

https://dspace.org/

# Cystic fibrosis: An update on the variant profile and carrier frequency in the Black South African population

**Smit, Ingrid**

**2024**

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/42995

Smit, Ingrid. (2024). Cystic fibrosis: An update on the variant profile and carrier frequency in the Black South African population [Master's dissertation, University of the Witwatersrand, Johannesburg]. WireDSpace.

Downloaded from a DSpace repository. https://dspace.org/

# Cystic fibrosis: An update on the variant profile and carrier frequency in the Black South African population

Student: Ingrid Smit
Student number: 2721074
Supervisor: Ms F. Essop
Co-supervisor: Ms. C Smal

A research report (in the format of a "submissible" paper)
submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg,
in partial fulfilment of the requirements for the Degree of Master of Science
in Medicine (Genomic Medicine) by Research and Coursework.
February 2024

NATIONAL HEALTH LABORATORY SERVICE
School of Pathology, University of the Witwatersrand
DIVISION OF HUMAN GENETICS
Hospital Street, Johannesburg, 2001 | PO Box 1038, Johannesburg, 2000
[T] +27 11 489 9223 | [M] +27 78 080 8841 | [F] +27 11 489 9226 | [E] human.genetics@nhls.ac.za

23 February 2024

**Re: Research report submitted to the Faculty of Health Sciences, University of the Witwatersrand, by Ms Ingrid Smit in partial fulfilment of the requirements for the degree of Master of Science in Medicine (Genomic Medicine)**

Dear examiner,

Thank you for agreeing to examine the research report submitted by Ms Ingrid Smit (student number 2721074) entitled "Cystic fibrosis: An update on the variant profile and carrier frequency in the Black South African population".

This research report is being submitted by the candidate in the format of a "submissible" paper as per the Style Guide for Thesis, Dissertation and Research Reports (updated March 2016) by the University of the Witwatersrand Faculty of Health Sciences. The research report contributes 33% towards the final mark of the MSc (Med) Genomic Medicine degree for which the candidate is enrolled. The candidate and her supervisors have chosen to submit the paper to the American Journal of Human Genetics. This international journal is open access. The submissible paper attached is therefore written in accordance with the author guidelines of the stipulated journal. These guidelines have been attached for your reference in the appendices section of this document. The only deviation from these guidelines is that the manuscript has been presented as one complete document (including al tables and figures) for ease of marking, rather than being submitted separately.

The research protocol has also been attached as an appendix for the purpose of providing an extended literature review and further contextualise the study. The protocol has already been assessed and approved by the Faculty of Health Sciences Post-Graduate Assessors committee and is therefore not for examination.

Please refer to the Table of Contents for a complete list of the contents provided in this research report.

Yours sincerely,

Ingrid Smit (Candidate)

23/02/2024

Ms Fahmida Essop (Supervisor)

Ms Clarice Smal (Co-supervisor)

This report is intended solely to record the observations and/or opinion of the writer. It does not constitute a medico-legal report

## Declaration

I, Ingrid Smit, declare that this Research Report (in the format of a "submissible" paper) is my own, unaided work. It is being submitted for the Degree of Master of Science in Medicine (Genomic Medicine) by Research and Coursework at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

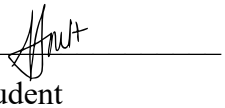


Ingrid Smit
2721074
23 February 2024

## Contribution of Candidate

**Declaration: Student's contribution to article(s) and agreement of co-author(s)**

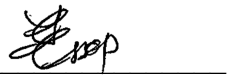
I, Ingrid Smit, student number 2721074, declare that this Research Report, entitled "Cystic fibrosis: An update on the variant profile and carrier frequency in the Black South African population" is my own work and that I contributed significantly towards research findings presented in the paper intended for publication below.



Student

23.02.2024

Date



Primary supervisor

23/02/2024

Date

**Agreement by co-authors:**

By signing this declaration, the co-authors listed below agree to the use of the article(s) by the student as part of her Research Report.

| Authors | Name | Signature | Date |
|---|---|---|---|
| 1st | Ms Ingrid Smit |  | 23.02.2024 |
| 2nd | Ms Clarice Smal |  | 23/02/2024 |
| 3rd | Ms Fahmida Essop |  | 23/02/2024 |

## Presentations arising from this research

1. University of the Witwatersrand Molecular Biosciences Research Thrust (MBRT) Postgraduate Symposium, 7th December 2023 (poster presentation).

## Abstract

Cystic fibrosis (CF) is an autosomal recessive disorder caused by pathogenic variants in the *CFTR* gene. Limited genetic research has been conducted on the Black South African population, and molecular testing is frequently the only way a diagnosis can be made. At the NHLS, testing is performed for the common 3120+1G>A (c.2988+1G>A) variant in the Black population, and other common European CF variants. Recent studies in the Division of Human Genetics show evidence of other recurrent *CFTR* variants. The aim of this study was to screen for these and other variants to update the *CFTR* variant profile and carrier frequency in the Black population. NGS data on 395 unaffected individuals was used for *CFTR* variant identification, annotation, prioritisation, and classification using the ACMG-AMP guidelines. The c.2988+1G>A variant accounted for 36.4% of CF alleles, which is less than previously reported (46%), suggesting that there are other common CF-causing variants in this population. The recurrent variants previously identified were not detected in this cohort, possibly due to limitations in NGS, the bioinformatic pipeline, or small sample size. Three novel likely pathogenic variants (c.3392T>C, c.3038C>G, and c.2594G>C) were identified, with carrier frequencies of 1 in 395 each, which could potentially be African-specific variants. Identifying these variants, not currently included in commercial panels, allows for targeted molecular testing in this population group. Additionally, a revised CF carrier rate of 1 in 36 was estimated which is consistent with literature, highlighting the accuracy of NGS data for carrier screening, leading to accurate risk counselling.

## Acknowledgements

I would like to thank my supervisors for their continuous support and guidance throughout this year. It has been a wonderful learning experience, and I am grateful to have had such knowledgeable, helpful, and kind supervisors. To both my supervisors, you had so much on your plate this year, but still set aside the time to prioritise me and this project. You were both always accessible, and I really appreciate your efforts. Ms Clarice Smal, thank you for going through a massive number of codes and classifications to ensure I did everything correctly. You were an incredible supervisor. Ms Fahmida Essop, thank you for being a constant source of advice and support, and for having patience with me as I was learning more about the topic at hand. I couldn't have asked for better guidance and for this I am grateful.

Thank you to the Human Genetics team and staff, I learned so much and I felt taken care of in this Division. I would also like to thank my fellow MSc Students for a wonderful year - I have made friends for life.

Lastly, thank you to my parents for continuous support and for providing me with a comfortable and relaxing environment in which I was able to study and work to the best of my abilities.

## Table of Contents

Declaration……iii
Contribution of Candidate……iv
Presentations arising from this research……v
Abstract……vi
Acknowledgments……vii
Table of Contents……viii
List of Appendices……ix
List of Figures and Tables……..x
List of Abbreviations……xi
Research report in the format of a "submissible" paper……..xii
1. Abstract for Journal Submission……xiii
2. Introduction……1
3. Methods
   3.1.Ethics……4
   3.2.Dataset……..……4
   3.3.Variant annotation……..……..4
   3.4.Variant prioritization……..……..5
   3.5.Variant classification……..……..6
   3.6.Carrier frequency calculations……..……..……7
4. Results
   4.1.Variant prioritization……8
   4.2.Variant classification……8
   4.3.Carrier frequency calculations……12
5. Discussion……..……..……..…...12
6. Declaration of Interests……..……..……..…...15
7. Acknowledgments……..……..……15
8. Web Resources…..……..……..……15
9. Author Contributions…..……..……..……..….16
10. Data and Code Availability…..……..……..……..….16
11. References……..……..……..……..……17
12. Supplementary Material...……..……..……21

## List of Appendices

Appendix A: Approved Project Protocol

Appendix B: Human Research Ethics Committee (Medical), University of the Witwatersrand (M230693)

Appendix C: Plagiarism Declaration and Turnitin Report

Appendix D: Journal Author Guidelines for Submissible Format

## List of Figures and Tables

Figure 1. Line-column chart depicting *CFTR* variant types and percentages after prioritization

Table 1. Three recent studies in the Division, with the identified *CFTR* variants (classified as either likely pathogenic and pathogenic), and the counts for each variant out of the total sample size (n)

Table 2. List of variants classified as pathogenic and likely pathogenic using ACMG codes

Table 3. Categorization of *CFTR* VUS, with their relevant ACMG codes and carrier frequencies

## List of Abbreviations

Cystic fibrosis (CF)

Cystic fibrosis transmembrane conductance regulator (CFTR)

Cystic fibrosis transmembrane conductance regulator gene (*CFTR*)

Next generation sequencing (NGS)

Inherited Disease Panel (IDP)

Variant Call Format (VCF)

Binary Alignment Map (BAM)

Variant Effect Predictor (VEP)

Human Genome Variation Society (HGVS)

Sorting Intolerant From Tolerant (SIFT)

Polymorphism Phenotyping v2 (Polyphen-2)

Combined Annotation Dependent Deletion (CADD)

Minor allele frequency (MAF)

Variants of Uncertain Significance (VUS)

Integrative Genomics Viewer (IGV)

American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP)

Untranslated region (UTR)

# Research report in the format of a "submissible" paper

To be submitted to the American Journal of Human Genetics

**Cystic fibrosis: An update on the variant profile and carrier frequency in the Black South African population**

Ingrid Smit[1], Clarice Smal[1], Fahmida Essop[1]

[1]Division of Human Genetics, National Health Laboratory Service and School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, JHB, South Africa

**Corresponding Author:**

Ingrid Smit

Division of Human Genetics, Corner Hospital and De Korte Streets, JHB, South Africa, 2001

[T]: 078 858 6376 [E]: ingridsmit17@gmail.com/2721074@students.wits.ac.za

## 1. Abstract for Journal Submission

Cystic fibrosis (CF) is an autosomal recessive disorder caused by pathogenic variants in the *CFTR* gene. Limited genetic research has been conducted on the Black South African population. CF is underdiagnosed due to its broad phenotype which overlaps with conditions frequently encountered in an African setting, and due to limitations of sweat testing, which is presently the gold standard for diagnosing CF. Molecular testing is therefore, in most circumstances, the only way in which a diagnosis can be confirmed. At the NHLS, testing is performed for the common 3120+1G>A (c.2988+1G>A) variant, which is the only common and unique variant known in the Black population, and other common European CF variants. Recent studies in the Division of Human Genetics show evidence of other recurrent *CFTR* variants in individuals of African ancestry. The aim of this study was therefore to screen for these and other recurrent *CFTR* variants and to update the *CFTR* variant profile and revise the CF carrier frequency in the Black South African population. Available NGS data on a cohort of 395 unaffected individuals was used for variant identification, annotation, prioritisation, and classification of *CFTR* variants using the ACMG-AMP guidelines. Variants that were classified as pathogenic or likely pathogenic were used to estimate the revised overall carrier frequency and predicted birth rate of CF, and the carrier frequency of the c.2988+1G>A variant. A revised CF carrier rate of 1 in 36 was estimated, which is consistent with literature (reported as 1 in 14-59), with a predicted birth rate of 1 in 5184. Furthermore, the 3120+1G>A variant was observed to account for 36.4% of CF alleles, which is less than previously reported (46%), which could be due to the small sample size. Additionally, the recurrent variants from recent studies were not detected in this cohort, indicative of possible limitations in sequencing, the bioinformatic pipeline, or small sample size. Three novel likely pathogenic variants (c.3392T>C, c.3038C>G, and c.2594G>C) were identified, with carrier frequencies of 1 in 395 each, which could potentially be African-specific variants. This study highlights the value of carrier screening, using routinely generated NGS data for the identification of novel, potentially African-specific variants in *CFTR*. Identifying these variants, not currently included in commercial testing kits, allows for targeted molecular testing in this population group. This study allowed for the updated estimation of the carrier frequency of c.2988+1G>A, and the findings suggest the presence of other common CF-causing variants in the Black population. Furthermore, the estimated carrier frequency of CF is concurrent with literature, highlighting the accuracy of this type of study for carrier screening, which is important for accurate risk counselling.

Keywords: Cystic fibrosis, *CFTR*, Black South African population, variant profile, c.2988+1G>A, carrier rate, birth rate, NGS, ACMG-AMP

## 2. Introduction

Cystic fibrosis (CF) is a fatal, autosomal recessive disorder, caused by pathogenic variants in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene (1). CFTR forms part of the ATP-binding cassette protein family (2), and acts as a chloride channel in epithelial/apical membranes in tissues found in the lungs, liver, intestinal tract, and pancreas (3). CFTR is responsible for the transport of electrolytes across these membranes, allowing for water to follow, which in turn leads to the thinning of mucus secretions (4). Pathogenic variants in *CFTR* lead to a decrease in CFTR protein function (2,5,6), resulting in a broad phenotypic manifestation, including respiratory disease, pancreatic insufficiency, failure to thrive and male infertility. (3). The disease therefore presents with variable expressivity and variable disease progression (5). Currently, there are over 2000 identified variants in *CFTR* with 719 disease-causing variants identified (5, http://cftr2.org/mutations_history accessed on 28 August 2023). The pathogenicity of a variant is dependent on its location within the gene, and the effect on CFTR function (6,7).

More than 15 years ago, research on CF in South Africa (and globally) was mostly focused on variants found in individuals of European descent (8), as it was believed that CF was far more common in this specific population group (9). However, it is now known that CF is more common in Black individuals than previously reported (10). The most common pathogenic variant found in the European population is p.Phe508del (c.1521_1523del or deltaF508), which accounts for 70% of CF variants globally (11). Common CF variants such as p.Phe508del are usually population specific and differ in frequency between different ethnic groups (3,12). In the South African White population, the p.Phe508del variant is reported in 81% of CF alleles, but this variant is not found to be common in the Black population (13–15).

Another common European variant is the R117H (c.350G>A) variant (16), and patients with this variant can present with a wide phenotypic range of typical CF-symptoms to milder symptoms such as the clinical absence of the vas deferens (17). The severity of the impact of the variant relies on the splicing of exon 9 in CFTR, which is directly influenced by the polypyrimidine tract found in intron 8 (IVS8) (17). The tract is polymorphic and can have 5, 7 or 9 thymidines (18). R117H must be in *cis* with a 5T allele (IVS8-5T) to be considered a pathogenic CF-causing variant (17).

CF is frequently misdiagnosed in an African setting, due to its broad phenotype, with symptoms that may be masked by common conditions encountered in Africa, such as HIV, chronic

pulmonary infections, protein energy malnutrition (PEM), and tuberculosis (14,19). The sweat chloride test is presently the gold standard for diagnosing CF (9,15). However, this testing method is for the most part inaccessible in a South African setting (13,15,19), and its reliability could be influenced by conditions such as PEM which has been shown to increase sweat chloride levels (20–22). Despite its limitations, the sweat test remains the gold standard testing method for CF (21).

Diagnostic testing is also limited by the unavailability of sequencing of the full *CFTR* gene in the state sector, resulting in some causative variants remaining undetected and unknown (12). However, in most circumstances, molecular testing is the only reliable way in which a CF diagnosis can be confirmed in South Africa. Commercially available genetic testing kits or assays are designed based on European data, which might also lead to variants in the Black population being unaccounted for (19). Currently, the Division of Human Genetics, NHLS uses the commercial Elucigene CF-EU2v1 kit (known as the CF50 panel), which screens for 50 known pathogenic variants (personal communication from Ms F. Essop). Even though the kit is designed based on European data, and the variant profile in the Black South African population has not yet been fully characterized (13,14), it is still used in a South African setting. Therefore, only 46% of *CFTR* variants are detected in the Black population using the Elucigene kit. This is due to the common 3120+1G>A variant (now referred to as c.2988+1G>A), which accounts for 46% of all CF alleles in the Black population with no other common *CFTR* variant identified to date (14). This drives the reasoning for using this kit in a South African setting, as a significant proportion of variants are still being covered by the kit, even though other variants are unaccounted for (personal communication from Ms F. Essop).

Thus far, limited data have been generated to investigate the profile of *CFTR* variants and their carrier frequencies, as well as the prevalence of CF in the South African Black population (12,14). Only one study has been conducted in the Division to determine the carrier frequency of the c.2988+1G>A variant, which was reported to be 1 in 34 (23). The carrier frequency of CF in the Black population is consequently estimated to be 1 in 14-59 (15,23). Since then, only a few other variants have been detected in this population group, including G1249E (p.Gly1249Glu), 3196del54, -94G>T, and 2183delAA (8,12,24).

Recent findings in the Division (listed in Table 1) indicate that there are other recurrent *CFTR* variants aside from the c.2988+1G>A variant in the African population. *CFTR* screening by using next generation sequencing (NGS) for 10 patients with a clinical diagnosis of CF and

heterozygous for the c.2988+1G>A variant was performed (Study A). Four different variants in exons 23 and 26 of the *CFTR* gene were identified in seven of the 10 patients. Sanger sequencing (Study B) of exons 23 and 26 was subsequently performed for 15 other suspected CF patients who were positive for only one *CFTR* variant (M/U patients), with three variants identified in this cohort. Following this, in 2022 NGS data generated within the Division using the Inherited Disease Panel (IDP), which is a custom designed NGS targeted panel, designed to test for single nucleotide variants in the 5' and 3' untranslated regions, the exons, and 10bp of intron-exon boundaries of ~500 genes (including *CFTR*) associated with Mendelian disorders, was used for *CFTR* heterozygous variant screening on a larger cohort of 227 individuals (Study C). The c.2988+1G>A variant and two other likely pathogenic variants, c.3983T>A (I1328T) and c.2594G>A (W865S) were identified. Frequency data for these two variants (Study C), as well as the variants identified on NGS (studies A and B) need to be obtained to determine its prevalence and review the classification thereof using updated data as likely pathogenic (from unpublished data; personal communication from Ms. F Essop and Ms. N. Botha).

Table 1. Three recent studies in the Division, with the identified *CFTR* variants (classified as either likely pathogenic and pathogenic), and the counts for each variant out of the total sample size (n)

| | **Study A** | | **Study B** | | **Study C** | |
|---|---|---|---|---|---|---|
| | **NGS** | **Count (n = 10)** | **Sanger sequencing** | **Count (n=15)** | **IDP dataset** | **Count (n=224)** |
| ***CFTR* variants identified** | c.3746G>A | 2 | c.3746G>A(G1249E) | 2 | c.2988+1G>A | 2 |
| | c.4242+1G>T | 3 | c.3763T>C | 1 | c.3983T>A | 2 |
| | c.4144C>T | 1 | c.4137-89A>G | 2 | c.2594G>C | 1 |
| | c.3773dupT | 1 | | | | |

Despite recent advances in knowledge, research on the genetics of CF in the Black South African population group remains limited, and therefore an incomplete variant profile exists (10). Updating the genetic profile in the Black population will direct molecular testing, which will lead to a more complete molecular diagnosis of CF, as well as improve carrier screening, recurrence risk predictions, and genetic counselling (10). The variant profile for the Black South African population can be expanded on and further characterized by screening additional datasets for variants identified in the recent NGS, Sanger and IDP studies (studies A, B, and C)

done in the Division, as well as for the identification of potential novel variants in the *CFTR* gene. A large amount of data is generated by routine NGS (using the IDP), and was therefore used in this study to identify heterozygous *CFTR* variants in the Black population. Together with the findings from the previous studies conducted in the Division, this study (Study D) was conducted to provide an estimation on how common the identified *CFTR* variants are. The aim of this study was therefore to characterize and update the genetic profile of CF and revise the carrier frequency in the Black South African population.

## 3. Methods

### *3.1. Ethics*

An application for ethics approval was submitted to the Human Research Ethics Committee (HREC), School of Pathology, at the University of Witwatersrand, for permission to conduct the study. Clearance was granted as a sub-study (clearance number M230693) under the parent study (clearance number M210989), to use the IDP data generated in the Division as part of routine testing, for further research and analyses. Approval from the Head of Division, Prof Amanda Krause, for permission to conduct the study, was granted.

### *3.2. Dataset*

The dataset consisted of Ion Torrent NGS data (IDP v3.0), from 168 unaffected individuals of African descent that were referred for routine testing for genetic disorders, excluding CF. The dataset, consisting of Variant Call Format (VCF) files and Binary Alignment Map (BAM) files generated from IDP was completely anonymized and de-identified upon receipt. Data from nine different runs were used.

### *3.3. Variant annotation*

The VCF files generated from sequencing each of the 168 individuals were annotated using the Variant Effect Predictor (VEP) tool (https://www.ensembl.org/Tools/VEP), which allocates metadata to the called variants (referenced to the Human Reference genome GRCh37/hg19). Annotations included HGVS nomenclature (descriptions for variants as recommended by the Human Genome Variation Society (25)), the variant type (nonsense, synonymous, missense, frameshift deletion or insertion, splice site etc.) and frequency data from gnomAD (https://gnomad.broadinstitute.org/). *In silico* pathogenicity prediction scores generated by tools such as SIFT (Sorting Intolerant From Tolerant) (26), Polyphen-2 (Polymorphism Phenotyping v2) (27), and CADD (Combined Annotation Dependent Deletion) (28) were also

included along with the predicted outcomes of the variant on protein function (pathogenic, likely pathogenic, benign, or likely benign). Splice predictor scores from SpliceAI (with a score ranging from 0 to 1, which represents the probability of the variant altering the splice site) (29) were included, and a combination of interpretations of the variants from different variant classification submissions on the genomic disease variant database, ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/). The Ensembl VEP output files (in Excel format) containing the called variants along with their annotations, were used for downstream filtering.

### 3.4. *Variant prioritization*

Only variants found in the *CFTR* gene were selected, and subsequently different filtering steps were applied to shortlist variants for classification. This included a minor allele frequency (MAF) < 0.05, which is defined as the frequency at which the less common allele occurs in the population (30,31). CF is a recessive disorder and individuals that are heterozygous carriers for pathogenic variants in *CFTR* could be present in the general population, and therefore a higher MAF was used (31). Furthermore, the variant type was considered and missense, frameshift, indels, splice site, and nonsense variants were prioritized. Missense variants were retained as this variant type accounts for the majority of variants in the *CFTR* gene (https://varsome.com/gene/hg38/CFTR accessed on 11 February 2024). Some intronic and synonymous variants were also retained due to their potential impact on correct splicing, and a possible lack of information rendering them Variants of Uncertain Significance (VUS). Even though the design of IDP only targets exons, intronic variants are frequently detected depending on the primers used, and due to the sequencing of intron-exon boundaries which may extend beyond 10bp.

If this data was available for the variant in question, Polyphen scores (0.85 – 1.00 is damaging), SIFT scores (0.0 is deleterious and 1.0 is tolerated), and CADD scores (1.0 – 99.0 with higher values indicative of a deleterious effect) were considered as indicators of a potential pathogenic effect. A quality control check was performed on the final list of variants obtained from the filtering steps, by viewing the BAM files of individuals in which the variants were present on the Integrative Genomics Viewer (IGV) (https://igv.org/) tool. This tool was used to investigate the position of the variant within *CFTR* and the quality of the call. A minimum depth of coverage of 30x with balanced reads had to be present to ensure confidence in the base calls and distinguish true variants from sequencing artefacts.

### 3.5. *Variant classification*

The prioritized variants that met the quality criteria, as well as the variants from Study C (Table 1) were individually classified according to the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP) guidelines (32). The guidelines use weighted pathogenic (very strong, strong, moderate, or supporting) and benign (stand-alone, strong, or supporting) criteria/codes to classify variants as pathogenic, likely pathogenic, VUS, benign or likely benign. It was decided as part of this project to further categorize the VUS into two classes, namely, VUS due to a lack of information/reports of the variant on public databases, and VUS due to conflicting interpretations (where both benign and pathogenic codes could be applied to the variant).

Evidence for the fulfilment of each code was collected from various resources, such as Varsome (https://varsome.com/) and Franklin by Genoox (https://franklin.genoox.com/), from which a region browser could be used to determine the location of the variant and whether it resides in a functional domain. Functional domains in the CFTR protein were interrogated through literature searches (33), as well as from the UniProt database (https://www.uniprot.org/). Previous submissions and classifications of the variant were retrieved from ClinVar, and variant frequencies from population variant frequency databases (gnomAD). Literature for *in vitro* or *in vivo* functional studies and gene/variant associated information was collected (32). Furthermore, pathogenicity prediction tools on Varsome were used in combination with the VEP pathogenicity annotations, and depending on how many tools were used, codes supporting either pathogenicity or benign variants could be applied.

### *3.6. Carrier frequency calculations*

After completion of the variant analysis on 168 samples, the dataset was combined with the cohort of 227 individuals from the previous NGS carrier screening study (Study C). A total cohort size of 395 individuals was thus used for carrier frequency and birth rate (which is the number of live births per thousand individuals per year) calculations. We define the number of individuals carrying the variant as $\chi$, the total cohort as 395, the total alleles as 790, and the total number of carriers for both pathogenic and likely pathogenic variants (or the combined carrier frequency) as $\gamma$:

The carrier frequency for each pathogenic and likely pathogenic variant was calculated as:

$$\chi \div 395$$

The allele frequencies were then calculated as:

$$\chi \div 790$$

A 95% confidence interval was calculated for both the allele and carrier frequencies of each variant, using the Wilson Score Interval in the R package *binom* (https://rdocumentation.org/packages/binom/versions/1.1-1.1) to increase statistical significance. The Fisher's exact test on R Studio was used to calculate a p-value to determine if there is a significant difference between the allele frequency of the c.2988+1G>A variant, as this is the only known common and unique variant in the Black population, compared to its allele frequency reported on gnomAD for African populations. A significance threshold of $p < 0.05$ was used.

The carrier rate for CF was calculated as:

$$\gamma \div 395$$

The birth rate was calculated as:

carrier rate × carrier frequency × ¼ (as there is a 25% chance of having an affected child if both parents are carriers)

Furthermore, due to it being the only common variant known, the birth rate for c.2988+1G>A was calculated as:

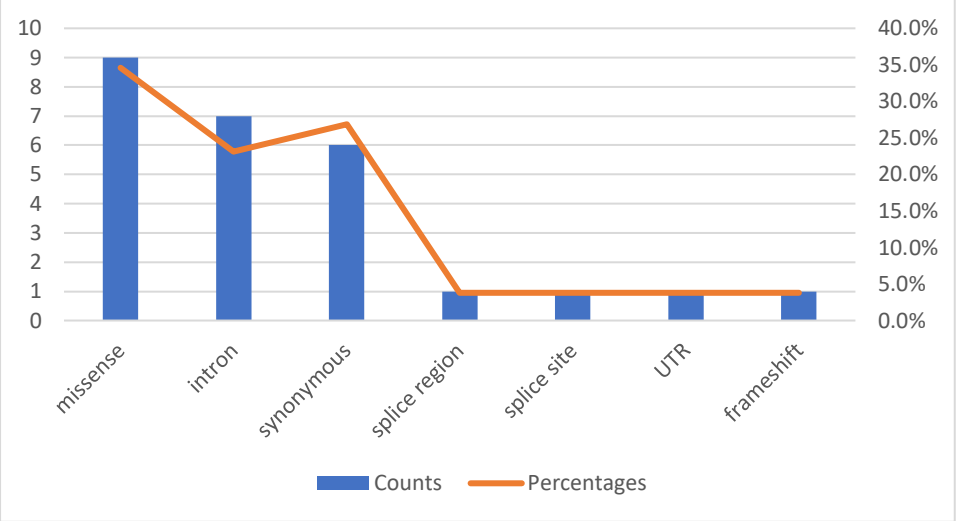
the carrier frequency × carrier frequency × ¼.

## 4. Results

### *4.1.Variant prioritization*

After variant annotation, filtering and prioritization, a final shortlist of 26 *CFTR* variants remained (Table S1), including missense, intronic, synonymous, splice region (within 3-10 bp of the exon), splice (1-2 bp from the exon), untranslated region (UTR), and frameshift variants as shown in Figure 1. Intronic and synonymous variants were initially retained for further investigation, as they may play a role in CF pathogenicity (34). Notably, the second most common variant type was intronic.

Figure 1. Line-column chart depicting *CFTR* variant types and percentages after prioritization




### *4.2.Variant classification*

By using the ACMG-AMP guidelines, three variants out of 26 were classified as pathogenic and two as likely pathogenic (Table 2). The three pathogenic variants, c.2988+1G>A c.3883_3886del, and c.350G>A are well known variants, however the latter two are not specific to the African population. The likely pathogenic variants c.3392T>C, c.3038C>G, and c.2594G>C, are novel variants. In addition to this, the classification of the c.3983T>C and c.2594G>C variants from the IDP dataset (Study C) were confirmed as likely pathogenic, and therefore included for the CF carrier frequency calculation, however, these two variants were not detected in this cohort (Study D). The relevant evidence for each ACMG code applied during the classification of the likely pathogenic and pathogenic variants can be found in Table S2. Furthermore, 18 variants were classified as VUS, and one as likely benign which was therefore removed from the analysis.

Table 2. List of *CFTR* variants identified and classified as pathogenic and likely pathogenic using ACMG codes

| HGVS nomenclature (nucleotide) | HGVS nomenclature (protein) | Legacy name | Variant type | ACMG codes applied[1] | ACMG classification | Carrier frequency | 95% CI | Allele frequency | 95% CI | Reported on gnomAD | Reported on ClinVar |
|---|---|---|---|---|---|---|---|---|---|---|---|
| c.2988+1G>A | | 3120+1G>A | splice variant | PVS1, PS3, PM2, PP3, PP5 | Pathogenic | 0,0101 (4/395) | 0.0039 to 0.026 | 0,0051 (4/790) | 0.0020 to 0.0129 | Yes | Yes |
| c.3883_3886del | p.Ile1295PhefsTer32 | I1295fs | frameshift | PVS1, PP2, PM5, PP5 | Pathogenic | 0,0025 (1/395) | 0.0004 to 0.0142 | 0,0013 (1/790) | 0.0002 to 0.0071 | No | Yes |
| c.350G>A | p.Arg117His | R117H | missense | PS3, PM1, PM2, PM5, PP2, PP3, PP5 | Pathogenic | 0,0025 (1/395) | 0.0004 to 0.0142 | 0,0013 (1/790) | 0.0002 to 0.0071 | Yes | Yes |
| c.3392T>C | p.Ile1131Thr | N/A | missense | PM1, PM2, PP2, PP3 | Likely Pathogenic | 0,0025 (1/395) | 0.0004 to 0.0142 | 0,0013 (1/790) | 0.0002 to 0.0071 | No | No |
| c.3038C>G | p.Pro1013Arg | N/A | missense | PM1, PM2, PP2, PP3 | Likely Pathogenic | 0,0025 (1/395) | 0.0004 to 0.0142 | 0,0013 (1/790) | 0.0002 to 0.0071 | No | No |
| c.3983T>C* | p.Ile1328Thr | N/A | missense | PM1, PM2, PP2, PP3 | Likely Pathogenic (reclassified) | 0,0051 (2/395) | 0.0014 to 0.0183 | 0,0025 (2/790) | 0.0007 to 0.0092 | Yes | Yes |
| c.2594G>C* | p.Trp865Ser | N/A | missense | PM1, PM2, PP2, PP3 | Likely Pathogenic (reclassified) | 0,0025 (1/395) | 0.0004 to 0.0142 | 0,0013 (1/790) | 0.0002 to 0.0071 | No | No |

[1]*PVS1* - Null variant such as frameshift or canonical +/−1 or 2 splice sites where loss of function is a known mechanism of disease. *PS3* - Well-established *in vitro* or *in vivo* functional studies show evidence of a damaging effect of variant. *PM1* – variant is located in a mutational hot spot or functional domain. *PM2* – The variant is at extremely low frequency on databases such as gnomAD. *PM5* - Novel missense change at an amino acid residue where a different pathogenic missense change has been observed. *PP2* - Missense variants are a common mechanism of disease with a low rate of benign missense variants observed. *PP3* - Multiple lines of computational evidence support a pathogenic effect. *PP5* - Dependable source recently reports variant as pathogenic but the evidence is not available to the laboratory to perform an independent evaluation.*Variants identified in Study C

The c.350G>A (R117H) pathogenic variant was investigated further, as this variant in *cis* with a 5T polypyrimidine tract of intron 8 (IVS8-5T) is disease-causing when coupled with a classic *CFTR* variant (De Nooijer et al., 2011; https://cftr2.org/mutation/general/R117H%253B7T accessed on 9 October 2023). To determine whether the NGS panel testing had the ability to detect the polypyrimidine tract, intron 8 of the *CFTR* gene in the individual carrying this variant was inspected on IGV. The tract could be observed (Supplementary Figure 5), however, the number of T's present in the tract and the chromosome phase could not be determined from IGV.

Any VUS that was observed to have evidence of a likely benign effect was excluded from the analysis. Synonymous variants with a SpliceAI score of 0.00 and low conservation scores were also excluded, due to a lack of African data and functional studies. Their predicted effects on protein function are not entirely accurate as it is reliant only on *in silico* tools. One synonymous variant (c.1584G>A) was very close to a splice site (splice distance -1) compared to the other synonymous variants; however, 52 homozygous individuals were reported on gnomAD in the general population, and the variant was therefore excluded as it could thus be a benign variation in the population. It was decided to categorize the 11 remaining VUS into one of two categories: either a VUS due to a lack of information on the variant on databases and in literature, or a VUS due to conflicting interpretations, that is, both benign and pathogenic codes were applied during classification (Table 3).

Table 3. Categorization of *CFTR* VUS, with their relevant ACMG codes and carrier frequencies

| HGVS nomenclature (nucleotide) | HGVS nomenclature (protein) | VUS type | ACMG codes applied[2] | Lack of information | Conflicting interpretations | Carrier frequency |
|---|---|---|---|---|---|---|
| c.2421A>G | p.Ile807Met | Missense | PM2, PP2, BS3 | | x | 1/395 |
| c.853A>T | p.Ile285Phe | | PM2, PP2, PP3, BP6 | | x | 3/395 |
| c.2079T>G | p.Phe693Leu | | PM2, PP2, BS3 | | x | 1/395 |
| c.224G>A | p.Arg75Gln | | PP2, BS1,BS3 | | x | 1/395 |
| c.1767-109G>A | | Intronic | PM2 | x | | 1/395 |
| c.2658-107G>A | | | PM2 | x | | 1/395 |
| c.273+38A>G | | | PM2 | x | | 1/395 |
| c.4242+49G>C | | | PM2 | x | | 1/395 |
| c.4242+10T>C | | | PM2 | x | | 1/395 |
| c.4242+13A>G | | | PM2, BP6 | | x | 1/395 |
| c.164+28A>G | | | NONE | x | | 1/395 |

[2]*BS1* – The variant allele frequency is higher than expected for disorder. *BS3* – Well-established *in vitro* or *in vivo* functional studies indicate no damaging effect of variant on protein function or splicing. *BP4* – Multiple lines of computational evidence support no impact on protein function. *BP6* – Reliable source has recently reported the variant as benign, but an evaluation cannot be made in the laboratory to confirm this as the evidence is not available.

4.3. *Carrier frequency calculations*

Carrier and allele frequencies for all the pathogenic and likely pathogenic variants were calculated (Table 2). Variants from the two other studies in the Division (Study A and B) were not detected in this cohort. The Fisher's exact test for the c.2988+1G>A variant showed that there is a significant difference (p-value = 0.02) in the carrier frequency reported on gnomAD in the African population (1 in 400; reported as 30 in 12486), compared to that found in this cohort (1 in 100). The carrier frequency in this study is therefore significantly higher. The birth rate of c.2988+1G>A was estimated as 1 in 40 000. Lastly, to calculate the overall carrier rate of CF, the pathogenic and likely pathogenic variant counts were combined (11/395), to give a carrier rate of 1 in 36, and the birth rate was estimated as 1 in 5184 individuals.

## 5. Discussion

Currently, the use of variant panels to diagnose CF patients in the Black South African population is not comprehensive using current panel designs based on European variant profiles, as it only accounts for 46% of CF alleles found in this population group. It is therefore imperative to determine the common variants in this population group to characterize the variant profile and direct molecular testing through updated panels. If all the pathogenic and likely pathogenic variants found in this study are assumed to be disease-causing (and correctly classified), it would imply that the c.2988+1G>A variant is observed to account for approximately one third (~36.4%) of CF alleles in this cohort. This is less than the 46% that has been reported in the literature thus far (23), and is therefore an indication that there are other common CF-causing variants in the Black South African population (albeit not necessarily the variants found in this study or in this cohort). The expected birth rate for this variant was calculated as 1 in 40000 individuals. Furthermore, the reported carrier frequency of c.2988+1G>A on gnomAD is significantly lower (1 in 400) than what was observed in this study (1 in 100). This is indicative of a possible geographic-specific difference, due to African data on gnomAD being based on West African, East African, and African American population groups. This once again confirms that gnomAD is not representative of the Southern African population. Furthermore, the small sample size could explain the difference seen in carrier frequencies, as the gnomAD database calculates frequencies using a cohort of approximately 31 500 individuals.

Three of the likely pathogenic variants (c.3392T>C, c.3038C>G, and c.2594G>C) found in this cohort have not yet been reported on ClinVar and gnomAD (Table 2). Therefore, these

variants are novel and could possibly be African-specific variants. This highlights the relevance of this study and how the use of IDP panel data for carrier screening is advantageous. It allows for the discovery/identification of variants that are specific to the Black population. Identifying these variants, not currently included in commercial panels, allows for targeted molecular testing in this population group. Furthermore, the classification of pathogenic and likely pathogenic variants may change as new information becomes available and as research on this population group advances. This study is investigating "healthy" individuals/carriers and not patients, which might add to the complexity of the classification process.

The c.3883_3886del variant (detected in one sample) was an interesting finding, even though this variant has previously been reported in Ashkenazi Jews, and the Lebanese population (36,37). This variant has not yet been observed in African populations. Furthermore, the R117H variant is known to lead to a variable phenotype ranging from severe, to CF-like lung disease, to the congenital absence of the vas deferens (CAVD). For it to be pathogenic it must be in *cis* with a 5T allele (17,35). Its interpretation is therefore challenging. Even though the polyT tract could be observed in individuals with this variant, it is not a completely accurate way of determining the number of thymidines and the chromosome phase. Furthermore, the tract consists of repeats, and due to limitations of Ion Torrent NGS in sequencing homopolymers, it might lead to sequencing artefacts and errors.

Unexpectedly, the variants from the previous studies conducted in the Division that were suspected to be recurrent variants in the Black population, were not detected in this cohort. This could be due to many factors such as limitations of sequencing, where coverage in these regions is too low. The coverage and depth of sequencing was inspected on IGV for randomly selected BAM files from the cohort, and the read depths appeared to be of good quality (> 100x). However, some strand bias/unbalanced reads were observed in select BAM files, especially in the region of the c.4242+1G>T variant. Ion Torrent sequencing can introduce strand bias during template amplification (38), which can lead to false negatives. This could be an explanation as to why the variant was not detected. Other reasons could include the variants being missed due to the small sample size, which remains the biggest limitation of this study. Also, low sensitivity of the IDP assay may have played a role, especially if the variants are located in regions that are difficult to sequence. Additionally, the bioinformatics pipeline used to analyze the IDP datasets could have been too stringent, therefore leading to the exclusion of these variants. Variants may also have been filtered out based on the quality cut-offs that the Ion torrent software uses for the primary analysis to compile the VCF files. A suggestion would

be to lower the quality control measures and minimum read depth used by Ion Torrent software, however, this increases the risk of including low quality variants/artefacts in the analysis.

As was expected, many intronic VUS were found in the cohort. This is due to the high diversity and a general lack of research in the Black South African population. Furthermore, ACMG-AMP guidelines are known to be biased to the classification of coding variants, and therefore non-coding variants in particular will, for the most part, be classified as VUS (32). Importantly, targeted testing/panels for CF have thus far not been able to detect deep-intronic variants, and a limited amount of research has been conducted on these variants (39). Intronic variants may play an important role in CF, as the *CFTR* gene has many *cis*-regulatory elements within the intronic regions that facilitate correct protein folding and expression (34). It is therefore important not to disregard these variants without further investigation.

For most of the intronic variants, the only ACMG code that could be applied was PM2 (the variant is at extremely low frequency on databases such as gnomAD). The application of this code is complicated by contradicting frequency cut-off values specified on Varsome and Franklin, with no specified allele frequency on ClinGen (https://www.clinicalgenome.org/). This leads us to question how applicable this code is for variant classifications in South Africa, given that the African population is not represented by frequency databases. This could therefore have consequences for many existing variant classifications, not only for VUS but for pathogenic and likely pathogenic variants as well. Furthermore, adding to the challenges of classification, there is a lack of guidelines and curated variants for *CFTR* on ClinGen and thus there is no standard approach or recommendations on how to classify certain variants in this gene.

Synonymous VUSs were excluded, due to no obvious effect on protein function through investigation of *in silico* pathogenicity tools. This, and SpliceAI scores, were the only available information on these variants due to a lack of African data. A synonymous variant would only be retained if the SpliceAI score was indicative of an effect on splicing. However, all the synonymous variants identified in this cohort had scores of 0.00. The remaining VUS were further categorized according to whether enough information is available on the variant and which codes were applied during the classification process. As research on the Black South African population progresses, these variants might be reclassified.

The CF disease carrier rate estimated in this study (1 in 36) is concurrent with what is previously reported in South Africa in literature to date (1 in 14-59) (23), meaning that the IDP

data was useful in determining the carrier rate from a cohort of unaffected individuals. The CF disease birth rate was therefore calculated as 1 in 5184 affected births. The findings from this study suggest that CF is the most common autosomal recessive Mendelian disorder in the Black South African population, which is currently thought to be albinism (40,41). This could have implications for new-born carrier screening in the state sector, such as changes in healthcare policies or practices, for example, improving or introducing new-born screening for early detection, treatment, etc. However, with the small sample size being a limitation of this study, it will have to be repeated in a bigger cohort to determine the true CF carrier rate and variant carrier frequencies for the pathogenic and likely pathogenic variants, which would give an indication of how common they actually are. A power calculation was done, and a sample size of more than 1000 individuals would be required as a minimum for optimal statistical power.

In conclusion, this study allowed for the identification of novel, potentially African-specific variants, and a proposed method to update the disease carrier frequency of CF as well as for the common c.2988+1G>A variant, if repeated in a bigger cohort. Given the limitations, this study showed the value of carrier screening using NGS data that is routinely generated in the Division. Expanding on the cohort size in future, will generate more accurate results. This study furthermore highlights the importance of the establishment of guidelines for the classification of variants in *CFTR*, as this process can be very subjective and challenging, especially with the lack of information in the Black population. This could be useful in expanding on the variant profile or updating the panel used for diagnostic testing in South Africa, which will have many benefits for this population group such as targeted molecular testing, therapies that are variant-specific, and accurate genetic counselling (42).

## 6. Declaration of Interests

The authors declare no competing interests.


## 7. Acknowledgments

Prof Amanda Krause (AK) for assistance with result interpretations and discussion points.


## 8. Web Resources

gnomAD (https://gnomad.broadinstitute.org/).

ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/)

Varsome (https://varsome.com/)

Franklin by Genoox (https://franklin.genoox.com/)

UniProt (https://www.uniprot.org/)

## 9. Author Contributions

Conceptualization: AK, FE. Data curation: all authors. Formal analysis: IS, CS. Writing—original draft: IS. Writing—review and editing: all authors. All authors approved the manuscript.

## 10. Data and Code Availability

The datasets and code supporting the current study have not been deposited in a public repository but are available from the corresponding author on request.

## 11. References


1. Riordan JR, Rommens JM, Kerem BS, Alon N, Rozmahel R, Grzelczak Z, et al. Identification of the Cystic Fibrosis Gene: Cloning and Characterization of Complementary DNA. Science (1979) [Internet]. 1989; Available from: www.sciencemag.org
2. Hyde S, Emsley P, Hartshorn M, Mimmack M, Gileadi U, Pearce S, et al. Structural model of ATP-binding proteins associated with cystic fibrosis, multidrug resistance and bacterial transport. Nature. 1990;346:362–5.
3. Bell SC, Mall MA, Gutierrez H, Macek M, Madge S, Davies JC, et al. The future of cystic fibrosis care: a global perspective. Lancet Respir Med. 2020 Jan 1;8(1):65–124.
4. Gentzsch M, Mall MA. Ion Channel Modulators in Cystic Fibrosis. Chest. 2018 Aug 1;154(2):383–93.
5. Ong T, Ramsey BW. Cystic Fibrosis: A Review. JAMA. 2023 Jun 6;329(21):1859–71.
6. Welsh MJ, Smith AE. Molecular Mechanisms of CFTR Chloride Channel Dysfunction in Cystic Fibrosis. Cell. 1993;73:1251–4.
7. Bareil C, Bergougnoux A. CFTR gene variants, epidemiology and molecular pathology. Archives de Pédiatrie [Internet]. 2020;27:8–12. Available from: www.sciencedirect.com
8. Des Georges M, Guittard C, Templin C, Altiéri JP, De Carvalho C, Ramsay M, et al. WGA allows the molecular characterization of a novel large CFTR rearrangement in a black South African cystic fibrosis patient. Journal of Molecular Diagnostics. 2008;10(6):544–8.
9. Zampoli M. Cystic fibrosis: What's new in South Africa in 2019. South African Medical Journal. 2019 Jan 1;109(1):16–9.
10. Krause A, Seymour H, Ramsay M. Common and Founder Mutations for Monogenic Traits in Sub-Saharan African Populations. Annu Rev Genomics Hum Genet [Internet]. 2018;19:149–75. Available from: https://doi.org/10.1146/annurev-genom-083117-
11. Schrijver I, Pique L, Graham S, Pearl M, Cherry A, Kharrazi M. The Spectrum of CFTR Variants in Nonwhite Cystic Fibrosis Patients: Implications for Molecular Diagnostic Testing. Journal of Molecular Diagnostics. 2016 Jan 1;18(1):39–50.
12. Van Rensburg J, Alessandrini M, Stewart C, Pepper MS. Cystic fibrosis in South Africa: A changing diagnostic paradigm. South African Medical Journal. 2018;108(8):624–8.
13. Goldman A, Graf C, Ramsay M. Molecular diagnosis of cystic fibrosis in South African populations. South African Medical Journal. 2003;93(7).
14. Masekela R, Zampoli M, Westwood AT, White DA, Green RJ, Olorunju S, et al. Phenotypic expression of the 3120+1G>A mutation in non-Caucasian children with cystic fibrosis in South Africa. Journal of Cystic Fibrosis. 2013 Jul;12(4):363–6.
15. Zampoli M, Verstraete J, Frauendorf M, Kassanjee R, Workman L, Morrow BM, et al. Cystic fibrosis in South Africa: spectrum of disease and determinants of outcome. ERJ Open Res. 2021 Jul 1;7(3).
16. Simon MA, Csanády L. Molecular pathology of the R117H cystic fibrosis mutation is explained by loss of a hydrogen bond. Structural Biology and Molecular Biophysics. 2021;1–19.
17. Massie R, Poplawski N, Wilcken B, Goldblatt J, Byrnes C, Robertson C, et al. Intron-8 polythymidine sequence in Australasian individuals with CF mutations R117H and R117C. Eur Respir J. 2001;17:1195–200.

18. Chu CS, Trapnell B, Curristin S, Cutting G, Crystal R. Genetic basis of variable exon 9 skipping in cystic fibrosis transmembrane conductance regulator mRNA. Nature. 1993;3:151–6.
19. Stewart C, Pepper MS. Cystic fibrosis on the African continent. Genetics in Medicine. 2016 Jul 1;18(7):653–62.
20. Rodrigues E, Melo M, Reis F, Penna F. Concentration of electrolytes in the sweat of malnourished children. Arch Dis Child. 1994;71:141–3.
21. Mutesa L, Bours V. Diagnostic challenges of cystic fibrosis in patients of African origin. J Trop Pediatr. 2009 Jul 22;55(5):281–6.
22. Mutesa L, Azad AK, Verhaeghe C, Segers K, Vanbellinghen JF, Ngendahayo L, et al. 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. 2009 May 1;135(5):1233–42.
23. Padoa C, Goldman A, Jenkins T, Ramsay M. Cystic fibrosis carrier frequencies in populations of African origin. J Med Genet. 1999;36:41–4.
24. Carles S, Desgeorges M, Goldman A, Thiart R, Guittard C, Kitazos CA, et al. First report of CFTR mutations in black cystic fibrosis patients of southern African origin. J Med Genet. 1996;33(9):802–4.
25. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, Mcgowan-Jordan J, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat. 2016 Jun 1;37(6):564–9.
26. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001 May;11(5):863–74.
27. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.
28. Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–5.
29. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019 Jan 24;176(3):535-548.e24.
30. Dashti MJS, Gamieldien J. A practical guide to filtering and prioritizing genetic variants. Biotechniques. 2017 Jan 1;62(1):18–30.
31. Seaby EG, Pengelly RJ, Ennis S. Exome sequencing explained: A practical guide to its clinical application. Brief Funct Genomics. 2016 Sep 1;15(5):374–84.
32. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine. 2015 May 8;17(5):405–24.
33. Liu F, Zhang Z, Csanády L, Gadsby DC, Chen J. Molecular Structure of the Human CFTR Ion Channel. Cell. 2017 Mar 23;169(1):85-95.e8.
34. Kerschner JL, Ghosh S, Paranjapye A, Cosme WR, Audrézet MP, Nakakuki M, et al. Screening for Regulatory Variants in 460 kb Encompassing the CFTR Locus in Cystic Fibrosis Patients. Journal of Molecular Diagnostics. 2019 Jan 1;21(1):70–80.
35. De Nooijer RA, Nobel JM, Arets HGM, Bot AG, van Berkhout FT, de Rijke YB, et al. Assessment of CFTR function in homozygous R117H-7T subjects. Journal of Cystic Fibrosis. 2011 Sep;10(5):326–32.

36. Farra C, Menassa R, Awwad J, Morel Y, Salameh P, Yazbeck N, et al. Mutational spectrum of cystic fibrosis in the Lebanese population. Journal of Cystic Fibrosis. 2010 Dec;9(6):406–10.
37. Shoshanl T, Augarten A, Yahav J, Gazlt E, Kerem B. Two novel mutations in the CFTR gene: W1089X in exon 17B and 4010delTATT in exon 21. Hum Mol Genet [Internet]. 1994;3(4):657–8. Available from: http://hmg.oxfordjournals.org/
38. Cheng C, Fei Z, Xiao P. Methods to improve the accuracy of next-generation sequencing. Frontiers in Bioengineering and Biotechnology. Frontiers Media S.A.; 2023;11.
39. Deignan JL, Astbury C, Cutting GR, del Gaudio D, Gregg AR, Grody WW, et al. CFTR variant testing: a technical standard of the American College of Medical Genetics and Genomics (ACMG). Genetics in Medicine. 2020 Aug 1;22(8):1288–95.
40. Manga P, Kromberg JGR, Box NF, Sturm RA, Jenkins T, Ramsay M. Rufous Oculocutaneous Albinism in Southern African Blacks Is Caused by Mutations in the TYRPI Gene. American Journal of Human Genetics 1997;61.
41. Kromberg JGR, Kerr R. Oculocutaneous albinism in southern Africa: Historical background, genetic, clinical and psychosocial issues. Afr J Disabil. 2022 Oct 14;11.
42. Cambraia A, Junior MC, Zembrzuski VM, Junqueira RM, Cabello PH, De Cabello GMK. Next-Generation Sequencing for Molecular Diagnosis of Cystic Fibrosis in a Brazilian Cohort. Dis Markers. 2021;2021.
43. Savant A, Lyman B, Bojanowski C, Upadia J. GeneReviews®. 2001. Cystic Fibrosis.
44. Donegà S, Rogalska ME, Pianigiani G, Igreja S, Amaral MD, Pagani F. Rescue of common exon-skipping mutations in cystic fibrosis with modified U1 snRNAs. Hum Mutat. 2020 Dec 1;41(12):2143–54.
45. Sharma N, Sosnay PR, Ramalho AS, Douville C, Franca A, Gottschalk LB, et al. Experimental Assessment of Splicing Variants Using Expression Minigenes and Comparison with In Silico Predictions. Hum Mutat. 2014 Oct 1;35(10):1249–59.
46. Sosnay PR, Siklosi KR, Van Goor F, Kaniecki K, Yu H, Sharma N, et al. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. Nat Genet. 2013 Oct;45(10):1160–7.
47. Yu YC, Sohma Y, Hwang TC. On the mechanism of gating defects caused by the R117H mutation in cystic fibrosis transmembrane conductance regulator. Journal of Physiology. 2016 Jun 15;594(12):3227–44.
48. Macek M, Mackova,' A, Hamosh A, Hilman BC, Selden RF, Lucottej G, et al. Identification of Common Cystic Fibrosis Mutations in African-Americans with Cystic Fibrosis Increases the Detection Rate to 75%. Am J Hum Genet. 1997;60:1122–7.
49. Cutting GR. Cystic fibrosis genetics: From molecular understanding to clinical application. Nature Reviews Genetics. 2015;16:45–56.
50. Rafeeq MM, Murad HAS. Cystic fibrosis: Current therapeutic targets and future approaches. J Transl Med. 2017 Apr 27;15(1):1–9.
51. Gillen AE, Harris A. Transcriptional regulation of CFTR gene expression. Vol. 4, Frontiers in Bioscience. 2012.
52. Ramalho AS, Clarke LA, Sousa M, Felicio V, Barreto C, Lopes C, et al. Comparative ex vivo, in vitro and in silico analyses of a CFTR splicing mutation: Importance of functional studies to establish disease liability of mutations. Journal of Cystic Fibrosis. 2016 Jan 1;15(1):21–33.
53. Fajac I, Girodon E. Genomically-guided therapies: A new era for cystic fibrosis [Internet]. Vol. 27, Archives de Pédiatrie. 2020. Available from: www.sciencedirect.com
54. Ratjen F, Döring G. Cystic fibrosis. Lancet. 2003:681–689.

55. Mishra A, Greaves R, Massie J. The Relevance of Sweat Testing for the Diagnosis of Cystic Fibrosis in the Genomic Era. Clinical Biochem Reviews. 2005; 26.
56. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: Public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862–8.
57. Verlander PC, Kaporis A, Liu Q, Zhang Q, Seligsohn U, Auerbach AD. Carrier Frequency of the IVS4 + 4 A-T Mutation of the Fanconi Anemia Gene FAC in the Ashkenazi Jewish Population. 1995.
58. Hanany M, Allon G, Kimchi A, Blumenfeld A, Newman H, Pras E, et al. Carrier frequency analysis of mutations causing autosomal-recessive-inherited retinal diseases in the Israeli population. European Journal of Human Genetics. 2018 Aug 1;26(8):1159–66.

## 12. Supplementary Material

Table 1. Short list of 26 *CFTR* variants after filtering and prioritization

| | HGVS nomenclature (nucleotide) | Variant type | Carrier Frequency | Allele Frequency |
|---|---|---|---|---|
| 1 | c.2988+1G>A | splice variant | 4/395 | 4/790 |
| **2** | c.3883_3886del | frameshift | 1/395 | 1/790 |
| **3** | c.3392T>C | missense | 1/395 | 1/790 |
| **4** | c.3038C>G | missense | 1/395 | 1/790 |
| **5** | c.350G>A | missense | 1/395 | 1/790 |
| **6** | c.3983T>C | missense | 2/395 | 2/790 |
| **7** | c.2594G>C | missense | 1/395 | 1/790 |
| **8** | c.2421A>G | missense | 1/395 | 1/790 |
| **9** | c.1365G>A | synonymous | 3/395 | 3/790 |
| **10** | c.2657+78G>A | intronic | 5/395 | 5/790 |
| **11** | c.1767-109G>A | intronic | 1/395 | 1/790 |
| **12** | c.853A>T | missense | 3/395 | 3/790 |
| **13** | c.2658-107G>A | intronic | 1/395 | 1/790 |
| **14** | c.2898G>A | synonymous | 3/395 | 3/790 |
| **15** | c.2571G>A | synonymous | 3/395 | 3/790 |
| **16** | c.273+38A>G | intronic | 1/395 | 1/790 |
| **17** | c.1584G>A | synonymous | 2/395 | 2/790 |
| **18** | c.4242+49G>C | intronic | 1/395 | 1/790 |
| **19** | c.164+28A>G | intronic | 1/395 | 1/790 |
| **20** | c.2079T>G | missense | 1/395 | 1/790 |
| **21** | c.4272C>T | synonymous | 1/395 | 1/790 |
| **22** | c.224G>A | missense | 1/395 | 1/790 |
| **23** | c.4242+10T>C | splice region variant | 1/395 | 1/790 |
| **24** | c.4242+13A>G | intronic | 1/395 | 1/790 |
| **25** | c.2820T>G | synonymous | 1/395 | 1/790 |
| **26** | c.-8G>C | UTR | 1/395 | 1/790 |

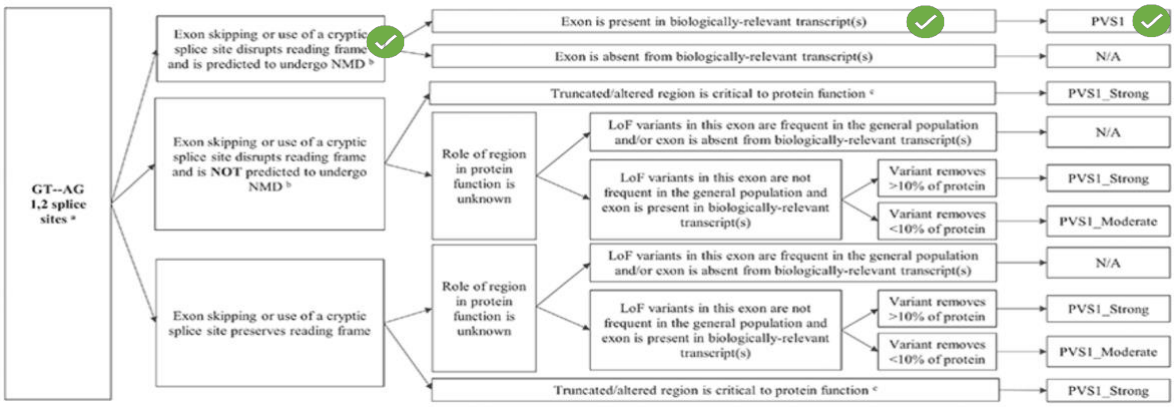
Table 2. Variant classification codes with relevant evidence for pathogenic and likely pathogenic variants in *CFTR*

| Variant | Legacy name | Classification | ACMG code applied | Evidence |
|---|---|---|---|---|
| **c.2988+1G>A** | 3120+1G>A | Pathogenic | PVS1 (very strong pathogenic) | A +1 splicing variant, with the known mechanism of disease for CF being a loss of function (43). The recommended flow chart (Figure 1) was followed. |
| | | | PS3 (strong) | Well-established functional studies on the variant show that the variant leads to the skipping of exon 18 (44), that has a negative/damaging impact on CFTR protein function, with the complete absence of the CFTR protein altogether (45). |
| | | | PM2 (moderate) | A low African frequency of 0.00142 is reported on gnomAD for this variant. |
| | | | PP3 (supporting) | Pathogenicity prediction tools on Varsome show 13 individual predictions of pathogenicity along with 6 pathogenic META scores. |
| | | | PP5 (supporting) | Submissions on ClinVar were investigated, with approximately 26 pathogenic submissions on this variant. One submission has also been reviewed by an expert panel (46) from the CFTR2 database (https://cftr2.org), where the variant is also described as CF-causing. This variant has been identified in multiple Black South African CF patients (14). |
| **c.3883_3886del** (p.Ile1295PhefsTer32) | 4010del4 | Pathogenic | PVS1 (very strong pathogenic) | Frameshift variant (therefore null variant), with the known mechanism of disease of CF being loss of function (43). The recommended flow chart (Figure 2) was followed. |
| | | | PM2 (moderate) | The variant is absent from the gnomAD database (with good coverage of in gnomAD exomes and genomes samples, 91.19% and 93.57% respectively). |
| | | | PP5 (supporting) | Submissions on ClinVar were investigated, and the variant has been reviewed by an expert panel (46), with two very recent reputable sources (Invitae and the Institute of Human Genetics, University of Leipzig Medical Center) reporting the variant as pathogenic. The variant is also found on the CFTR2 database, which reports the variant as disease-causing (https://cftr2.org/mutation/general/4010del4/). |
| **c.350G>A** (p.Arg117His) | R117H | Pathogenic | PS3 (strong) | Well-established functional studies on the variant show that the variant has a damaging effect on protein function, surface expression, and channel opening (16,35,47). |
| | | | PM1 (moderate) | Functional domains on UniProt (https://www.uniprot.org/) indicate that this variant is located at amino acid position 117, and therefore lies within the ABC transmembrane type 1-1 domain which is a functional domain (Figure 3) (33). No benign variation is reported close to this position. |
| | | | PM2 (moderate) | A very low African frequency of 0.000369 is reported on gnomAD for this variant. |
| | | | PM5 (moderate) | Different missense changes at this position have previously been described as pathogenic which has been reviewed by an expert panel (46). |

| | | | | |
|---|---|---|---|---|
| | | | PP2 (supporting) | Varsome (https://varsome.com/) reports that out of the total missense variants found within *CFTR*, 563 are pathogenic and 48 are benign, meaning that there is a low rate of benign missense variation within this gene (Figure 4). |
| | | | PP3 (supporting) | Pathogenicity prediction tools on Varsome shows 15 individual predictions of pathogenicity along with 10 pathogenic META scores. An aggregated moderate pathogenic score of 0.846 was recorded on the Franklin by Geenox database (https://franklin.genoox.com/). |
| | | | PP5 (supporting) | There are approximately 34 pathogenic submissions on ClinVar, one of which has also been reviewed by an expert panel. |
| **c.3392T>C** (p.Ile1131Thr) | | Likely Pathogenic | PM1 (moderate) | Functional domains on UniProt (https://www.uniprot.org/) indicate that this variant is located at amino acid position 1131, and therefore lies within the ABC transmembrane type 1-2 domain which is a functional domain (Figure 3) (33). No benign variation is reported close to this position. |
| | | | PM2 (moderate) | The variant is absent from the gnomAD database (with good coverage of in gnomAD exomes and genomes samples, 99.86% and 93.25% respectively). |
| | | | PP2 (supporting) | Varsome (https://varsome.com/) reports that out of the total missense variants found within *CFTR*, 563 are pathogenic and 48 are benign, meaning that there is a low rate of benign missense variation within this gene (Figure 4). |
| | | | PP3 (supporting) | Pathogenicity prediction tools on Varsome shows 17 individual predictions of pathogenicity along with 11 pathogenic META scores. An aggregated moderate pathogenic score of 0.867 was recorded on the Franklin by Geenox database (https://franklin.genoox.com/). |
| **c.3038C>G** (p.Pro1013Arg) | P1013L | Likely Pathogenic | PM1 (moderate) | Functional domains on UniProt (https://www.uniprot.org/) indicate that this variant is located at amino acid position 1031, and therefore lies within the ABC transmembrane type 1-2 domain which is a functional domain (Figure 3) (33). No benign variation is reported close to this position. |
| | | | PM2 (moderate) | The variant is not reported in the African population on gnomAD with a total frequency of 0.00000737. |
| | | | PP2 (supporting) | Varsome (https://varsome.com/) reports that out of the total missense variants found within *CFTR*, 563 are pathogenic and 48 are benign, meaning that there is a low rate of benign missense variation within this gene (Figure 4). |
| | | | PP3 (supporting) | Pathogenicity prediction tools on Varsome shows 22 individual predictions of pathogenicity along with 13 pathogenic META scores. An aggregated moderate pathogenic score of 0.859 was recorded on the Franklin by Geenox database (https://franklin.genoox.com/). |
| **c.3983T>C** (p.Ile1328Thr) (reclassification) | I1328T | Likely Pathogenic | PM1 (moderate) | Functional domains on UniProt (https://www.uniprot.org/) indicate that this variant is located at amino acid position 1328, and therefore lies within the ABC transporter 2 domain which is a functional domain (Figure 3) (33). No benign variation is reported close to this position. |

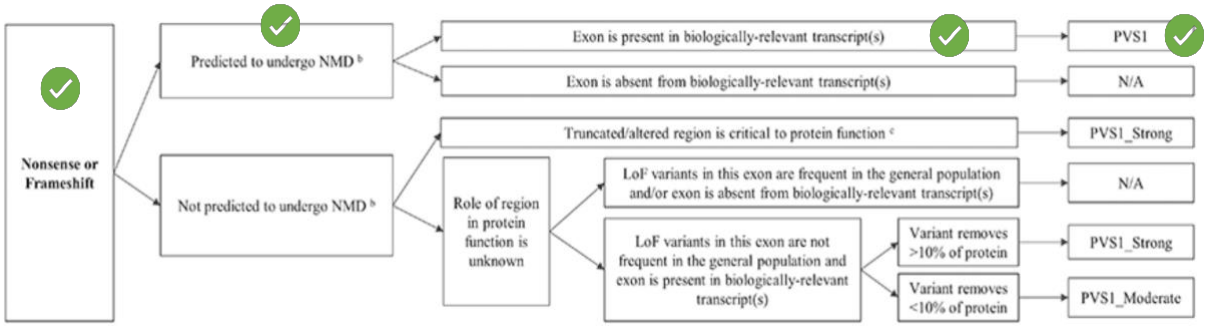
| | | | | |
|---|---|---|---|---|
| | | | PM2 (moderate) | The variant is reported on gnomAD with a total frequency of 0.0000159, with no African frequency reported. |
| | | | PP2 (supporting) | Varsome (https://varsome.com/) reports that out of the total missense variants found within *CFTR*, 563 are pathogenic and 48 are benign, meaning that there is a low rate of benign missense variation within this gene (Figure 4). |
| | | | PP3 (supporting) | Pathogenicity prediction tools on Varsome shows 15 individual predictions of pathogenicity along with 11 pathogenic META scores. An aggregated strong pathogenic score of 0.984 was recorded on the Franklin by Geenox database (https://franklin.genoox.com/). |
| **c.2594G>C** (p.Trp865Ser) (reclassification) | | Likely Pathogenic | PM1 (moderate) | Functional domains on UniProt (https://www.uniprot.org/) indicate that this variant is located at amino acid position 865, and therefore lies within the ABC transmembrane type 1-2 domain which is a functional domain (Figure 3) (33). No benign variation is reported close to this position. |
| | | | PM2 (moderate) | The variant is absent from the gnomAD database (with good coverage of in gnomAD exomes and genomes samples, 99.63% and 88.04% respectively). |
| | | | PP2 (supporting) | Varsome (https://varsome.com/) reports that out of the total missense variants found within *CFTR*, 563 are pathogenic and 48 are benign, meaning that there is a low rate of benign missense variation within this gene (Figure 4). |
| | | | PP3 (supporting) | Pathogenicity prediction tools on Varsome showed conflicting interpretations, however, the CADD tool (https://cadd.gs.washington.edu/snv) generated a score of 25.8, Mutation Taster (https://www.mutationtaster.org/) indicated that the variant is disease-causing, and the Mendelian Clinically Applicable Pathogenicity (M-CAP) score of 0.448 is indicative of a likely pathogenic outcome for this variant. |

Figure 1. Recommended flow chart by Abou Tayoun et al., 2018, with the fulfilment of each criterion for the c.2988+1G>A variant marked in green




The c.2988+1G>A variant is found in intron 18 and leads to exon skipping (44,48) (and therefore it is not found in the most 3' exon, which means that it is predicted to undergo nonsense-mediated decay). The skipped exon (exon 18) is found in the biologically-relevant transcript, which is the MANE select transcript of this gene (NM_000492.4) is curated by the HUGO Gene Nomenclature Committee (https://www.genenames.org/data/gene-symbol-report/#!/hgnc_id/HGNC:1884), and therefore PVS1 can be applied to this variant.

Figure 2. Recommended flow chart by Abou Tayoun et al., 2018, with the fulfilment of each criterion for the c.3883_3886del variant marked in green




For the c.3883_3886del variant (p.Ile1295PhefsTer32), the premature stop codon caused by the variant is not found in the most 3' exon, as it occurs in exon 24 of 27 in the *CFTR* gene, and it is known that this variant causes a premature stop-codon (37). Furthermore, the MANE select transcript of this gene (NM_000492.4) is curated by the HUGO Gene Nomenclature Committee (https://www.genenames.org/data/gene-symbol-report/#!/hgnc_id/HGNC:1884) meaning that exon 24 is present in the biologically relevant transcript, and therefore PVS1 could be applied.

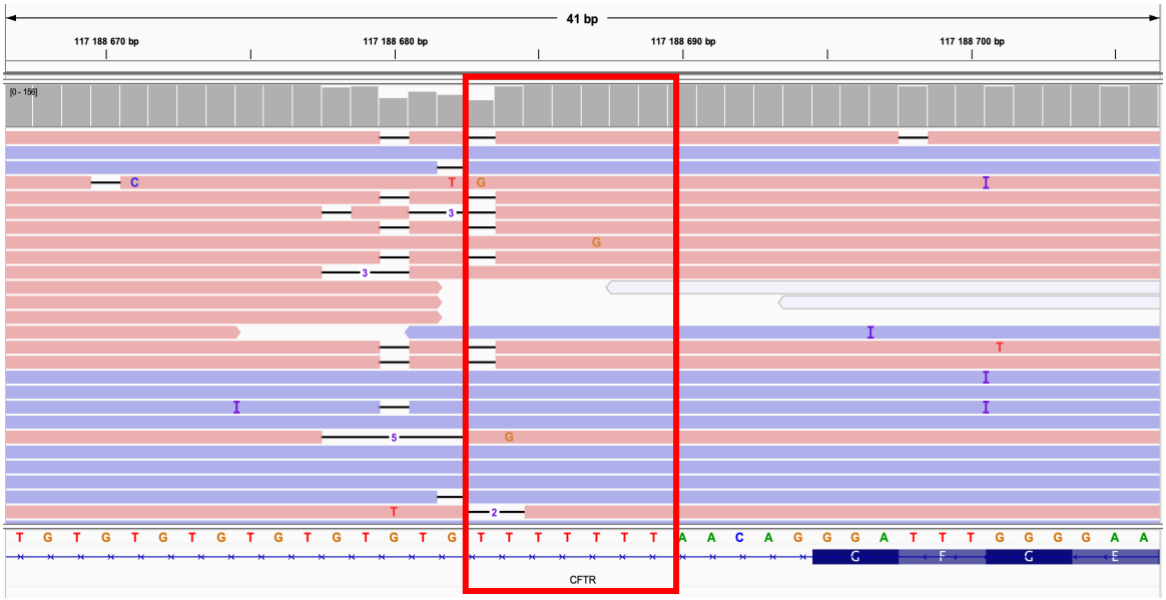
Figure 3. Functional domains within the *CFTR* gene as illustrated on UniProt

| | | | |
|---|---|---|---|
| ▶ Domain | 81-365 | ABC transmembrane type-1 1 PROSITE-ProRule Annotation | BLAST Add |
| ▶ Domain | 423-646 | ABC transporter 1 PROSITE-ProRule Annotation | BLAST Add |
| ▶ Region | 654-831 | Disordered R region 1 Publication | BLAST Add |
| ▶ Domain | 859-1155 | ABC transmembrane type-1 2 PROSITE-ProRule Annotation | BLAST Add |
| ▶ Domain | 1210-1443 | ABC transporter 2 PROSITE-ProRule Annotation | BLAST Add |
| ▶ Region | 1386-1480 | Interaction with GORASP2 1 Publication | BLAST Add |
| ▶ Region | 1452-1480 | Disordered Automatic Annotation | BLAST Add |
| ▶ Compositional bias | 1463-1480 | Basic and acidic residues Automatic Annotation | BLAST Add |
| ▶ Motif | 1478-1480 | PDZ-binding 2 Publications | BLAST Add |

Figure 4. Variant types and counts found in *CFTR* as reported on Varsome

| Coding impact | Pathogenic | Likely Pathogenic | Uncertain Significance | Likely Benign | Benign | Total |
|---|---|---|---|---|---|---|
| Synonymous | 8 | 3 | 94 | 616 | 13 | 734 |
| Missense | 563 | 125 | 1762 | 48 | 44 | 2542 |
| Nonsense | 388 | 55 | 38 | 0 | 0 | 481 |
| Start loss | 9 | 1 | 4 | 0 | 0 | 14 |
| Stoploss | 0 | 2 | 3 | 0 | 0 | 5 |
| Frameshift | 442 | 87 | 37 | 0 | 0 | 566 |
| Inframe Indel | 49 | 5 | 46 | 3 | 0 | 103 |
| Splice junction loss | 173 | 34 | 14 | 1 | 0 | 222 |
| Non-coding | 46 | 8 | 120 | 114 | 14 | 302 |
| Total | 1678 | 320 | 2118 | 782 | 71 | 4969 |

Figure 5: screenshot of IGV analysis of intron 8 in the individual with the c.350G>A variant, to assess the poly-T tract (red block) and the ability of NGS to detect the number of T's present



# Appendix A:
# Approved Project Protocol

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Private Bag 3 Wits, 2050
Fax: 027117172119
Tel: 02711 7172076

Reference: Mrs Sandra Benn
E-mail: sandra.benn@wits.ac.za

17 July 2023
Person No: 2721074
PAG

Miss IM Smit
23 The Braids Road
Emmarentia
2195
South Africa

Dear Miss Ingrid Smit

**Master of Science in Medicine: Approval of Title**

We have pleasure in advising that your proposal entitled *Cystic fibrosis: An update on the variant profile and carrier frequency in the Black South African population* has been approved. Please note that any amendments to this title have to be endorsed by the Faculty's higher degrees committee and formally approved.

Yours sincerely

Mrs Sandra Benn
Faculty Registrar
Faculty of Health Sciences

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

| | | | |
|---|---|---|---|
| CANDIDATE'S SURNAME: SMIT | | FIRST NAME/S: INGRID | STUDENT NUMBER: 2721074 |
| CURRENT QUALIFICATIONS: BSc HONS GENETICS | | | |
| TEL: | CELL: 0788586376 | E-MAIL: 2721074@students.wits.ac.za | FAX: |
| DEGREE FOR WHICH PROTOCOL IS BEING SUBMITTED: MSc (MED) GENOMIC MEDICINE | | | |
| PART-TIME OR FULL-TIME: FULL-TIME | | | |
| FIRST REGISTERED FOR THIS DEGREE: | TERM : 2023 | | YEAR: 1 |
| DEPARTMENT: HUMAN GENETICS | | | |
| TITLE OF PROPOSED RESEARCH: Cystic fibrosis: An update on the variant profile and carrier frequency in the Black South African population | | | |
| CANDIDATE'S SIGNATURE: | | | DATE: 25/04/2023 |
| SUPERVISOR 1 (NAME & SURNAME): FAHMIDA ESSOP | | | % Supervision |
| SUPERVISOR'S QUALIFICATIONS: MSc (MED) HUMAN GENETICS | | | |
| SUPERVISOR'S DEPARTMENT: HUMAN GENETICS | | | |
| SUPERVISOR'S ADDRESS / TEL / E-MAIL: fahmida.essop@nhls.ac.za or fahmida.essop@wits.ac.za | | | |
| SUPERVISOR 2 (NAME & SURNAME): NADINE BOTHA | | | % Supervision |
| SUPERVISOR'S QUALIFICATIONS: MSc (MED) HUMAN GENETICS | | | |
| SUPERVISOR'S ADDRESS / TEL / E-MAIL: nadine.botha@nhls.ac.za | | | |
| SUPERVISOR 3 (NAME & SURNAME): N/A | | | % Supervision |
| SUPERVISOR'S QUALIFICATIONS: N/A | | | |
| SUPERVISOR'S ADDRESS / TEL / E-MAIL: N/A | | | |
| *SYNOPSIS OF RESEARCH:(Brief summary of proposed research project; between 200-300 words only; with sub-headings: an introduction and justification for study, aim/s, proposed methodology and expected outcome/s)* [Use reverse side of this page if more space is required] | | | |

| | | |
|---|---|---|
| WITS ETHICS NOT REQUIRED: Yes No<br>WITS ETHICS PENDING: Yes No<br>WITS ETHICS APPROVED: Yes No<br>(circle appropriate symbol)*<br><br>***Please note the final human ethics clearance certificate or animal ethics certificate must be available prior to starting research** | | IF Y SUPPLY ETHICS CLEARANCE CERTIFICATE AS ATTACHMENT AND INCLUDE ETHICS NUMBER HERE: |
| **As supervisor/s, I/we confirm that I have read the protocol which has been submitted for assessment.** | | |
| SIGNATURE OF SUPERVISOR/S: | …… | …… |
| SIGNATURE PG OFFICE STAFF<br><br>…… | REGISTERED<br><br>YES….. NO….. | STAMP |

***SYNOPSIS OF RESEARCH CONTINUED***

***Introduction and Justification:***

Cystic Fibrosis (CF) is an autosomal recessive disorder caused by pathogenic variants in the *CFTR* gene. In the Black South African (SA) population, limited research has been conducted on the genetics of this disease. CF is underdiagnosed due its broad spectrum of symptoms being similar to conditions frequently encountered in Africa, and due to limitations of diagnostic testing for CF. The sweat chloride test, which is the gold standard for diagnosing CF, is generally inaccessible in South Africa and in most circumstances molecular testing is the only way in which a CF diagnosis can be confirmed. At the NHLS, the Elucigene CF50 kit is used to test for the 3120+1G>A variant and other common CF variants found in the European population. Except for the 3120+1G>A variant, other variants unique to the Black South African population remain unknown and unaccounted for. However, recent studies in the Division show evidence of other recurrent *CFTR* variants in individuals of African ancestry. These variants and others could be further explored in this population group by using NGS data available from the Division. The aim of this study is therefore to expand on the variant profile and revise the carrier frequency of CF for individuals of African ancestry in South Africa. NGS data generated for unaffected individuals will be used to review previously identified pathogenic variants and identify novel pathogenic variants in the *CFTR* gene. From this, the variant carrier frequencies will be estimated, as well as the overall carrier frequency and prevalence of the disease in the general Black population. Ultimately, updating the genetic profile will direct molecular testing, which will lead to a more complete molecular diagnosis of CF, as well as improved carrier screening, recurrence risk predictions, and genetic counselling that Black CF patients will receive.

***Aim:***

To characterize and update the genetic profile of CF and revise the carrier frequency and prevalence in the Black South African population.

***Methodology:***

Methodology will include *CFTR* variant identification (previously identified variants as well as novel variants) from NGS data available from the Division, variant annotation, prioritization, and classification of variants as pathogenic or likely pathogenic according to ACMG guidelines.

***Expected outcome:***

What we expect to find when screening for variants in *CFTR,* is that CF is more common in Black South Africans than previously identified, with a higher carrier frequency, and other common/recurrent variants than previously reported.

25 April 2023/MP

# Protocol

**Ingrid Smit**

**MSc (Med) Genomic Medicine**

**2721074**

---

**Cystic fibrosis: An update on the variant profile and carrier frequency in the Black South African population**

---

Supervisors: Ms. F. Essop and Ms. N. Botha

**Table of Contents**


**1. Title** ..... **7**

**2. Investigators** ..... **7**

**3. Introduction** ..... **7**

***3.1. Cystic Fibrosis*** ***7***
***3.2. Variants in the CFTR gene*** ***7***
***3.3. CF diagnosis in South Africa*** ***8***
***3.4. Research of CF variants in Black South African individuals*** ***9***
***3.5. Study rationale*** ***10***

**4. Aims & Objectives** ..... **10**

***4.1. Aim*** ***10***
***4.2. Objectives*** ***11***

**5. Methods** ..... **11**

***5.1. Dataset*** ***11***
***5.2. Objective 1*** ***11***
*5.2.1. Variant annotation* ..... *11*
*5.2.2. Variant prioritization* ..... *12*
***5.3. Objective 2*** ***13***
*5.3.1. Variant classification* ..... *13*
***5.4. Objective 3*** ***13***
*5.4.1. Carrier frequency estimation* ..... *13*

**6. Ethics** ..... **13**

**7. Timeline** ..... **14**

**8. Funding** ..... **14**

**9. Limitations** ..... **14**

**10. References** ..... **14**

## 1. Title

Cystic fibrosis: An update on the variant profile and carrier frequency in the Black South African population

## 2. Investigators

Ms. Fahmida Essop (MSc (Med) Human Genetics, Medical Scientist)

Ms. Nadine Botha (MSc (Med) Human Genetics, Medical Scientist)

## 3. Introduction

### *3.1. Cystic Fibrosis*

Cystic fibrosis (CF) is a fatal, autosomal recessive disorder, affecting multiple organs in the body that have mucus-secreting functions, such as the lungs, liver, intestinal tract, and pancreas (as reviewed by Cutting, 2015; Van Rensburg et al., 2018). CF is caused by mutations (variants) in both copies of the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene (as reviewed by Bareil & Bergougnoux, 2020). The *CFTR* gene is 189 Kb in size and encodes a protein, CFTR, which acts as a chloride channel in epithelial membranes throughout the body. The protein is responsible for transporting electrolytes across these membranes, allowing for water to follow, thereby thinning out mucus secretions (50,51).

### *3.2. Variants in the CFTR gene*

Currently, there are 2114 identified variants (both benign and pathogenic), (http://www.genet.sickkids.on.ca/cftr accessed on 15 March 2023) in *CFTR*. Pathogenicity of variants is dependent on its position within *CFTR* and the effect on the function of the CFTR protein, for example, if a variant is located in the CFTR domain that is involved in protein folding, this will lead to instability, having severe consequences and ultimately leading to a CF manifestation (as reviewed by Bareil & Bergougnoux, 2020). There are six different classes (Class I to VI) of pathogenic variants found in *CFTR*, grouped according to variant type and effect (50). The classes are indicative of the level of severity of the effect on the protein function, from most severe (Class I) to least (Class VI) (Rafeeq & Murad, 2017; Ramalho et al., 2016). The classes can be used to direct CFTR therapies, for example potentiators, such as Ivacaftor, can be used in instances where the CFTR protein is present in a small amount (Class V) or is dysfunctional (Class III and IV), to improve channel opening. Correctors, such as Lumacaftor, can be used for Class II variants that lead to trafficking defects, to essentially improve transportation of the protein to the epithelial membrane (53).

The most common pathogenic variant found in the European Caucasian population, is p.Phe508del (c.1521_1523del), a Class II variant, leading to defective localization/trafficking of CFTR due to a 3bp deletion at codon 508 (Ramalho et al., 2016). This variant accounts for 70% of CF variants globally (Mutesa & Bours, 2009; Schrijver et al., 2016). Common CF variants such as p.Phe508del are usually population specific and differ in frequency between different ethnic groups (12,54).

### *3.3. CF diagnosis and clinical testing in South Africa*

More than 15 years ago, it was believed that CF only affected Caucasian individuals (8), and therefore research in South Africa (and globally) was mostly focused on variants found in this specific population group (9). However, reports of CF in Black South Africans started increasing since 1954, and it is now known to be much more common than previously thought (10,24). CF is frequently misdiagnosed in South Africa (14) due to the broad phenotype of CF with symptoms that may be masked by conditions commonly found in Africa, such as HIV, chronic pulmonary infections, protein energy malnutrition (PEM), and tuberculosis (19,21).

Sweat chloride testing is the gold standard for diagnosing CF (9), however, it is for the most part inaccessible in a South African setting, due to the technical skills and training with experienced staff required to perform the test (13,19). The reliability of the sweat chloride test could be influenced by conditions such as PEM that has been shown to increase sweat chloride levels (20), and in a study by Mutesa et al. conducted in 2009, 37 out of 52 patients that tested positive for a sweat chloride test (>60 mmol/L), did not have any variants causative of CF in *CFTR* (21,22). Elevated sweat levels from disorders that are difficult to differentiate from CF will therefore lead to false positives and is insufficient in a CF diagnosis (55). The sweat test remains the gold standard despite its limitations, yet it is still not accessible in many African countries (21). A pancreatic insufficiency (PI) screening test is usually done when sweat testing is not available, as PI occurs in about 83% of all CF patients (Masekela et al., 2013; Zampoli, 2019).

### *3.4. Genetic testing in South Africa*

Molecular testing in South Africa is, in most circumstances, the only way in which a CF diagnosis and be confirmed, but diagnostic testing is ultimately limited due to a lack of sequencing of the full *CFTR* gene for causative variants in the state sector, meaning that other causative variants remain undetected and unknown (12). With commercially available genetic testing kits or assays designed based on European data, this might also lead to variants in the

Black population being unaccounted for (19). For example, in the South African Caucasoid population, the p.Phe508del variant is reported in 81% of CF alleles, but this variant is not found to be common in the Black population (13,14), further demonstrating that CF variants are population-specific (12).

Currently, the commercial Elucigene CF-EU2v1 kit (known as the CF50 panel) is used at the NHLS, which screens for 50 known pathogenic variants (Zampoli, 2019; personal communication from Ms F. Essop). Even though the kit is based on Caucasian data, it is still used in a South African setting, as some variants may be shared between Caucasians and other ethnic groups due to high levels of admixture in the South African population (11,24). Using the Elucigene kit, 46% of *CFTR* variants are accounted for in the Black population, and this is due to the common 3120+1G>A variant (c.2988+1G>A) accounting for 46% of all CF alleles with no other common *CFTR* variant identified to date, as the variant profile in this population group has not yet been fully characterized (13,14). This further drives the reasoning behind using this kit in a South African setting, as a significant proportion is still being covered by the kit, even though other variants are unaccounted for (Zampoli, 2019; personal communication from Ms F. Essop).

### *3.5. Research of CF in Black South African individuals*

Currently, limited data have been generated to investigate the profile of *CFTR* variants and their carrier frequencies, as well as the prevalence of CF in the Black population (12,14). Thus far, only one study has been conducted to determine the carrier frequency of the common 3120+1G>A variant, which was reported to be 1 in 34 (23). Since then, only a few other variants have been detected in this population group, including G1249E (p.Gly1249Glu), 3196del54, -94G>T, and 2183delAA (8,12,24).

Recent findings in the Division (listed in *Table 1* from left to right) indicate that there are other recurrent *CFTR* variants aside from the 3120+1G>A variant in the African population: next generation sequencing (NGS) for 10 patients with a clinical diagnosis of CF, and heterozygous carriers for the common 3120+1G>A variant, was performed. Four different variants in exons 23 and 26 of the *CFTR* gene were identified in seven of the 10 patients. This led to the decision to perform Sanger sequencing in exons 23 and 26 for 15 other suspected CF patients who were positive for only one *CFTR* variant (M/U patients), with three variants identified in this cohort.

Following this, NGS data from the Inherited Disease Panel (IDP) was used for *CFTR* heterozygous variant screening on a larger cohort of 224 individuals, with the common

3120+1G>A variant being identified, including two other likely pathogenic variants, c.3983T>A and c.2594G>A. Frequency data for these two variants needs to be obtained to determine its prevalence and confirm the classification thereof as likely pathogenic (from unpublished data; personal communication from Ms. F Essop and Ms. N. Botha).

*Table 1*: Three recent studies in the Division, with the variants that were found listed (classified as either likely pathogenic and pathogenic), along with the counts/frequencies for each variant out of the total sample size in brackets.

| **Study:** | **NGS** | **Sanger sequencing** | **IDP dataset** |
|---|---|---|---|
| **Variants identified:** | c.3746G>A (2/10)<br>c.4242+1G>T (3/10)<br>c.4144C>T (1/10)<br>c.3773dupT (1/10) | c.3746G>A(G1249E) (2/15)<br>c.3763T>C (1/15)<br>c.4137-89A>G (2/15) | c.2988+1G>A (2/224)<br>**c.3983T>A** (2/224)<br>**c.2594G>C** (1/224) |

### *3.6.Study rationale*

At present, despite recent increases in knowledge, CF genetics in the Black South African population group is insufficiently researched, and therefore an incomplete variant profile exists (10). The variant profile for the Black South African population can be expanded on and further characterized by screening additional datasets for previously identified variants from the recent NGS, Sanger and IDP studies done in the Division, as well as for the identification of potential novel variants in the *CFTR* gene. A large amount of data is generated by routine NGS testing offered by the NHLS that can be used to identify CF variants in the *CFTR* gene in the Black population. By combining these results with results from larger cohorts from previous studies, this will provide an estimation on how common the identified *CFTR* variants are, as well as an estimation of their carrier frequencies and overall disease prevalence. Ultimately, by updating the genetic profile, this will direct molecular testing, which will lead to a more complete molecular diagnosis of CF, as well as improved carrier screening, recurrence risk predictions, and genetic counselling that Black CF patients will receive (10).

## 4. Aims & Objectives

### *4.1.Aim*

To characterize and update the genetic profile of CF and revise the carrier frequency and prevalence in the Black South African population.

### *4.2. Objectives*

4.2.1. To review previously identified variants and identify novel pathogenic heterozygous variants in the *CFTR* gene using NGS data from individuals of African ancestry.

4.2.2. To classify the identified (novel and previously identified) variants using ACMG guidelines.

4.2.3. To estimate carrier frequencies of identified likely pathogenic and pathogenic variants and review the carrier frequency of cystic fibrosis in Black South Africans.

## 5. Methods

### *5.1. Dataset*

Next generation sequencing (NGS) data from an additional 104 unaffected individuals of African descent that were referred for routine testing for disorders, excluding CF, using the IDP version 3 (IDP v3.0) will be used in this study. The IDP v3.0 is a custom designed NGS targeted panel, designed to test for single nucleotide variants in the 5' and 3' untranslated regions, the exons, and 10bp of intron-exon boundaries of ~500 genes (including *CFTR*) associated with Mendelian disorders.

The dataset, consisting of Variant Call Format (VCF) files and Binary alignment MAP (BAM) files generated from IDP will be completely anonymized and de-identified upon receival. Upon completion of the methodology, this cohort of 104 individuals will be combined with 224 individuals from the previous study undertaken in the Division. By using the full dataset consisting of 328 individuals, the carrier frequency and prevalence of CF can be estimated (Objective 3).

### *5.2. Objective 1*

To achieve this objective, the identification of previously identified pathogenic variants and novel pathogenic variants in the *CFTR* gene will be done through variant annotation and prioritization:

#### *5.2.1. Variant annotation*

Variant annotation will be performed using online tools such as IonReporter software (https://ionreporter.thermofisher.com/ir/) and the Variant Effect Predictor (VEP) (https://www.ensembl.org/info/docs/tools/vep/index.html) using VCF files that were generated after sequencing as input. These tools allocate metadata to the called variants (30), such as

HGVS nomenclature (descriptions for variants as recommended by the Human Genome Variation Society (25)), variant location within the gene (coding, non-coding, splice position), variant type (nonsense, synonymous, missense, frameshift deletion or insertion, splice site etc.), allele frequencies from databases such as GnomAD (https://gnomad.broadinstitute.org/), as well as a combination of interpretations of variants from different variant classification submissions from ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/). Whether the variant has been previously reported or not (thus novel) (from GnomAD), and *in silico* pathogenicity prediction scores with outcomes from pathogenicity prediction tools, will also be added to the variants. The data gathered during variant annotation will be used for variant prioritization.

### *5.2.2. Variant prioritization*

To prioritize the called variants, variants will be filtered using GnomAD (https://gnomad.broadinstitute.org/), based on the population variant frequency or Minor Allele Frequency (MAF), which is the reported frequency of the variant in a given population (30,31). Variants will be retained according to a MAF < 5%, and this cut-off is chosen because CF is an autosomal recessive disorder and unaffected individuals that are carriers of a heterozygous variant may be present in the general population (31). Other filtering factors will include variant type, if the variant has been reported in association with the disease or not (thus potentially novel), and pathogenicity prediction scores.

For the variant type (from the VEP tool), retained variants will include missense, frameshift, indels, splice site, and nonsense variants, as these variants are known to have a greater impact on the protein than noncoding and synonymous variants, which will be excluded (30). ClinVar is a publicly available database that has information regarding a history of interpretations of variants throughout the human genome (56), and by using this database, identified variants (from variant annotation) that are not available in the database could potentially be novel. Lastly, pathogenicity prediction tools such as Polyphen and SIFT, will help identify likely candidates for deleterious effects, and to identify those that are predicted to have the highest impact on the protein by using pathogenicity prediction scores, taking into consideration the effect of the variant on different levels, such as DNA, mRNA, and protein level (31).

After variant prioritization, a quality control check of variants will be performed by visualizing variants on binary alignment map (BAM) files using the Integrative Genomics Viewer (IGV) (https://igv.org/). A quality control check of variants will also be performed, and variants of low quality will be excluded from the analysis (31).

### *5.3. Objective 2*

#### *5.3.1. Variant classification*

The ACMG guidelines will be used to classify the prioritized variants (Richards et al., 2015). The guidelines use a weighted pathogenic (very strong, strong, moderate, or supporting) and benign (stand-alone, strong, or supporting) criteria to classify variants as pathogenic, likely pathogenic, variant of uncertain significance (VUS), benign or likely benign. Evidence for the fulfillment of each criterion would be collected from various resources, such as publicly available disease variant databases (HGMD (Human Gene Mutation Database (http://www.hgmd.cf.ac.uk/ac/index.php)) and ClinVar) and population variant frequency databases (GnomAD), as well as literature for functional studies and gene/variant associated information (32).

### *5.4. Objective 3*

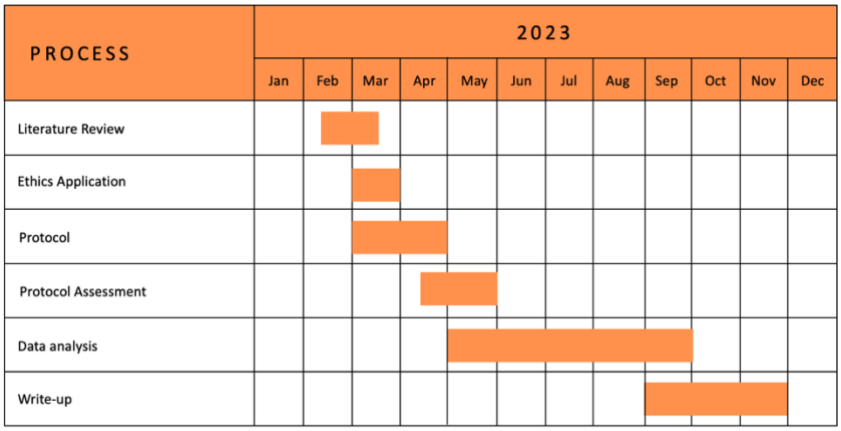
#### *5.4.1. Carrier frequency estimation*

Variant carrier frequencies will be estimated from the GnomAD database, and a combination of these frequencies will be used to estimate the overall disease carrier frequency. By using the principle of Hardy-Weinberg equilibrium, the disease prevalence will be estimated. A 95% confidence interval will be estimated, using the exact binomial method as described in Verlander et al. (1995) for both the disease carrier frequency and prevalence, to increase the level of statistical significance (57). The variant carrier frequencies from the cohort of this study can potentially be compared to reported frequencies found on GnomAD, because we know that frequency information available for the African population on this database is not representative of South Africans.

## 6. Ethics

An application for ethical clearance as a sub study has already been submitted to the University of the Witwatersrand Human Research Ethics Committee (Medical). Once approval has been granted, the diagnostic runs will be made available for use. Approval from the Head of Department for permission to conduct the study, and approval from the PI of the parent study for the use of the IDP dataset (ethics clearance number: **M210989**) has already been granted.

## 7. Timeline




## 8. Funding

No funding will be needed for this project, as data from an existing database will be analyzed.

## 9. Limitations

The main limitation of this study is the small sample size. There is a risk that variants may not be identified, and it might not be reflective of the real carrier frequencies in this population group. The cohort will, however, be combined with a larger sample from a previous study, which will increase the statistical significance of the study. Furthermore, if the prioritization method used is too stringent, along with limited information on variants, this can result in VUS classifications, meaning that once again there is a risk of not identifying pathogenic or likely pathogenic variants. HWE calculations for carrier frequencies are based on certain assumptions in a given population, which is rarely the case when studying data from a human population (58). Also, due to the IDP dataset consisting of information from exons only, any relevant causative variants found in deep intronic regions of the *CFTR* gene will be missed. In addition, scores generated by pathogenicity prediction tools only assess the probability of pathogenicity, and therefore a combination of different tools and different types of information will be used to prevent the inclusion of variants not likely to be pathogenic (30).

## 10. References

1. Riordan JR, Rommens JM, Kerem BS, Alon N, Rozmahel R, Grzelczak Z, et al. Identification of the Cystic Fibrosis Gene: Cloning and Characterization of Complementary DNA. Science (1979) [Internet]. 1989; Available from: www.sciencemag.org
2. Hyde S, Emsley P, Hartshorn M, Mimmack M, Gileadi U, Pearce S, et al. Structural model of ATP-binding proteins associated with cystic fibrosis, multidrug resistance and bacterial transport. Nature. 1990;346:362–5.

3. Bell SC, Mall MA, Gutierrez H, Macek M, Madge S, Davies JC, et al. The future of cystic fibrosis care: a global perspective. Lancet Respir Med. 2020 Jan 1;8(1):65–124.
4. Gentzsch M, Mall MA. Ion Channel Modulators in Cystic Fibrosis. Chest. 2018 Aug 1;154(2):383–93.
5. Ong T, Ramsey BW. Cystic Fibrosis: A Review. JAMA. 2023 Jun 6;329(21):1859–71.
6. Welsh MJ, Smith AE. Molecular Mechanisms of CFTR Chloride Channel Dysfunction in Cystic Fibrosis. Cell. 1993;73:1251–4.
7. Bareil C, Bergougnoux A. CFTR gene variants, epidemiology and molecular pathology. Archives de Pédiatrie [Internet]. 2020;27:8–12. Available from: www.sciencedirect.com
8. Des Georges M, Guittard C, Templin C, Altiéri JP, De Carvalho C, Ramsay M, et al. WGA allows the molecular characterization of a novel large CFTR rearrangement in a black South African cystic fibrosis patient. Journal of Molecular Diagnostics. 2008;10(6):544–8.
9. Zampoli M. Cystic fibrosis: What's new in South Africa in 2019. South African Medical Journal. 2019 Jan 1;109(1):16–9.
10. Krause A, Seymour H, Ramsay M. Common and Founder Mutations for Monogenic Traits in Sub-Saharan African Populations. Annu Rev Genomics Hum Genet [Internet]. 2018;19:149–75. Available from: https://doi.org/10.1146/annurev-genom-083117-
11. Schrijver I, Pique L, Graham S, Pearl M, Cherry A, Kharrazi M. The Spectrum of CFTR Variants in Nonwhite Cystic Fibrosis Patients: Implications for Molecular Diagnostic Testing. Journal of Molecular Diagnostics. 2016 Jan 1;18(1):39–50.
12. Van Rensburg J, Alessandrini M, Stewart C, Pepper MS. Cystic fibrosis in South Africa: A changing diagnostic paradigm. South African Medical Journal. 2018;108(8):624–8.
13. Goldman A, Graf C, Ramsay M. Molecular diagnosis of cystic fibrosis in South African populations. South African Medical Journal. 2003;93(7).
14. Masekela R, Zampoli M, Westwood AT, White DA, Green RJ, Olorunju S, et al. Phenotypic expression of the 3120+1G>A mutation in non-Caucasian children with cystic fibrosis in South Africa. Journal of Cystic Fibrosis. 2013 Jul;12(4):363–6.
15. Zampoli M, Verstraete J, Frauendorf M, Kassanjee R, Workman L, Morrow BM, et al. Cystic fibrosis in South Africa: spectrum of disease and determinants of outcome. ERJ Open Res. 2021 Jul 1;7(3).
16. Simon MA, Csanády L. Molecular pathology of the R117H cystic fibrosis mutation is explained by loss of a hydrogen bond. Structural Biology and Molecular Biophysics. 2021;1–19.
17. Massie R, Poplawski N, Wilcken B, Goldblatt J, Byrnes C, Robertson C, et al. Intron-8 polythymidine sequence in Australasian individuals with CF mutations R117H and R117C. Eur Respir J. 2001;17:1195–200.
18. Chu CS, Trapnell B, Curristin S, Cutting G, Crystal R. Genetic basis of variable exon 9 skipping in cystic fibrosis transmembrane conductance regulator mRNA. Nature. 1993;3:151–6.
19. Stewart C, Pepper MS. Cystic fibrosis on the African continent. Genetics in Medicine. 2016 Jul 1;18(7):653–62.
20. Rodrigues E, Melo M, Reis F, Penna F. Concentration of electrolytes in the sweat of malnourished children. Arch Dis Child. 1994;71:141–3.
21. Mutesa L, Bours V. Diagnostic challenges of cystic fibrosis in patients of African origin. J Trop Pediatr. 2009 Jul 22;55(5):281–6.
22. Mutesa L, Azad AK, Verhaeghe C, Segers K, Vanbellinghen JF, Ngendahayo L, et al. 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. 2009 May 1;135(5):1233–42.
23. Padoa C, Goldman A, Jenkins T, Ramsay M. Cystic fibrosis carrier frequencies in populations of African origin. J Med Genet. 1999;36:41–4.
24. Carles S, Desgeorges M, Goldman A, Thiart R, Guittard C, Kitazos CA, et al. First report of CFTR mutations in black cystic fibrosis patients of southern African origin. J Med Genet. 1996;33(9):802–4.
25. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, Mcgowan-Jordan J, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat. 2016 Jun 1;37(6):564–9.
26. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001 May;11(5):863–74.
27. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.
28. Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–5.
29. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019 Jan 24;176(3):535-548.e24.
30. Dashti MJS, Gamieldien J. A practical guide to filtering and prioritizing genetic variants. Biotechniques. 2017 Jan 1;62(1):18–30.
31. Seaby EG, Pengelly RJ, Ennis S. Exome sequencing explained: A practical guide to its clinical application. Brief Funct Genomics. 2016 Sep 1;15(5):374–84.
32. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine. 2015 May 8;17(5):405–24.
33. Liu F, Zhang Z, Csanády L, Gadsby DC, Chen J. Molecular Structure of the Human CFTR Ion Channel. Cell. 2017 Mar 23;169(1):85-95.e8.
34. Kerschner JL, Ghosh S, Paranjapye A, Cosme WR, Audrézet MP, Nakakuki M, et al. Screening for Regulatory Variants in 460 kb Encompassing the CFTR Locus in Cystic Fibrosis Patients. Journal of Molecular Diagnostics. 2019 Jan 1;21(1):70–80.
35. De Nooijer RA, Nobel JM, Arets HGM, Bot AG, van Berkhout FT, de Rijke YB, et al. Assessment of CFTR function in homozygous R117H-7T subjects. Journal of Cystic Fibrosis. 2011 Sep;10(5):326–32.
36. Farra C, Menassa R, Awwad J, Morel Y, Salameh P, Yazbeck N, et al. Mutational spectrum of cystic fibrosis in the Lebanese population. Journal of Cystic Fibrosis. 2010 Dec;9(6):406–10.
37. Shoshanl T, Augarten A, Yahav J, Gazlt E, Kerem B. Two novel mutations in the CFTR gene: W1089X in exon 17B and 4010delTATT in exon 21. Hum Mol Genet [Internet]. 1994;3(4):657–8. Available from: http://hmg.oxfordjournals.org/
38. Cheng C, Fei Z, Xiao P. Methods to improve the accuracy of next-generation sequencing. Vol. 11, Frontiers in Bioengineering and Biotechnology. Frontiers Media S.A.; 2023.
39. Deignan JL, Astbury C, Cutting GR, del Gaudio D, Gregg AR, Grody WW, et al. CFTR variant testing: a technical standard of the American College of Medical Genetics and Genomics (ACMG). Genetics in Medicine. 2020 Aug 1;22(8):1288–95.

40. Manga P, Kromberg JGR, Box NF, Sturm RA, Jenkins T, Ramsay M. Rufous Oculocutaneous Albinism in Southern African Blacks Is Caused by Mutations in the TYRPI Gene. Vol. 61, Am. J. Hum. Genet. 1997.
41. Kromberg JGR, Kerr R. Oculocutaneous albinism in southern Africa: Historical background, genetic, clinical and psychosocial issues. Afr J Disabil. 2022 Oct 14;11.
42. Cambraia A, Junior MC, Zembrzuski VM, Junqueira RM, Cabello PH, De Cabello GMK. Next-Generation Sequencing for Molecular Diagnosis of Cystic Fibrosis in a Brazilian Cohort. Dis Markers. 2021;2021.
43. Savant A, Lyman B, Bojanowski C, Upadia J. GeneReviews®. 2001. Cystic Fibrosis.
44. Donegà S, Rogalska ME, Pianigiani G, Igreja S, Amaral MD, Pagani F. Rescue of common exon-skipping mutations in cystic fibrosis with modified U1 snRNAs. Hum Mutat. 2020 Dec 1;41(12):2143–54.
45. Sharma N, Sosnay PR, Ramalho AS, Douville C, Franca A, Gottschalk LB, et al. Experimental Assessment of Splicing Variants Using Expression Minigenes and Comparison with In Silico Predictions. Hum Mutat. 2014 Oct 1;35(10):1249–59.
46. Sosnay PR, Siklosi KR, Van Goor F, Kaniecki K, Yu H, Sharma N, et al. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. Nat Genet. 2013 Oct;45(10):1160–7.
47. Yu YC, Sohma Y, Hwang TC. On the mechanism of gating defects caused by the R117H mutation in cystic fibrosis transmembrane conductance regulator. Journal of Physiology. 2016 Jun 15;594(12):3227–44.
48. Macek M, Mackova,' A, Hamosh A, Hilman BC, Selden RF, Lucottej G, et al. Identification of Common Cystic Fibrosis Mutations in African-Americans with Cystic Fibrosis Increases the Detection Rate to 75%. Am J Hum Genet. 1997;60:1122–7.
49. Cutting GR. Cystic fibrosis genetics: From molecular understanding to clinical application. Vol. 16, Nature Reviews Genetics. Nature Publishing Group; 2015. p. 45–56.
50. Rafeeq MM, Murad HAS. Cystic fibrosis: Current therapeutic targets and future approaches. J Transl Med. 2017 Apr 27;15(1):1–9.
51. Gillen AE, Harris A. Transcriptional regulation of CFTR gene expression. Vol. 4, Frontiers in Bioscience. 2012.
52. Ramalho AS, Clarke LA, Sousa M, Felicio V, Barreto C, Lopes C, et al. Comparative ex vivo, in vitro and in silico analyses of a CFTR splicing mutation: Importance of functional studies to establish disease liability of mutations. Journal of Cystic Fibrosis. 2016 Jan 1;15(1):21–33.
53. Fajac I, Girodon E. Genomically-guided therapies: A new era for cystic fibrosis [Internet]. Vol. 27, Archives de Pédiatrie. 2020. Available from: www.sciencedirect.com
54. Ratjen F, Döring G. Cystic fibrosis. In: Lancet. Elsevier B.V.; 2003. p. 681–9.
55. Mishra A, Greaves R, Massie J. The Relevance of Sweat Testing for the Diagnosis of Cystic Fibrosis in the Genomic Era. Vol. 26, Clin Biochem Rev. 2005.
56. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: Public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862–8.
57. Verlander PC, Kaporis A, Liu Q, Zhang Q, Seligsohn U, Auerbach AD. Carrier Frequency of the IVS4 + 4 A-T Mutation of the Fanconi Anemia Gene FAC in the Ashkenazi Jewish Population. 1995.
58. Hanany M, Allon G, Kimchi A, Blumenfeld A, Newman H, Pras E, et al. Carrier frequency analysis of mutations causing autosomal-recessive-inherited retinal diseases

in the Israeli population. European Journal of Human Genetics. 2018 Aug 1;26(8):1159–66.

# Appendix B:
# Human Research Ethics Committee (Medical), University of the Witwatersrand (M230693)

| UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG | HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL) |
|---|---|

**Office of the Deputy Vice-Chancellor (Research and Innovation)**

**TO:** Ms IM Smit
School of Pathology
Division of Human Genetics
Medical School
University

E-mail: 2721074@students.wits.ac.za

**CC:** Supervisor: Mlles F Essop and N Botha
<Fahmida.Essop@nhls.ac.za>
and <HREC-Medical Research Office@wits.ac.za>

**FROM:** Mr Iain Burns
Human Research Ethics Committee (Medical)
Tel: 011 717 1252

E-mail: Iain.Burns@wits.ac.za

**DATE:** 2023/07/28

**REF:** R14/49

**PROTOCOL NO:** **M230693** (This is your ethics application reference number. Please quote it in all enquiries, oral or written, relating to this study.)

**PROJECT TITLE:** *Cystic fibrosis: an update on the variant profile and carrier frequency in the Black South African population*

Please find attached the Clearance Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to Government funding of the University.



MSWorks2000/Iain0007/Clearscan.wps

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R49 Ms IM Smit

# HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
## CLEARANCE CERTIFICATE NO. M230693

| | |
|---|---|
| **NAME:** (Principal Investigator) | Ms IM Smit |
| **DEPARTMENT:** | School of Pathology<br>Division of Human Genetics<br>Medical School<br>University |
| **PROJECT TITLE:** | *Cystic fibrosis: an update on the variant profile and carrier frequency in the Black South African population* |
| **DATE CONSIDERED:** | Ad hoc |
| **DECISION:** | Approved unconditionally |
| **CONDITIONS:** | Sub-study under M21/09/89 |
| **NOTE:** | If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za> |
| **SUPERVISOR:** | Mlles F Essop and N Botha |
| **APPROVED BY:** | Dr M Vorster, Co-Chairperson, HREC (Medical) |
| **DATE OF APPROVAL:** | 2023/07/28 **EXPIRY DATE:** 2028/07/27 |

**This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.**

---

**DECLARATION OF INVESTIGATORS**

To be completed in duplicate and **ONE COPY** returned to the Research Office secretariat on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **June** and therefore reports and re-certification will be due in the month of **June** each year. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator _____ Date _____

# Appendix C:
# Plagiarism Declaration and Turnitin Report




**PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS**

SENATE PLAGIARISM POLICY: APPENDIX ONE

I _____ (Student number: _____)
am a student registered for the degree of _____ in the academic year _____.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature:_____ Date:_____

## Ingrid Smit Research Report.docx

ORIGINALITY REPORT

| 10% | 6% | 9% | 4% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | www.frontiersin.org<br>Internet Source | 1% |
| 2 | Khadijat Abubakar Bobbo, Umar Ahmad, De-Ming Chau, Norshariza Nordin, Syahril Abdullah. "A comprehensive review of cystic fibrosis in Africa and Asia", Saudi Journal of Biological Sciences, 2023<br>Publication | 1% |
| 3 | Saumya E. Samaraweera, Paul P. S. Wang, Ka Leung Li, Debora A. Casolari et al. "Childhood acute myeloid leukemia shows a high level of germline predisposition", Blood, 2021<br>Publication | 1% |
| 4 | "Abstracts from the 52nd European Society of Human Genetics (ESHG) Conference: Posters", European Journal of Human Genetics, 2019<br>Publication | 1% |
| 5 | Submitted to The University of Manchester<br>Student Paper | 1% |
| | jhir.library.jhu.edu | |

| | | |
|---|---|---|
| 6 | Internet Source | <1 % |
| 7 | iris.unige.it<br>Internet Source | <1 % |
| 8 | Teresa Sullivan, Eswary Thirthagiri, Chan-Eng Chong, Stacey Stauffer et al. " Epidemiological and ES cell-based functional evaluation of variants identified in families with breast cancer ", Human Mutation, 2020<br>Publication | <1 % |
| 9 | handwiki.org<br>Internet Source | <1 % |
| 10 | pure.rug.nl<br>Internet Source | <1 % |
| 11 | www.nshg.no<br>Internet Source | <1 % |
| 12 | Submitted to Imperial College of Science, Technology and Medicine<br>Student Paper | <1 % |
| 13 | Samantha N. Hartin, Waheeda A. Hossain, David Francis, David E. Godler, Sangjucta Barkataki, Merlin G. Butler. "Analysis of the Prader–Willi syndrome imprinting center using droplet digital PCR and next-generation whole-exome sequencing", Molecular Genetics & Genomic Medicine, 2019<br>Publication | <1 % |

| | | |
|---|---|---|
| 14 | Submitted to University of Witwatersrand<br>Student Paper | <1 % |
| 15 | rep.bioscientifica.com<br>Internet Source | <1 % |
| 16 | ro.uow.edu.au<br>Internet Source | <1 % |
| 17 | journals.lww.com<br>Internet Source | <1 % |
| 18 | Gabriela A Vasques, Mariana F A Funari, Frederico M Ferreira, Miriam Aza-Carmona et al. "IHH gene mutations causing short stature with non-specific skeletal abnormalities and response to growth hormone therapy", The Journal of Clinical Endocrinology & Metabolism, 2017<br>Publication | <1 % |
| 19 | Kyoung-Jin Park, Woochang Lee, Sail Chun, Won-Ki Min. "The Frequency of Discordant Variant Classification in the Human Gene Mutation Database: A Comparison of the American College of Medical Genetics and Genomics Guidelines and ClinVar", Laboratory Medicine, 2020<br>Publication | <1 % |
| 20 | www.mdpi.com<br>Internet Source | <1 % |

| | | |
|---|---|---|
| 21 | Submitted to University of Melbourne<br>Student Paper | <1 % |
| 22 | Submitted to University of Nottingham<br>Student Paper | <1 % |
| 23 | pesquisa.bvsalud.org<br>Internet Source | <1 % |
| 24 | Yuewu Tang, Yi Luo. "Identification of a novel mutation in complement receptor 2 in Chinese familial systemic lupus erythematosus", Archives of Rheumatology, 2022<br>Publication | <1 % |
| 25 | eprints.whiterose.ac.uk<br>Internet Source | <1 % |
| 26 | mdpi-res.com<br>Internet Source | <1 % |
| 27 | Submitted to Queensland University of Technology<br>Student Paper | <1 % |
| 28 | Anya T. Joynt, Erin W. Kavanagh, Gregory A. Newby, Shakela Mitchell et al. "Protospacer modification improves base editing of a canonical splice site variant and recovery of CFTR function in human airway epithelial cells", Molecular Therapy - Nucleic Acids, 2023<br>Publication | <1 % |

| | | |
|---|---|---|
| 29 | Guffanti, Federica. "DNA Repair and Response to Therapy: Exploring Their Relation in Epithelial Ovarian Cancer Models", Open University (United Kingdom), 2021<br>Publication | <1 % |
| 30 | Haseena Sait, Somya Srivastava, Manmohan Pandey, Deepak Ravichandran et al. "Neurodegeneration with brain iron accumulation: a case series highlighting phenotypic and genotypic diversity in 20 Indian families", neurogenetics, 2023<br>Publication | <1 % |
| 31 | Karen S. Raraigh, Kathleen C. Paul, Jennifer L. Goralski, Erin N. Worthington et al. "CFTR bearing variant p.Phe312del exhibits function inconsistent with phenotype and negligible response to ivacaftor", JCI Insight, 2022<br>Publication | <1 % |
| 32 | Mateja Smogavec, Maria Gerykova Bujalkova, Reinhard Lehner, Jürgen Neesen et al. "Singleton exome sequencing of 90 fetuses with ultrasound anomalies revealing novel disease-causing variants and genotype–phenotype correlations", European Journal of Human Genetics, 2022<br>Publication | <1 % |
| 33 | Zhang, Nana, Haijing Liu, Guanjun Yue, Yan Zhang, Jiangfeng You, and Hua Wang. | <1 % |

| | | |
|---|---|---|
| | "Molecular Heterogeneity of Ewing Sarcoma as Detected by Ion Torrent Sequencing", PLoS ONE, 2016.<br>Publication | |
| 34 | academic.oup.com<br>Internet Source | <1 % |
| 35 | assets.researchsquare.com<br>Internet Source | <1 % |
| 36 | core.ac.uk<br>Internet Source | <1 % |
| 37 | journals.plos.org<br>Internet Source | <1 % |
| 38 | www.medrxiv.org<br>Internet Source | <1 % |
| 39 | www.ncbi.nlm.nih.gov<br>Internet Source | <1 % |
| 40 | www.omicsdi.org<br>Internet Source | <1 % |
| 41 | www.researchsquare.com<br>Internet Source | <1 % |
| 42 | Abdelkader Heddar, Micheline Misrahi. "Genetics of primary ovarian insufficiency: a careful step-by-step approach based on solid foundations to bring new knowledge", Fertility and Sterility, 2022<br>Publication | <1 % |

| | | |
|---|---|---|
| 43 | Kuan-lin Huang, R. Jay Mashl, Yige Wu, Deborah I. Ritter et al. "Pathogenic Germline Variants in 10,389 Adult Cancers", Cell, 2018<br>Publication | <1 % |
| 44 | Sarah E. Brnich, Edgar A. Rivera-Muñoz, Jonathan S. Berg. "Quantifying the potential of functional evidence to reclassify variants of uncertain significance in the categorical and Bayesian interpretation frameworks", Human Mutation, 2018<br>Publication | <1 % |
| 45 | Elena Doménech, Gonzalo Gómez-López, Daniel Gzlez-Peña, Mar López et al. "New Mutations in Chronic Lymphocytic Leukemia Identified by Target Enrichment and Deep Sequencing", PLoS ONE, 2012<br>Publication | <1 % |
| 46 | Richards, Sue, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, and Heidi L. Rehm. "Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology", Genetics in Medicine, 2015.<br>Publication | <1 % |

# Appendix D:

# Journal Author Guidelines for Submissible Format

## Manuscript organization

*AJHG* manuscripts generally contain the following sections, in this order: title, author list, affiliations, correspondence email(s), additional footnotes (if any), abstract, main text (which, in articles, should be separated into introduction; material and methods [or subjects and methods]; results; and discussion); appendices (if any); acknowledgments; author contributions; declaration of interests; web resources (if any); references; figure titles and legends; and table titles and legends. All of these sections should be provided as one document. Figures, supplemental information, and the graphical abstract should be provided separately.

### Title

The title should convey the conceptual significance of the paper to a broad readership. Titles may occupy no more than three lines of type. Each line must contain no more than 54 characters, including spaces.

### Authors and affiliations

Author names should be spelled out rather than set in initials and should not include professional titles. Authors should be footnoted with numerical superscripts to their corresponding affiliations and shared authorship roles. Commas should precede numerical superscripts (e.g., John B. Smith,[1] Jane C. Doe,[2] etc.).

Affiliations should contain the following core information in this order: department(s) or subunit(s), institution, city, state or region, zip code or postal code, country. Please check author names carefully; we cannot amend or correct these sections after publication without publishing a formal correction.

### Footnotes in the author list

Footnotes are only allowed on page 1 of the text (and in tables). Footnotes may note a present address or may indicate equally contributing or senior authors. For more on designations of author contributions, please see the authorship section.

### Contact info

Corresponding authors should be noted with an asterisk in the author e-mail address(es) of the corresponding author(s) and should be listed after the author list footnotes, e.g., "*Correspondence:john_doe@cell.com."

## Abstract

The abstract consists of a single paragraph of 250 words or fewer. It should clearly convey the conceptual advance and significance of the work to a broad readership. In particular, the abstract should contain a brief background of the question, a description of the results without extensive experimental detail, and a brief summarization of the significance of the findings. References should not be cited in the abstract.

## Graphical abstract

Authors may choose to submit a graphical abstract, which is an image that summarizes the main findings of a paper. It adds a rich, visual component to the start of a paper, and helps readers to quickly appreciate and understand the central message. The image should be 1,200 pixels square at 300 dpi, and should use Arial font with a size of 12–16 points; smaller fonts will not be legible online. Please refer to our graphical-abstract guidelines for more details and recommendations.

## Introduction

The introduction should be succinct, with no subheadings, and should present the background information necessary to provide a biological context for the results.

## Material and methods (or subjects and methods)

The material and methods section needs to include sufficient detail so that readers can understand how the experiments were performed and so that all procedures can be repeated in conjunction with cited references. This section should also include a description of any statistical methods employed in the study.

## Results

The results section may be divided with subheadings.

## Discussion

The discussion should explain the significance of the results and place them into a broader context. It should not be redundant with the Results section. This section may contain subheadings.

## Appendices

Detailed results of statistical analyses may be presented as an appendix. Appendices may contain subheadings.

## Declaration of interests

This section is required for all papers. Please use it to disclose any competing interests, in accordance with Cell Press's "declaration of interests" policy. If there are no interests to declare, please note that with the following wording: "The authors declare no competing interests." The text in this section should match the text provided in the declaration of interests form.

## Acknowledgments

Use this section to acknowledge contributions from non-authors and to list funding sources. Because this section contains important information and many funding bodies require inclusion of grant numbers here, please check it carefully. There is a 250-word limit for the acknowledgments section. If you are not able to pare your acknowledgements section down to 250 words, you may include acknowledgements as supplemental information.

## Author contributions

Authors may include a concise description of each author's contributions, using initials to indicate each author's identity. We encourage you to use the CRediT taxonomy, but you can also use a traditional format (e.g., "A.B. and C.D. conducted the experiments; E.F. designed the experiments and wrote the paper").

## Web resources

Web-based resources (e.g., database, online computer program, etc.) may be listed along with their URLs in a separate section entitled "web resources," following the acknowledgments. Alternatively, URLs may be provided in the text where appropriate.

## Data and code availability

For publication, we require a "data and code availability" section that includes a statement describing the availability of new datasets and/or code associated with the paper. This includes any conditions for access to datasets and/or code not publicly available. This section should also include any accession numbers, DOIs or unique identifiers, or web links to deposited datasets. Examples of the types of appropriate "data and code availability statements" are below. Statements with multiple types of datasets may use a combination of statements.

- The [datasets/code] generated during this study are available at [name of repository] [accession code/web link].
- The published article includes all [datasets/code] generated or analyzed during this study.
- This study did not generate/analyze [datasets/code].
- There are restrictions to the availability of [dataset/code] due to [reason for restrictions].
- Original/source data for [figures/datatype] in the paper is available [e.g., Mendeley Data DOI].
- The [datasets/code] supporting the current study have not been deposited in a public repository because [reason data are not public] but are available from the corresponding author on request.

## References

References to journal articles should include only papers that are published or in press. We encourage the inclusion of DOIs in all journal article references. For references to in press articles, please confirm with the cited journal that the article is in fact accepted and in press and include a DOI number and scheduled online

publication date. Posted preprints may also be included in the references list with appropriate identification information and an independent persistent identifier such as a digital object identifier (DOI).

Unpublished data, submitted manuscripts, and personal communications should be cited within the text only and not included in the references list. Personal communication should be documented by a letter of permission. Submitted articles should be cited as unpublished data, data not shown, or personal communication.

All datasets, program code, and methods used in your manuscript must be appropriately cited in the text. For online material that does not have a DOI (e.g., software hosted on GitHub or on its own website), we ask you to include the URL either in the text or in the web resources section. If there is an original study that introduced the online resource, you may include that publication in the reference list in addition to the URL in the text or web resources section.

Reference citations in the text should be superscript numerals. The references section should list the numbered citations in the order in which they were cited in the text or tables. Please use the following styles shown below for references. Note that "et al." should only be used after ten authors.

***Article in a periodical***

1. Sondheimer. N. and Lindquist, S. (2000). Rnq1: An epigenetic modifier of protein function in yeast. Mol. Cell *5*, 163–172. 10.1016/S1097-2765(00)80412-8.

***Article on a preprint server or other repository***

De Virgilio, C., Hatakeyama, R., Péli-Gulli, M.-P., Hu, Z., Jaquenoud, M., Osuna, G.M.G., Sardu, A., and Dengjel, J. (2018). Spatially distinct pools of TORC1 balance protein homeostasis. Mendeley Data, 10.17632/m9s42s94fc.1.

***Article in a book***

King, S.M. (2003). Dynein motors: structure, mechanochemistry and regulation. In Molecular Motors, M. Schliwa, ed. (Wiley-VCH Verlag GmbH), pp. 45–78.

***An Entire book***

Cowan, W.M., Jessell, T.M., and Zipursky, S.L. (1997). Molecular and Cellular Approaches to Neural Development (Oxford University Press).

## Figure titles and legends

Each figure must be numbered consecutively with whole numbers. In other words, figures must be numbered as Figure 1, Figure 2, Figure 3, etc., rather than as Figure 1a, Figure 1b, Figure 1c, etc.

Figure titles and legends should consist of a brief title that describes the entire figure without citing specific panels and a subsequent description of each panel. Figure titles may not contain parenthetical information, reference citations, or footnotes.

All reference citations within a figure must also be included in the figure legend.

For any figures presenting pooled data, the measures should be defined in the figure legends (for example, data are represented as the mean ± SEM). Please also clearly outline the number of biological and technical replicates for each experiment.

### Tables

Include tables in the submitted manuscript after the figure titles and legends. Tables should not be saved as figures, i.e., as .jpg or .tif files. All tables intended for print should be incorporated into the end of the manuscript Word or LaTeX file. Tables should not be uploaded individually.

When creating a table, please use the Microsoft Word table function. Do not place an Excel table into a Word document. Tables not created with the Microsoft Word table function will be sent back for revision. Do not submit a table in PDF format.

- Word tables should not be tab or space delineated and should not include colored text or shading, but embedded graphics with color are OK.
- Do not use paragraph returns to separate data within a cell.
- Tables should include a title, and footnotes and/or legends should be concise.
- Table titles may not contain parenthetical information, reference citations, or footnote citations.
- Use superscripted lowercase letters (beginning with “a”) for footnotes in tables. Do not use numbers or symbols.
- Tables must be numbered as Table 1, Table 2, Table 3, etc., rather than as Table 1a, Table 1b, Table 1c, etc.
- If italic font is used within a table to indicate some feature of the data, an explanation of its meaning must be given in the table legend. Bold text may not be used in tables.
- If a referenced paper or study is mentioned within a table, it must be included in the References list and must be followed by its appropriate citation number (e.g., “Author et al.1”) within the table.
- All abbreviations within a table must be defined in the table legend or footnotes.

### Permissions

Please provide proof of permission to include any work cited as “personal communication.” This may be in the form of an e-mail communication, letter, or other appropriate form of permission.

For figures that have been reprinted from an outside source, please provide proof of permission for their use.

## Organization of the supplemental information

Supplemental files are restricted to (1) figures that cannot be rendered in print with enough detail to be informative, (2) tables that have too many columns and/or rows to fit across two printed pages, (3) clinical descriptions, tables, and figures that would substantially lengthen the print version of the manuscript, (4) movies, and (5) supplemental methods.

For full-length articles, supplemental material and methods should be limited to those materials or methods related to the supplemental display items, detailed analytical methods, and tabular presentations of primers or other information that would not fit well in the main text for formatting reasons.

Reports, however, may include a complete methods section. If included, such a section should present a comprehensive description of the methods, reagents, and statistics required to reproduce the experiments in the paper. We encourage authors to provide complete descriptions rather than referring to previous publications. Please note that the main paper also needs to give brief explanations of all the methods with enough detail to allow readers to understand the general experimental design and the results of the experiments.

Any tables, such as a list of primers, included in the supplemental material and methods should not be numbered.

Supplemental information should be provided with the original submission. Please follow the figure guidelines below for preparing figures. All figures and tables should have titles and legends.

Please provide a single PDF that contains all supplemental case reports, supplemental figures and legends, supplemental tables, and supplemental references (in this order). If a supplemental table cannot fit onto two 8.5" x 11" pages, please instead supply the table as an Excel file. **Please do NOT include the title or author list in the PDF**; we will add a coversheet with this information. **Please also do not include movie titles and legends**; leave those in the main text, and we will move them to a separate online page that links to the movies. In addition, **please do not include page numbers in your final PDF**. We strongly recommend that the final size of this PDF be less than 10 MB in order to ensure successful downloads for all readers. In addition, this PDF should be considered the FINAL version; it will be published as is, except for the coversheet that we will add. Scientific errors detected in the supplemental information after publication will require that a correction be published (as with errors in the main paper).

Please follow the following style preferences to ensure that your supplemental PDF is consistent with the copyedited version of your main text:

- Case reports should be titled "Supplemental note: Case reports." These descriptions should fully describe the phenotypes of the affected individuals and are not subject to a word limit.
- Figures should be titled Figure S1, Figure S2, etc. (NOT Supplemental Figure 1, Supplemental Figure 2, etc.)
- Similarly, tables should be titled Table S1, Table S2, etc. (NOT Supplemental Table 1, Supplemental Table 2, etc.)

- Please use the word “Supplemental” rather than “Supplementary” in headings (e.g., “Supplemental Material and Methods,” “Supplemental References,” etc.)

As with main-text figures, supplemental figures should include error bars where appropriate, and these error bars should be clearly defined in the figure legends

### Supplemental movies and Excel spreadsheets

Supplemental movies may be submitted through EM. Our preferred format is .mp4, but we also accept .mov, .avi, and .mpg files. Please note that we cannot accept movie files that require the reader to download particular codecs; the files must be playable on computers with standard media players, such as QuickTime or Windows Media Player. To create high-quality files with maximum compression, as well as ensure that your video can be played on our website and ScienceDirect's flash media player, the following specifications are strongly recommended:

- File size: <150 MB
- Frame rate: 30 frames per second
- Field order: none (progressive, not interlaced)
- Aspect ratio: widescreen 16:9
- Video codec: H.264
- Video bitrate: 2 Mbps
- Audio codec: AAC
- Audio bitrate: 128 kbps

If you choose the submission item "supplemental movies and spreadsheets," the PDF builder will embed links within the PDF where editors and reviewers will be able to download files. This also works for Excel files that do not display properly once converted to a PDF.

## Figure organization, format, and style

Digital figure files submitted through EM must conform to our digital figure guidelines, or authors will be asked to revise them.

If you have any questions about digital files, please contact *AJHG* senior production editor, Kerry Evans, at kevans@cell.com.