684AsnfsX38/p.Gln685ThrsX4	p.Gln1463_IIe1464ins41			
CF1658976	Mixed	U/U	U/U	N/A						
CF1697504	White	DF508/394delTT	DF508/394delTT		p.Trp57Leu	p.Leu88IlefsTer22	p.Gly451Ter	DF508	c.1680-1G>T	p.Gly1125Ter
CF1782680	No Data	No Data	No Data		p.Met1Thr	DF508	p.Gln1463_IIe1464ins41			
CF2173052	Black	U/U	Leu383X/Arg709X	Leu383X (Not Validated By Cheryl)/Arg709X (Not Validated For This Patient)	p.Leu383Ter	p.Arg709X				
CF2349244	Black	U/U	U/U	N/A	p.Trp57Leu	c.1680-1G>T	p.Gly1125Ter	p.Gln1411Pro	p.Gln1463_IIe1464ins41	
CF2433640	Mixed	U/U	3120+1G>A/Tyr577X	Confirmed	p.Tyr577Ter	3120+1G>A	p.Gln1463_IIe1464ins41			
CF2843425	No Data	No Data	No Data		DF508	p.Arg1158X				
CF2954129	Mixed	U/U	U/U	N/A	c.*1043A>C					
CF3019852	Mixed	3272-26A>G/U	3272-26A>G/Ser877PhefsX29	Confirmed	p.Phe17SerfsX8	p.Ser877PhefsTer29	p.Gln1463_IIe1464ins41			
CF3115703	White	DF508/DF508	DF508/DF508	N/A	p.Trp57Leu	DF508	c.1680-1G>T			

Sample	Ethnicity	NHLS genotype	NGS genotype (C. Stewart)	Confirmed By C. Stewart	NGS Variant 1	NGS Variant 2	NGS Variant 3	NGS Variant 4	NGS Variant 5	NGS Variant 6
CF3239825	Mixed	3120+1G>A /U	No Data		p.His856SerfsTer5	3120+1G->A	p.Gln1382X	p.Gln1463_IIe1464ins41		
CF3512286	White	DF508/U	DF508/Leu206Trp	Confirmed	p.Leu206Trp	DF508	c.3963+9G>C	p.Gln1463_IIe1464ins41		
CF3594271	Black	U/U	U/U	N/A	p.Phe17SerfsX8	c.3963+9G>C				
CF3719491	Black	3120+1G>A /U	3120+1G>A /U	N/A	3120+1G->A	p.Gln1463_IIe1464ins41				
CF3796568	No Data	No Data	No Data		DF508	p.Ser1118Ph e	p.Gln1463_IIe1464ins41			
CF3803349	No Data	DF508/U	DF508/394delTT*	Confirmed	p.Leu88IlefsTer22	DF508	p.Tyr627MetfsTer36	p.Gln1463_IIe1464ins41		
CF4062212	Indian	U/U	No Data		p.Leu218Ter	p.Arg303AlafsTer16	1525-1G->A,orc.1393-1G>A	p.His856SerfsTer5	p.Ala1465AspfsTer91	p.Gln1463_IIe1464ins41
CF4223536	Indian	U/U	U/n.166+3321T>G	Not Confirmed: C. Stewart Filtered Out With GMAF	p.Gln1463_IIe1464ins41					
CF4283433	Black	U/U	Gly458Val/Ser466X	Gly458Val (Not Confirmed)/Ser466X (Confirmed)	p.Gly458Val	p.Ser466Ter	p.Gln1463_IIe1464ins41			
CF4379523	Mixed	U/U	No Data		p.Gln1463_IIe1464ins41					
CF4471587	White	3120+1G>A /U	3120+1G>A /Arg352Gln	Confirmed	p.Arg352Gln	3120+1G->A	c.3963+9G>C	p.Gln1463_IIe1464ins41		
CF4495056	Mixed	DF508/U	No Data		DF508	p.Ser877PhefsTer29	p.Gln1463_IIe1464ins41	c.*1043A>C		
CF4544212	White	DF508/DF508	No Data		p.Trp57Leu	p.Gly451Ter	DF508	c.1680-1G>T		
CF4602380	No Data	No Data	No Data		p.Arg75Ter	DF508				
CF4832869	Mixed	U/U	U/U	N/A	p.Gln1463_IIe1464ins41					
CF4833948	Black	3120+1G>A /3120+1G>A	3120+1G>A /3120+1G>A	N/A	p.Trp57Leu	p.Gly451Ter	c.1680-1G>T	3120+1G->A	c.3963+9G>C	p.Gln1463_IIe1464ins41
CF4869626	Mixed	U/U	No Data		p.Gln1463_IIe1464ins41					
CF5107567	Mixed	DF508/U	DF508/Trp846X*	Confirmed	DF508	p.Trp846X	p.Gln1463_IIe1464ins41			
CF5158167	Black	3120+1G>A /3120+1G>A	3120+1G>A /3120+1G>A	N/A	p.Gly451Ter	3120+1G->A	p.Gly1125Ter	c.3963+9G>C	p.Gln1463_IIe1464ins41	
CF5181003	White	DF508/DF508	DF508/DF508	N/A	DF508	p.Gly1125Ter	c.3963+9G>C	p.Gln1463_IIe1464ins41		
CF5365245	Black	U/U	U/Leu183Ile	Confirmed	p.Leu183Ile	c.1680-1G>T				
CF5384911	Mixed	DF508/U	DF508/Ser158IlefsX2	Confirmed: Ser158IlefsX2 (Confirmed); *1043C>A was also confirmed for this patient	p.Ser158IlefsTer2	DF508	p.Gln1463_IIe1464ins41	c.*1043A>C		
CF5830853	Black	3120+1G>A /S945L*	3120+1G>A /S945L	Confirmed: S945L	p.Trp57Leu	c.1680-1G>T	p.Ser945Leu	3120+1G->A	p.Gly1125Ter	p.Gln1463_IIe1464ins41
CF5865254	Mixed	U/U	N/N	N/A	p.Gln1463_IIe1464ins41					
CF5980227	White	DF508/U	DF508/n.166+3321T>G	Not Confirmed: C. Stewart	DF508					

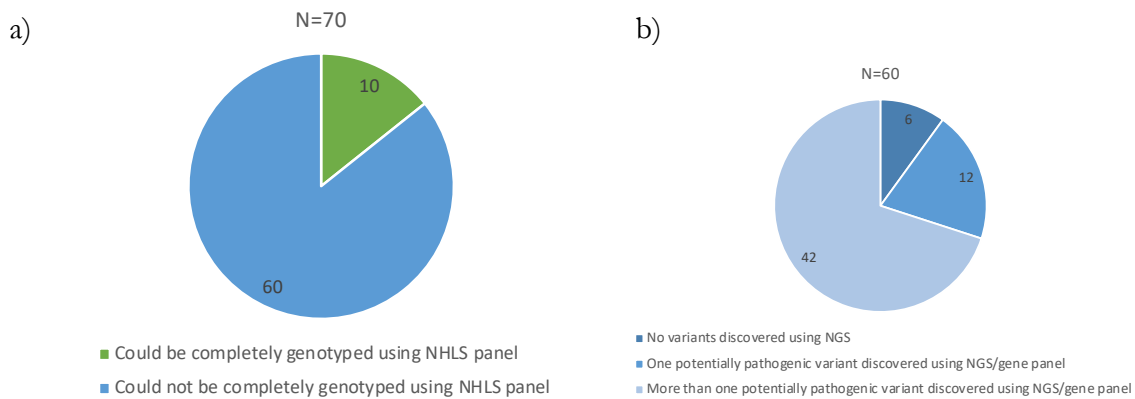
Sample	Ethnicity	NHLS genotype	NGS genotype (C. Stewart)	Confirmed By C. Stewart	NGS Variant 1	NGS Variant 2	NGS Variant 3	NGS Variant 4	NGS Variant 5	NGS Variant 6
				Filtered Out With GMAF						
CF6004268	Black	U/U	N/n.166+3400C>T	Invalidated						
CF6175627	Mixed	U/U	N/N	N/A	p.Lys684AsnfsX38/p.Gln685ThrfsX4	p.Gln1463_Ile1464ins41				
CF6188367	Mixed	U/U	N/N	N/A	p.Ser1297LeufsTer31	c.3963+9G>C	p.Gln1463_Ile1464ins41			
CF6268769	No Data	No Data	No Data		p.Ser158IlefsTer2	DF508	p.Gln1463_Ile1464ins41	c.*1043A>C		
CF6724226	Mixed	U/U	U/U	N/A	p.Lys684AsnfsX38/p.Gln685ThrfsX4					
CF6746590	Black	3120+1G>A/U	3120+1G>A/Arg792X	Confirmed	p.Arg792X	3120+1G->A	p.Gln1463_Ile1464ins41			
CF6757915	Black	3120+1G>A/U	No Data		3120+1G->A	4374+1G->T	p.Gln1463_Ile1464ins41			
CF6803591	Black	U/U	U/U	N/A	p.Gly1125Ter	p.Gln1463_Ile1464ins41				
CF7465992	Mixed	U/U	U/U	N/A						
CF7527369	White	DF508/U	DF508/394delTT(a.k.a p.Leu88IlefsTer22)*	Confirmed	p.Leu88IlefsTer22	DF508	p.Lys684AsnfsX38/p.Gln685ThrfsX4	c.3963+9G>C		
CF7600423	Indian	3848+10kbC>T/U	3848+10kbC>T/I502T	Confirmed	p.Ile502Thr	p.Gln1463_Ile1464ins41				
CF7760687	Mixed	U/U	N/DF508	Not confirmed	DF508					
CF7778750	Black	3120+1G>A/U	3120+1G>A/G1249E	Confirmed	G1249E	3120+1G->A	p.Gln1463_Ile1464ins41			
CF7930867	White	DF508/E585X*	DF508/Glu585X	Confirmed	p.Phe17SerfsX8	DF508	p.Glu585X	p.Gln1463_Ile1464ins41		
CF8095908	Mixed	U/U	N/N	N/A						
CF8213552	Mixed	DF508/U	DF508/Gly1173GlnfsX21	Confirmed	DF508	p.Gly1173GlnfsTer21	c.3963+9G>C			
CF8219823	Mixed	U/U	U/n.166+3400C>T	Invalidated						
CF8754900	Mixed	3120+1G>A/U	3120+1G>A/U	N/A	p.Met1Thr	3120+1G->A	p.Gln1463_Ile1464ins41			
CF9111494	Mixed	W1282X/U	Trp1282X/Arg1158X	Confirmed	p.Arg1158X	W1282X	c.3963+9G>C	p.Gln1463_Ile1464ins41		
CF9295572	White	DF508/U	DF508/U	N/A	p.Phe17SerfsX8	DF508				
CF9442098	Black	3120+1G>A/U	3120+1G>A/U	N/A	3120+1G->A	p.Gln1463_Ile1464ins41				
CF9830825	Black	3120+1G>A/DF508	3120+1G>A/DF508	N/A	p.Trp57Leu	DF508	3120+1G->A	p.Gly1125Ter	c.3963+9G>C	p.Gln1463_Ile1464ins41
CF9862557	White	DF508/U	No Data		DF508					

\*Variant recorded in database as having been identified using gene panel screening, but this is unlikely since the variant is not present on the NHLS panel.

\*Discrepancy in NHLS screening results: 394delTT and Trp846X are present on the NHLS panel but were not identified in any of the patients that were screened using the panel, even though they were found in NGS.

As can be seen in Table 3.10 and **Figure 3.11** (below), genotype information was available for some of the patients in the cohort as gene panel screening had been performed for some patients through the NHLS. Some of the patients (7/70) were fully genotyped using this panel (either homozygous for one variant or heterozygous with two or more variants). Three additional patients should have been completely genotyped using the gene panel as variants were found in NGS that are present on the panel, but were not recorded as having been identified using the panel even

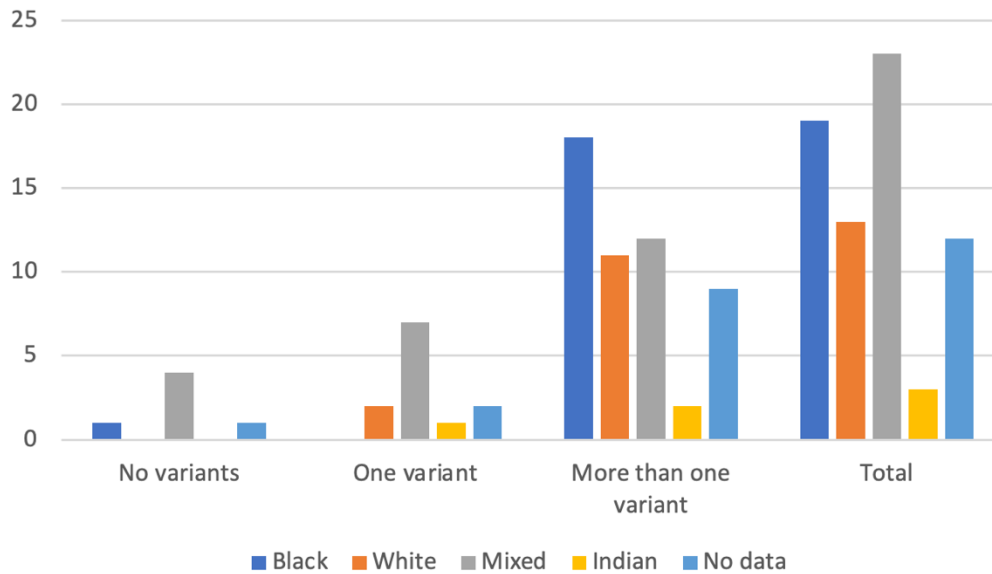
though the screening was performed. Thus, 10/70 patients have two variants that are present on the gene panel and could have been completely genotyped using the panel alone. An additional five patients would have been incompletely genotyped (only one variant was found) if the gene panel had been done for these patients. This is because the patients were not screened with the panel but the NGS results identified variants that can be detected using the current panel.



**Figure 3.11:** Visual representation of the genotyping results for the cohort. a) Distribution of samples that could or could not be fully genotyped using the NHLS panel. b) Distribution of samples that had variants discovered using NGS (excluding samples that were completely genotyped using the NHLS panel). Charts prepared using Microsoft Excel.

Thus, 60/70 patients could not be fully genotyped using gene panel screening and thus constitute the focus of this study. Of these, 6/60 did not have any potentially pathogenic variants that could be identified using either a gene panel or NGS, with one being the parent of a patient and another having had the CF diagnosis negated since the start of the study. 12/60 patients only had one potentially pathogenic variant discovered using NGS, and 54/60 patients had at least one potentially pathogenic variant discovered using NGS. Finally, 42/60 patients had more than one potentially pathogenic variant discovered using NGS.

The ethnicity distributions for the cohort are displayed in **Figure 3.12**. Some of the patients did not have ethnicity recorded in the databases, and so these are recorded as “no data”. Although a large proportion of the patients who could not be fully genotyped using either a gene panel or NGS were black or of mixed ancestry, the overall count was low.



**Figure 3.12:** Ethnicity distribution of potentially pathogenic variants found in samples using NGS or gene panel screening before confirmation with Sanger sequencing. Charts prepared using Microsoft Excel.

### 3.7. Sanger Sequencing confirmation results

#### 3.7.1. PCR Results

Images of the 1% agarose gel electrophoresis visualisation of the PCR optimisations can be found in Supplementary 2. Some primers only produced dim bands at this stage, but when increased to 35 cycles, optimal results were achieved. Images of the 1% agarose gel electrophoresis visualisation of the PCR products used for validation with Sanger sequencing can be found in Supplementary 2. Some reactions had to be repeated as there were no bands visible.

#### 3.7.2. Sanger sequencing results

The variants confirmed using Sanger sequencing are listed in Table 3.11 (below). The variants that were invalidated by Sanger sequencing can be found in Supplementary 1. The missing samples are also recorded in Supplementary 1. Eleven variants that were not in Dr. C Stewart's investigation have been confirmed, seven of which have been annotated as "causative of CF" in the CFTR2 database. The remaining two variants are p.Met1Thr (start-lost variant) and p.Leu383Ter (stop-gain variant). Variants for which there is no CFTR2 annotation in Table 3.11 have not yet been listed in the database as having been functionally validated as CF-causing but have been identified by this study as *potentially* pathogenic. Only two variants were identified in more than ten patients: p.Phe508del and 3120+1G->A. These variants are common in the European and African American CF populations, respectively. The p.Leu88IlefsTer22 variant was identified in three

patients and c.\*1043A>C was identified in four patients. The c.\*1043A>C variant may have a mildly pathogenic effect, according to literature (Amato, Seia et al. 2013). The remaining variants were each identified in fewer than three samples. All Sanger analysis figures (trace files, alignments to CFTR, and variant positions) can be found in Supplementary 3.

**Table 3.11: Confirmed pathogenic variants.**

RS#	Variant	NHLS panel	CFTR2 annotation	No. of patients	Consequence	IMPACT	Confirmed (C. Stewart)	Confirmed (O. le Grange)	Overall confirmation
rs754147777/ rs121908769	p.Leu88IlefsTer22	*	This variant causes CF when combined with another CF-causing variant.	3	frameshift_variant	HIGH	YES	YES	Confirmed
.	p.Ser158IlefsTer2			2	frameshift_variant	HIGH	YES	YES	Confirmed
rs121908760	p.Arg709X		This variant causes CF when combined with another CF-causing variant.	2	stop_gained	HIGH	YES	YES	Confirmed
.	p.Ser877PhefsTer29			2	frameshift_variant	HIGH	YES	YES	Confirmed
rs79850223	p.Arg1158X		This variant causes CF when combined with another CF-causing variant.	2	stop_gained	HIGH	YES	YES	Confirmed
rs10234329	c.*1043A>C			4	3_prime_UTR_variant	MODIFIER	YES	YES	Confirmed
rs397508476	p.Met1Thr			2	start_lost	HIGH		YES	Confirmed
rs121908749	p.Arg75Ter		This variant causes CF when combined with another CF-causing variant.	1	stop_gained	HIGH		YES	Confirmed
rs397508777	p.Leu218Ter		This variant causes CF when combined with another CF-causing variant.	1	stop_gained	HIGH		YES	Confirmed
.	p.Leu383Ter			1	stop_gained	HIGH		YES	Confirmed
rs397508200	c.1393-1G>A		This variant causes CF when combined with another CF-causing variant.	1	splice_acceptor_variant	HIGH		YES	Confirmed
rs1297060838	<b>p.Phe508del</b>	NHLS	This variant causes CF when combined with another CF-causing variant.	24	inframe_deletion	MODERATE		YES	Confirmed
rs146521846	p.Ser1118Phe		This variant causes CF when combined with another CF-causing variant.	1	missense_variant	MODERATE		YES	Confirmed
rs397508684	p.Gln1382X		This variant causes CF when combined with another CF-causing variant.	1	stop_gained	HIGH		YES	Confirmed
rs150177304	p.Gln1411Pro		Q1411X, p.Gln1411X, c.4231C>T; This variant causes CF when combined with another CF-causing variant.	1	missense_variant	MODERATE		YES	Confirmed
rs372227120	c.4242+1G>T		This variant causes CF when combined with another CF-causing variant.	1	splice_donor_variant	HIGH		YES	Confirmed
.	p.Leu1258PhefsX7		This variant causes CF when combined with another CF-causing variant.	1	frameshift_variant	HIGH		YES	Confirmed
rs121909009	p.Gly458Val		1504delG, p.Gly458AspfsX11, c.1373delG: This variant causes CF when combined with another CF-causing variant.	1	missense_variant	MODERATE		YES	Confirmed
.	p.Gly1173GlnfsTer21			2	frameshift_variant	HIGH	YES	N/A	Confirmed
rs267606722	p.Trp846X	NHLS	This variant causes CF when combined with another CF-causing variant.	1	stop_gained	HIGH	YES	N/A	Confirmed
rs121908752	p.Leu206Trp		This variant causes CF when combined with another CF-causing variant.	1	missense_variant	MODERATE	YES	N/A	Confirmed



RS#	Variant	NHLS panel	CFTR2 annotation	No. of patients	Consequence	IMPACT	Confirmed (C. Stewart)	Confirmed (O. le Grange)	Overall confirmation
rs121908753	p.Arg352Gln		This variant causes CF when combined with another CF-causing variant.	1	missense_variant	MODERATE	YES	N/A	Confirmed
rs121909040	G1249E	*	G1249R, p.Gly1249Arg, c.3745G>A: This variant causes CF when combined with another CF-causing variant.	1	missense_variant	MODERATE	YES	N/A	Confirmed
rs397508751	p.Leu183Ile			1	missense_variant	MODERATE	YES	N/A	Confirmed
rs121908805	p.Ser466Ter		This variant causes CF when combined with another CF-causing variant.	1	stop_gained	HIGH	YES	N/A	Confirmed
rs397508222	p.Ile502Thr		This variant causes CF when combined with another CF-causing variant.	1	missense_variant	MODERATE	YES	N/A	Confirmed
rs55928397	p.Tyr577Ter		This variant causes CF when combined with another CF-causing variant.	1	stop_gained	HIGH	YES	N/A	Confirmed
rs397508296	p.Glu585X		This variant causes CF when combined with another CF-causing variant.	1	stop_gained	HIGH	YES	N/A	Confirmed
rs145449046	p.Arg792X		This variant causes CF when combined with another CF-causing variant.	1	stop_gained	HIGH	YES	N/A	Confirmed
rs397508442	p.Ser945Leu		This variant causes CF when combined with another CF-causing variant.	1	missense_variant	MODERATE	YES	N/A	Confirmed
rs75096551	<b>3120+1G-&gt;A</b>	NHLS	This variant causes CF when combined with another CF-causing variant.	13	splice_donor_variant	HIGH		N/A	N/a
rs77010898	p.Trp1282X	NHLS	This variant causes CF when combined with another CF-causing variant.	1	stop_gained	HIGH		N/A	N/a
rs121908755	p.Ser549Asn	*	This variant causes CF when combined with another CF-causing variant.	1	missense_variant	MODERATE		YES	Confirmed
.	p.Ser427ThrfsTer16			1	frameshift_variant	HIGH	YES	N/A	Confirmed
rs76151804	3272-26A>G	NHLS	This variant causes CF when combined with another CF-causing variant.	1	intron_variant	MODIFIER	N/A	N/A	N/a

\* These variants have been identified as potential variants that could be included in a South African population-specific CF gene panel (Goldman, Graf et al. 2003).

### 3.8. Amendment of genotype information following confirmation with Sanger sequencing

The results following amendment of the genotype information for the patients, after validation of true-positive variants, are presented below (Table 3.12). Feedback for three patients without any pathogenic variants revealed that they had since had the CF diagnoses negated.

Table 3.12: Updated genotype information with corresponding variants identified using NGS and confirmed with Sanger.

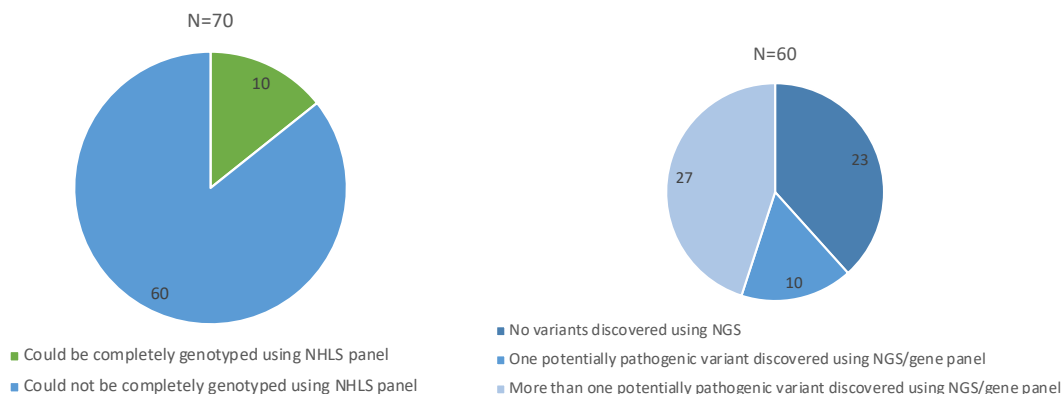
Sample	Ethnicity	NHLS genotype	NGS genotype (C. Stewart)	Confirmed By C. Stewart	Confirmed NGS Variant 1	Confirmed NGS Variant 2	Confirmed NGS Variant 3
CF3239825	Mixed	3120+1G>A/U	No Data		3120+1G->A	p.Gln1382X	
CF3719491	Black	3120+1G>A/U	3120+1G>A/U	N/A	3120+1G->A		
CF4833948	Black	3120+1G>A/3120+1G>A	3120+1G>A/3120+1G>A	N/A	3120+1G->A (homozygous)		
CF5158167	Black	3120+1G>A/3120+1G>A	3120+1G>A/3120+1G>A	N/A	3120+1G->A (homozygous)		
CF6757915	Black	3120+1G>A/U	No Data		3120+1G->A	4374+1G->T	
CF9442098	Black	3120+1G>A/U	3120+1G>A/U	N/A	3120+1G->A		
CF2954129	Mixed	U/U	U/U	N/A	c.*1043A>C		

Sample	Ethnicity	NHLS genotype	NGS genotype (C. Stewart)	Confirmed By C. Stewart	Confirmed NGS Variant 1	Confirmed NGS Variant 2	Confirmed NGS Variant 3
CF1133987	White	DF508/Ser549Asn*	No Data		DF508	Ser549Asn	
CF1478689	Mixed	DF508/U	DF508/Gly1173GlnfsX21	Confirmed	DF508	p.Gly1173GlnfsX21	
CF2843425	No Data	No Data	No Data		DF508	p.Arg1158X	
CF3115703	White	DF508/DF508	DF508/DF508	N/A	DF508 (homozygous)		
CF3796568	No Data	No Data	No Data		DF508	p.Ser1118Phe (homozygous)	
CF4495056	Mixed	DF508/U	No Data		DF508	p.Ser877PhefsTer29	c.*1043A>C
CF4544212	White	DF508/DF508	No Data		DF508 (homozygous)		
CF5107567	Mixed	DF508/U	DF508/Trp846X	Confirmed	DF508	p.Trp846X	
CF5181003	White	DF508/DF508	DF508/DF508	N/A	DF508 (homozygous)		
CF5980227	White	DF508/U	DF508/n.166+3321T>G	Not Confirmed: c. Stewart filtered out with GMAF	DF508		
CF7760687	Mixed	U/U	N/DF508	Not confirmed	DF508 (unconfirmed)		
CF8213552	Mixed	DF508/U	DF508/Gly1173GlnfsX21	Confirmed	DF508	p.Gly1173GlnfsTer21	
CF9295572	White	DF508/U	DF508/U	N/A	DF508		
CF9830825	Black	3120+1G>A/DF508	3120+1G>A/DF508	N/A	DF508	3120+1G->A	
CF9862557	White	DF508/U	No Data		DF508		
CF7778750	Black	3120+1G>A/U	3120+1G>A/G1249E	Confirmed	G1249E	3120+1G->A	
CF9111494	Mixed	W1282X/U	Trp1282X/Arg1158X	Confirmed	p.Arg1158X	W1282X	
CF4471587	White	3120+1G>A/U	3120+1G>A/Arg352Gln	Confirmed	p.Arg352Gln	3120+1G->A	
CF4602380	No Data	No Data	No Data		p.Arg75Ter	DF508	
CF6746590	Black	3120+1G>A/U	3120+1G>A/Arg792X	Confirmed	p.Arg792X	3120+1G->A	
CF2349244	Black	U/U	U/U	N/A	p.Gln1411Pro		
CF4283433	Black	U/U	Gly458Val/Ser466X	Gly458Val (Not confirmed)/Ser466X (Confirmed)	p.Gly458Val	p.Ser466Ter	
CF7600423	Indian	3848+10kbC>T/U	3848+10kbC>T/I502T	Confirmed	p.Ile502Thr		
CF0014912	No Data	No Data	No Data		p.Leu1258PhefsX7		
CF5365245	Black	U/U	U/Leu183Ile	Confirmed	p.Leu183Ile		
CF3512286	White	DF508/U	DF508/Leu206Trp	Confirmed	p.Leu206Trp	DF508	
CF4062212	Indian	U/U	No Data		p.Leu218Ter	c.1393-1G>A	
CF2173052	Black	U/U	Leu383X/Arg709X	Leu383X (Not validated by C. Stewart)/Arg709X (Not validated by C. Stewart)	p.Leu383Ter	p.Arg709X	
CF1697504	White	DF508/394delTT	DF508/394delTT		p.Leu88IlefsTer22	DF508	
CF3803349	No Data				p.Leu88IlefsTer22	DF508	
CF7527369	White	DF508/U	DF508/394delTT (a.k.a. p.Leu88IlefsTer22)	Confirmed	p.Leu88IlefsTer22	DF508	
CF1782680	No Data	No Data	No Data		p.Met1Thr	DF508	
CF8754900	Mixed	3120+1G>A/U	3120+1G>A/U	N/A	p.Met1Thr	3120+1G->A	
CF7930867	White	DF508/E585X*	DF508/Glu585X	Confirmed	p.Phe17SerfsX8	DF508	p.Glu585X
CF5384911	Mixed	DF508/U	DF508/Ser158IlefsX2	Confirmed: Ser158IlefsX2 (Confirmed); *1043C>A was also confirmed for this patient	p.Ser158IlefsTer2	DF508	c.*1043A>C
CF6268769	No Data	No Data	No Data		p.Ser158IlefsX2	DF508	c.*1043A>C
CF0018616	Black	U/U	R709X/p.Ser427ThrfsX16	Confirmed	p.Ser427ThrfsX16	p.Arg709X	
CF3019852	Mixed	3272-26A>G/U	3272-26A>G/Ser877PhefsX29	Confirmed	p.Ser877PhefsX29		
CF5830853	Black	3120+1G>A/S945L*	3120+1G>A/S945L	Confirmed: P.S945L	p.Ser945Leu	3120+1G->A	
CF2433640	Mixed	U/U	3120+1G>A/Tyr577X	Confirmed	p.Tyr577X	3120+1G->A	
CA0144930	No Data	No Data	No Data				
CA1615190	No Data	No Data	No Data				
CA4542870	No Data	No Data	No Data				
CA4932026	No Data	No Data	No Data				
CA8443975	No Data	No Data	No Data				
CF0235490	White	U/U	N/N	N/A			
CF1323468	Black	U/U	U/U	N/A			
CF1658976	Mixed	U/U	U/U	N/A			

Sample	Ethnicity	NHLS genotype	NGS genotype (C. Stewart)	Confirmed By C. Stewart	Confirmed NGS Variant 1	Confirmed NGS Variant 2	Confirmed NGS Variant 3
CF3594271	Black	U/U	U/U	N/A			
CF4223536	Indian	U/U	U/n.166+3321T>G	Not confirmed: C. Stewart filtered out with GMAF			
CF4869626	Mixed	U/U	No Data				
CF5865254	Mixed	U/U	N/N	N/A			
CF6004268	Black	U/U	N/n.166+3400C>T	Invalidated			
CF6188367	Mixed	U/U	N/N	N/A			
CF6803591	Black	U/U	U/U	N/A			
CF8219823	Mixed	U/U	U/n.166+3400C>T	Invalidated			
CF1534048	Black	U/U	U/U	N/A			
CF4379523	Mixed	U/U	No Data				
CF4832869	Mixed	U/U	U/U	N/A	CF excluded		
CF6175627	Mixed	U/U	N/N	N/A			
CF6724226	Mixed	U/U	U/U	N/A	CF excluded		
CF7465992	Mixed	U/U	U/U	N/A	CF excluded		
CF8095908	Mixed	U/U	N/N	N/A			

\*Variant recorded in database as having been identified using gene panel screening, but this is unlikely since the variant is not present on the NHLS panel.

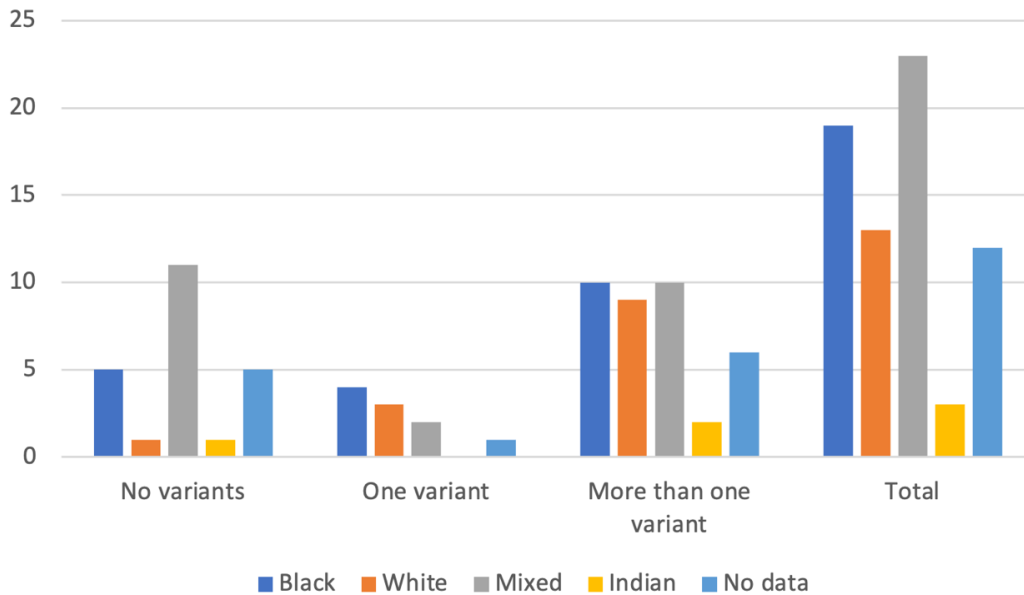
After validation of true-positive variants using Sanger, 23/60 patients did not have any confirmed variants. Of these, three have had the CF diagnosis negated and five are parents of patients. 10/60 samples had only one potentially pathogenic variant identified using NGS or the NHLS gene panel. 27/60 patients had two or more potentially pathogenic variants identified using NGS and/or gene panel; thus, 37/70 patients were able to be completely genotyped using NGS and NHLS gene panel screening. These results are displayed in **Figure 3.13** (below).



**Figure 3.13:** Visual representation of the genotyping results for the cohort. a) Distribution of samples that could or could not be fully genotyped using the NHLS panel. b) Distribution of samples that had variants discovered using NGS (excluding samples that were completely genotyped using the NHLS panel). Charts prepared using Microsoft Excel.

Two individuals are 3120+1G->A homozygous and four are p.Phe508del homozygous. One patient (CF3796568) is homozygous for p.Ser118Phe and heterozygous for p.Phe508del. Furthermore, two individuals are heterozygous for 3120+1G->A and have not had another

pathogenic variant confirmed. Four individuals are heterozygous for p.Phe508del but have not had a second pathogenic variant confirmed. Out of 19 black individuals in this cohort, 10 did not have the 3120+1G->A variant. The updated ethnicity distributions for the cohort are displayed in **Figure 3.14**.



**Figure 3.14:** Ethnicity distribution of potentially pathogenic variants found in samples using NGS or gene panel screening after confirmation with Sanger sequencing. Charts prepared using Microsoft Excel.

### 3.9. Concluding remarks and summary of results

Next-generation sequencing of this cohort enabled 27 individuals that were lacking a complete molecular diagnosis to be fully genotyped. Overall, 37 individuals have thus been completely genotyped to have pathogenic variants. Ten individuals remain incompletely genotyped with only one variant confirmed and 23 individuals have not had any variants confirmed. 23 of 34 variants have been functionally tested and validated as causative of CF in the CFTR2 database. Only one non-coding variant, c.\*1043A>C, was considered potentially pathogenic and this information was provided by literature rather than pathogenicity score. Four variants are present on the NHLS gene panel screening test, an additional three have previously been suggested for inclusion, and the remaining variants are not included in gene panel screening of this population. Eleven of the 34 variants are yet to be functionally validated and added to the CFTR2 database, and so are considered *potentially* pathogenic based on scores of predicted pathogenicity. Ten variants are present in more than one individual, whereas the remaining 24 variants were only found in one individual. Sanger sequencing confirmed 34 of 48 potentially pathogenic variants, with 14 variants identified as false positives. NGS corroborated the NHLS screening results, where data was available.

### 3.10. References:

- Amato, F., M. Seia, S. Giordano, A. Elce, F. Zarrilli, G. Castaldo and R. Tomaiuolo (2013). "Gene mutation in microRNA target sites of CFTR gene: a novel pathogenetic mechanism in cystic fibrosis?" *PLoS One* **8**(3): e60448.
- Andrews, S. (2010). "FASTQC. A quality control tool for high throughput sequence data."
- Bardou, P., J. Mariette, F. Escudie, C. Djemiel and C. Klopp (2014). "jvenn: an interactive Venn diagram viewer." *BMC Bioinformatics* **15**: 293.
- Ewels, P., M. Magnusson, S. Lundin and M. Käller (2016). "MultiQC: summarize analysis results for multiple tools and samples in a single report." *Bioinformatics* **32**(19): 3047-3048.
- Goldman, A., C. Graf, M. Ramsay, F. Leisegang and A. T. Westwood (2003). "Molecular diagnosis of cystic fibrosis in South African populations." *S Afr Med J* **93**(7): 518-519.
- Okonechnikov, K., A. Conesa and F. García-Alcalde (2015). "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data." *Bioinformatics* **32**(2): 292-294.

## *Chapter 4 & 5: Discussion and Conclusion*

Diversity in the CFTR variants in South African patients with Cystic Fibrosis.

#### 4.1. Introduction

The under-representation of African genomics is pervasive and extends to the study of CF. There is considerable bias towards diagnosis of CF in patients of European ancestry (Mutesa and Bours 2009). The gene panels used to provide a molecular diagnosis of CF are also biased toward the spectrum of variants identified in European patients, resulting in a high variant detection rate in these populations. However, the variant detection rate using gene panels is markedly lower in the South African population (Stewart and Pepper 2016). A significant percentage of South African patients do not receive a complete molecular diagnosis of CF using gene panels (Zampoli, Verstraete et al. 2021). This study sought to identify and investigate variants in patients who could not be fully genotyped using a gene panel. 34 variants were discovered in this cohort, with only six being included in CF gene panel screening protocols and eleven having no functional annotation in CFTR2 yet.

#### 4.2. QC, Mapping and Variant detection

Variant detection has been the subject of much discussion and evaluation since its introduction. Unfortunately, the concordance between different variant discovery methods has been disappointingly low. Furthermore, there seems to be little consensus on a single, best method for discovery of variants in all circumstances. Hwang *et al.* used gold standard personal exome variants to perform systematic comparison of variant calling pipelines (Hwang, Kim et al. 2015). They found that a pipeline combining BWA-MEM and Samtools, as well as Freebayes with any aligner, showed the best performance for SNP calling on *Illumina* data sets. They also found that the GATK HaplotypeCaller performed better than the other callers on indels from *Illumina* data sets regardless of aligner. When doing concordance analysis, the attributes of the data sets (such as exome regions, coverage, sequencing quality, etc.) need to be considered in addition to the pipeline used, and performing concordance comparison on a single data set is cautioned (Hwang, Kim et al. 2015). The conclusion is that the BWA-MEM pipeline with GATK-HaplotypeCaller should be used for indel calling, and the BWA-MEM and Samtools pipeline should be used for SNP calls. O'Rawe and colleagues also found low concordance of multiple variant-calling pipelines but noted that more recent versions of GATK HaplotypeCaller have shown great improvement in variant detection, particularly with its accuracy of indel calling (O'Rawe, Jiang et al. 2013). Pirooznia and colleagues also evaluated the accuracy of different variant calling pipelines and used both Sanger sequencing and array genotyping for validation (Pirooznia, Kramer et al. 2014). They found that GATK was more accurate than Samtools and that the HaplotypeCaller algorithm was more accurate than the UnifiedGenotype algorithm. The authors also found that mapping quality, read

depth and allele balance influenced the accuracy of variant detection, but that following the advised best practices for the pipelines eliminated the need to filter based on these parameters. Kumaran and colleagues performed a similar analysis on WES and simulated exome data and used a high confidence variant callset from GiaB for validation (Kumaran, Subramanian et al. 2019). They found that the combination of BWA or Novoalign with DeepVariant and Samtools was more accurate when calling SNPs. For indel calling, they found that BWA or Novoalign with DeepVariant and GATK was more accurate. Lastly, they found that merging the most accurate variant calling pipelines provided the most accurate set of calls.

With this in mind, the analysis of four sets of variant calls was approached with an informed consensus approach, to attempt to maximise the discovery of true-positive variants. The comparison of two aligners (BWA and Bowtie2) confirmed that BWA provided superior mapping quality. Furthermore, variant calls from the BWA-GATK method proved to have concordance with many variants from the other pipelines. Ideally, all variants from all discovery methods would be evaluated for pathogenicity further downstream. However, this complicates the methodology for publication as the reproducibility of the CLC Genomics and CASAVA pipelines is questionable and the variant quality filtering remains unknown. It is particularly important to be able to clearly indicate which variant was discovered using which method, in conjunction with the NGS platform (Lee, Kweon et al. 2021).

#### *4.3. In silico validation*

*In silico* validation of variants following variant detection using current best practices is an uncommon step in mainstream evaluation of NGS. However, when evaluating multiple sets of variants, it may be a useful addition as it provides a measure of probability of variants being true-positive calls (Cantarel, Weaver et al. 2014). The confidence in many of the GATK calls was improved; however, this tool may be introducing more uncertainty than alleviating it. This is because BAYSIC scores were not predictably concordant with any one method and further complicated the evaluation of variants. Thus, considering the BAYSIC scores proved useful but should not be used as the sole qualifying or disqualifying criterion when selecting variants for further evaluation.

#### *4.4. Variant annotation and effect prediction*

Initial summary statistics provided by Ensembl's VEP tool have rather limited utility other than giving an overview of the annotation of the variant calls. This is also true of heatmaps constructed



before prioritizing potentially pathogenic variants, as these only provide a visual representation of variant calls per patient in each set. However, the comparison of heatmaps and summary statistics between the different call sets does provide an indication of the sensitivity and quality filtering employed by each method. For example, the CASAVA summary statistics and heatmap indicate a vast majority of variants being called in almost all patients, whereas the GATK summary statistics and heatmap indicate that only a few variants are found in almost all patients.

The master variant list provided the starting point for prioritization of potentially pathogenic variants. The CFTR2 database is the most useful source of information for this purpose, as variants have already undergone extensive functional evaluation and confirmation (CFTR2 2011). However, there were eleven variants identified in this study that do not have functional annotation in the database at present. CFTR2 has done an extensive probing of the functional consequences of a great number of variants (485 at present). As such, many of the variants discovered in our cohort of patients had already been functionally tested. This improves the ease and confidence of variant prioritization greatly as there is no need to evaluate the various pathogenicity scoring tools for these variants as they have already been functionally tested. Furthermore, the ambiguity of missense variants is alleviated. However, the CFTR2 database is biased towards European CF patients (Lim, Silver et al. 2016). The inclusion of a diverse cohort will likely help to make this database more applicable and representative of the global spectrum of CF variants. This will improve the reach of the database and its utility in populations with variants that are less common in the European population.

ClinVar is a useful annotation tool for known phenotypic effects of human variants. It is included in the VEP package and is an impactful tool for identifying likely pathogenic variants and their clinical significance (Landrum, Lee et al. 2018). However, for those variants for which there were no annotations in CFTR2 or ClinVar, predicted pathogenicity scores provide the best means of identifying variants with the potential of causing CF when homozygous or combined with another variant. Each tool has its own method for evaluating deleteriousness of a variant and this provides a level of confidence when the scores collectively agree on the prediction. This is particularly useful in missense variants. For other variants, the pathogenicity can largely be predicted from the variant type: stop-gain, start-lost and frameshift variants are likely pathogenic as they affect the final protein or truncation thereof.

#### 4.5. CFTR variants in a diverse population

Overall, 34 variants in this cohort were confirmed by Sanger sequencing. As has been thoroughly discussed, CF populations of European ancestry typically contain a few variants that are shared between individuals with high frequency. The variants discovered in this cohort are individually infrequent except for p.Phe508del and 3120+1G->A. These two variants are common to European and African CF patients, respectively (Bobadilla, Macek et al. 2002). They were found to be common in a few patients but were mostly present as heterozygous variants or were found in the controls that could be fully genotyped using panel screening. It is surprising that 52% of the black South African patients in this cohort do not have even one 3120+1G->A variant, since it has previously been identified as a frequent variant in this population (Goldman, Graf et al. 2003, Stewart and Pepper 2016, Zampoli, Verstraete et al. 2021). 24 of 34 potentially pathogenic variants were only found in a single patient, which is indicative of large-scale diversity within the South African CF population. However, this study is not representative of *all* South African patients with CF. The samples were selected because they could not be fully genotyped using gene panel screening. Thus, the cohort is more representative of a group of patients in this population that have variants not present on the gene panels. This means that diversity within this cohort is expected. Extrapolation to the whole population would be ill-advised, but these findings add to the case of alternative evolutionary events to that of European CF populations (Bobadilla, Macek et al. 2002) and highlight the need for alternate screening methods in patients in African ethnolinguistic groups.

Regarding the less frequent variants found in this study, the p.Leu88IlefsTer22 (a.k.a. 394delTT) variant was found in three patients of European origin. This aligns with previous studies of a South African cohort, which found the variant in 3.7% of patients of European origin using a population-specific gene panel (Goldman, Graf et al. 2003). This variant was also found in 2% (18 patients) of the patients in the SACFR; all are of European origin (Zampoli, Verstraete et al. 2021). The p.Trp1282X variant was identified in one patient of mixed ancestry. This variant was previously found to have a frequency of 1% in South African patients of European origin (Goldman, Graf et al. 2003), and less than 1% in the SACFR (Zampoli, Verstraete et al. 2021). The p.Ser549Asn variant was found in one patient of European origin. This variant was previously found in South African patients of European origin at a frequency of 0.25% (Goldman, Graf et al. 2003), and less than 1% in the SACFR (Zampoli, Verstraete et al. 2021). The 3272-26A>G variant was genotyped using the NHLS panel in one patient of mixed ancestry. This variant was found with a frequency of 2.6% in the SACFR: 8 patients of mixed ancestry and 15 of European origin (Zampoli,

Verstraete et al. 2021). It was previously found in 4% and 1.2% of patients of European origin and mixed ancestry patients, respectively (Goldman, Graf et al. 2003). The G1249E variant was found in one black patient. This aligns with previous studies that found it at a frequency of 3.6% in black South African patients (Goldman, Graf et al. 2003). The p.Trp846X variant was genotyped in one patient of mixed ancestry using the NHLS gene panel. This variant was reported as having less than 1% frequency in the SACFR and was not suggested for inclusion in a population-specific gene panel (Goldman, Graf et al. 2003, Zampoli, Verstraete et al. 2020). It is present in the CFTR2 database at a frequency of 0.00039 (CFTR2 2011). Only six variants found in this cohort are present on the NHLS gene panel, and a further two are part of a “population-specific” gene panel (Goldman, Graf et al. 2003).

This study did not seek to quantify the variant detection rate of gene panels in the South African population, but rather to identify variants that were not being found using these panels. The current variant detection rate is estimated to be between 70-79% in the SA population. However, about 70% of the patients in the SACFR are of European origin. The variant detection rate in this population is biased towards these patients whose variants are included on the gene panel. The diagnosis of CF is also biased towards these patients and the “classic clinical presentation of CF” (Lim, Silver et al. 2016). This is likely to be the reason leading to more South African patients of European origin being diagnosed with CF. Additionally, the burden of malnutrition, HIV, and TB in South Africa may also be leading to misdiagnosis of CF in this population (Zampoli, Verstraete et al. 2021). This considered, diversity in the Northern and Southern regions of Italy posed a great challenge to effective variant detection of CFTR variants using gene panel screening (Lucarelli, Porcaro et al. 2017). It was suggested that an NGS approach be used instead, where 188 known population-specific variants are first screened, and the rest of the data is probed if a patient is still incompletely genotyped (Lucarelli, Porcaro et al. 2017). This illustrates an example of diversity in the CFTR variants within a country, supporting the results obtained in this study, as well as an approach to addressing molecular diagnosis in a diverse population. However, inequality of access to resources and the expense of NGS may be challenging in the South African context.

Lastly, there are a few limitations that have become evident upon retrospective review, many of which are likely the result of how much time passed between the selection of participants and the conclusion of the study. The selection of participants for this study included adult carriers as well as patients with inconclusive sweat test results. Ideally, the results reported in this dissertation would have been limited to only those with definitive sweat test results as this would have

simplified the interpretation of the results and given them greater impact. The study could also have been improved by including a population analysis and the inclusion of controls. Another limitation is the lack of clinical data, such as sweat test results and phenotypic presentation. Unfortunately, a single robust database was not used to record this information and as such the clinical data was inconsistent and unreliable. Inclusion of this information would have made the genetic results more difficult to interpret and ultimately affect the integrity of the study. Finally, since the patients with multiple heterozygous variants were all found to have *cis* configuration, their genotyping and clinical presentation requires further analysis as they would likely have milder disease due to the presence of a functional copy of the CFTR gene.

### 5.1. Conclusion and future work

Knowing that the African population is ethnically diverse, that gene panels for CF have a lower variant detection rate in ethnically diverse populations and that many variants are found in only one individual (the results of this investigation), it is likely that a single gene panel will not be successful in this population. While it may help as a first step in the diagnostic protocol for CF, it should not be used in isolation (Shum, Bennett et al. 2021). Ideally, NGS of the entire CFTR gene region should be performed routinely if a CF diagnosis is suspected (Lucarelli, Porcaro et al. 2017). However, the implementation of NGS in many South African hospitals and clinics is currently constrained by a lack of resources. The solution may be to continue using population-specific gene panels for the molecular diagnosis of CF, but that NGS must be conducted if one or no variant is found (Shum, Bennett et al. 2021). NGS of the CFTR gene has become much more affordable in recent years through the invention of targeted library preparation kits and streamlined sequencing protocols.

Finally, a challenge that needs to be addressed is the current bias towards diagnosis of CF in patients of European origin (Lim, Silver et al. 2016), as well as the misdiagnosis and potential under-estimation of incidence of CF in African ethnolinguistic groups. Continued use of limited gene panel screening in isolation will only perpetuate the problem. These panels have great clinical utility in populations of European descent but will perpetuate the bias if used as an exclusionary device (Shum, Bennett et al. 2021). The lack of a complete molecular diagnosis using a gene panel does not exclude the possibility of CF in any patient, particularly so in ethnically diverse patients. Future research will also need to be devoted to treatments that are effective and affordable for patients with a diverse range of variants, as the current therapies are designed for treating patients

with common CFTR variants and are not accessible in South Africa (Zampoli, Verstraete et al. 2021).

This study has provided valuable insight into the spectrum of CFTR variants in South African patients with CF and sheds light on the extent of variation that can be found in countries with diverse population ancestries. This will have a significant impact on the improvement of genetic diagnosis of South African patients and will help to inform new protocols that are more inclusive of the patients within this population. In addition, the results of this study have great clinical utility for the patients who now have confirmed variants, as the treatment strategy can be adjusted to a personalised approach. Not only will the patients know which targeted treatments they could benefit from, but they will also be able to avoid treatments designed for variants they do not have. This will help to make the process of finding the best treatment more efficient and minimise the waste of time and resources. Furthermore, this list of variants may serve as a starting point for an extended gene panel that can be used for screening and the development of novel treatments. As South Africa starts to evaluate the viability of NBS and efforts are made to improve access to diagnosis and treatment for all, there is undoubtedly an abundance of hope for patients with CF in South Africa.

## References:

- Bobadilla, J. L., M. Macek, Jr., J. P. Fine and P. M. Farrell (2002). "Cystic fibrosis: a worldwide analysis of CFTR mutations--correlation with incidence data and application to screening." *Hum Mutat* **19**(6): 575-606.
- Cantarel, B. L., D. Weaver, N. McNeill, J. Zhang, A. J. Mackey and J. Reese (2014). "BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity." *BMC Bioinformatics* **15**(1): 104.
- CFTR2 (2011). "The Clinical and Functional TRanslation of CFTR (CFTR2)."
- Goldman, A., C. Graf, M. Ramsay, F. Leisegang and A. T. Westwood (2003). "Molecular diagnosis of cystic fibrosis in South African populations." *S Afr Med J* **93**(7): 518-519.
- Hwang, S., E. Kim, I. Lee and E. M. Marcotte (2015). "Systematic comparison of variant calling pipelines using gold standard personal exome variants." *Sci Rep* **5**: 17875.
- Kumaran, M., U. Subramanian and B. Devarajan (2019). "Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data." *BMC Bioinformatics* **20**(1): 342.
- Landrum, M. J., J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J. B. Holmes, B. L. Kattman and D. R. Maglott (2018). "ClinVar: improving access to variant interpretations and supporting evidence." *Nucleic Acids Res* **46**(D1): D1062-D1067.
- Lee, J. H., S. Kweon and Y. R. Park (2021). "Sharing genetic variants with the NGS pipeline is essential for effective genomic data sharing and reproducibility in health information exchange." *Sci Rep* **11**(1): 2268.
- Lim, R. M., A. J. Silver, M. J. Silver, C. Borroto, B. Spurrier, T. C. Petrossian, J. L. Larson and L. M. Silver (2016). "Targeted mutation screening panels expose systematic population bias in detection of cystic fibrosis risk." *Genet Med* **18**(2): 174-179.
- Lucarelli, M., L. Porcaro, A. Biffignandi, L. Costantino, V. Giannone, L. Alberti, S. M. Bruno, C. Corbetta, E. Torresani, C. Colombo and M. Seia (2017). "A New Targeted CFTR Mutation Panel Based on Next-Generation Sequencing Technology." *J Mol Diagn* **19**(5): 788-800.
- Mutesa, L. and V. Bours (2009). "Diagnostic challenges of cystic fibrosis in patients of African origin." *J Trop Pediatr* **55**(5): 281-286.
- O'Rawe, J., T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W. E. Johnson, Z. Wei, K. Wang and G. J. Lyon (2013). "Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing." *Genome Med* **5**(3): 28.
- Pirooznia, M., M. Kramer, J. Parla, F. S. Goes, J. B. Potash, W. R. McCombie and P. P. Zandi (2014). "Validation and assessment of variant calling pipelines for next-generation sequencing." *Hum Genomics* **8**: 14.
- Shum, B. O. V., G. Bennett, A. Navilebasappa and R. K. Kumar (2021). "Racially equitable diagnosis of cystic fibrosis using next-generation DNA sequencing: a case report." *BMC Pediatr* **21**(1): 154.
- Stewart, C. and M. S. Pepper (2016). "Cystic fibrosis on the African continent." *Genet Med* **18**(7): 653-662.
- Zampoli, M., J. Verstraete, M. Frauendorf, R. Kassanjee, L. Workman, B. M. Morrow and H. J. Zar (2021). "Cystic fibrosis in South Africa: spectrum of disease and determinants of outcome." *ERJ Open Res* **7**(3).
- Zampoli, M., J. Verstraete, M. Frauendorf and L. Workman (2020). South African Cystic Fibrosis Patient Registry Annual Report 2018.