Genetic susceptibility to mycetoma in Sudan

Rayan Saif Eldin Yousif Ali

A thesis submitted in partial fulfilment of the requirements of the

The University of Brighton and the University of Sussex for a programme

of study undertaken at Brighton and Sussex Medical School for the

degree of Doctor of Philosophy

In collaboration with the

Mycetoma Research Centre, University of Khartoum

Khartoum, Sudan

February 2024

To the soul of my wonderful father.

Your love and legacy are my guiding lights. With gratitude and the hope that you're proud of your baby girl, I dedicate this thesis to you.

Acknowledgements

First and foremost, I extend my deepest gratitude to the Almighty God for granting me the strength to overcome the challenges and complete this endeavour. My heartfelt appreciation goes to the resilient mycetoma patients, whose strength in facing stigma, disabilities, and neglect was a powerful motivation throughout this research, driving me to contribute to improving their health outcomes.

I sincerely thank the NIHR Global Health Research Unit for Neglected Tropical Disease at Brighton & Sussex Medical School for their generous funding and invaluable opportunity to pursue my doctoral research.

A special acknowledgement is reserved for my supervisor, Prof. Melanie Newport, whose unwavering commitment to my academic growth has been a driving force. Her time, energy, and support surpassed all expectations, contributing not only to the development of this project but also to my personal growth. My sincere appreciation also extends to my second supervisor, Dr Sahar Bakhiet, who provided me with exceptional guidance and support that were vital to the completion of my research project.

I want to extend my deepest thanks to Prof. Muntaser Ibrahim for his belief in my capabilities and his kind mentorship throughout my PhD journey. I also thank Prof. Ahmed Hassan Fahal for his invaluable guidance and mentorship.

I am indebted to Prof. Heather Cordell for her indispensable advice and guidance during the data analyses of familial aggregation and population-based studies. Additional appreciation goes to Mrs Nahid Eid (Institute of Endemic Diseases, University of Khartoum) for her invaluable training on RT-qPCR, Dr Emmanuel Edwar Siddig (Mycetoma Research Centre, University of Khartoum) for his guidance on molecular identification of mycetoma causative organisms, and Mr Mahmoud Hussein and Mr Mohamed Abu Gesesa (Mycetoma Research Centre, University of Khartoum) for their dedicated work in sample collection. I am also grateful to Mr. Mohammad Maher (Siemens Healthineers, Egypt) for providing the 96-well PCR plates for the population-based study.

My sincere thanks go to my supportive friends and colleagues at the Institute of Endemic Diseases and the Mycetoma Research Centre. I offer special thanks to Dr

Rowa Hassan and Dr Mohammed Elshaikh, my fellow PhD companions, who began this journey with me and successfully concluded their own.

Finally, I would like to express my heartfelt gratitude to the people who have been my greatest source of support throughout the past few years. These include my mother, my husband (who has been my informal supervisor), my sweet baby Leen, my kind in-laws and my beloved sisters and brothers. Their unwavering support and prayers have given me the strength to keep going.

Abstract

Mycetoma is a chronic inflammatory infection that can lead to severe disability by damaging the skin, soft tissues, and bones if left untreated. It is one of the neglected tropical diseases (NTDs) caused by either bacteria (actinomycetoma) or fungi (eumycetoma). Sudan has one of the highest disease burdens globally, with eumycetoma responsible for 70% of infections.

Genetic factors likely influence susceptibility, evidenced by familial clustering. However, to date, genetic studies have focused on a pre-defined set of functional polymorphisms in 13 candidate genes selected for their involvement in immunity. All previous studies were among small sample-sized cohorts, and positive findings have not been reproduced. This thesis, therefore, employed a systematic approach beginning with estimating heritability and sibling recurrence risk ratio. Subsequently, two hypothesis-free genomic approaches were utilised: family-based whole-genome sequencing (WGS) and population-based genome-wide association studies (GWAS).

The family-based study involved 46 pedigrees, estimating mycetoma heritability at approximately 23%, with a higher likelihood of occurrence among siblings as indicated by a recurrence risk ratio ranging from 18.4 to 28.7. To identify rare genetic variants present in affected individuals but not in their healthy family members, 22 individuals from four families with multiple cases of mycetoma were selected for WGS. This analysis identified four candidate genes: *DEFA3, CASP8, PLEKHO2*, and *RAPGEF2*. Subsequently, network and pathway analyses emphasised the candidate genes' interconnectedness within crucial pathways such as the immune system and signal transduction. The population-based GWAS identified 15 suggestive associations, with rs16861260, rs319883, and rs10520048 associated with lower susceptibility, while in silico functional analyses of the identified genetic markers revealed pathways related to lymphocyte function and migration regulation. The genotyping data also showed a significant proportion of the disease's heritability attributed to common genetic variants, which accounted for nearly 80%. The high rate of consanguinity within the Sudanese population, as evidenced by a genomic inbreeding coefficient of 0.0317, highlighted the persistent impact of cultural practices on genetic diversity. Finally, The study emphasised the superiority of African ancestry reference panels for genotype imputation of Sudanese datasets.

This thesis marks the first-ever attempt to investigate the genetics of susceptibility to mycetoma systematically. It collectively advances our understanding of mycetoma's genetic basis, providing valuable insights for future investigations and potentially informing targeted interventions.

Table of Contents

List of Figures

List of Tables

Abbreviations

1KG	1000 Genomes
3' UTRs	3' Untranslated regions
ABO	ABO, Alpha 1-3-N-acetylgalactosaminyltransferase and Alpha 1-3-galactosyltransferase
ACMG	American College of Medical Genetics
AGR	African Genome Resource
AIRE	Autoimmune regulator
AK6	Adenylate kinase 6
ALG14	Asparagine-linked glycosylation 14 homolog
ALPS	Autoimmune lymphoproliferative syndrome
AMCase	Acidic mammalian chitinase
APCs	Antigen-presenting cells
APECED	Autoimmune Polyendocrinopathy Candidiasis Ectodermal Dystrophy
ARNTL2	Aryl hydrocarbon receptor nuclear translocator-like 2
ATP2B4	ATPase plasma membrane Ca2+ transporting 4
ATP5J2	ATP synthase membrane subunit F
BAM	Binary Alignment Map
BCL10	B Cell Lymphoma 10
BEGAIN	Brain-enriched guanylate kinase-associated
BMDMs	Bone Marrow-Derived Macrophages
BTLA	B- and T-lymphocyte attenuator
C1D	C1D nuclear receptor corepressor
C4BPB	Complement component 4 binding protein beta
CAAPA	Consortium on Asthma Among African Ancestry Populations in the Americas
CARD9	Caspase recruitment domain-containing protein 9
CASP8	Caspase 8
CBM	Chromoblastomycosis
CCL5	C-C chemokine ligand 5
CCND1	Cyclin D1
CD200	Cluster of differentiation 200
CD200R1	CD200 receptor 1

CF	Cystic fibrosis
CFTR	Cystic fibrosis transmembrane conductance regulator
CGD	Chronic granulomatous disease
CHIT1	Chitotriosidase
ClinGen	Clinical Genome Resource
CLRs	C-type lectin receptors
CMC	Chronic mucocutaneous candidiasis
CNN3	Calponin 3
CNVs	Copy number variants
COMT	Catechol-O-methyltransferase
COX19	Cytochrome C oxidase assembly factor COX19
CPGR	Centre for Proteomic and Genomic Research
CPSF4	Cleavage and polyadenylation specific factor 4
CR1	Complement receptor 1
CSV	Comma Separated Value
CXCL10	C-X-C motif ligand 10
CXCL8	Interleukin 8
CXCR2	C-X-C motif chemokine receptor 2
CYP17	Cytochrome P450 subfamily 17
CYP19	Cytochrome P450 subfamily 19
CYP1B1	Cytochrome P450 subfamily 1B1
CYP3A5	Cytochrome P450 family 3 subfamily A member 5
DAG	Diacylglycerol
DCs	Dendritic Cells
DEFA1	Defensin alpha 1
DEFA3	Defensin alpha 3
DOMINO	Detection Of Mode of Inheritance from Non-heritable Observations
DPH6-DT	DPH6 divergent transcript
DPYD	Dihydropyrimidine dehydrogenase
EBV	Epstein-Barr Virus
ELISA	Enzyme-linked immunosorbent assay
EOS	Eosinophils
eQTL	Expression Quantitative Trait Loci

EVD	Ebola Virus Disease
FNAC	Fine-needle aspiration cytology
FOXP1	Forkhead Box P1
FUMA	Functional Mapping and Annotation of Genome-Wide Association Studies
GCSAM	Germinal centre-associated signalling and motility
GCTA	Genome-Wide Complex Trait Analysis
GDP	Gross domestic product
GDP	Guanosine diphosphate
GEF	Guanine nucleotide exchange factor
GMM	Generalised mixed model
GTEx	Genotype-Tissue Expression
GTP	Guanosine triphosphate
GWAS	Genome-wide association studies
Hb	Haemoglobin
HBB	Haemoglobin subunit beta
HbS	Haemoglobin S
HICs	High-income countries
HLA	Human leukocyte antigen
HNPs	Human neutrophil peptides
HRC	Haplotype Reference Consortium
HSCT	Hematopoietic Stem Cell Transplantation
HSD3B	Hydroxysteroid dehydrogenase 3B
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
IDAT	Intensity Data Files
iDCs	Immature dendritic cells
IFN	Interferon
Ig	Immunoglobulin
IL	Interleukin
IL12Rβ1	Interleukin-12 receptor Beta 1
InDels	Insertion/Deletions
Ins(1,4,5)P3	Inositol-(1,4,5)-trisphosphate

IOM DTM	International Organization For Migration Displacement Tracking Matrix
IQS	Imputation quality score
ITS	Internal transcribed spacer
KCNJ6	Potassium inwardly-rectifying channel subfamily J member 6
KCNS3	Potassium voltage-gated channel modifier subfamily S member 3
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LD	Linkage disequilibrium
LFA-1	Lymphocyte function-associated antigen-1
LMICs	Low- and middle-income countries
LMM	Linear mixed model
lncRNAs	Long non-coding RNAs
LOF	Loss of function
LSU	rDNA partial large subunit
MAF	Minor allele frequency
MAGMA	Multi-Marker Analysis Of Genomic Annotation
Malaria GEN	Malaria Genomic Epidemiology Network
MALT1	Mucosa-associated lymphoid tissue 1
MBL	Mannose-binding lectin
MCP-1	Macrophage chemoattractant protein-1
M-CSF	Macrophage-colony stimulating factor
MDS	Multidimensional scaling
MIR	MicroRNA
MIS	Michigan Imputation Server
ML	Maximum likelihood
MLMA	Mixed Linear Model Association Analysis
MM9-2	Metalloproteinases-2
MMP-9	Metalloproteinases-9
MRC	Mycetoma Research Centre
MSigDB	Molecular Signatures Database
Mtb	Mycobacterium tuberculosis
Mφs	Macrophages
ncRNA	Noncoding-RNA
NDR	Non-reference discordance rate

NGS	Next generation sequencing
NK	Natural killer
NOD	Non-Obese Diabetic
NOS2	Nitric oxide synthase 2
NTDs	Neglected Tropical Diseases
OMIM	Online Mendelian Inheritance In Man
OR	Odds ratio
PAMPs	Pathogen-associated molecular patterns
PBMC	Peripheral blood mononuclear cells
PCR	Polymerase chain reaction
PID	Primary immunodeficiency
PKC	Protein kinase C
PKD	Protein Kinase D
PLC-γ	Phospholipase C- γ
PLEKHO2	Pleckstrin homology domain-containing family O member 2
PLG	Plasminogen
PPFIBP1	PTPRF interacting protein, binding protein 1
PPP3R1	Protein phosphatase 3 regulatory subunit B, alpha
PRRs	Pathogen recognition receptors
PTBP2	Polypyrimidine tract-binding protein 2
PTCD1	Pentatricopeptide repeat domain 1
PtdIns(4,5)P2	Phosphatidylinositol-(4,5)-bisphosphate
PTX3	Pentraxin 3
qPCR	Quantitative polymerase chain reaction
QQ	Quantile-Quantile
r2	Imputation r-squared
RAPGEF2	Rap guanine nucleotide exchange factor 2
RASGRP1	RAS guanyl-releasing protein 1
RIPK1	Receptor interaction protein kinase 1
RMSE	Root mean squared error
RNA-seq	RNA sequencing
RNU6-274P	U6 small nuclear 274
ROH	Runs of homozygosity

ROS	Reactive oxygen species
RR	Relative Risk
RWDD3	RWD domain-containing protein 3
SAGE	Statistical Analysis of Genetic Epidemiology
SAIGE	Scalable and Accurate Implementation of Generalised mixed model
SIDT1	SID1 transmembrane family member 1
SMCO2	Single-pass membrane protein with coiled-coil domains 2
SNPs	Single nucleotide polymorphisms
SNX7	Sorting nexin 7
SOLAR	Sequential Oligogenic Linkage Analysis Routines
SPRED2	Sprouty-related EVH1 domain-containing protein 2
STAT6	Signal transducer and activator of transcription 6
STK38L	Serine/Threonine kinase 38 like
SUN1	Sad1 and UNC84 domain containing 1
SVs	Structural variants
TAF9	TATA-box binding protein associated factor 9
TB	Tuberculosis
T-cells	T-lymphocytes
TCR	T-Cell Receptor
Th	T-helper
THBS4	Thrombospondin4
TIMP-1	Tissue inhibitor of metalloproteinases
TIS	TOPMed Imputation Server
TLRs	Toll-like receptors
TNF-α	Tumour necrosis factor-A
TOPMed	Trans-Omics for Precision Medicine
UCF3	Upstream transcription factor family member 3
VCF	Variant Call Format
VEP	Variant effect predictor
VUS	Variants of uncertain significance
WES	Whole exome sequencing
WGS	Whole genome sequencing
WHO	World health organization

ZSCAN25	Zinc finger and Scan domain containing 25

Author’s declaration

I declare that the research contained in this thesis, unless otherwise formally indicated within the text, is the author's original work. The thesis has not been previously submitted to these or any other university for a degree and does not incorporate any material already submitted for a degree.

Signature: RSYA Date: 4 December 2023

Thesis layout

The thesis is structured into chapters, each with a background or overview. Chapter One offers an extensive background on mycetoma pathogenesis, epidemiology, immunological response, risk factors, diagnosis, and treatment. It further investigates the Sudanese population structure and outlines the methodologies for identifying susceptibility genes related to infectious diseases. Chapter Two provides an in-depth review of genetic susceptibility to fungal infections, including mycetoma (the corresponding published review article is included in Appendix 1). Chapter Three details the systematic methods used for investigating mycetoma host genetics. The outcomes of this investigation are presented in chapters Four and Five. Chapter Four reveals the results of the family-based data analyses, encompassing familial aggregation estimates and whole genome sequencing. Chapter Five showcases population-based association studies aimed at identifying common genetic markers associated with mycetoma and explores the genetic characteristics of the Sudanese population. The final chapter is a comprehensive discussion summarising all findings, addressing their impact and implications, and outlining potential future directions.

CHAPTER 1 - Unveiling host genetic predisposition to mycetoma in Sudan: An Introduction

1.1 Mycetoma

1.1.1 Background

Neglected tropical diseases (NTDs) are a group of conditions strongly associated with impoverished populations in tropical and subtropical regions (1). The concept of NTDs was formulated by the World Health Organization (WHO) with a list of 20 diseases affecting more than one billion people worldwide (2). Mycetoma, an inflammatory granulomatous skin and underlying tissue infection, was recognised as one of the NTDs in May 2016 (3). This disease is characterised by a triad of painless tumour-like subcutaneous swellings, draining sinuses, and compact aggregates of the causative organism known as grains in the pus or tissues (***Figure 1.1***) (4). If left untreated, pathogens have the potential to spread through both muscular and lymphatic pathways into muscles and bones, resulting in potentially life-threatening consequences such as deformity, disability, amputation, and, in some instances, death (5).

Aetiologically, this disease has two forms, eumycetoma and actinomycetoma, caused by fungi and bacteria, respectively (4). Globally, mycetoma is endemic in tropical and subtropical regions in the so-called “mycetoma belt” between the latitudes of 15°S and 30°N (6). This belt includes Sudan, Senegal, Yemen, India, Mexico, and Venezuela (6,7). Like other NTDs, mycetoma is heavily concentrated in low- and middle-income countries where a lack of resources and adequate healthcare infrastructure contributes to poor outcomes (8). It is considered an occupational disease of individuals who work in rural areas, such as farmers and herders. Thus, the foot and hand account for over 80% of the sites affected by mycetoma (9,10).

Currently, diagnosis of mycetoma depends on imaging, cytological, molecular, histopathological, serological and grain culture techniques (11–13). Accurate diagnosis is vital for mycetoma treatment since actinomycetoma patients are treated with courses of antibiotics (e.g. amikacin and co-trimoxazole) (14,15) while the best treatment for eumycetoma is antifungal agents (from the azole group) along with surgery (15,16). There is no vaccine.

Figure 1.1 *Mycetoma caused by Madurella mycetomatis. The three hallmarks of mycetoma are draining sinuses, painless subcutaneous tissue swelling, and grains in the pus or tissues. The picture originates from the Mycetoma Research Centre, Sudan.*

1.1.2 Mycetoma epidemiology

Eumycetoma is most commonly reported in tropical and subtropical climate zones of Africa and Asia, with the fungus *Madurella mycetomatis* being the most prevalent causative agent (12,17). In contrast, actinomycetoma is more endemic in Central and South America, with *Nocardia brasiliensis* as the primary etiological organism (10,13).

Among countries endemic with mycetoma, Sudan, India, and Mexico have the highest reported cases (10,17). Sudan had the highest number of reported cases per year, with an estimated prevalence of 1.81 cases per 100,000 inhabitants (**Figure 1.2**) in a meta-analysis done by van de Sande and colleagues in 2013 (17). Hassan and associates recently conducted an epidemiological survey covering 60 villages in Eastern Sennar Locality, Sennar State, one of Sudan's highly endemic states (18). The study reported an overall prevalence of 0.87% in the study area, with the age group 31-54 years having the highest prevalence. Although the predominance of mycetoma in males has been known for a long time, this study found a slightly higher disease prevalence of 0.92% in females compared to 0.83% in males (18). The authors proposed that the estimate above can be deemed more precise than the data gathered from previous hospital-based research, as females tend to delay seeking medical assistance and consequently present themselves at a later stage of the disease. A study conducted in West Bengal, India, between 1981 and 2000 analysed 264 cases of mycetoma, focusing on epidemiological aspects (19). The study revealed

a diverse spectrum of causative organisms, with actinomycetes being the most common (74.6%), and the majority of cases occurred in males and the age group of 21-30 years. In 2019, a hospital-based study investigated the epidemiological profile and spectrum of thirteen eumycetoma cases in Delhi, North India (20). The investigators found that eumycetoma primarily affected males, with a peak incidence in the age group of 21-40. Various fungal species were identified as causative agents, with *M. mycetomatis* being the most prevalent. In Mexico, one study extensively reviewed 3933 mycetoma cases to determine the epidemiological characteristics of mycetoma in the Mexican population (21). That study highlighted the predominance of mycetoma cases in males (75.6%) and the age group of 16-50. Actinomycetoma was substantially the most common form, with a frequency of 96.52%. No comprehensive prospective research has been conducted to determine the disease's true global incidence, prevalence, or burden.

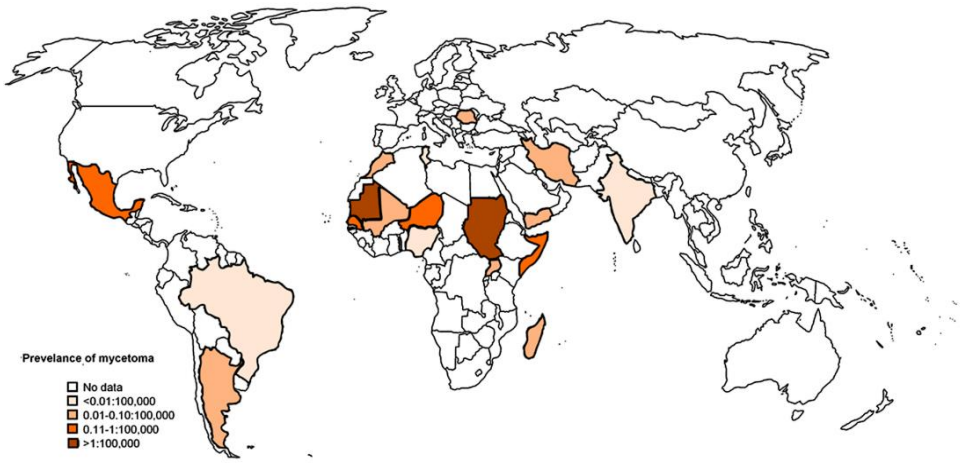


Figure 1.2 *Global prevalence of mycetoma. The average prevalence of mycetoma is estimated by dividing the number of a country's reported cases per year by that country's total population in the same year. Adapted from van de Sande* (17).

1.1.3 Mode of transmission

One of the basic unanswered research questions regarding mycetoma relates to the disease's transmission mode. With more than 70 different microorganisms (bacteria and fungi) known to cause mycetoma, 70% of infections in Sudan are attributable to *M. mycetomatis* (7,15). Presently, three theories explaining the transmission route are suggested; the most favourable one suggests traumatic inoculation of the causative organism into the skin and subcutaneous tissue through contaminated Acacia tree

thorns. This theory was supported by the presence of *M. mycetomatis* DNA on Acacia thorns or in soil samples (22). Samy and his associates also indicated an Acacia-mycetoma association using ecological niche modelling that linked Acacia geographic distribution to mycetoma case distribution (23). The second theory suggests cattle dung contaminated with *M. mycetomatis* as an adjuvant of inoculation. To support this theory, phylogenetic analysis was performed to predict the natural habitat of *Madurella* species using neighbouring taxa with previously known habitats (24). This was done with 89 strains belonging to 4 species of the genus *Madurella* and representative genera from the family of *Chaetomiaceae* using DNA sequence data from the rRNA partial large subunit (*LSU*) gene and the internal transcribed spacer (ITS) region. *Madurella* species were nested within the *Chaetomiaceae* family, often found in animal dung and soil. This theory is further supported by a) the abundance of cattle among other animals in villages endemic with mycetoma in Sudan, b) the inhabitants' houses are usually made of mud mixed with animal dung, and c) the majority of animals' sheds in endemic regions are located in the grounds of the inhabitants' houses (8). The third theory proposes the presence of transmission vectors after screening tick samples collected from domestic animals raised within households of inhabitants of two endemic villages in Eastern Sennar Locality (25). The researchers used the fungal barcode ITS and *M. mycetomatis*-specific PCR primers, successfully detecting the fungus in one sample. It must be noted that the finding does not offer definitive proof of ticks being the sole agent responsible for transmission.

1.1.4 Immunological response to mycetoma

Despite the presence of 70 different causative organisms of mycetoma, research on the host immune response has been limited, targeting only a few of these organisms (7,26). Following the inoculation of the causative agent into the subcutaneous tissue and grain formation, innate immunity plays a significant role in early response via neutrophil infiltration, as demonstrated by previous studies (27,28). The adherence and degranulation of those neutrophils represent the first zone of tissue reaction (type 1) against mycetoma, resulting in grain disintegration. The second zone, representing type 2 reactions, consists mainly of macrophages, which engulf the grain and neutrophil debris, while zone 3 (type 3 reactions) contains lymphocytes and plasma cells in addition to giant cells, as demonstrated in **Figure 1.3** (28).

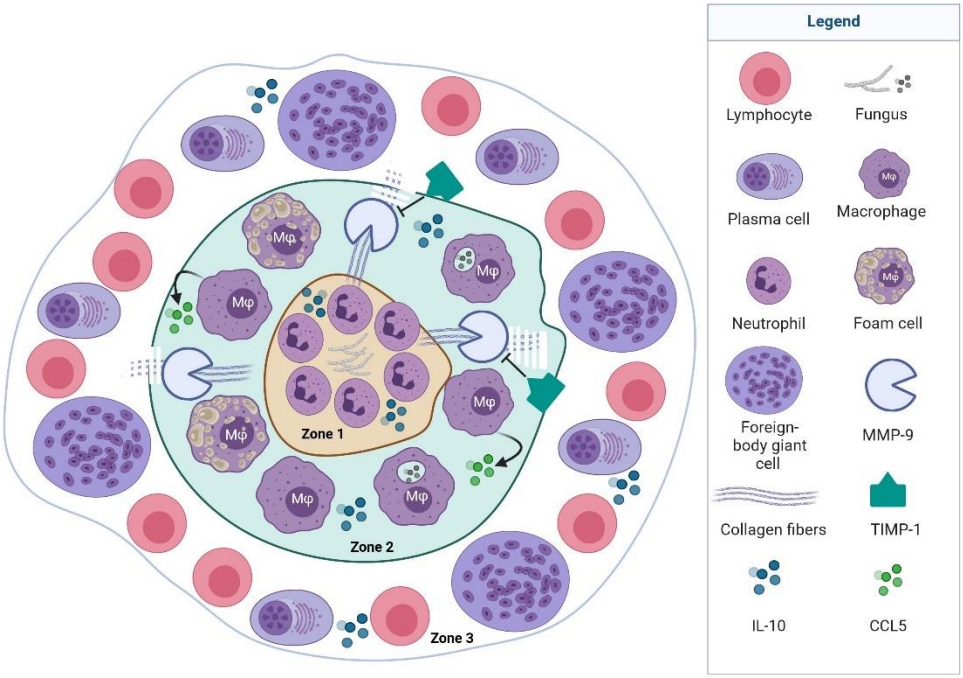


Figure 1.3 *Mycetoma granuloma. In host tissue, the fungal grain is surrounded by three zones of inflammatory cells through which different cytokines, chemokines and enzymes are dispersed. Zone 1 consists of neutrophils, zone 2 contains macrophages, while in zone 3 reside giant cells, lymphocytes, and plasma cells. The grain is encapsulated by collagen fibres located in zone 2. Tissue remodelling around the grains involves the presence of matrix metalloproteinases-9 (MMP-9) and its inhibitor, tissue inhibitor of metalloproteinases-1 (TIMP-1) (both present in zone 2) in addition to IL-17A (zones 1 and 2) which enhances the expression of MMP-9. IL-10 is present in all three zones, whereas C-C chemokine ligand 5 (CCL5), expressed by macrophages, is found in zone 2. Figure created using BioRender.com.*

On the other hand, cell-mediated immunity is also required for an effective response to mycetoma, with T-lymphocytes playing a central role in defence. Genetic defects in underlying cell-mediated immunity were suggested by Mahgoub and associates in 1977 to be linked with the pathogenesis of eumycetoma (29). Protective immunity against mycetoma is conferred by T-helper (Th) type 1 lymphocytes, while disease progression is associated with Th2 immunity as previously demonstrated by the significantly elevated levels of Th2 cytokines, including interleukin (IL)-4, IL-5, IL-6 and IL-10 in the sera or stained tissue sections taken from lesions of mycetoma patients (30–32). In addition, it has been observed that patients with eumycetoma who underwent surgical treatment exhibited markedly elevated levels of Th1 cytokines, including interferon (IFN)-γ, tumour necrosis factor-α (*TNF-α*) and IL-1β, as compared

to those who received medical treatment (31). This suggests that surgical intervention may trigger a more robust immune response, potentially improving outcomes in patients with eumycetoma. In light of the linkage between IL-1β and eumycetoma's pathogenesis, the research team conducted further investigations encompassing a cohort of 70 eumycetoma patients and 70 matched healthy controls (33). The objective was to elucidate the roles played by IL-1 family members (IL-1β and IL-37) and IL-12 family members (IL-12 and IL-35) in the sera of the study cohort. The study revealed a positive correlation between the levels of immunosuppressive cytokines, IL-35 and IL-37, and the size of the lesions and disease duration. This was in contrast to the serum levels of pro-inflammatory cytokines, IL-1β and IL-12, which were observed to significantly decrease as the size of the lesions and disease duration increased (33).

In a more recent study, Siddig and colleagues demonstrated the expression of IL-17A in the granuloma induced by different mycetoma causative agents, indicating the importance of Th17 in immunity against mycetoma infection (34). The importance of IL-17 was also indicated in defence against *Aspergillus fumigatus* and *Fusarium oxysporum,* which are both causative agents of mycetoma (17,35), by Taylor and associates in murine models of fungal keratitis (36). In cooperation, IL-17-producing neutrophils and Th17 cells conferred protective immunity against fungal hyphae and regulated the severity of the corneal disease.

Regarding the humoral arm of acquired immunity against mycetoma, Wethered and his colleagues used an enzyme-linked immunosorbent assay (ELISA) to quantify immunoglobulin levels in sera collected from *M. mycetomatis* eumycetoma cases and matched controls (37). High levels of immunoglobulin (Ig)M were observed in most patients, whereas low levels of specific IgG were found in some patients. Moreover, IgM, IgG and complement factors were detected on the surface of the grains collected from actinomycetoma lesions caused by *Streptomyces somaliensis* (30).

Co-infection with schistosomiasis was found to be correlated with susceptibility to mycetoma in a small case-control study of 84 subjects in Sudan (38). In that study, the authors hypothesised that the prolonged Th2 response induced by schistosomiasis, which elevated the anti-inflammatory cytokine IL-10 expression, predisposed individuals to develop mycetoma.

1.1.5 Risk factors

Risk factors for mycetoma can be categorised into several key areas, including socioeconomic, host, environmental, and behavioural factors.

1.1.5.1 Socioeconomic factors

Individuals living in mycetoma endemic areas often work as farmers, shepherds, and labourers, which puts them at risk of exposure to the causative organisms of mycetoma. Based on a recent household survey in Sennar State, Sudan, it has been determined that mycetoma is prevalent among individuals involved in agricultural activities (8). Sudan practices traditional rain-fed and modern irrigated farming, but impoverished rural and isolated communities engaged in manual agriculture are the most affected by this condition. These individuals are at a heightened risk of injury while working without proper protective footwear in the field (**Figure 1.4**).

Figure 1.4 *Snapshots of life in mycetoma endemic areas. This figure depicts the absence of infrastructure, ubiquitous presence of animals and their dung, the harsh acacia thorns underfoot, individuals navigating the terrain without proper footwear, and the overall challenging environment emblematic of poverty prevalent in these communities. Adapted from Bakhiet et al. (39), Fahal et al. (9), and Fahal and Bakhiet (40).*

Furthermore, the survey found that 64.6% of those afflicted with mycetoma are involved in subsistence farming. Prior research has indicated that certain occupations, including animal grazing and lumberjacking, may contribute to developing mycetoma (41,42). In regions where mycetoma is prevalent, various domestic animals pose a significant risk, including cattle, goats, sheep, and donkeys. Rural residents often construct animal enclosures using thorny tree branches, exposing themselves to the risk of thorn pricks during construction and maintenance. This is a crucial risk factor for mycetoma, as noted in a recent population-based case-control study of sociodemographic risk factors for mycetoma (43).

Poor hygiene practices resulting from underdeveloped infrastructure to provide clean water and sanitation and lack of awareness and education increase the risk of mycetoma. Additionally, healthcare facilities are frequently situated at considerable distances and are challenging to reach for those residing in regions where mycetoma is prevalent. These individuals usually have limited financial resources and must embark on long journeys to obtain medical assistance. Furthermore, due to religious beliefs, many individuals opt for traditional and herbal remedies instead of seeking medical treatment, which may increase the chances of secondary bacterial infections and late presentation (44,45).

1.1.5.2 Host risk factors

Understanding host risk factors is vital for identifying individuals at a heightened risk of mycetoma and implementing preventive measures, especially in endemic regions. An individual's immune status is determined by a combination of immunological factors, including leukocytes, antibodies, cytokines, nutrients, and genetic factors, collectively shaping their ability to defend against infections. Certain conditions, such as HIV/AIDS or immunosuppressive therapy, can weaken the immune system, increasing susceptibility to mycetoma. Furthermore, differences in immune responses between individuals can also impact susceptibility, with some people exhibiting less effective immune responses against mycetoma pathogens. Many such inter-individual differences in immune responses to infection are genetically regulated. This has been well-demonstrated for many infectious diseases both in animal models and humans (46–48). For example, research has revealed various genetic polymorphisms that influence an individual's vulnerability or resistance to malaria (49), while variations in

human leukocyte antigen (HLA) genes can impact the progression of HIV (50). The contribution of genetic factors to mycetoma susceptibility remains a topic of ongoing research, with evidence suggesting that genetics may indeed play a role (51). Evidence includes previous studies detecting the presence of antibodies against mycetoma causative agents in unaffected residents of endemic areas, suggesting exposure to the causative organisms (52–54). Additionally, familial clustering of mycetoma patients (9,55–57) and the broader observation of host genetic factors determining susceptibility to other fungal subcutaneous infections (58–60) further support the role of genetics in mycetoma susceptibility. Lastly, several studies have examined the potential association, all between candidate genes involved in the host immune response and genetic predisposition to the disease (55,61–65). The studies were generally small and inconclusive or have not been replicated, and none of them took a genome-wide approach. A thorough review of the immunogenetics of mycetoma was undertaken as the first phase of my PhD research, and my findings are presented in the next chapter.

Nutrient consumption supports various bodily functions, including cell growth, metabolism, and immune response to pathogens. Both macro and micronutrients are essential for the proper functioning of the cells that help fight infections. Inadequate intake of proteins or fats can negatively impact collagen synthesis, angiogenesis, and the production of cytokines and chemokines that fight infections. Similarly, deficiency in vitamins such as vitamin A or minerals such as zinc can impair the functions of B and T cells as well as natural killer cells, making it challenging for the immune system to defend against infections (66–68). Malnutrition can further weaken the immune system in areas where mycetoma is endemic, making individuals more susceptible to infections caused by bacteria and fungi (69).

1.1.5.3 Environmental factors

Mycetoma is a disease prevalent in tropical and subtropical areas with particular weather patterns. These regions usually have low annual rainfall, a dry winter, a brief yet intense rainy season, and high daytime temperatures (45-60°C) all year round. These extreme environmental factors are thought to create ideal conditions for the survival of microorganisms that cause mycetoma (70). Through analysing soil samples from highly affected areas in Khartoum State, Sudan, researchers discovered *M.*

mycetomatis DNA fragments identical to those in mycetoma patients (71). This fungus is responsible for approximately 70% of cases in Sudan and is commonly found in soil, suggesting it may be the primary source of infection. Researchers recently conducted a study utilising a metabarcoding technology-based soil sampling survey to identify fungal species in soil samples collected from seven villages in Sennar state (72). Metabarcoding is a high-throughput DNA sequencing technique that identifies and quantifies microorganisms such as bacteria, fungi, archaea, and protists in an environmental sample (73). This technique amplifies and sequences specific DNA markers conserved across a group of organisms (e.g., the 16S ribosomal RNA gene for bacteria or the ITS region for fungi. The study focused on the ITS region and identified twelve eumycetoma causative organisms, including *M. mycetomatis*, in 83% of the collected soil samples.

The Eastern Sennar Locality survey reported the majority of mycetoma infections (83%) occurred during the rainy seasons (summer and autumn), suggesting a preference by the causative organisms for wet soil (8). This survey also indicated that 70% of farms in endemic areas were surrounded by trees, of which 90% of species had thorns. These trees with thorns serve multiple purposes for the locals, including building enclosures for livestock (39), using them as a primary source of cooking fuel (8), and even as fences for their houses (9). According to another study by Hassan and associates, mycetoma is more prevalent in arid areas near water sources, soil with low calcium and sodium levels, and certain species of thorny trees (74).

1.1.5.4 Behavioural Factors

Failing to wear protective footwear in areas where mycetoma is expected, mainly when walking or working outside, raises the likelihood of injury and the subsequent development of mycetoma infection (41). Additionally, improper wound care and hygiene practices after an injury may result in secondary infections that can progress to mycetoma.

1.1.6 Diagnosis of mycetoma

Since numerous agents can cause mycetoma, identifying the causative organism is essential to differentiate between the disease's two forms and develop an appropriate treatment plan. A combination of techniques is required to reach an accurate diagnosis of mycetoma. The recommended protocol in endemic areas with limited resources

suggests starting with fine-needle aspiration cytology (FNAC). If the FNAC result is negative, it is recommended to perform a grain culture and histopathological examination with a sample obtained from a deep surgical biopsy (**Figure 1.5**). Molecular identification of the causative organisms by polymerase chain reaction (PCR) and whole genome sequencing (WGS) are the most accurate techniques. Still, both are considered expensive in regions where mycetoma is endemic (75). This raises the need for a reliable, fast, and cheap diagnostic tool for mycetoma since all existing techniques are invasive, time-consuming and inaccessible in endemic areas (11,51,76).

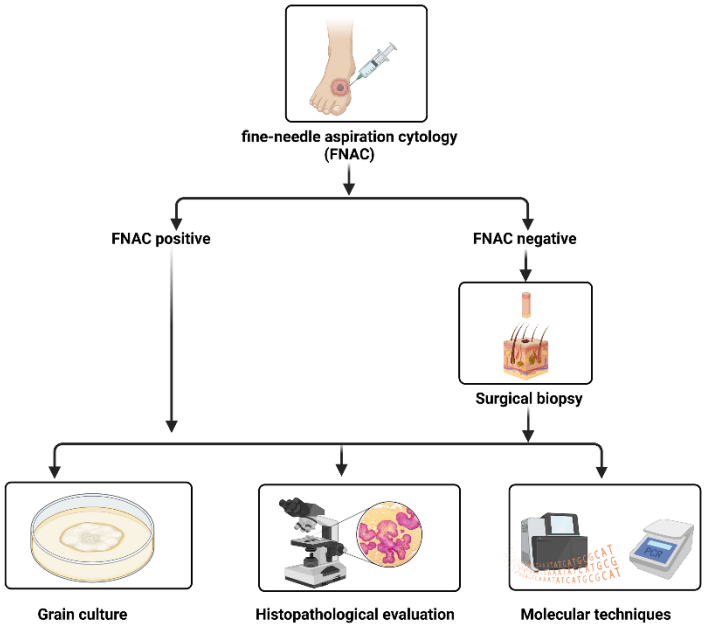


Figure 1.5 *The recommended protocol for diagnosis of mycetoma. Adapted and modified from Ahmed AA (75). Figure created using BioRender.com.*

1.1.7 Mycetoma treatment

Regarding treatment, the regimen used to treat actinomycetoma (combination therapy of amikacin sulphate and trimethoprim-sulfamethoxazole) has a relatively high cure rate of about 90% (14). On the other hand, the current eumycetoma treatment (azole therapy, typically itraconazole, before and after surgical excision) has a low cure rate of 25–35% at best and frequent recurrence, which could eventually lead to limb amputation (9,77). The long duration of treatment is suboptimal, especially for patients with a late diagnosis, which is the case for most mycetoma patients (15,78). A vicious cycle leads to poor disease management; the patient presents late owing to fear of

amputation, which is the conclusive choice of therapy for late-presenting patients. Another reason for late presentation is that nearly 42% of mycetoma patients first seek herbal and traditional treatment (44,45), which is relatively cheaper compared to an estimated average cost of treatment of 567 USD per year (8).

In 2017, a double-blind, randomised clinical trial was conducted at the Mycetoma Research Centre (MRC), a WHO-collaborative centre in Sudan, to evaluate fosravuconazole for eumycetoma treatment. This drug was initially developed for onychomycosis with more favourable pharmacokinetic properties when compared to itraconazole and low toxicity (79). A recently published results of the clinical trial indicate that fosravuconazole and itraconazole have comparable efficacy rates (80). Fosravuconazole demonstrated rates of 65% and 85% in the 300-milligram and 200-milligram arms, respectively, while itraconazole exhibited 80% in the 400-milligram arm, which exceeded expectations. Although the differences were not statistically significant, fosravuconazole presents significant advantages over itraconazole. Notably, it only necessitates one pill per week, whereas itraconazole requires four tablets daily. This substantial reduction in pill burden, from 120 to 4 tablets per month, enhances patient adherence to the treatment regimen. Furthermore, fosravuconazole's once-a-week dosing, lack of food effect, and minimal drug interactions render it a promising alternative treatment for mycetoma (80). Concurrently, there are several attempts to repurpose existing drugs for mycetoma treatment.

1.2 Sudanese population structure

1.2.1 Background on Sudan

Sudan, a nation that straddles Sub-Saharan Africa and the Middle East, boasts a vast land area of approximately 2.5 million square kilometres, equating to nearly one million square miles. Due to its extensive land area, it ranks as the third-largest country in Africa (81). The country shares its borders with seven neighbouring countries: Egypt to the north, Eritrea and Ethiopia to the east, South Sudan to the south, the Central African Republic to the southwest, Chad to the west, and Libya to the northwest. Moreover, the country is bounded by the Red Sea to the northeast (82). Sudan's topography is diverse, encompassing desert and mountain regions. The Nile is Sudan's most prominent geographical feature and primary water source (83). It is

formed by two major tributaries: the White Nile, which originates from the Great Lakes region in central Africa, and the Blue Nile, which starts in the Ethiopian Highlands. The two tributaries converge at the capital city, Khartoum, where the river is then named the Nile and flows northward into Egypt.

Sudan experiences a tropical climate with varying temperatures depending on the region and season. Temperatures in the arid northern region can exceed 45 degrees Celsius, and rainfall is negligible. In the central region, semi-arid conditions persist with high temperatures and low humidity. The southern border, adjacent to South Sudan, experiences a semi-humid climate. The Red Sea coast has a different climate from the rest of the country, associated with the formation of Lake Nubia behind the Aswan High Dam near the border with Egypt (84). The country is divided into 18 states and 189 localities, each with smaller administrative units. North Darfur is the largest state, but despite being the smallest province in the country, Khartoum has the largest population of 7.38 million (82,85).

Sudan has an estimated population of 48.29 million, but the distribution is not evenly spread across the vast territory (86). Most of the population resides in the central and northern regions, particularly along the Nile River and its tributaries. The urban and rural populations coexist in Sudan, with cities like Khartoum and Port Sudan boasting modern infrastructure, including schools, hospitals, and transportation networks, resulting in higher population densities (86). Conversely, rural areas tend to be more traditional and rely on agriculture, pastoralism, and traditional crafts for livelihoods (87). Nomadic and semi-nomadic communities in the Western regions, particularly in Darfur and Kordofan, lead a mobile lifestyle and often follow traditional migration routes in search of grazing land and water sources for their livestock (88). Disparities in healthcare access and educational opportunities exist between urban and rural areas (89). Access to quality healthcare and education can vary significantly based on geographic location and socioeconomic status (90,91).

Sudan's economy has experienced significant transformations in recent years. Before the discovery of oil in 1999, agriculture was the primary source of income for the country since its independence in 1956. However, after the commencement of commercial oil exports, Sudan's economy became heavily dependent on oil revenues. From 2000 to 2011, oil revenues contributed an average of 20% and 95% to the

country's Gross Domestic Product (GDP) and export earnings, respectively. Nonetheless, when South Sudan seceded in 2011, the country lost 76% of its oil revenues, resulting in economic challenges such as negative growth, high inflation, budget deficits, and exchange rate instability (92,93). As per the World Bank's report for 2023, the country's GDP has been declining since 2018, with an annual average decrease of -2.3% from 2018 to 2022 (94). During the same timeframe, per capita GDP has also witnessed an average decline of 5.4% yearly. Since South Sudan's secession, the per capita GDP has continuously declined, averaging 3.0% yearly (94). The primary reliance on agriculture accounted for around 40% of total GDP and employed approximately 70% of the population (95). Sudan's economy heavily relies on animal husbandry, which has seen a boost in production due to improved veterinary treatment, lenient credit policies, and higher market prices. This sector supports approximately 17% of the population's livelihoods. Livestock products from Sudan met the country's domestic meat demands in 2015, with enough excess for exports, which comprised around 25% of total exports.

As per the official estimates based on the 2014-2015 National Budget and Household Survey, about 36.1% of the population had per capita spending below the national poverty line (96). The current estimate is on the rise, primarily due to the heightened instability and insecurity brought about by the ongoing power conflict between the Sudanese armed forces and the Rapid Support Forces militia, which commenced in mid-April 2023. The conflict has spilt over to various parts of the country, reigniting hostilities in traditionally volatile regions such as Darfur and Kordofan. Additionally, Sudan’s already dire economic situation is worsening further, making economic and financial stabilisation prospects unlikely in the foreseeable period (97).

1.2.2 Ethnolinguistic groups

Africa is the birthplace of anatomically modern humans, who originated there within the past 300,000 years (98). The origins of our human lineage can be traced back to modern-day Sudan over a million years ago (99). Interestingly, this predates the estimated emergence of *Homo sapiens* by roughly 700,000 years (99). This continent is where humans began migrating out of around 60,000 to 80,000 years ago (100–102). Africa is known for its diverse range of ethnic and linguistic groups, with over 2,000 ethnolinguistic groups representing around one-third of the world’s languages

(103). Within this vibrant tapestry, four major ethnolinguistic groups stand out, each with unique characteristics and historical significance. Afro-Asiatic is one of the largest language families in Africa, with 382 languages spoken by agropastoralist and agriculturalist populations in North Africa and parts of the Horn of Africa (104). Arabic, a prominent Afro-Asiatic language, is spoken across the Arab world and serves as the liturgical language of Islam. Other Afro-Asiatic languages include Amharic in Ethiopia and Hausa in West Africa (103). The Niger-Kordofanian (interchangeably named Niger-Congo) language family is the largest in Africa (1554 languages) and is spoken by a vast number of people in Sub-Saharan Africa residing across West, Central, and Southern Africa (105). The diversity within Niger-Congo encompasses tonal languages like Yoruba, the Bantu languages spoken across much of Sub-Saharan Africa, and many more (103). Nilo-Saharan includes 210 languages predominantly spoken by pastoralists in North and East Africa, often in regions around the Nile River and its tributaries (106). Dinka-Nuer in South Sudan, Kanuri in Nigeria and Chad, and Zaghawa in Sudan are examples of Nilo-Saharan languages (103). The linguistic diversity within the group mentioned above reflects the vast array of landscapes it inhabits. This familial unit is present across extensive regions situated east and north of Lake Victoria in East Africa while extending as far west as the Niger Valley in Mali, West Africa (106). Khoisan, a unique and relatively small ethnolinguistic group, is characterised by click consonants and is found primarily in Southern Africa (107). The Khoisan peoples are known for their hunter-gatherer lifestyles and distinct click languages (108).

Sudan is home to 19 major ethnic groups and over 597 ethnic subgroups communicating in more than 100 languages and dialects (109). The predominant ethnic group in the area is comprised of Arabic-speaking Muslims, constituting around 70% of the entire populace. Their daily lives are greatly shaped by the significance of Islam, which plays a pivotal role in their cultural, legal, and social practices (86,109). Arabic is the official language used in urban settings and for official purposes. The languages spoken in Sudan belong to three major African language families: Afro-Asiatic, Niger-Kordofanian, and Nilo-Saharan. Some putative language isolates are also found in the South Kordofan region (103).

1.2.3 Genetic population structure

The connection between ethnolinguistic groups and population genetic structure is complex, encompassing the impact of culture, language, and genetics. Despite Africa's crucial role in the emergence and evolution of anatomically modern humans, there is a conspicuous absence of African representation in human genomic research, databases, and reference panels (110). This deficiency impedes our understanding of human ancestry and the evolution of complex traits and related diseases. Nevertheless, this underrepresentation contradicts that African genetic diversity is substantial, with compelling evidence of biological, environmental, and climate-related adaptations within African genomes (102). Several studies using Africans' WGS data reported millions of previously undescribed variants identified in underrepresented ethnolinguistic groups, particularly in North and East African populations (110–112). Therefore, there is a pressing need for a full repertoire of African genomic variation to represent better the distribution of variants in both the African diaspora and global populations (112). Furthermore, it has been observed that some genetic variants considered potentially pathogenic by ClinVar, a public database collecting information on genetic variants and their clinical significance in human health and disease (74), may be benign and occur frequently in specific African populations (110). As such, it is essential to incorporate diverse ethnic groups in genetic studies to avoid relying solely on rarity as a criterion for evaluating a variant's potential pathogenicity in clinical studies.

The linguistic diversity in Sudan mirrors genetic diversity, as people within the same linguistic group share common genetic markers due to their historical and ancestral connections (113). A recent study explored the fine-scale genetic structure and historical admixture (the blending of genetic traits from different ethnic or regional groups due to interbreeding, migration, or other forms of gene flow) within African populations (114). The study scrutinised data from 1387 individuals, predominantly hailing from Cameroon, the Republic of the Congo, Ghana, Nigeria, and Sudan, to relate genetic data to various questions, particularly those grounded in anthropology and archaeology. Notably, among Sudanese populations along the Nile, belonging to Arabic and Nubian ethnic groups, genetic variation displayed minimal correspondence with ethnicity, indicating the Nile's role in promoting intermixing among these Sudanese groups (115). In contrast, the Nuba mountains in southern Sudan exhibited

fine-scale genetic structure correlating with ethnolinguistic groups, suggesting the mountains' role as a historical refuge and differentiation between Niger-Congo (Kordofanian) and Nilo-Saharan speakers (114). The study also identified multiple waves of admixture in Sudanese populations, particularly related to Arabian-like sources. For instance, the Beni-Amer, a coastal Beja ethnic group, exhibited more significant inferred genetic variation related to Ethiopian Afro-Asiatic groups, possibly stemming from interactions during the period of the first millennium in Common Era (CE) when the Kingdom of Aksum extended across northern Ethiopia, coastal Sudan, and Yemen. Additionally, two Sudanese clusters, primarily containing Nile inhabitants, showed admixture events with sources related to present-day Arabians and East African Nilo-Saharan speakers, confirming a previous study (116) that justified the findings as a reflection of the historical collapse of the Kingdom of Makuria, which allowed Arabic groups to expand down the Nile into Sudan (117).

Marriage is a prevalent custom in Sudan, with a vast majority of both genders (95%) tying the knot by the time they reach 50 years of age (118). However, the age at which marriages occur tends to vary depending on whether they are in a rural or urban setting. Typically, women in rural areas get married at 19, whereas their urban counterparts tend to wait until they are 24 years old. Conversely, men usually get married at the median age of 27 in rural areas and 32 in urban areas (118). The Sudanese population has one of the highest rates of consanguineous marriages globally, ranging from 40-49% (119). This rate has remained consistently above 40%, indicating its deep-rooted presence in our population (120,121). One of the earliest studies on the general population in Khartoum state, dating back to 1988, found a consanguinity rate of 52% using data from 4833 marriages (120). The most common form of consanguinity in that study was first-cousin unions, constituting approximately half of all consanguineous marriages with an inbreeding coefficient of 0.03. A recent national household survey collected data from 2272 participants between December 2022 and March 2023 (121). The survey revealed that a significant proportion of respondents (42.8%) were married to consanguineal partners with various levels of relatedness. Additionally, 32.1% of single respondents planned to marry close relatives, again indicating the persistence of consanguineous marriages. Religious and social factors can explain the high percentage of first-cousin marriages. More than 90% of the Sudanese population follows Islam (122), and the marriage regulations in

Islam allow for first-cousin and double first-cousin marriages (123). However, the Quran prohibits uncle-niece or aunt-nephew marriages. That being said, there is no encouragement of consanguinity within Islam, but the strong preference for first-cousin marriage in most Muslim countries reflects perceived socioeconomic advantages (123). These advantages include more accessible marriage arrangements, greater compatibility with future in-laws, more secure marital relationships, and maintaining family holdings' integrity (124).

Consanguinity was reported as a significant risk factor for human susceptibility to infectious diseases such as tuberculosis and hepatitis (125). Moreover, the link between consanguinity and genetic disorders, especially autosomal recessive conditions, has been documented in numerous studies conducted in Sudan, including haemoglobinopathies, sexual development and neuropsychiatric disorders (126–133). The high prevalence of the haemoglobin S (HbS) mutation in the human haemoglobin (Hb) subunit beta among the Bagara (nomadic pastoralists of Arab descent) of western Sudan highlights the influence of culture on disease burden and the accumulation of deleterious mutations. The frequency is among the highest in Sudan and Africa, apparently due to consanguinity (126,127). Their migration and admixture within present-day Sudan from the region around Lake Chad are estimated to have occurred over the last 4,000 years (126). It is speculated that they introduced the HbS mutation, with its unique variety of typical and atypical haplotypes, to the local Nilo-Saharan populations who rarely practice cultures of internal marriages (within-tribe and/or consanguineous marriages), resulting in a much lower frequency of carriers and disease (102). Furthermore, Daak and associates reported a high carrier (HbAS) percentage in western Kordofan State (127), which they linked to the heterozygote genotype advantage against fatal malaria (134,135). Another study investigated the ethnic differences concerning visceral leishmaniasis in two villages in eastern Sudan (136). This study employed genome-wide linkage scans of multicase families to reveal associations with markers on chromosomes 1 and 6. Additionally, the study uncovered a concealed population structure and village-specific genetic lineage, which were attributed to a founder effect and consanguinity (136).

It is imperative to acknowledge the significant influence of the ongoing turmoil and political unrest on Sudan's population structure, particularly in demographics and distribution. Countless individuals have lost their lives or suffered injuries, and millions

have been forced to flee their homes within Sudan and beyond its borders. Over 5.1 million people have been internally and externally displaced as of mid-April 2023. According to the International Organization for Migration Displacement Tracking Matrix (IOM DTM), as of September 5th 2023, more than 4.1 million people have been internally displaced, which is twice that before the conflict began (137). People have fled from eight states and have sought shelter in 3,733 locations across all 18 states. Additionally, the UN Refugee Agency (UNHCR) has reported that as of September 6th, over one million people have crossed into neighbouring countries, including the Central African Republic, Chad, Egypt, Ethiopia, and South Sudan (138).

1.3 Approaches to identifying infectious diseases susceptibility genes

Several sources of evidence confirm the role of host genetic variations in susceptibility and response to infectious diseases (46,48). Various complementary approaches have been adopted to identify genetic variation predisposing to infectious diseases. I explored how these approaches can help in studying mycetoma host genetics. Some examples will be discussed in the following sections.

1.3.1 Twin and adoption studies

Several clues signifying the role of host genetics in susceptibility to infections have come from twin and adoptee studies. Both study types are used to untangle a particular trait's genetic and environmental influences. Twin studies are based on comparing concordance for particular infectious diseases between monozygous (identical) and dizygous (non-identical) twins. Monozygous twins share 100% of their genetic makeup while dizygous twins only share 50%; therefore, higher concordance in monozygous twins than dizygous twins suggests a genetic component to a given trait. This type of study was used to pinpoint the role of host genetics in predisposition to tuberculosis, leprosy, and hepatitis B (139–141). Adoption studies, alternately, examine individuals adopted early in life by unrelated parents. Perhaps one of the most prominent adoption studies was in 1988, when the researchers reported a relative risk of death due to infection of 5.81 in adoptees if one of their biological parents prematurely died due to an infection (142). The following approaches typically use family studies and the former two types.

1.3.2 Familial aggregation studies

Conventional approaches based on familial aggregation of a given trait include heritability and sibling recurrence risk (K_s). Heritability is the proportion of phenotypic variance attributed to genetics (143). Narrow-sense heritability (h^2) is linked to the additive genetic component and denotes the degree to which genes transmitted by parents determine the offspring phenotype. Broad sense heritability (H^2) is the ratio of total genetic variance, including dominance and gene-gene interaction, divided by the total phenotypic variance (143). Meanwhile, sibling recurrence risk refers to the probability of developing the disease in a sibling of an affected individual (144). Both estimates were successfully used to confirm the contribution of the genetic component in another NTD named podoconiosis (145). The study was based on 59 pedigrees (family trees) collected from an endemic region in Ethiopia and found a relatively high heritability of 63% (SE 0.069, $P = 1 \times 10^{-7}$) and a sibling recurrence risk of 27.7% (over one-quarter of siblings of probands were themselves affected). That study formed the basis for further investigations that identified an association between HLA class II loci variants and podoconiosis using genome-wide association studies (GWAS) (146,147).

The utilisation of consanguineous families can prove to be a valuable tool in identifying single genes that play a role in immunity towards complex infections phenotypes. For instance, although mycobacterial infections, including tuberculosis (TB), are complex diseases influenced by a combination of genetic, environmental, and socio-economic factors, certain severe paediatric forms are caused by single-gene mutations (148). A study conducted in areas where consanguineous marriages are prevalent, found that two children with severe TB from unrelated consanguineous families from Iran and Turkey had two different homozygous loss-of-function (LOF) mutations in the interleukin-12 receptor beta 1 (*IL12Rβ1*) gene (149). Both patients had complete IL-12Rβ1 deficiency, which impaired the IL-12 signalling pathway, a crucial component in host immunity against tuberculosis infection (150).

To date, no studies have been undertaken to estimate familial aggregation, investigate the disease's mode of inheritance or confirm the genetic component of susceptibility towards mycetoma based on the observed familial clustering (mentioned in section 1.1.5.2).

1.3.3 Candidate gene approach

Until the early 2000s, this was the most common approach to looking for an association between complex diseases, including infections and candidate genes (151). Using case-control studies, this approach focuses on genotyping polymorphisms in biologically plausible candidate genes. All the previous studies on genetic susceptibility to mycetoma followed this approach as described in the following chapter. The drawbacks of this approach are a) their reliance on a predefined hypothesis based on often incomplete biological knowledge, b) small sample sizes that result in inadequate study power, and c) poor degree of replication (151,152). Coupling candidate gene studies with hypothesis-free GWAS has identified large-effect variants conferring susceptibility to many infectious diseases (153).

1.3.4 Genome-Wide Association Studies

To overcome the limitations of the candidate gene approach, GWAS was developed to identify shared common genetic variation with local correlation patterns known as linkage disequilibrium (LD) (101). The term LD refers to the non-random coinheritance of alleles at different genomic regions (loci) in a population due to their physical proximity on a chromosome and lack of recombination over many generations (154). GWAS utilised this LD phenomenon along with the human genome sequence, the International HapMap project and the developments in microarray-based high-throughput genotyping technology to enable genotyping up to millions of polymorphisms across the genome, with no prior assumption about the location of causal variants (155).

This approach identifies variants with a possible association with the disease trait by using large, well-characterised cohorts of cases and controls to compare allele frequencies of genetic variants between these two groups. Thus, this requires large sample sizes to get sufficient statistical power to detect true associations after correction for multiple testing. The statistical significance threshold of a GWAS requires a P value of $< 5\times10^{-8}$ (156).

The genotyping arrays used for GWAS can capture only a fraction of all genetic variants, estimated to reach up to 10 million common variants (157). Genotype imputation is a statistical method utilised to overcome this problem via estimation of the missing genotypes using reference panels (158) such as the 1000 Genomes

Project, the Haplotype Reference Consortium (HRC), and Trans-Omics for Precision Medicine (TOPMed) reference panels (159–161). Another drawback of GWAS is that they only identify loci likely to harbour disease-causing genes without pinpointing the causal variant. Fine mapping solves this drawback by determining the causal variants by estimating the probability of each variant being causal in a correlated set compared to others in the same set (162).

Most large-scale GWAS have been conducted in European populations; thus, the outcomes may not apply to African populations suffering the most from infectious diseases (163). Additionally, African populations are genetically highly diverse and have shorter regions of LD (157). This led to the creation of genotyping arrays such as the H3Africa consortium array to provide genome-wide denser coverage for African populations (164). Malaria is an excellent example of how the selection of population-specific SNPs that are relevant and polymorphic in African populations, in addition to enhancing the density of genotyping arrays, has increased the power to detect causal variants. In 2009, Jallow and associates used the Affymetrix 500K GeneChip to carry out a GWAS in 1,060 children with severe malaria and 1,500 controls (165). The investigators failed to detect genome-wide association signals corresponding to previously reported malaria associations, particularly the locus that contains the HbS causal variant in the haemoglobin subunit beta (*HBB)* gene, which confers a malaria-protective effect (134,135). However, they managed to increase the genome-wide association signal of the HbS locus from $P = 3.9 \times 10^{-7}$ to 4.5×10^{-14} by imputation using a reference panel of 62 randomly selected Gambian controls. This study highlighted the need for denser genotyping arrays to overcome the low LD present in African populations. Ten years later, the malaria genomic epidemiology network (Malaria GEN) combined GWAS data of 17000 individuals from Africa, Asia and Oceania, genotyped using the Illumina Omni 2.5 M platform, to confirm loci previously associated with resistance to severe malaria such as *HBB*, ABO, alpha 1-3-N-acetylgalactosaminyltransferase and alpha 1-3-galactosyltransferase (*ABO)*, ATPase plasma membrane Ca2+ transporting 4 (*ATP2B4)* and the glycophorin region on chromosome 4 in addition to a new locus on chromosome 6 (166).

1.3.5 Next generation sequencing

In the past few years, advances in DNA sequencing technologies have revolutionised biomedical research and facilitated the discovery of genetic variations associated with complex diseases. Next generation sequencing (NGS) processes allow massively parallel sequencing of biological macromolecules such as nucleic acids and proteins. The main advantage of sequencing over genotyping is the ability to identify any variant in a genome rather than using a pre-specified few million common variants on a genotyping array (167). The power of NGS has broadened the scale of many applications, including the identification of genetic variation, new genomes or transcriptomes assembly, quantitative RNA-sequencing analysis, and epigenetic change detection (168). In the following sections, I will review some of these advanced applications and how they were utilised to unravel the complexities of the genetics of some infectious diseases.

1.3.5.1 Whole-genome and whole-exome sequencing

Whole genome sequencing (WGS) is the process in which the entire genomic DNA from a single sample is digested into small fragments that are then sequenced simultaneously (massively parallel sequencing) (168). This process generates millions of short reads (~150 base pairs) that are aligned to a reference genome for detecting potential genetic variation sites. Whole exome sequencing (WES) is the selective capture and sequencing of the protein-coding (exonic) regions of the genome. These exonic regions constitute approximately 1-2 % of the human genome, yet it is estimated that 85% of known variations causing disease-related traits are found within these regions (169). This is because the variations are typically clinically relevant and potentially actionable (i.e., likely pathogenic or pathogenic variants in genes associated with diseases that could lead to altering the medical management such as treatment or surveillance) compared to those that occur in non-coding regions (170). Compared to WGS, WES is more cost-effective and provides better coverage and higher power to detect low-frequency and rare disease-causing variants by deep sequencing all known exons (171,172). However, WGS provides much more comprehensive information on genes by sequencing the noncoding regions, which capture 98%–99% of the genome and include promoters and enhancers, among other regulatory elements. Genetic variants identified by WGS include single nucleotide polymorphisms (SNPs), insertions/deletions (indels), structural variants (duplications,

deletions, and inversions), and copy number variants (CNVs). In contrast, WES can only detect variants most likely to affect protein structure, such as nonsense, start-loss, stop-loss, splice-site variants, and frameshift indels (173).

The success of sequencing in studying monogenic diseases is reflected, for example, in a study done on chronic granulomatous disease (CGD) which is a primary immunodeficiency (PID) resulting from defects in the NADPH oxidase subunits (*CYBB, CYBA, NCF-1, NCF-2* and *NCF-4*) that are essential for the generation of reactive oxygen species (ROS) within phagocytes (174). The defective CGD phagocytes increase susceptibility to severe recurrent fungal infections, including aspergillosis (in one-third of all CGD cases) (174) and candidiasis (175). WGS was recently employed to identify a homozygous LOF variant in the *CYBC1* gene in two brothers diagnosed with CGD, which led to the discovery of a novel function for CYBC1 as a chaperone for one of the subunits of the NADPH oxidase (176).

1.3.5.2 Dual RNA sequencing

Similar to how GWAS was preferred over candidate-gene studies based on hypothesis-free versus hypothesis-based approaches, RNA sequencing (RNA-seq) has become the first choice in transcriptomic analysis, replacing hypothesis-driven microarrays. RNA-seq, or whole transcriptome sequencing, involves sequencing the entire RNA population (reverse-transcribed into cDNA) in a sample (177). In addition to studying the expression levels of the transcripts similar to microarrays, RNA-seq can provide information about novel transcripts and offers higher sensitivity and the ability to detect non-coding RNAs (such as microRNAs) (177,178).

Dual RNA-seq is an emerging technology that allows simultaneous sequencing of the transcriptomes of both host and pathogens in vitro. Subsequent isolation of each transcriptome is conducted in silico by aligning the sequencing reads to the respective genomes of the organisms (179). This technique allows the identification of the molecular mechanisms underlying host-pathogen interaction at the site of infection. A recent study of dual transcriptomes on leprosy patients' skin lesions helped the researchers to identify a link between *Mycobacterium leprae* heat shock proteins and the host humoral immunity (180). This approach can be utilised to identify critical determinants of molecular pathogenesis for several poorly understood diseases, such as mycetoma.

1.3.5.3 Multi-omics analysis

Another approach to reveal host-pathogen interactions is multi-omics, where different omics technologies are integrated to gain a comprehensive understanding of the biological system of interest. Examples of omics include genomics, transcriptomics, epigenomics, proteomics, metabolomics and lipidomics, representing high throughput characterisation of genes, RNA, DNA methylation or histone modifications in chromosomes, proteins, metabolites and lipids, respectively (181). This approach was successfully used after the West African Ebola virus disease (EVD) outbreak of 2013-2016 to understand how the difference in host response contributed to disease severity (182). The researchers used peripheral blood mononuclear cells (PBMC) and plasma from 9 patients with fatal EVD, 11 EVD survivors and 10 healthy controls to integrate metabolomics, proteomics, lipidomics and transcriptomics data. They were able to identify several biomarkers, such as L-threonine and the vitamin D binding protein (GC), that could predict disease fatality. Additionally, they compared their PBMC transcriptome with that of patients with septic shock and identified a common signalling pathway between EVD fatalities and patients with sepsis (182).

We are in an era where various approaches may help predict disease risk or lead to more personalised interventions. However, many challenges remain in utilising the abovementioned approaches to increase our understanding of human genetics and genomics and their relation to infectious diseases, including mycetoma, and the specific challenges of working in low-resource settings.

1.4 Rationale and Objectives

Some evidence suggests that host genetic factors regulate susceptibility to mycetoma. Still, mycetoma is a complex trait where the interaction of multiple genes with each other and environmental factors, as well as the pathogen, causes the disease. It is thus challenging to identify the causal gene variants. Family-based analyses are better suited for identifying rare high-risk causal variants and offer insights into critical immune response pathways that can be further explored at the population level. Co-segregation of such variants with the disease in affected family members, using extended pedigrees, can lend evidence to pathogenicity predictions and aid in discovering rare variants (minor allele frequency of <1%) with large effect sizes. Adopting NGS technology such as WGS can underpin a new strategy to interrogate

the genome for variants associated with mycetoma. In parallel, at the population level, GWAS can facilitate the detection of common genetic variants (minor allele frequency of >5%) based on the “common disease, common variant” hypothesis (183).

I postulated a crucial role of host genetic variation in determining susceptibility to mycetoma. My project used several methods to determine the genetic basis of predisposition to the disease. The study objectives were:

1. To critically assess the role of genes in the development of mycetoma using published literature.
2. To estimate the relative contribution of genetic factors to mycetoma aetiology using familial aggregation analysis.
3. To identify rare variants conferring susceptibility to mycetoma using family-based WGS.
4. To detect common genetic variants associated with mycetoma using population-based GWAS.

The original contribution of this thesis is to provide insight into the role of host genetic variations as risk factors for mycetoma in Sudan.

CHAPTER 2 - Exploring host genetic susceptibility to mycetoma and fungal infections: An in-depth review

2.1 Overview

The development and progression of diseases are heavily influenced by host factors, as observed in various infections. This is also true for mycetoma, a condition that significantly burdens affected communities. Unfortunately, our current understanding of this disease is far from complete, hindering progress in developing essential medications and vaccines. Hence, my research is motivated by the need to drive progress in this area. This chapter explores the intricate relationship between host genetic factors and susceptibility to mycetoma and other fungal infections. Grounded in epidemiological evidence, the understanding emerges that genetic predisposition to these infections is a complex interplay of various genes, environmental factors, and pathogens that cause disease. A review of existing knowledge of genetic susceptibility to fungal infections, with particular relevance to mycetoma, constitutes the cornerstone of this chapter. This narrative review aimed to provide a profound and multidimensional understanding of the genetic underpinnings governing susceptibility to mycetoma and analogous fungal infections.

2.2 Methods

An exhaustive search was conducted by utilising various electronic databases such as PubMed/MEDLINE and Google Scholar to conduct a comprehensive narrative review of the existing literature. The search included combinations of the following keywords: “mycetoma,” “polymorphisms,” and “genetic susceptibility” for reviewing articles regarding susceptibility to mycetoma. For genetic susceptibility to fungal infections, the keywords “fungal,” “infection,” “mycoses,” and “genetic susceptibility” were used as search terms.

All articles identified through the two electronic databases were analysed to ensure they were within the scope of this review. Only papers written in English were included, and the year of publication was not restricted.

2.3 Host genetic susceptibility to mycetoma and other fungal infections

Host risk factors that could influence susceptibility to mycetoma include immunological status, genetic predisposition, nutritional status, immunosuppression from HIV, co-infections, and using drugs such as antibiotics or steroids. Although there is a clear link to low socioeconomic status, only a proportion of individuals exposed to mycetoma-causing organisms develop clinical disease, notwithstanding a shared environment and the ubiquitous environmental presence of the pathogen in endemic areas. This observation raised the hypothesis that host genetic factors have a role in determining the development of infection. Several lines of evidence support this hypothesis, including:

i. Three previous studies used ELISA to demonstrate the presence of antibodies against mycetoma causative agents in the sera of unaffected residents in endemic areas, indicating exposure to the causative organisms (52–54).

ii. Familial clustering of mycetoma patients, which Al Dawi and her colleagues first observed in a hospital-based study at the MRC that included 53 eumycetoma patients and 31 healthy controls (55). This study, conducted in 2013, showed that more than 62% of the enrolled patients had at least one relative affected by mycetoma. A year later, a community-based study also suggested a role for genetic predisposition, with 52% of mycetoma patients having a family member with the disease (9). A retrospective study in 2015 reported a family history of mycetoma in 12% of 6,792 patients seen at the MRC from 1991 to 2014 (56). In a more recent community-based study involving 41,176 individuals surveyed in Eastern Sennar Locality, Sennar State, Hassan and associates reported a family history of 35.7% from 359 mycetoma patients (57).

   It is important to note here that although it is known that families share their environment and their genes, in communities affected by mycetoma, multi-case families live near other families with no mycetoma cases, suggesting gene-environment interaction is of significant consequence.

iii. There are other skin fungal infections in which host genetic factors determine disease susceptibility. Examples include chromoblastomycosis (CBM) and phaeohyphomycosis which are caused by a group of dematiaceous (melanin-pigmented) fungi that also includes causative organisms of eumycetoma (184). CBM is a subcutaneous fungal infection similar to mycetoma in that the

microorganism *Phialophora verrucose* can be a causative organism for both diseases (185,186). Pérez-Blanco and associates suggested familial inheritance in CBM after finding a 3.5 times higher risk of developing CBM among members of common ancestry in an endemic state in Venezuela (58). In another study in Brazil involving 32 CBM patients and 77 healthy matched controls, Tsuneto and colleagues found that HLA-A29 was significantly associated with susceptibility to the disease (p=0.03) with an estimated 10-fold higher relative risk in individuals having the HLA-A29 antigen (59). Other examples of monogenic and polygenic disorders and susceptibility to fungal infections will be discussed thoroughly in the following two sections.

2.3.1 Monogenic disorders and fungal infections

Several single-gene defects have been linked to the pathogenesis of some infectious diseases. These rare, large-effect variants found in essential immune-related genes have paved the way to understanding the pathways directly related to host defences against pathogens and subsequently affecting susceptibility to infectious diseases (152).

Human CARD9 deficiency is caused by mutations in the gene encoding the caspase recruitment domain-containing protein 9 (CARD9). This protein is an intracellular adaptor protein downstream from C-type lectin receptors (CLRs), which are pathogen recognition receptors (PRRs) that have a major role in fungal recognition (187). In humans, CLRs (e.g., dectin-1, dectin-2, and MINCLE) recognise fungal pathogen-associated molecular patterns (PAMPs) and, along with their adaptor molecule CARD9, have a critical role in antifungal host defence. Several studies highlighted a greater fungus-specific infection susceptibility due to mutations in the genes encoding CLRs or CARD9 (187,188). In one study carried out on a consanguineous family with chronic mucocutaneous candidiasis (CMC), dermatophytosis, and Candida meningitis, Glocker and colleagues found an autosomal recessive mutation in *CARD9* to be the underlying cause with significant defects in the Th17 response due to decreased proportions of circulating IL-17^{+} T lymphocytes (189). Other compound heterozygous mutations in *CARD9* were found to confer susceptibility to CMC and Candida meningitis and trigger impaired neutrophil killing and diminished production of IL-1β and IL-6, which are essential for priming Th17 cell differentiation (190). Interestingly, two compound heterozygous and one homozygous frameshift mutations

in *CARD9* were also detected in patients with phaeohyphomycosis, which is another subcutaneous fungal infection similar to mycetoma and caused by *P. verrucose* (60). The three identified mutations did not affect gene expression but resulted in a lack of wild-type mature CARD9 protein, consequently decreasing the proportions and cytokine production of Th17 cells and weakening patients' immune responses (60). Fortunately, Queiroz-Telles and associates recently used hematopoietic stem cell transplantation (HSCT) to treat two unrelated patients with deep/invasive dermatophytosis caused by CARD9 deficiency from Brazil and Morocco (191).

Another monogenic disorder is CMC, an immunodeficiency of cell-mediated immunity that develops from a defect in IL-17 and IL-22 immunity required for defence against fungal infections (192). CMC is transmitted in autosomal dominant or recessive inheritance patterns and affects patients' skin, nails, and mucous membranes (175). The autosomal dominant form of CMC is primarily caused by gain-of-function missense mutations in the CC-domain of the *STAT1 gene* that encodes a critical transcription factor downstream from IFN-α/β and IFN-γ signalling pathways. Consequently, this leads to defective Th1 and Th17 lymphocyte responses and reduced production of IL-17, IL-22 and IFN-γ, explaining the increased susceptibility to fungal infection (193). Mutations in *IL17RA* and *IL17F* have also been associated with autosomal dominant CMC that causes impaired IL-17 immunity (193). Autosomal recessive CMC results from mutations in the autoimmune regulator (*AIRE)* gene, which leads to a rare condition known as autoimmune polyendocrinopathy candidiasis ectodermal dystrophy (APECED) (174). *AIRE* encodes a transcriptional regulator expressed by medullary thymic epithelial cells that regulate the expression of peripheral tissue-specific self-antigens and promote central tolerance by deleting self-reactive T cells (194). Loss-of-function mutations in *AIRE* produce neutralising autoantibodies against specific cytokines with antifungal properties, such as IL-17E, IL-17F and IL-22 (195).

A similar scenario needs to be investigated in mycetoma pathogenesis, where both rare and common variants in the same molecular pathway may be interplaying. The nature and extent of such interactions between these variants are still unknown, but common variants may modify the penetrance of rare variants.

2.3.2 Polygenic disorders and fungal infections

Although very informative, single-gene mutations leading to fungal infections are relatively rare in the general population, where a complex trait genetic mode of inheritance is more likely. Recent advances in genome interrogation technologies have enabled researchers to detect genetic variations predisposing susceptibility to fungal infections and the interplay between innate and adaptive immune cells and other effector cells constituting the host immune response to fungal invasion (175,196). For example, genetic susceptibility to candidiasis was investigated using GWAS, which revealed a 19.4-fold increased risk in individuals with two or more SNPs in the CD58, LCE4A-C1orf68 and TAGAP loci (197).

Genetic susceptibility to aspergillosis is also polygenic (198). *Aspergillus spp.* are seldom pathogenic in immunocompetent hosts due to innate immune responses mediated mainly through alveolar macrophages and neutrophils (199). Most cases of aspergillosis occur in the context of downregulation of the immune system by immunosuppressive therapies. Susceptibility to different forms of aspergillosis, including invasive allergic (IA) and chronic non-invasive syndromes, is associated with polymorphic variants in Toll-like receptors *(TLR)1*, *TLR6*, *TLR4*, *TLR9* and *TLR3* (200–202). One study identified variants in the chemokine (C-X-C motif) ligand 10 (*CXCL10*), an inflammatory mediator that stimulates the directional migration of Th1 in addition to increasing T cell adhesion to the endothelium (203), to be associated with invasive *A. fumigatus* infection (204). In that study, the presence of a haplotype in *CXCL10* (rs1554013 (+11101 **C**/T), rs3921 (+1642 C/**G**) and rs4257674 (-1101 **A**/G)) was associated with the inability of immature dendritic cells (iDCs) to express *CXCL10* in patients with IA after allogeneic HSCT, which led to an increased risk of developing IA. Susceptibility to IA has also been linked to deficiencies in PRRs with opsonic activity, such as mannose-binding lectin (*MBL*) and long pentraxin 3 (*PTX3*, which mediates fungal uptake and killing by phagocytes) (187,205,206). Another molecule with an opsonic activity associated with susceptibility to IA in allogeneic HSCT patients is plasminogen (*PLG*) (207). Using a novel candidate gene polymorphism identification method, Zaas and colleagues used computational genetic mapping analysis of suppressed murine model survival data to identify a variant in human *PLG* that increased the risk for IA in HSCT recipients.

The elucidation of the genetic underpinnings of susceptibility to fungal infections, encompassing both monogenic and polygenic disorders, is crucial in understanding genetic predisposition to a neglected fungal infection such as eumycetoma.

2.3.3 Genetic susceptibility to mycetoma

The present section aims to provide a comprehensive overview of the existing literature on genetic susceptibility to mycetoma. To this end, previous studies have focused on a specific set of functional polymorphisms in candidate genes, which were selected based on their potential involvement in immunity against the disease (**Figure 2.1**) (208).

2.3.3.1 Polymorphisms in innate immunity genes

The first indication of host genetic susceptibility to mycetoma was demonstrated in 2007 by van de Sande and her colleagues in research involving 125 Sudanese mycetoma patients and 140 matched controls (64). Their work focused on genetic variations in the genes involved in innate immunity owing to the significance of this system, led by neutrophils, in early defence against mycetoma (27,28). The studies were undertaken on eleven SNPs in eight genes: complement receptor 1 (*CR1*), interleukin 8 (*CXCL8*), C-X-C motif chemokine receptor 2 (*CXCR2*), nitric oxide synthase 2 (*NOS2*), *MBL*, *TNFα*, macrophage chemoattractant protein-1 (*MCP-1*), and thrombospondin4 (*THBS4* also known as *TSP4*). Those genes were selected based on their role in neutrophil function. Significant differences in allele distribution were detected in six SNPs in *CR1*, *CXCL8*, its receptor *CXCR2, THBS4* and *NOS2.* CR1 is expressed, not exclusively, on neutrophils' surfaces to mediate phagocytosis via regulation of the complement cascade (209). In the study, the Sl2 allele of rs17047661 and the McC^{a} allele of rs17047660 had higher distribution in mycetoma patients when compared to matched endemic controls, and the investigators suggested that those two alleles caused conformational changes in CR1 that eventually led to defects in phagocytosis of the mycetoma-causative agents. Moreover, the -251A allele of *CXCL8*, 29929C allele of *THBS4* and +785C allele of *CXCR2* had significantly higher allelic distributions in mycetoma patients. All these alleles are previously known to be associated with increased production of CXCL8, a neutrophil chemoattractant (210–213). On the other hand, the allele frequency of the G954C allele of *NOS2* was found to be significantly higher among the healthy endemic

controls and this allele is known to result in a 7-fold higher nitric oxide (NO) activity, which is essential in mediating tumouricidal and bactericidal actions (214). Hence, induced CXCL8 and reduced NO production were risk factors for developing mycetoma. One drawback of this study is the multiple tests done on a small cohort with no correction for multiple testing, which may have produced false positives or overlooked true positive signals in the remaining SNPs in the other THREE genes. Another drawback is the selection of a limited number of variants that were probably first described in European populations and might not be relevant in the Sudanese sample (e.g., minor allele frequency might be very low). A key aspect of the genetics of African populations, including Sudanese, is the high level of genetic diversity within individual populations (215). This diversity is believed to reflect the long history of human evolution in Africa after the human migrations out of Africa to colonise other parts of the world around 60,000 years ago (99,113,215–218).

2.3.3.2 Polymorphisms in adaptive immunity genes

The association between several HLA alleles (both class I and class II) and susceptibility or resistance to infectious diseases has been previously confirmed (219–221). In 2013, the frequencies of *HLA-DRB1* and *HLA-DQB1* alleles among confirmed eumycetoma patients admitted to the MRC and matched controls were analysed by Al Dawi and co-workers (55). For *HLA-DRB1,* the *HLA-DRB1*13* allele was significantly associated with eumycetoma infection (p=0.044, OR=2.629). Conversely, the *HLA-DRB1*02* allele had a high frequency in the control group (9.8%, p=0.047), while it was absent in the eumycetoma patients, suggesting a protective role against eumycetoma. Of the *HLA-DQB1* alleles, the *HLA-DQB1*5* allele showed a significantly higher frequency in mycetoma patients when compared to healthy controls (p=0.029, OR=3.471), indicating a possible association between this allele and the development of clinical mycetoma. This hospital-based study enrolled only 53 eumycetoma patients and 31 healthy matched controls. Hence, further studies are needed on larger sample sizes and among communities living in mycetoma endemic areas.

In the same year, another study by Mhmoud and associates identified an association between polymorphisms in cytokine genes C-C chemokine ligand 5 (*CCL5)* and *IL-10* (both involved in acquired immune responses) and mycetoma (222). CCL5 has an essential role in attracting T-cells, dendritic cells, eosinophils and natural killer (NK)

cells (223) to sites of inflammation, while IL-10 is a potent Th2 response-inducing cytokine (224). The study explored the role of three functional SNPs in *CCL5* and two SNPs in *IL-10* in mycetoma granuloma formation. A significant difference in allele frequencies for two SNPs in *CCL5* (rs2280788 and rs280789) and one SNP in *IL-10* (rs1800872) were found between 149 mycetoma patients and 206 matched controls. This difference in allele frequencies was linked to elevated serum levels of both CCL5 and IL-10 in mycetoma patients. It should be noted that the two alleles in *CCL5* which were significantly associated with mycetoma have opposite functional effects: the G-allele at SNP rs2280788 induces higher *CCL5* transcription, whereas the C-allele in rs280789 is associated with reduced transcription of *CCL5*. Accordingly, the exact role of CCL5 in mycetoma granuloma formation remains to be elucidated.

2.3.3.3 Polymorphisms in host enzyme genes

Two studies linked certain host enzymes to susceptibility to mycetoma. The first, reported in 2014, investigated the role of matrix metalloproteinases-2 (MMP-2), matrix metalloproteinases-9 (MMP-9) and tissue inhibitor of metalloproteinases (TIMP-1) in the formation of the mycetoma granuloma around the fungal grains of *M. mycetomatis* in eumycetoma patients (65). These enzymes are hypothesised to be involved in the deposition of collagen around the grains to encapsulate them within the granulomas, resulting in excessive collagen formation that may contribute to treatment failure in mycetoma patients. The study addressed three aspects: (i) measuring absolute serum levels of MMP-2 and MMP-9 in the sera of 36 *M. mycetomatis* infected patients and 36 healthy controls using ELISA, (ii) detection of MMP-2 and MMP-9 around fungal grains using immunohistochemical staining of eight tissue sections taken from eumycetoma patients, and (iii) genotyping three functional SNPs in the promoter regions of *MMP-2* (rs243865), *MMP-9* (rs3918242) and *TIMP-1 (*rs4898*)* using genomic DNA from 125 Sudanese eumycetoma patients and 103 healthy endemic controls. Active MMP-9 was found in the sera of 36% of eumycetoma patients. However, that cannot be attributed solely to mycetoma infection since no data about co-infection was recorded during the sample collection process. No genetic differences were identified between patients and healthy controls for *MMP-2* and *MMP-9*. Yet, there was a higher frequency of the T allele of the rs4898 SNP in *TIMP-1* in 77 eumycetoma male patients compared to 44 healthy male controls. This allele is associated with a lower TIMP-1 protein expression in inflammatory bowel disease,

and the investigators suggested that it could explain the previously reported male predominance in mycetoma (225).

The second study, in 2015, investigated the presence of a structural polymer found in fungal cell walls, known as chitin, in the cell wall of *M. mycetomatis* (226). This polysaccharide triggers macrophages to release chitin-degrading enzymes such as chitotriosidase (CHIT1) and acidic mammalian chitinase (AMCase, also known as CHIA). Four functional polymorphisms in *CHIT1* and *CHIA* were investigated in 112 eumycetoma patients and 103 matched controls from the same cohort used for the previous metalloproteinases study. Both human chitinases were expressed in mycetoma lesions. Furthermore, a 24-bp insertion in *CHIT1* was found to be significantly associated with eumycetoma, and this polymorphism is known to impair enzyme activity (227,228). Since both chitotriosidase and AMCase were detected in mycetoma lesions, further studies on polymorphisms affecting the activity of these enzymes are still needed with larger sample sizes.

2.3.3.4 Polymorphisms in sex hormone biosynthesis genes

A male predominance of mycetoma has been indicated in several studies, signifying a possible role of hormonal influence in mycetoma susceptibility (6,15,65,77). Based on this, a study was done in 2015 on 125 eumycetoma patients and 103 matched controls (the same metalloproteinases cohort) to investigate changes in hormonal levels that could result from polymorphisms within genes encoding enzymes involved in sex hormone biosynthesis (229). Five polymorphisms were selected including rs4680 in catechol-O-methyltransferase (*COMT*), rs743572 in cytochrome p450 subfamily 17 (*CYP17*), rs1056836 in cytochrome p450 subfamily 1B1 (*CYP1B1*), rs700518 in cytochrome p450 subfamily 19 (*CYP19*) and rs6203 in hydroxysteroid dehydrogenase 3B (*HSD3B*). Only alleles of rs4680 SNP in *COMT* and rs700518 in *CYP19* had significant differences in distribution among mycetoma patients compared to healthy endemic controls.

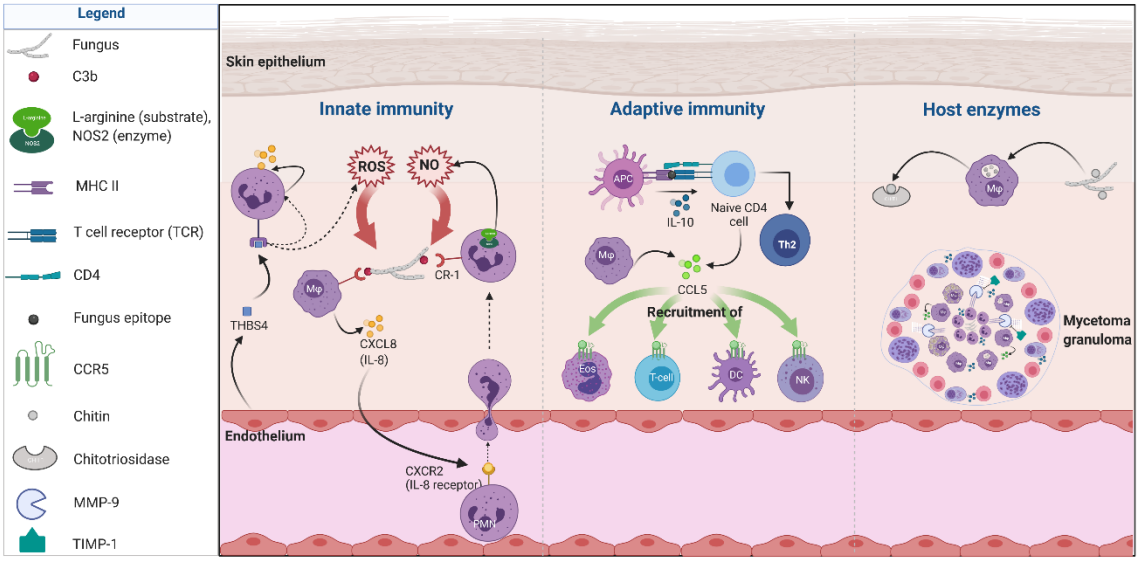


Figure 2.1 *Candidate genes' roles in immunity against mycetoma. The innate immune response towards mycetoma is triggered by traumatic inoculation of the fungus into the subcutaneous tissue. The fungal cell wall molecules activate the complement system, producing the opsonin C3b that binds to the fungal surface. C3b is then recognised by the complement receptor CR1, expressed by macrophages (Mφs) and neutrophils (PMNs). Macrophages produce several cytokines, including CXCL8 (IL-8), attracting more neutrophils from the bloodstream into the infection site. This attraction occurs via the binding of CXCL8 to its receptor CXCR2 expressed on neutrophils' surface. CXCL8 also induces neutrophils to generate nitric oxide (NO) species via the enzyme nitric oxide synthase 2 (NOS2). Other reactive oxygen species (ROS) are also generated through thrombospondin4 (THBS4), secreted by endothelial cells. On the one hand, the adaptive immunity against mycetoma, antigen presentation through HLA class II (MHC II), which is expressed by antigen-presenting cells (APCs), leads to the transformation of naïve CD4 cells into T-helper type 2 lymphocytes under the influence of IL-10. On the other hand, CCL5 released by macrophages and CD4 T cells have an essential role in the formation of mycetoma granuloma in addition to the recruitment of other immune cells such as eosinophils (EOS), T-lymphocytes (T-cells), dendritic cells (DCs) and natural killer (NK) cells. The contribution of host enzymes in immunity against mycetoma involves MMP-9 and chitotriosidase. MMP-9, expressed by the immune cells surrounding the grain, is responsible for collagen deposition around the grain during mycetoma granuloma formation. Activated macrophages release Chitotriosidase to degrade chitin, a polysaccharide expressed on the cell wall of M. mycetomatis. Figure created using BioRender.com.*

The studies mentioned above are generally small, take a candidate gene approach, and have yet to be replicated. Moreover, no systematic genome-wide studies on susceptibility to mycetoma have been reported to date.

In brief, 13 genes had allelic variants found to be associated with susceptibility to mycetoma and lie in different pathways and systems (**Table 2.1**).

Table 2.1 *A summary of genes with allelic variants significantly associated with mycetoma.*

Gene	**SNP**	**Rs-number**	***P* value**	**Reference**
CCL5	–28C/G	rs2280788	<.0001	(32)
	1648044C>T*	rs280789	<.0001	(32)
CHIT1	24bp insertion	rs3831317	0.004	(226)
COMT	19951271G>A*	rs4680	0.006	(229)
CR1	207782889A>G*	rs17047661	0.039	(64)
	207782856A>G*	rs17047660	0.001	(64)
CXCL8	−251T/A	rs4073	0 .008	(64)
CXCR2	219000310C>T*	rs2230054	0 .037	(64)
CYP19	51529112T>C*	rs700518	0 .004	(229)
HLA-DRB1	HLA-DRB1*13	-	0.044	(55)
	HLADRB1*02	-	0.047	(55)
HLA-DQB1	HLA-DQB1*5	-	0.029	(55)
IL-10	-592A/C	rs1800872	0.0005	(32)
NOS2	−954 G>C*	rs1800482	0 .0006	(64)
TIMP-1	372T/C	rs4898	0.0004 (males) 0.53 (females)	(65)
THBS4	79361265G>C*	rs1866389	0.030	(64)

* Nomenclature according to GRCh37 human reference genome.

To study the interaction between susceptibility genes, a network analysis was conducted using the GeneMANIA plugin (230) within version 3.8.2 of Cytoscape software (231). Several parameters were included: co-expression, shared protein domains, physical interactions, common pathways, and co-localisation (**Figure 2.2**). This network revealed the interaction between mycetoma susceptibility genes and other candidate genes to be considered potential contributors to susceptibility to mycetoma. The results showed that biological interactions were primarily co-expression, accounting for approximately 49% of the interactions studied. Physical interactions accounted for 39.5% of the total interactions, while shared protein domains, co-localisation, and pathway interactions accounted for 4.56%, 3.66%, and 3.53%, respectively. Understanding the underlying network connectivity would enhance our ability to identify crucial genes that can potentially become valuable targets for future therapeutics.

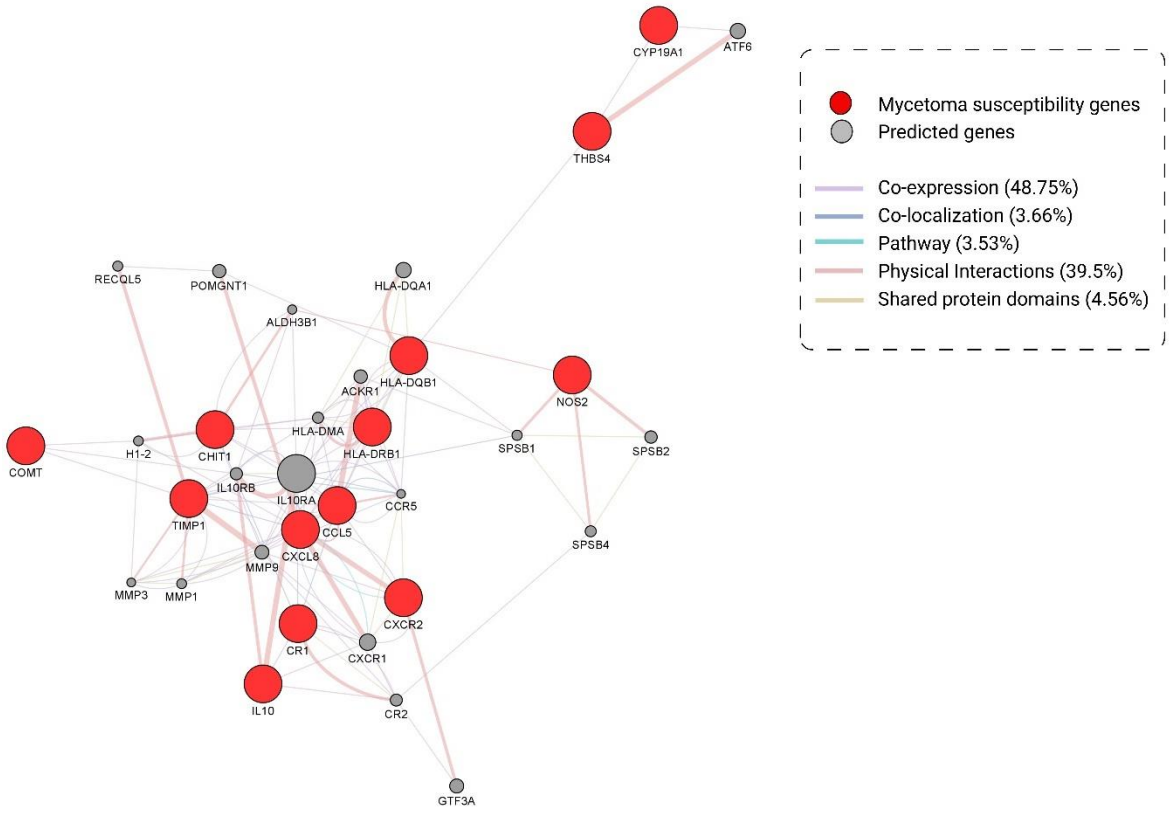


Figure 2.2 *A network demonstrating the 13 genes associated with susceptibility to mycetoma (red nodes) and their interaction with each other and with other candidate genes (grey nodes). The underlying molecular links were attributed to the co-expression associations, co-localization, pathway links, physical interactions, and shared protein domains.*

2.4 Concluding remarks and future perspectives

The studies on genetic susceptibility to mycetoma described above suggest a role for the host's underlying genetic profile. However, the studies are generally small, take a candidate gene approach, and none have been replicated. Recent progress in genomic science and corresponding technological advancements have opened avenues for conducting comprehensive investigations into the genome to identify variations linked to mycetoma through GWAS or NGS technologies. Furthermore, several genomes for the most common causative organisms, including *Nocardia brasiliensis* (232)*, Streptomyces somaliensis* (233)*, Actinomadura madurae* (234)*, Nigrograna mackinnonii* (235) and *Madurella mycetomatis* (236) are now publicly available. Parallel analysis of pathogen and host genomes could give valuable insights into host-pathogen interactions in mycetoma (237). This would eventually aid in devising better strategies for risk assessment, treatment, and prognosis for mycetoma patients.

It is essential to include large cohorts to avoid the multiple challenges of genetic association studies, including population stratification, genetic heterogeneity and insufficient replication power (238,239). Additionally, genotype-phenotype correlation analyses are mandatory to correlate the relative contribution of genetic findings to specific clinical outcomes.

Dissecting the contribution that host genetic variation has on susceptibility to mycetoma will enable the identification of pathways that are potential targets for new treatments for mycetoma and will also enhance our ability to stratify "at-risk" individuals, allowing the potential to develop preventive and personalised clinical care strategies in the future.

CHAPTER 3 - Methods employed for genetic analysis of mycetoma susceptibility

3.1 Choice of method

Despite the existing literature supporting a role for host genetic variation in determining predisposition to mycetoma, there has not been a systematic approach to interrogate the presence of causal variants. Therefore, this PhD project was undertaken to address this gap and included three main approaches:

i. Determining the extent of familial aggregation for mycetoma by estimating heritability and sibling recurrence risk. This was done by analysing extended pedigrees, with more than three generations, collected from mycetoma patients either in their endemic villages or during their visits to the MRC clinic.
ii. Conducting family-based whole genome sequencing (WGS) to identify rare genetic variants in mycetoma cases' genomes that were not present in their healthy relatives.
iii. Undertaking a population-based case-control study to detect common genetic differences between mycetoma cases and unaffected controls, using genome-wide association studies (GWAS).

3.2 Research sites and access

Sudan is divided into 18 states (wilayat), each with its own administrative structure (85). These states are further divided into districts (mudiriyat), and districts are divided into localities (baladiyat), which are the lowest administrative units in Sudan (85). The study participants were recruited at three sites, as outlined in **Figure 3.1**:

i. The MRC clinic in the capital city, Khartoum, which has more than 10,000 registered patients.
ii. The Wad Onsa mycetoma clinic in Eastern Sennar Locality (baladiyat), Sennar State (wilayat), located about 250 km southeast of Khartoum. This clinic was set up in a rural area where mycetoma is endemic.
iii. Wad El Nimar village in Eastern Sennar Locality, Sennar State, adjacent to Wad Onsa village. Both villages are part of the Wad Onsa administrative unit.

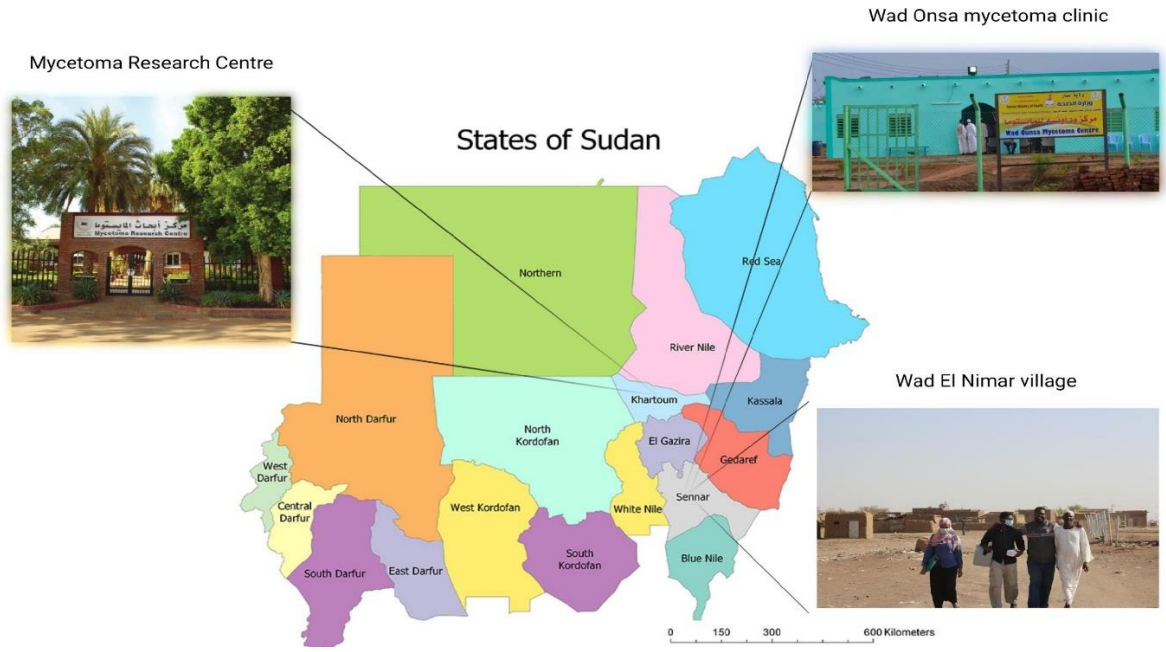


Figure 3.1 *Research sites. Original map adapted and modified from https://www.mappr.co/counties/states-of-sudan/*(85). *Figure created using BioRender.com.*

Eastern Sennar locality comprises five administrative units: Wad Onsa, Dooba, Wad Al-Abbas, Wad Taktook and Alreef Alsharqi. It includes 292 villages with a population of 353,196 residents (240). The reason for targeting this locality was high mycetoma endemicity in the area, as reported by the surveillance database at the MRC (240).

3.3 Ethical considerations

Ethics permissions were obtained from the BSMS Research Governance and Ethics Committee (RGEC ER/BSMS435/2) in the United Kingdom and the Soba University Hospital Ethics Committee in Sudan.

Written or thumbprint-signified informed consent was obtained from all study participants or their guardians (in case of patients below 18 years old). Data anonymisation was carried out after sample collection, and participants' study records and medical information were kept confidential, identified by a study code number known only to the study staff under the terms and conditions of the study design.

Although genotype and genome sequence data were generated in this study, the data analyses focused solely on genetic variants related to genetic susceptibility to mycetoma and data was anonymised before analyses. This research did not uncover unexpected and possibly clinically relevant findings in this context.

3.4 Study participants recruitment and data collection

3.4.1 Familial aggregation analyses

Eligibility to participate was according to the following criteria:

3.4.1.1 Inclusion criteria

a. Families with more than one case of eumycetoma.
b. All ages and both genders were eligible to participate.

3.4.1.2 Exclusion criteria

a. Singletons, i.e., pedigrees that contained only one affected individual.
b. Families with actinomycetoma cases.
c. Consent was refused.

A total of 46 families were recruited for this part of the study. The ideal sample size for studying familial aggregation in complex diseases like mycetoma is challenging to specify precisely, as it depends on factors such as disease prevalence, mode of inheritance, the actual number of affected cases in the family and the total number of family members we could collect information about. Enrolling 46 pedigrees was pragmatic and would serve as a reasonable starting point, especially since all pedigrees spanned three generations and mostly consisted of a large number of individuals, providing a foundation for further investigations.

I collected the pedigrees from probands (the first person who sought medical attention for the disease) who attended the MRC, Wad Onsa clinic, or from visiting the patients' houses in Wad El Nimar endemic village. During my visits to the village, I was accompanied by Mr Siddig Ibrahim Abu Sin, a community activist who identified all the families affected by mycetoma.

In most cases during pedigree collection, the proband arrived with other family members, which increased the reliability of the family structure information. It was possible that the probands may not have known if their grandparents' generation had mycetoma since public awareness about this disease is recent. As a precaution, I labelled individuals whose disease status was unknown as "NA".

The demographic and pedigree data of families with at least one case of mycetoma was collected after getting consent from the proband. The data was collected using a

pedigree collection datasheet adapted from Williams-Blangero S, 2006 (241) (Appendix 2). Data collected on the family history of mycetoma in the proband's family ranged from simple questions about whether the proband's parents had mycetoma to more comprehensive questions about affection status in siblings, cousins, and other members of the proband's extended family.

3.4.2 Whole genome sequencing study

From the pedigrees collected using the abovementioned approach, I identified four families from Wad El Nimar village with at least four cases of mycetoma in each family. Those families were selected to be enrolled in the family-based WGS. Demographic and pedigree data collection was done using Phenotips, an open-source clinical phenotype and genotype data collection tool (242), by which the pedigrees in **Figure 3.2** were created.

This work was exploratory, without knowledge of the mode of inheritance. There are high levels of consanguinity (mainly between first cousins) in these communities where mycetoma is a complex trait. I hypothesised that specific variants had become concentrated in those affected families. Studying isolated communities and consanguineous families with a higher prevalence of infectious diseases has provided valuable information about the genetic factors influencing susceptibility, resistance, or treatment response (section 1.3.2).Thus sample size calculations are irrelevant for studying single gene disorders in consanguineous families.

3.4.3 Genome-wide association studies

Eligibility to participate was according to the following criteria:

3.4.3.1 Inclusion criteria

Individuals diagnosed with eumycetoma who gave consent were eligible to be enrolled as cases. Control subjects, matched by gender, were eligible for inclusion with no evidence of history or clinical examination of mycetoma, and they were residents in mycetoma endemic areas for a minimum of 20 years. Control subjects were over the age of 40 years. We deliberately did not match on age because mycetoma prevalence peaks in the 20-40 year age, and younger controls may be susceptible but have yet to develop the disease, thus confounding the study. This approach has been previously

adopted for conditions with a slow progression that requires prolonged environmental exposure (146).

3.4.3.2 Exclusion criteria

a. Patients registered at the MRC as having actinomycetoma (most cases in Sudan are eumycetoma, and there may be differences in the host response to the two different disease forms).
b. Controls under 40 years.
c. Consent was refused.

Demographic data were collected using a collection datasheet administered to the study participants with questions including socio-demographic variables (e.g., ethnicity, age, sex, and occupation) and clinical data for cases (family history of disease, age of onset and site of eumycetoma infection). The English version of the collection datasheet can be found in Appendix 3.

Using the online version of the Genetic Association Study power calculator (243), I estimated a sample size of 600 mycetoma patients and 500 healthy controls to give 80% power to detect any significant association ($P \leq 5e\text{-}08$) between a given single nucleotide polymorphism and mycetoma, assuming a disease prevalence of 5%, a minor allele frequency of 0.2 and a genotype relative risk (RR) of 1.8. In practice, collecting as many cases and controls as possible was essential because it was recognized that assumptions based on a genotype RR of 1.8 may lean towards optimism. To detect effects with lower RRs robustly, it was imperative to have a larger pool of cases and controls. Therefore, an initial genotyping of 960 samples was strategically conceived, comprising of 560 cases and 400 controls. This approach was developed in agreement with the Centre for Proteomic and Genomic Research (CPGR) in Cape Town, South Africa, where the genotyping was outsourced. The samples were planned to be shipped in two batches, in 96-well PCR plates.

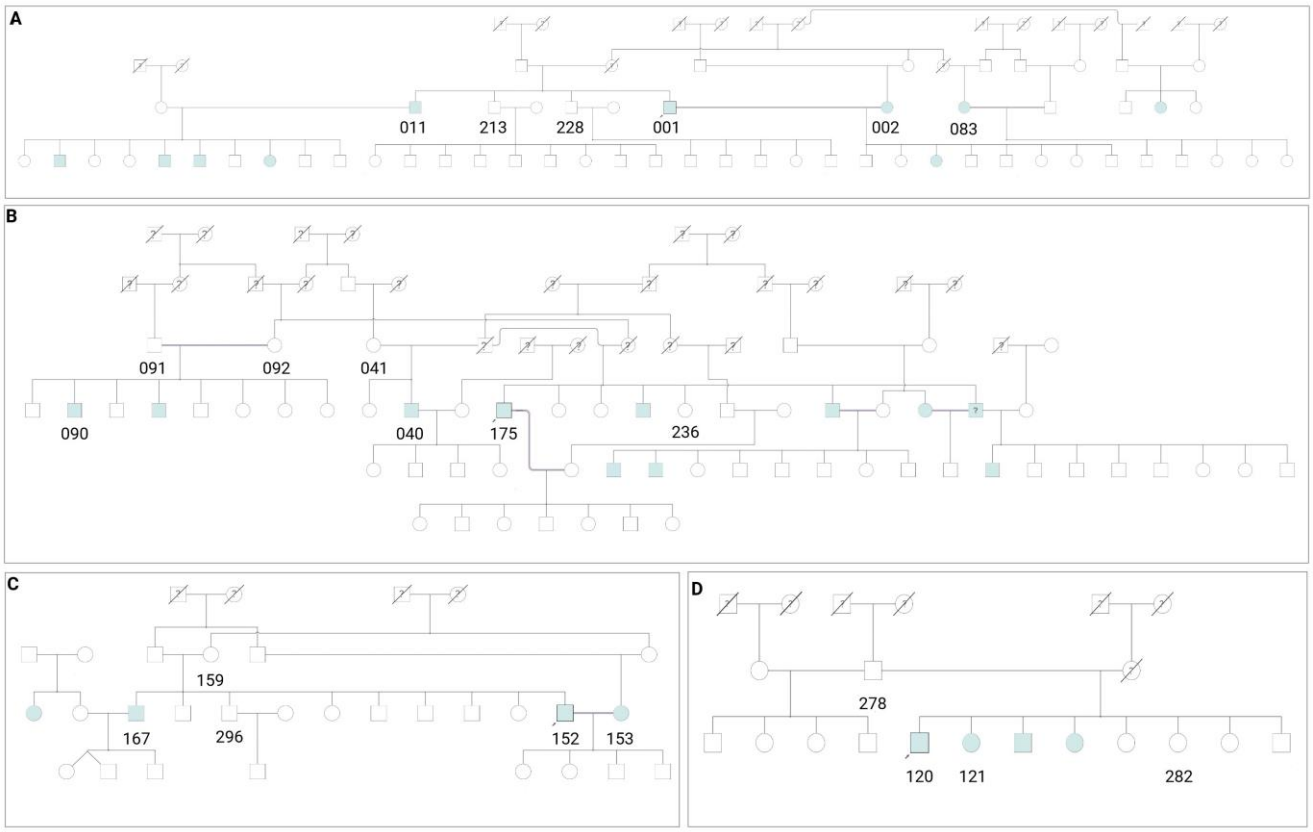


Figure 3.2 *Pedigrees of the four families recruited for WGS study from Eastern Sennar locality, Sudan. Round symbols indicate female; square, male; diamond, unspecified gender; slash through circle or square: dead female or male; blue-filled symbols, mycetoma; unfilled symbols, unaffected; question mark within symbol, unknown affection status; number immediately under the symbol, identity number of an individual within pedigree; double lines, consanguinity; arrow, proband. (A) Family 1 pedigree; (B) Family 2 pedigree; (C) Family 3 pedigree and (D) Family 4 pedigree.*

3.5 Laboratory methods

3.5.1 Whole genome sequencing study

3.5.1.1 Sample collection and DNA extraction

A 2-ml saliva sample was obtained from 22 participants representing the four consanguineous families using Oragene™ (OG-600) DNA kits (DNA Genotek Inc., Ottawa, ON, Canada). These kits were chosen based on their convenience in our setting where healthy individuals are hesitant to provide blood samples. Additionally, there could be a delay of up to three hours before the samples are stored at the Wad Onsa Mycetoma Clinic laboratory in Sennar state and transported to the MRC in Khartoum state. Oragene DNA kits allow shipping and storage of samples for years at ambient temperature without DNA degradation and yield a median of 110µg DNA of high molecular weight (> 23kb) per 2ml saliva collected (244). Saliva samples were collected according to the manufacturer's instructions (Appendix 4).

DNA extraction was done at the MRC laboratory following the prepIT®.L2P manual protocol provided by the manufacturer (Appendix 5). DNA quantity and quality were checked using an Implen Nanophotometer (Implen GmbH, Munich, Germany) and the standard agarose gel electrophoresis (Appendix 6).

3.5.1.2 Whole genome sequencing

50 µl of DNA solution from 22 samples were sent for WGS at BGI Genomics in Hong Kong. Following library preparation, massively parallel sequencing was done on a DNBseq™ sequencing platform (BGI Genomics, Hong Kong).

3.5.1.3 Sanger sequencing

Sanger sequencing is the gold-standard method to confirm the presence of WGS candidate variants, ruling out the possibility of these variants being sequence artefacts. Polymerase chain reaction (PCR) was done to amplify the desired regions in our candidate genes following the protocol in Appendix 7. Primers for the PCR were designed on the Primer3Plus (245). I confirmed the absence of any SNPs in the primer binding regions that could hinder annealing using the SNPCheck V3 online tool (246). Primer melting temperatures were calculated theoretically using the UCSC In-Silico PCR which is part of the UCSC Genome Browser (247). Moreover, primer specificity was checked using the Primer-BLAST online tool (248).

A total of 20 µl of PCR products and 30 µl of the primers were sent to Macrogen Europe (Netherlands) for Sanger sequencing.

3.5.2 Genome-wide association studies

3.5.2.1 Sample collection and DNA extraction

A 2-ml saliva sample was collected from all participants using Oragene^{TM} (OG-600) DNA kits (DNA Genotek Inc., Ottawa, ON, Canada). I used the same protocol described in section 3.5.1.1. Extracted DNA samples that showed good quality with A260/A280 ratios between 1.8 and 2.0 were normalised (50 ng/µl) and shipped in 96-well PCR plates with a sample volume of 20 µl.

3.5.2.2 Genotyping

Genotyping of the first batch of samples was outsourced to CPGR, Cape Town, South Africa. Genotyping was performed using the Illumina Infinium H3Africa Consortium Array v2 containing 2,271,503 SNPs specifically selected based on its representation of genetic variation in African populations (249). The genotypes obtained from genotyping arrays were mapped to specific genomic positions in the human reference genome using GRCh37 (build 37) coordinates (250). Sample and SNP process report files were received for initial quality information. The resulting raw data was in the format of Intensity Data files *(*IDAT).

3.6 Data analysis

3.6.1 Familial aggregation analyses

3.6.1.1 Pedigree statistics

Basic descriptive statistics of the pedigrees were generated using the PEDINFO program in Statistical Analysis of Genetic Epidemiology (SAGE) software version 6.4.2 (251).

3.6.1.2 Sibling recurrence risk

Sibling recurrence risk (K_s) was calculated by Professor Heather Cordell using an adaptation of Olson and Cordell's formula described previously (144). Under the special case that there is precisely one proband per family, the estimate was calculated using the following equation:

$$\widehat{K}_s = \frac{\sum_{s=1}^{\infty}\sum_{a=1}^{s} w_a\,(a-1)n_{s(a)}}{\sum_{s=1}^{\infty}\sum_{a=1}^{2} w_a\,(s-1)n_{s(a)}},$$

In this formula, $n_{s(a)}$is the number of sibships of size s and a the number of affected cases. The weights w_a (and thus the estimated sibling risk) depend on what one assumes about the ascertainment scheme. Ascertainment adjustment was essential because families were recruited via a proband and the varying size of sibships. I considered single and complete ascertainment as two ends of a spectrum, thus providing a lower and upper bound for the risk to a sibling of an affected individual. The sibling recurrence risk ratio (λ_s) for mycetoma was calculated according to the formula $\lambda_s = \frac{K_s}{K}$, where K was the population prevalence of mycetoma that was estimated at 0.87% according to a previous study done in the same endemic area where most of our pedigrees were collected from (57).

3.6.1.3 Heritability estimation

Heritability (h^2) was quantified using three software programs selected based on their ability to a) estimate pedigree-based heritability and b) handle discrete traits –such as mycetoma– in pedigree datasets. The first was the Sequential Oligogenic Linkage Analysis Routines (SOLAR) software package (252). The estimation was done under a polygenic model using the discrete trait algorithms implemented in SOLAR Eclipse version 8.5.1. Age and sex were included as covariates in the model. Mendel software package (253) was also used to verify the results, employing analysis option 19, the Variance Components (POLYGENIC-QTL option). The third was the ASSOC program implemented in SAGE version 6.4.2. The ASSOC program utilises a linear mixed model (LMM) and estimates its parameters through the maximum likelihood (ML) method (254).

3.6.2 Whole genome sequencing study

The initial bioinformatics analysis, including quality control, mapping against GRCh37, and variant calling, was provided by BGI Genomics due to limitations in our server at MRC. The data sent from BGI included Binary Alignment Map (BAM), Variant Call Format (VCF) and Comma Separated Values (CSV) files. The subsequent analysis included the following steps:

3.6.2.1 Subsetting samples according to shared variants

The CSV files were used for subsequent analysis by R version 4.0.2 (255) to select variants shared among the cases but not found in their related controls. R scripts used in this analysis were all executed from RStudio version 2022.07.2. The following R scripts are examples run for each family:

1. For single nucleotide polymorphisms (SNPs) files:

    a. Set the working directory in RStudio
    setwd (“F:/WGS_data/All_samples/F2/SNPs”)
    b. Load the samples csv files into RStudio
    G1 <- read.csv(“M175.snp.annot.csv”, header = TRUE, stringsAsFactors = FALSE)
    G2 <- read.csv(“M040.snp.annot.csv”, header = TRUE, stringsAsFactors = FALSE)
    G5 <- read.csv(“M236.snp.annot.csv”, header = TRUE, stringsAsFactors = FALSE)
    G6 <- read.csv(“M041.snp.annot.csv”, header = TRUE, stringsAsFactors = FALSE)
    c. Merge the controls’ csv files
    combined.controls <- rbind(G5,G6)
    d. Load plyr R package
    library(plyr)
    e. Match the cases files based on shared variants
    F <- match_df(G1,G2, on =“Start”)
    f. Subset the matched dataset from combined controls
    finaldf<- subset (F, !(Start %in% combined.controls$Start))
    g. Export the subsetted file containing variant found in cases only
    write.csv(finaldf, file = “finaldf_F2a_snp.csv”)
    h. Subset the resulting variants based on impact
    unique(finaldf$Impact)
    HIGH <- subset(finaldf, Impact == “HIGH”)
    MODERATE <- subset(finaldf, Impact == “MODERATE”)
    HiModr <- rbind(HIGH,MODERATE)
    write.csv(HiModr, file = “HiMod.snp.csv”)
    LOW <- subset(finaldf, Impact == “LOW”)
    MODIFIER <- subset(finaldf, Impact == “MODIFIER”)

LoMODF <- rbind(LOW,MODIFIER)

i. Export the resulting file

write.csv(LoMODF, file = “LowMod.snp.csv”)

2. For insertion/deletions (InDels) files:

a. Set the working directory in RStudio

setwd (“F:/WGS_data/All_samples/F1/InDels”)

b. Load the samples csv files into RStudio

G1 <- read.csv(“M001.indel.annot.csv”, header = TRUE, stringsAsFactors = FALSE)

G2 <- read.csv(“M002.indel.annot.csv”, header = TRUE, stringsAsFactors = FALSE)

G3 <- read.csv(“M011.indel.annot.csv”, header = TRUE, stringsAsFactors = FALSE)

G4 <- read.csv(“M083.indel.annot.csv”, header = TRUE, stringsAsFactors = FALSE)

G5 <- read.csv(“M213.indel.annot.csv”, header = TRUE, stringsAsFactors = FALSE)

G6 <- read.csv(“M228.indel.annot.csv”, header = TRUE, stringsAsFactors = FALSE)

c. Merge the controls’ csv files

combined.controls <- rbind(G5,G6)

d. Subset controls based on homozygosity

unique(combined.controls$Otherinfo)

hom_com_con <- subset(combined.controls, Otherinfo == “hom”)

e. Load plyr R package

library(plyr)

f. Match the cases files based on shared variants

I <- match_df(G1,G2, on =“Start”)

H <- match_df(I,G3, on =“Start”)

G <- match_df(H,G4, on =“Start”)

g. Subset cases based on homozygosity

unique(G$Otherinfo)

hom_G <- subset(G, Otherinfo == “hom”)

j. Subset the matched homozygous dataset from combined controls

finaldf<- subset (hom_G, !(Start %in% hom_com_con$Start))

k. Export the subsetted file containing variant found in cases only

write.csv(finaldf, file = “finaldf_F1_hom_indel.csv”)

l. Subset the resulting variants based on impact

```
unique(finaldf$Impact)
HIGH <- subset(finaldf, Impact == “HIGH”)
MODERATE <- subset(finaldf, Impact == “MODERATE”)
HiModr <- rbind(HIGH,MODERATE)
write.csv(HiModr, file = “HiMod.indel_hom.csv”)
LOW <- subset(finaldf, Impact == “LOW”)
MODIFIER <- subset(finaldf, Impact == “MODIFIER”)
LoMODF <- rbind(LOW,MODIFIER)
```

m. Export the resulting file

```
write.csv(LoMODF, file = “LowMod.indel_hom.csv”)
```

3. For copy number variants (CNVs):

a. Set the working directory in RStudio

```
setwd("F:/WGS_data/All_samples/F3/CNV")
```

b. Load the samples csv files into RStudio

```
G1 <- read.table("M152.CNV.vep.anno.xls", header = FALSE,
stringsAsFactors = FALSE)
G2 <- read.table("M153.CNV.vep.anno.xls", header = FALSE,
stringsAsFactors = FALSE)
G3 <- read.table("M167.CNV.vep.anno.xls", header = FALSE,
stringsAsFactors = FALSE)
G4 <- read.table("M159.CNV.vep.anno.xls", header = FALSE,
stringsAsFactors = FALSE)
G5 <- read.table("M296.CNV.vep.anno.xls", header = FALSE,
stringsAsFactors = FALSE)
```

c. Merge the controls’ csv files

```
combined.controls <- rbind(G4, G5)
```

d. Load plyr R package

```
library(plyr)
```

e. Match the cases files based on shared variants

```
H <- match_df(G1,G2, on ="V4")
G <- match_df(H,G3, on ="V4")
```

f. Subset the matched dataset from combined controls

```
finaldf<- subset (G, !(V4 %in% combined.controls$V4))
```

g. export the subsetted file containing variant found in cases only

write.csv(finaldf, file = "finaldf_F3_CNV_V4_FINAL_23.csv"

4. For structural variants (SVs) files:

    a. Set the working directory in RStudio
    setwd("F:/WGS_data/All_samples/F4/SV")
    b. Load the samples csv files into RStudio
    G1 <- read.table("M120.sv.vep.anno.xls", header = FALSE, stringsAsFactors = FALSE)
    G2 <- read.table("M121.sv.vep.anno.xls", header = FALSE, stringsAsFactors = FALSE)
    #Load Controls
    G3 <- read.table("M278.sv.vep.anno.xls", header = FALSE, stringsAsFactors = FALSE)
    G4 <- read.table("M282.sv.vep.anno.xls", header = FALSE, stringsAsFactors = FALSE)
    c. Load plyr R package
    library(plyr)
    d. Match the cases files based on shared variants
    F <- match_df(G1,G2, on ="V4")
    e. Merge the controls' csv files
    combined.controls <- rbind(G3, G4)
    f. Subset the matched dataset from combined controls
    finaldf<- subset (F, !(V4 %in% combined.controls$V4))
    g. export the subsetted file containing variant found in cases only
    write.csv(finaldf, file = "finaldf_F4_sv_V4_Final.csv")

The above scripts were modified several times to capture both homozygous and heterozygous variants, as the disease's mode of inheritance is unknown.

3.6.2.2 Variant filtering

The resulting CSV files were then filtered based on the following criteria:

1. Frequency:
    I selected alleles with a minor allele frequency (MAF) of less than 1% according to gnomAD genomes v2.1.1 frequency database (256) with an additional focus on allele frequencies in African populations. Rare variants (MAF< 1%) are more likely to be functional or have a larger effect size on phenotypic outcomes than common variants (257). They can have a higher likelihood of being deleterious

or disease-causing due to their recent appearance or purifying selection that kept them at low frequency (258). By focusing on rare variants with MAF< 1%, I increased the chances of identifying variants that could play a significant role in mycetoma development.

2. Impact:

   To prioritise variants based on their impact on protein function, expression, or sequence conservation, only high-level variants (such as frame-shifts, splicing, large deletions, or duplications) and moderate-level variants (such as missense or in-frame-shifts) were chosen for further consideration.

3. Pathogenicity:

   According to the American College of Medical Genetics (ACMG) classification (259), only pathogenic, likely pathogenic or variants of uncertain significance (VUS) were chosen for further filtering.

4. Loss of function:

   I also considered the Loss of Function (LOF) variants as identified by the Ensembl Variant Effect Predictor (VEP) tool (260). The LOF variants are defined as genetic alterations in the genome that are predicted to disrupt the normal function of a gene, potentially leading to the loss or reduction of gene activity and may contribute to disease susceptibility or phenotypic changes.

5. Relevance to mycetoma:

   The Human Protein Atlas (261) and Genotype-Tissue Expression (GTEx) (262) databases were used to investigate whether the candidate genes were expressed in tissues or immune cells related to mycetoma. Reactome database (263) was also used to explore if the candidate genes were involved in biological pathways associated with mycetoma pathogenesis.

   In addition, I conducted a literature search to determine whether the identified variants/genes were associated with increased susceptibility to mycetoma, other fungal infections, or any other infectious diseases.

3.6.2.3 Variant prioritisation

For the final filtering of candidates in each family, the following databases were used:

1. Varsome search engine (264), which aggregates human genetic variation information from several databases, including:

i. ACMG for pathogenicity classification.
ii. Detection of Mode of Inheritance from Non-heritable Observations (DOMINO), a computational method that predicts the mode of inheritance (autosomal dominant, autosomal recessive, or X-linked) of genes based on the analysis of non-heritable observational data, such as gene expression levels, without relying on traditional genetic segregation data (265).
iii. GTEx for tissue-specific gene expression,
iv. GnomAD genomes allele frequency, and
v. Several integrated algorithms to predict whether a genetic variant is pathogenic (pathogenicity predictors) and assesses its conservation across different species (conservation predictors), aiding in the interpretation of genetic variants in a clinical context.

2. The Clinical Genome Resource (ClinGen) (266), a publicly accessible resource that provides curated and standardised information on the clinical relevance and interpretation of genetic variants. It provides data on gene-disease relationships, variant interpretation, and clinical validity, aiding in the accurate interpretation of genetic variants and their associations with various medical conditions.
3. Alliance of Genome Resources (267), a collaborative platform that provides a comprehensive source of information about candidate genes' functions. It integrates data from diverse model organisms and human studies to offer insights into gene function, biological pathways, expression patterns, and associated diseases.
4. Online Mendelian Inheritance in Man (OMIM) catalogue (268), a comprehensive resource for information about human genes and genetic disorders. It provides detailed insights into the genetic basis of diseases, including gene function, inheritance patterns, clinical characteristics, and associated references, enabling clinicians and researchers to better comprehend the relationships between genes and diseases for diagnostic and research purposes.

To validate this analysis, a senior colleague annotated, filtered, and prioritised the families' VCF files independently using VarAFT software (269).

3.6.2.4 Validation by Sanger Sequencing

The raw sequencing data, known as AB1 files, were analysed using MEGA version 7.0.26 software (270).

3.6.2.5 Identity by descent analysis

In order to verify identity by descent (IBD) for the identified candidate variants in Family 2 and Family 3, an IBD analysis was performed on the regions within a 400 kb window of each variant. Only Family 2 and Family 3 were chosen for this analysis since the identified variants were detected in third degree relatives, therefore I had to confirm that the shared variants came from a recent common ancestor. The analysis was conducted on VCF files of the affected individuals using --genome option in PLINK version 1.9 (271). PLINK --genome determines the proportion of loci shared between two individuals as zero alleles (Z0), one allele (Z1), or two alleles (Z2). Another estimate resulting from this analysis is the PI-HAT (π-HAT) value, which represents the proportion of $IBD = P(IBD = 2) + 0.5*P(IBD = 1)$ between two individuals. It is another measure used to estimate pairwise genetic relatedness between individuals based on their genotypic data. PI-HAT values close to zero indicate unrelated individuals, while values close to 1 suggest a high degree of relatedness. Using R version 4.0.2 (255), the proportions of loci shared between individual pairs of the same family were plotted.

3.6.2.6 Predicting alterations in protein stability upon mutation

Missense mutations can affect protein stability and interaction with other biological molecules, eventually leading to protein dysfunction. The changes in protein stability are determined by calculating the differences in unfolding Gibbs free energy ($\Delta\Delta G$) between wild-type and mutant proteins. DynaMut web server (272) was used to computationally predict the effect of the candidate missense variants on their corresponding proteins' stabilities, conformations, and flexibilities.

3.6.2.7 Network analysis

A candidate genes interaction network was created using the GeneMANIA plugin (230) within Cytoscape software version 3.8.2 (231). Using the identified candidate genes, the plugin can detect other genes with similar roles and establish links by integrating existing genomics and proteomics information, including co-expression

associations, co-localization, genetic interactions, pathway links, physical interactions, and shared protein domains.

The generated network was then analysed by another plugin in Cytoscape software named Centiscape (273). This tool calculates various centrality measures in biological networks to identify and prioritise essential genes (nodes) based on their network connectivity and potential functional significance. I chose two parameters of centrality:

1. Degree which evaluates the regulatory importance of a gene (node) by measuring the number of its connections (edges) and
2. Betweenness, which measures a node's biological intermediary role in communication between all other nodes. It is calculated by counting the shortest paths that pass through the node.

3.6.2.8 Pathway analysis

Reactome analysis tools (263) were utilised to gain insights into the biological significance of the candidate genes. By systematically mapping these candidates to curated biological pathways, Reactome enables the identification of pathways that exhibit statistically significant enrichment among the provided gene set. This approach helps prioritise key pathways potentially relevant to the biological context and offers a comprehensive view of the functional implications and interactions between the candidate genes. Reactome's capacity to analyse overrepresented pathways aids in elucidating the underlying molecular mechanisms driving observed genetic association and contributing to a deeper understanding of disease aetiology.

3.6.3 Genome-wide association studies

Raw IDAT files were converted from the GenomeStudio Final Report file into the standard PLINK version 1.9 (271) format for subsequent analysis. The commands and R scripts used in the GWAS analyses were all executed from the Unix command line. The following commands were used for the format conversion:

```
#Convert GenomeStudio output file into the standard plink1.9 format
plink1.9 --file MYC --make-bed --allow-no-sex --out myc_gwa_1
#add sex information
plink1.9 --update-sex myc_sex_2.txt --bfile myc_gwa_1 --no-sex --no-pheno --no-
parents --make-bed --out myc_gwa_2
#add phenotype information
```

```
plink1.9 --bfile myc_gwa_2 --pheno myc_pheno_2.txt --make-bed --out myc_gwa_3
```

3.6.3.1 Data quality control

Standard GWAS quality control (274) was applied using PLINK and R version 4.0.2. This included:

1. Removal of individuals and SNPs with >2% missing genotype count using the following commands:

```
# Investigate missingness per individual and SNP
plink1.9 --bfile myc_gwa_3 --missing
# Generate plots to visualise the missingness results
Rscript --no-save hist_miss.R
# Delete SNPs with missingness >0.2.
plink1.9 --bfile myc_gwa_3 --geno 0.2 --make-bed --out myc_gwa_3_1
# Delete individuals with missingness >0.2.
plink1.9 --bfile myc_gwa_3_1  --mind 0.2 --make-bed --out myc_gwa_3_2
# Delete SNPs with missingness >0.02.
plink1.9 --bfile myc_gwa_3_2 --geno 0.02 --make-bed --out myc_gwa_3_3
# Delete individuals with missingness >0.02.
plink1.9 --bfile myc_gwa_3_3 --mind 0.02 --make-bed --out myc_gwa_3_4
```

2. Removal of individuals with conflicting sex information (discordance between self-reported and genotyped sex based on X chromosome heterozygosity/ homozygosity rates). This was done using the following commands:

```
plink1.9 --bfile myc_gwa_3_4 --check-sex
Rscript --no-save gender_check.R
plink1.9 --bfile myc_gwa_3_4 --remove sex_discrepancy.txt --make-bed –out
myc_gwa_4
```

   The sex of three individuals was wrongly assigned and therefore imputed based on the genotype information using the next command:

```
plink1.9 --bfile myc_gwa_4 --impute-sex --make-bed --out myc_gwa_5
```

3. Deletion of SNPs with a low MAF of less than 0.05 with the following commands:

```
# Select autosomal SNPs only (i.e., from chromosomes 1 to 22).
awk '{ if ($1 >= 1 && $1 <= 22) print $2 }' myc_gwa_5.bim > snp_1_22.txt
plink1.9 --bfile myc_gwa_5 --extract snp_1_22.txt --make-bed --out myc_gwa_6
# Generate a plot of the MAF distribution.
plink1.9 --bfile myc_gwa_6 --freq --out MAF_check
```

```
Rscript --no-save MAF_check.R
# Remove SNPs with low MAF frequency.
plink1.9 --bfile myc_gwa_6 --maf 0.05 --make-bed --out myc_gwa_7
```

4. Deletion of SNPs which were not in Hardy-Weinberg equilibrium (HWE). Since mycetoma is a binary trait, I excluded HWE p-value <1e−10 in cases and <1e−6 in controls. A less strict case threshold avoids discarding disease-associated SNPs under selection. This step was done using the following commands:

```
plink1.9 --bfile myc_gwa_7 --hardy
# Selecting SNPs with HWE p-value below 0.00001.
awk '{ if ($9 <0.00001) print $0 }' plink.hwe>plinkzoomhwe.hwe
Rscript --no-save hwe.R
plink1.9 --bfile myc_gwa_7 --hwe 1e-6 --make-bed --out myc_hwe_filter_step1
plink1.9 --bfile myc_hwe_filter_step1 --hwe 1e-10 --hwe-all --make-bed --out
myc_gwa_8
```

5. Removal of individuals with an outlying heterozygosity rate (deviating more than three standard deviations from the mean of the heterozygosity rate). The next commands were used:

```
plink1.9 --bfile myc_gwa_8 --exclude inversion.txt --range --indep-pairwise 50 5 0.2 --
out indepSNP
plink1.9 --bfile myc_gwa_8 --extract indepSNP.prune.in --het --out R_check
#Plot of the heterozygosity rate distribution
Rscript --no-save check_heterozygosity_rate.R
# performing statistical calculations in R:
Rscript --no-save heterozygosity_outliers_list.R
# Output of the command above: fail-het-qc.txt .
# Adapt this file to make it compatible with PLINK, by removing all quotation marks
from the file and selecting only the first two columns.
sed 's/"// g' fail-het-qc.txt | awk '{print$1, $2}'> het_fail_ind.txt
# Remove heterozygosity rate outliers.
plink1.9 --bfile myc_gwa_8 --remove het_fail_ind.txt --make-bed --out  myc_gwa_9
```

6. Duplicates with a pihat (π-hat) = 1 were removed using the following command:

```
plink1.9 --bfile myc_gwa_9 --remove fail_ibd_qc_T3.txt --make-bed --out
myc_gwa_10
```

Pi-hat is a measure used to estimate pairwise genetic relatedness between individuals based on their genotypic data. It is calculated using common SNPs shared between pairs of individuals and quantifies the proportion of alleles that are

identical by descent (IBD) between two individuals. Pi-hat values close to zero indicate unrelated individuals, while values close to 1 suggest a high degree of relatedness. To eliminate related subjects, the typically threshold is pi-hat >0.2 (274). However, since this dataset included families, I revised the command to remove only duplicates having pi-hat=1.

7. Population outliers were checked by running the multidimensional scaling (MDS) approach incorporated in PLINK using the 1000 Genomes (1KG) Project reference panel (August 2010 release) (275). The 1KG dataset includes 629 individuals descending from European, African, Admixed American, and Asian ancestries. The commands used for this step were:

```
# Extract the variants present in my dataset from the 1000 genomes dataset.
awk '{print$2}' myc_gwa_11.bim > myc_gwa_11_SNPs.txt
plink1.9 --bfile 1kG_MDS5 --extract myc_gwa_11_SNPs.txt --make-bed --out
1kG_MDS6
# Extract the variants present in the 1000 Genomes dataset from my dataset.
awk '{print$2}' 1kG_MDS6.bim > 1kG_MDS6_SNPs.txt
plink1.9 --bfile myc_gwa_11 --extract 1kG_MDS6_SNPs.txt --recode --make-bed --
out myc_MDS
# Change the 1000 Genomes data build.
awk '{print$2,$4}' myc_MDS.map > buildmyc.txt
plink1.9 --bfile 1kG_MDS6 --update-map buildmyc.txt --make-bed --out 1kG_MDS7
# Set reference genome before merging 1000 Genomes data with our dataset.
awk '{print$2,$5}' 1kG_MDS7.bim > 1kg_ref-list.txt
plink1.9 --bfile myc_MDS --reference-allele 1kg_ref-list.txt --make-bed --out myc-adj
# Check for potential strand issues.
awk '{print$2,$5,$6}' 1kG_MDS7.bim > 1kGMDS7_tmp
awk '{print$2,$5,$6}' myc-adj.bim > myc-adj_tmp
sort 1kGMDS7_tmp myc-adj_tmp |uniq -u > all_differences.txt
# Flip SNPs for resolving strand issues.
# Print SNP-identifier and remove duplicates.
awk '{print$1}' all_differences.txt | sort -u > flip_list.txt
# Flip the non-corresponding SNPs.
plink1.9 --bfile myc-adj --flip flip_list.txt --reference-allele 1kg_ref-list.txt --make-bed --
out corrected_myc
# Check for SNPs which are still problematic after they are flipped.
awk '{print$2,$5,$6}' corrected_myc.bim > corrected_myc_tmp
```

```
sort 1kGMDS7_tmp corrected_myc_tmp |uniq -u  > uncorresponding_SNPs.txt
# Remove the problematic SNPs from both datasets.
awk '{print$1}' uncorresponding_SNPs.txt | sort -u > SNPs_for_exlusion.txt
plink1.9 --bfile corrected_myc --exclude SNPs_for_exlusion.txt --make-bed --out
myc_MDS2
plink1.9 --bfile 1kG_MDS7 --exclude SNPs_for_exlusion.txt --make-bed --out
1kG_MDS8
#Perform MDS on our data anchored by 1000 Genomes data using a set of pruned
SNPs.
plink1.9 --bfile MDS_merge2 --extract indepSNP.prune.in --genome --out
MDS_merge2
plink1.9 --bfile MDS_merge2 --read-genome MDS_merge2.genome --cluster --mds-
plot 10 --out MDS_merge2
# Create a racefile of our data.
awk '{print$1,$2,"OWN"}' myc_MDS.fam>racefile_own.txt
# Concatenate racefiles
cat race_1kG14.txt racefile_own.txt | sed -e '1i\FID IID race' > racefile.txt
# Generate population stratification plot.
Rscript MDS_merged.R
```

3.6.3.2 Data analysis

3.6.3.2.1 Estimating the genomic inbreeding coefficient (F_{ROH})

Estimations of consanguinity based on family pedigrees have limitations, such as the inability to reveal information about close-kin marriages in past generations, leading to an underestimation of cumulative inbreeding effects. To overcome this challenge, the quality-controlled GWAS data was utilised to determine individual autozygosity (F_{ROH}) by identifying uninterrupted runs of homozygosity (ROH). The term ROH refers to stretches of consecutive homozygous genotypes along the genome. Whereas F_{ROH} is a calculation that provides insight into the degree of recent inbreeding or relatedness within a population or individual. Higher F_{ROH} values can indicate higher levels of inbreeding, leading to an increased risk of recessive genetic disorders due to the expression of deleterious alleles.

The estimation was performed using PLINK version 1.9 and detectRUNS R package version 0.9.6 (276). The following commands were used:

1. Calculation of the ROH using PLINK

```
plink1.9 --bfile myc_gwa_10 --homozyg --homozyg-group
```

2. Estimation of F_{ROH} in R:

```
library(detectRUNS)
# Dataset was uploaded using the following command:
t2homrun <- readExternalRuns(inputFile = "plink.hom", program = "plink")
#FROH table was generated using the following command:
Froh_inbreeding(runs = t2homrun, mapFile = "MYC.map")
#FROH plots were generated using the following command:
plot_InbreedingChr(runs = t2homrun, mapFile = "MYC.map", groupSplit = FALSE,
style = "All")
```

3.6.3.2.2 Association analysis

The quality-controlled dataset was used to run association analyses using two methods: mixed linear model association analysis (MLMA) implemented in Genome-wide Complex Trait Analysis (GCTA) software package (277) and generalised mixed model (GMM) implemented in Scalable and Accurate Implementation of Generalised mixed model (SAIGE) R package (278). The commands used for both methods are detailed in the following two sections.

3.6.3.2.2.1 Association testing using Genome-wide Complex Trait Analysis

Association analysis was performed using GCTA version 1.94.1 and R version 4.2.2. In this analysis, I used the LMM approach to perform association analysis while allowing for relatedness between individuals. The next commands were used for the analysis:

```
gcta-1.94.1 --mlma --bfile myc_gwa_10 --pheno phenos.txt --out GCTAresults
#Calculate the genomic control inflation factor (in R):
source("qqmanHJCupdated.R")
res<-read.table("GCTAresults.mlma", header=T)
head(res)
chi<-(qchisq(1-res$p,1))
lambda=median(chi)/0.456
lambda
#Produce QQ and Manhattan plots:
new<-data.frame(res$SNP, res$Chr, res$bp, res$p)
```

```
names(new)<-c("SNP", "CHR", "BP", "P")
head(new)
png("qq.png")
qq(new$P)
dev.off()
png("mh.png")
Manhattan(new, pch=20, suggestiveline=F, genomewideline=F, ymin=2,
cex.x.axis=0.65, colors=c("black","dodgerblue"), cex=0.5)
dev.off()
#Export results:
write.table(new[-log10(new$P)>5,], "SNPs_GCTA", row.names=T, col.names=T,
quote = T)
```

3.6.3.2.2.2 Association testing using Scalable and Accurate Implementation of Generalised mixed model

In the second approach, association analysis was done using SAIGE R package version 1.2.0 and R version 4.3.1. SAIGE was convenient for the mycetoma dataset since it allowed conducting association testing using a GMM and adjusting for sex and principal components 1–5. The next commands were used for the analysis:

```
#Prepare the phenotype file:
#Sex information extracted from myc_gwa_10.fam
write.table(myc_gwa_10[, c(1,2,5)], "sex_covar", row.names=F, col.names=F, quote
= F)
#Read in phenotypes and covariates in R:
library(data.table)
library(magrittr)
phenotype <- fread("phenos.txt")
covariate <- fread("sex_covar")
pcs <- fread("pcs.eigenvec")
colnames(pcs) <- c("FID", "IID", paste0("PC", 1:5))
colnames(covariate) <- c("FID","IID", "sex")
pheno <-merge(phenotype, covariate) %>% merge(., pcs)
write.table(pheno, "phenoFile", row.names=F, col.names=T, quote = F)
#STEP 1 in SAIGE:
#Fitting the null model using a full GRM that will be calculated on-the-fly using
genotypes in myc_gwa_10
```

```
library(SAIGE)
Rscript step1_fitNULLGLMM.R --plinkFile=myc_gwa_10 --phenoFile=phenoFile.txt --phenoCol=MYC --covarColList=sex,PC1,PC2,PC3,PC4,PC5 --qCovarColList=sex --sampleIDColinphenoFile=IID --traitType=binary --outputPrefix=MYC_model --nThreads=1 --skipVarianceRatioEstimation=FALSE --IsOverwriteVarianceRatioFile=TRUE
#Step 2 in SAIGE
#Perform association testing for each genetic marker by applying SPA to obtain more accurate p-values, and Firth's Penalized Likelihood to reduce bias in the estimated coefficients and p-values.
Rscript step2_SPAtests.R --bedFile=myc_gwa_10.bed --bimFile=myc_gwa_10.bim --famFile=myc_gwa_10.fam --SAIGEOutputFile=MYC_Assoc.txt --minMAF=0 --minMAC=20 --GMMATmodelFile=MYC_model.rda --varianceRatioFile=MYC_model.varianceRatio.txt --is_Firth_beta=TRUE --pCutoffforFirth=0.01 --is_output_moreDetails=TRUE --LOCO=FALSE
#calculate the genomic control inflation factor:
res<-read.table("MYC_Assoc.txt", header=T)
chi<-(qchisq(1-res$p.value,1))
lambda=median(chi)/0.456
lambda
#produce QQ and Manhattan plots:
source("qqmanHJCupdated.R")
SAIGE_new<-data.frame(res$MarkerID, res$CHR, res$POS, res$p.value)
names(SAIGE_new)<-c("SNP", "CHR", "BP", "P")
png("qqT6.png")
qq(new$P)
dev.off()
png("mh_T6.png")
manhattan(new, pch=20, suggestiveline=F, genomewideline=F, ymin=2, cex.x.axis=0.65, colors=c("black","dodgerblue"), cex=0.5)
dev.off()
#export top SNPs full info:
write.table(res[-log10(res$p.value)>5,], "SNPs_SAIGE_RAW_T6", row.names=F, col.names=T, quote = F)
```

3.6.3.2.3 Regional association

Regional association plots for the lead SNPs ($-\log_{10} P$ above 5) in our analysis were generated using LocusZoom (279). For this analysis, I considered all the SNPs within a 400 kb window of each top SNP. The resulting text files were generated using the following command as an example for one of our lead SNPs, rs167677:

```
awk '{ if ($1 == 1 && $3 > 96068117 && $3 < 96868117) print }' GCTAresults.mlma>
rs167677_LD.txt
```

These text files were uploaded to the LocusZoom website (https://my.locuszoom.org/). The African LD population was chosen on the region page, and results were exported as PNG images.

3.6.3.2.4 Annotation and functional analysis

Using GWAS summary statistics, functional mapping and annotation of genome-wide association studies (FUMA) online platform version 1.5.4 (280) was utilised to pinpoint regulatory elements, protein-coding genes, or pathways linked to the associated variants. The parameters used for this analysis can be found in Appendix 8. Multi-marker analysis of genomic annotation (MAGMA) version 1.08 (281), implanted in FUMA, was utilised to perform gene-based association analysis. This is done by aggregating the SNP-level association statistics within each gene to identify potential candidate genes significantly associated with mycetoma.

3.6.3.2.5 SNP heritability

GCTA was also used to estimate the heritability accounted for by all autosomal genome-wide SNPs via the following commands:

```
gcta-1.94.1 --bfile myc_gwa_10 --autosome --make-grm-bin --out GCTAgrm
gcta-1.94.1 --reml --grm-bin GCTAgrm --pheno phenos.txt --out GCTAherit
```

3.6.3.3 Imputation performance evaluation

Genotype imputation is a statistical method utilised to estimate the missing genotypes in GWAS datasets, thereby increasing the statistical power of association analysis using reference panels such as the 1000 Genomes Project, the Haplotype Reference Consortium (HRC), and Trans-Omics for Precision Medicine (TOPMed) reference panels (159,160,282). These panels contain sets of reference genotypes from which unobserved or missing genotypes in study sets can be inferred. The reference panels

are predominantly composed of data from populations of European ancestry, with limited data available for African populations and admixed populations of African ancestry (283). Recent years have witnessed the emergence of several reference panels representing African populations, such as the Consortium on Asthma among African ancestry populations in the Americas (CAAPA) (284) and the African Genome Resource (AGR, not publicly available). Northeastern African populations often have unique genetic variations not well represented in reference panels. Thus, this imputation accuracy assessment was done to identify how well our Sudanese dataset is imputed and whether the imputed genotypes align with their true genetic makeup represented by the WGS available data.

For this analysis, I used 17 samples from the families with multiple cases of mycetoma, which had high-coverage WGS (30x coverage) and quality-controlled GWAS data. It was performed to assess the performance of two reference panels, Trans-Omics for Precision Medicine (TOPMed; n = 97,256) and the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA; n = 883), to identify which is the optimal panel for genotype imputation of Sudanese GWAS datasets. The analysis pipeline is summarised in **Figure 3.3**.

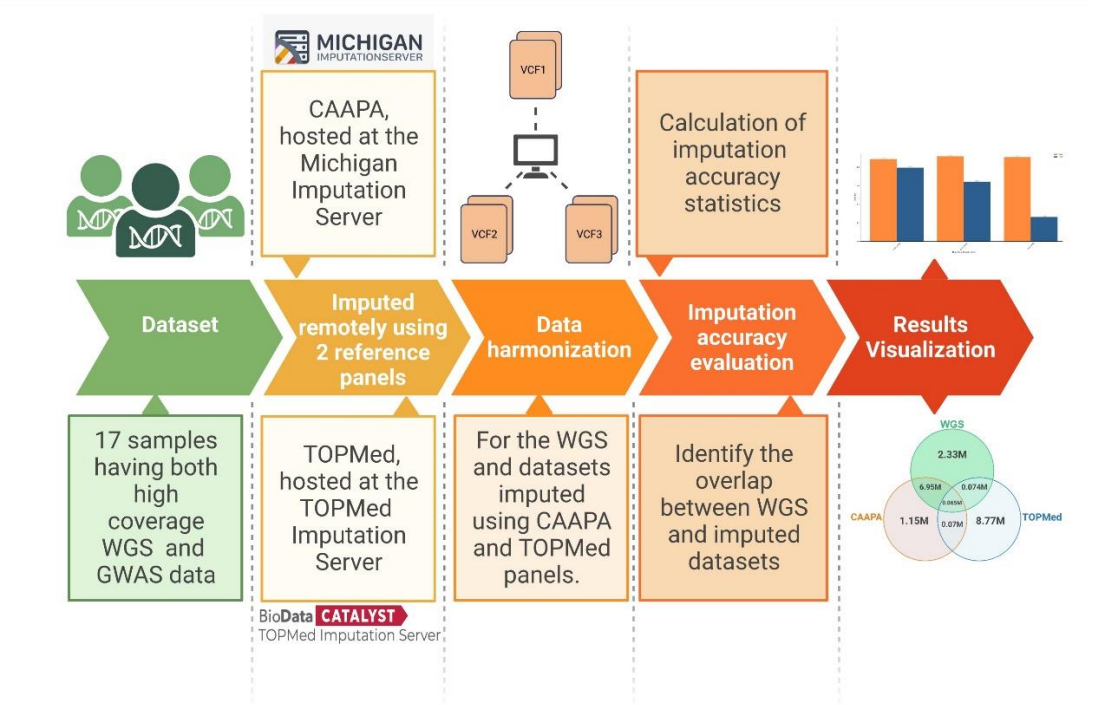


Figure 3.3 *Imputation accuracy evaluation pipeline. Figure created using BioRender.com.*

Imputation accuracy evaluation involves various metrics, including concordance rates, imputation R-squared (r^2), non-reference discordance rate (NDR), and more, which will be discussed in detail in sections 3.6.3.3.7and 3.6.3.3.10. The steps taken for this analysis were:

3.6.3.3.1 Preparing input files for imputation

i. Extract 20 WGS samples from the QCed PLINK file

```
plink1.9 --bfile myc_gwa_9 --keep Fam_ids --make-bed --out myc_gwa_fam
```

ii. compute allele frequencies

```
plink1.9 --bfile myc_gwa_fam --allow-no-sex --freq --out myc_gwa_fam.freq
```

iii. In R, extract a list of palindromic SNPs (SNPs with alleles on the positive strand that could match the alleles on the negative strand) for subsequent exclusion

```
#read in myc_gwa_fam.bim file
bim = read.table("myc_gwa_fam.bim", header = F)
#get indices of A/T and G/C SNPs
w = which ((bim$V5=="A" & bim$V6=="T") |
 (bim$V5=="T" & bim$V6=="A") |
 (bim$V5=="C" & bim$V6=="G") |
 (bim$V5=="G" & bim$V6=="C"))
```

```
#extract A/T and G/C SNPs
at.cg.snps = bim[w,]
#save A/T and G/C SNPs into a file at-cg-snps.txt
write.table(at.cg.snps$V2, "at-cg-snps.txt", row.names = F, col.names = F, quote = F)
```

iv. Run the following PERL script (available from Will Raynor (http://www.well.ox.ac.uk/~wrayner/tools/) for validating the SNPs against the HRC reference panel for strand, id names, positions and alleles

- For CAAPA:

```
perl HRC-1000G-check-bim.pl -b myc_gwa_fam.bim -f myc_gwa_fam. freq.frq -r
all.caapa.sorted.txt -h
```

- For TOPMed:

```
perl HRC-1000G-check-bim.pl -b myc_gwa_fam.bim -f myc_gwa_fam.freq.frq -r
PASS.Variants.TOPMed_freeze3_hg19_dbSNP.tab.gz -h
```

v. Run the Unix script Run-plink.sh, which contains PLINK commands, to create a VCF file myc_gwa_fam-updated-chr*.vcf (The first command below makes the script executable, while the 2nd command runs it).

```
chmod a+x Run-plink.sh
./Run-plink.sh
```

vi. Compress the VCF files using bgzip

```
for chr in {1..22}; do bgzip -c myc_gwa_fam-updated-chr$chr.vcf > myc_gwa_fam-
updated-chr$chr.impute.vcf.gz; done
```

3.6.3.3.2 Uploading samples to imputation servers

A. This step was done using Michigan imputation server for phasing and imputation:
   i. Upload the compressed files to Michigan imputation server.
   ii. Select Minimac4 as the imputation engine, Eagle 2.4 phasing and CAAPA as the reference panel.
   iii. Retain variants with imputation quality (Rsq or estimated R2) of 0.3 or greater.

B. Using TOPMed imputation server for phasing and imputation:
   i. Upload the compressed files to TOPMed imputation server.
   ii. Select Minimac4 as the imputation engine, Eagle 2.4 phasing, TOPMed reference panel “TOPMed r2” and population “vs TOPMed panel”.

iii. Retain variants with imputation quality (Rsq or estimated R2) of 0.3 or greater.

3.6.3.3.3 Extracting imputation results

i. To extract the results files

```
for zip_file in chr_*.zip; do
    unzip $zip_file
done
```

ii. Indexing dose files to combine them using bcftools

```
for chr_vcf in chr*.dose.vcf.gz; do
    bcftools index $chr_vcf
done
```

iii. Concatenate all chromosome VCF files into a single VCF file

```
bcftools concat chr1.dose.vcf.gz chr2.dose.vcf.gz chr3.dose.vcf.gz chr4.dose.vcf.gz
chr5.dose.vcf.gz chr6.dose.vcf.gz chr7.dose.vcf.gz chr8.dose.vcf.gz chr9.dose.vcf.gz
chr10.dose.vcf.gz chr11.dose.vcf.gz chr12.dose.vcf.gz chr13.dose.vcf.gz
chr14.dose.vcf.gz chr15.dose.vcf.gz chr16.dose.vcf.gz chr17.dose.vcf.gz
chr18.dose.vcf.gz chr19.dose.vcf.gz chr20.dose.vcf.gz chr21.dose.vcf.gz
chr22.dose.vcf.gz -Oz -o combined_imputed_data.vcf.gz
```

iv. Indexing the combined file

```
bcftools index combined_imputed_data.vcf.gz
```

3.6.3.3.4 Preparing the WGS files for comparison

To merge WGS VCF files from different samples into one VCF file for comparison with the imputed VCF file:

i. Indexing the VCF files for subsequent merge

```
for vcf_file in M001.snp.vcf.gz M002.snp.vcf.gz M040.snp.vcf.gz M041.snp.vcf.gz
M083.snp.vcf.gz M090.snp.vcf.gz M091.snp.vcf.gz M120.snp.vcf.gz M121.snp.vcf.gz
M159.snp.vcf.gz M167.snp.vcf.gz M213.snp.vcf.gz M228.snp.vcf.gz M236.snp.vcf.gz
M278.snp.vcf.gz M282.snp.vcf.gz M296.snp.vcf.gz; do
    bcftools index $vcf_file
done
```

ii. Merging and indexing the merged file

```
bcftools merge M001.snp.vcf.gz M002.snp.vcf.gz M040.snp.vcf.gz M041.snp.vcf.gz
M083.snp.vcf.gz M090.snp.vcf.gz M091.snp.vcf.gz M120.snp.vcf.gz M121.snp.vcf.gz
```

M159.snp.vcf.gz M167.snp.vcf.gz M213.snp.vcf.gz M228.snp.vcf.gz M236.snp.vcf.gz M278.snp.vcf.gz M282.snp.vcf.gz M296.snp.vcf.gz -Oz -o merged_wgs_data.vcf.gz

#Indexing

bcftools index merged_wgs_data.vcf.gz

3.6.3.3.5 Harmonising the WGS and imputed VCF files before comparison

A. For the WGS VCF

i. Remove sex chromosomes from WGS VCF file

bcftools view --exclude 'CHROM="chrX" || CHROM="chrY"' merged_wgs_data.vcf.gz -O z -o wgs.noXY.vcf.gz

bcftools index wgs.noXY.vcf.gz

ii. Sort the WGS file

bcftools sort wgs.noXY.vcf.gz -o wgs.noXY.sorted.vcf.gz

bcftools index wgs.noXY.sorted.vcf.gz

iii. Split multi-allelic sites WGS VCF (none were found in Imputation VCF)

bcftools norm -m -any wgs.noXY.sorted.vcf.gz -O z -o wgs.final.vcf.gz

bcftools index wgs.final.vcf.gz

iv. Create a WGS file with annotated variant IDs to match CAAPA variant IDs

bcftools annotate --rename-chrs CHR wgs.final.vcf.gz -Oz -o corrected_wgs.final.vcf.gz

bcftools annotate --set-id '%CHROM\:%POS\:%REF\:%ALT' corrected_wgs.final.vcf.gz -Oz -o wgs2.final.vcf.gz

bcftools index wgs2.final.vcf.gz

B. For the imputation VCF files

i. Keep only the 17 samples that have WGS data

bcftools view -S ^Fam_ids_rm -Oz -o imputed_filtered.vcf.gz combined_imputed_data.vcf.gz

ii. Use the reheader subcommand to change the sample names in the imputation VCF file

bcftools reheader -s change_ids_impute imputed_filtered.vcf.gz -o imputed_data.vcf.gz

iii. check if the names were correctly changed

bcftools view -h imputed_data.vcf.gz

iv. Rearrange samples order to match WGS file

bcftools view --samples-file sample_order.txt -Oz -o rearranged_imputed.vcf.gz imputed_data.vcf.gz

v. Sort and index

bcftools sort -Oz -o imputed_rearranged_sorted.vcf.gz rearranged_imputed.vcf.gz

3.6.3.3.6 Comparing the files

i. For CAAPA imputation file, caapa_sorted.vcf.gz:

Use bcftools isec to identify common and unique variants between the WGS and imputed VCF files:

bcftools isec -p vcf_comparison2 -Oz wgs2.final.vcf.gz caapa_sorted.vcf.gz

The output files are:

0000.vcf: Variants unique to WGS file.

0001.vcf: Variants unique to CAAPA imputation file.

0002.vcf: Variants in the wgs2.final.vcf.gz shared by both files.

0003.vcf: Variants in the caapa_sorted.vcf.gz shared by both files.

ii. For TOPMed imputation file, topmed_sorted.vcf.gz:

bcftools isec -p vcf_comparison2 -Oz wgs.final.vcf.gz topmed_sorted.vcf.gz

The output files are:

0000.vcf: Variants unique to WGS file.

0001.vcf: Variants unique to TOPMed imputation file.

0002.vcf: Variants in the wgs2.final.vcf.gz shared by both files.

0003.vcf: Variants in the topmed_sorted.vcf.gz shared by both files.

3.6.3.3.7 Using the imputation accuracy calculator

Imputation accuracy was evaluated using the imputation accuracy calculation software available on GitHub (285). Only SNPs common to the imputed and sequence data were considered in this evaluation; therefore, 0003.vcf.gz output files from the previous step were used as the imputation input files. The following steps were taken to calculate the imputation accuracy:

i. Remove missing genotypes from the WGS file prior to comparison

bcftools filter -e 'GT="./."' -o wgs.final.nomiss.vcf.gz wgs2.final.vcf.gz

ii. Indexing WGS input file using tabix

tabix -p vcf wgs.final.nomiss.vcf.gz

iii. Running the comparison Python script provided on the GitHub page

python3 Compare_imputation_to_WGS.py --wgs wgs.final.nomiss.vcf.gz --imputed vcf_comparison/0003.vcf.gz

The output files (per_sample_results.txt and per_variant_results.txt) consisted of comprehensive reports that include the concordance rate, square correlation (r^2), precision, recall, Imputation Quality Score (IQS) and Root Mean Squared Error (RMSE). The concordance rate measures the proportion of concordant genotypes between imputed and true genotypes (derived from the WGS data), while the square correlation, also known as r^2, refers to the Pearson correlation coefficient between the imputed and genotyped dosage of every SNP, squared. Precision is the ratio of true positive predictions to the total predicted positives and recall is the ratio of true positive predictions to the actual positive samples. The IQS parameter provides an overall assessment of imputation accuracy by combining various metrics, such as concordance, correlation, and accuracy, into a single score, providing an overall assessment of imputation performance. Additionally, RMSE quantifies the difference between imputed and true genotypes, further evaluating imputation performance.

3.6.3.3.8 Visualising the comparison results

To accurately assess IQS and r^2 results, I excluded all variants with MAF equal to zero from the analysis. These two parameters are not indicative of accuracy or imputation quality when MAF is equal to zero.

All graphs were created using the following codes in Python 3 (1):

i. For squared correlation (r^2) versus root mean squared error (RMSE) plot:
To demonstrate how MAF affects imputation accuracy and error rates, I plotted RMSE against r^2 using the following Python script:

```
import pandas as pd
import matplotlib.pyplot as plt
# Load data from the text file into a pandas DataFrame
data = pd.read_csv( '0003.vcf_per_variant_results.txt', sep='\t')
# Extract r2 and error columns from the DataFrame
r2 = data['r2']
error = data['RMSE']
# Filter out variants with MAF = 0
data_filtered = data[data['WGS_MAF'] > 0]
# Calculate MAF bins based on the error values
maf_bins = [0, 0.05, 0.1, 0.2, 0.5, 1]
```

```
data_filtered['MAF_bin'] = pd.cut(data_filtered['RMSE'], bins=maf_bins,
labels=maf_bins[:-1])
# Set up colors for different MAF bins
colors = plt.cm.tab10.colors[:len(maf_bins)-1]
# Create the scatter plot with different colors for each MAF bin
for i, (bin_start, bin_end) in enumerate(zip(maf_bins[:-1], maf_bins[1:])):
    subset_data = data_filtered[(data_filtered['RMSE'] >= bin_start) &
(data_filtered['RMSE'] < bin_end)]
    plt.scatter(subset_data['RMSE'], subset_data['r2'], s=20, alpha=0.5, label=f'MAF
[{bin_start}, {bin_end})', color=colors[i])
plt.xlabel('Root Mean Squared Error (RMSE)')
plt.ylabel('r2')
plt.legend()
plt.grid(True)
plt.show()
```

Then the following command was used to run the Python script on the terminal:

```
python3 r2VSerror.py
```

ii. For comparing the mean concordance rate by MAF stratification of TOPMed and CAAPA imputed variants, the following code was used:

```
import pandas as pd
import matplotlib.pyplot as plt
# Load concordance data for TOPMed from the text file
topmed_data = pd.read_csv('TOPMed0003.vcf_per_variant_results.txt', sep='\t')

# Load concordance data for CAAPA from the text file
caapa_data = pd.read_csv('CAAPA0003.vcf_per_variant_results.txt', sep='\t')
# Define the MAF bins
maf_bins = [0.01, 0.05, 0.25, 0.5]
# Bin the data based on WGS MAF values for TOPMed
topmed_data['MAF_Bin'] = pd.cut(topmed_data['WGS_MAF'], bins=maf_bins)
# Bin the data based on WGS MAF values for CAAPA
caapa_data['MAF_Bin'] = pd.cut(caapa_data['WGS_MAF'], bins=maf_bins)
# Group data by MAF bins and calculate the mean concordance rate in each bin for
TOPMed
binned_topmed_data =
topmed_data.groupby('MAF_Bin')['concordance_P0'].mean().reset_index()
```

```
# Group data by MAF bins and calculate the mean concordance rate in each bin for
CAAPA
binned_caapa_data =
caapa_data.groupby('MAF_Bin')['concordance_P0'].mean().reset_index()
# Create the bar plot with grouped bars
fig, ax = plt.subplots(figsize=(10, 6))
bar_width = 0.2  # Adjust the bar width for closer bars
topmed_bars = plt.bar(range(len(binned_topmed_data)),
binned_topmed_data['concordance_P0'], width=bar_width, align='center',
label='TOPMed')
caapa_bars = plt.bar([x + bar_width for x in range(len(binned_caapa_data))],
binned_caapa_data['concordance_P0'], width=bar_width, align='center',
label='CAAPA')
plt.xlabel('Minor allele frequency bins')
plt.ylabel('Mean Concordance Rate')
plt.title('Mean Concordance Rate between Imputed and WGS Genotypes by MAF
Bins')
plt.xticks([x + bar_width / 2 for x in range(len(binned_topmed_data))],
binned_topmed_data['MAF_Bin'])
plt.legend()
# Add text annotations for the mean concordance values on the bars
for bar, concordance_rate in zip(topmed_bars,
binned_topmed_data['concordance_P0']):
    plt.text(bar.get_x() + bar.get_width() / 2, concordance_rate,
f'{concordance_rate:.2f}', ha='center', va='bottom', color='#2C5E90')
for bar, concordance_rate in zip(caapa_bars,
binned_caapa_data['concordance_P0']):
    plt.text(bar.get_x() + bar.get_width() / 2, concordance_rate,
f'{concordance_rate:.2f}', ha='center', va='bottom', color='#FF8A3B')
plt.tight_layout()
plt.show()
```

Then the following command was used to run the Python script on the terminal:

```
python3 3_TopMed_CAAPA_conc.py
```

iii. For MAF bins against:

a) Root mean squared error (RMSE):

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
# Load concordance data for CAAPA from the text file
caapa_data = pd.read_csv('CAAPA0003.vcf_per_variant_results.txt', sep='\t')
# Load concordance data for TOPMed from the text file
topmed_data = pd.read_csv('TOPMed0003.vcf_per_variant_results.txt', sep='\t')
# Define the MAF bins
maf_bins = [0.01, 0.05, 0.25, 0.5]  # Customize the MAF bins as needed
# Bin the data based on WGS MAF values for CAAPA
caapa_data['MAF_Bin'] = pd.cut(caapa_data['WGS_MAF'], bins=maf_bins)
# Bin the data based on WGS MAF values for TOPMed
topmed_data['MAF_Bin'] = pd.cut(topmed_data['WGS_MAF'], bins=maf_bins)
# Group data by MAF bins and calculate the mean RMSE in each bin for CAAPA
binned_caapa_data = caapa_data.groupby('MAF_Bin')['RMSE'].mean().reset_index()
# Group data by MAF bins and calculate the mean RMSE in each bin for TOPMed
binned_topmed_data =
topmed_data.groupby('MAF_Bin')['RMSE'].mean().reset_index()
# Create the bar plot comparing mean RMSE between CAAPA and TOPMed
plt.figure(figsize=(10, 6))
plt.bar(range(len(binned_caapa_data)), binned_caapa_data['RMSE'], align='center',
width=0.4, label='CAAPA', color='#FF8A3B')
plt.bar([x + 0.4 for x in range(len(binned_topmed_data))],
binned_topmed_data['RMSE'], align='center', width=0.4, label='TOPMed',
color='#2C5E90')
plt.xticks([x + 0.2 for x in range(len(binned_caapa_data))],
binned_caapa_data['MAF_Bin'], rotation=45)
plt.xlabel('Minor allele frequency bins')
plt.ylabel('Mean RMSE')
plt.title('Mean RMSE between Imputed and WGS Genotypes by MAF Bins (CAAPA
vs TOPMed)')
plt.legend()
# Add text annotations for the mean RMSE values on the bars
for x, y in enumerate(binned_caapa_data['RMSE']):
    plt.text(x, y, f'{y:.4f}', ha='center', va='bottom', color='black')
for x, y in enumerate(binned_topmed_data['RMSE']):
    plt.text(x + 0.4, y, f'{y:.4f}', ha='center', va='bottom', color='black')
plt.tight_layout()
plt.show()
```

Then the following command was used to run the Python script on the terminal:

```
python3 MAFvsRMSE2.py
```

b) Imputation quality score (IQS):

```
import pandas as pd
import matplotlib.pyplot as plt
# Load concordance data for CAAPA from the text file
caapa_data = pd.read_csv('CAAPA0003.vcf_per_variant_results.txt', sep='\t')
# Load concordance data for TOPMed from the text file
topmed_data = pd.read_csv('TOPMed0003.vcf_per_variant_results.txt', sep='\t')
# Define the MAF bins
maf_bins = [0.01, 0.05, 0.25, 0.5]  # Customize the MAF bins as needed
# Bin the data based on WGS MAF values for CAAPA
caapa_data['MAF_Bin'] = pd.cut(caapa_data['WGS_MAF'], bins=maf_bins)
# Bin the data based on WGS MAF values for TOPMed
topmed_data['MAF_Bin'] = pd.cut(topmed_data['WGS_MAF'], bins=maf_bins)
# Group data by MAF bins and calculate the mean IQS in each bin for CAAPA
binned_caapa_data = caapa_data.groupby('MAF_Bin')['IQS'].mean().reset_index()
# Group data by MAF bins and calculate the mean r2 in each bin for TOPMed
binned_topmed_data = topmed_data.groupby('MAF_Bin')['IQS'].mean().reset_index()
# Create the bar plot comparing mean r2 between CAAPA and TOPMed
plt.figure(figsize=(10, 6))
plt.bar(range(len(binned_caapa_data)), binned_caapa_data['IQS'], align='center',
width=0.4, label='CAAPA', color='#FF8A3B')
plt.bar([x + 0.4 for x in range(len(binned_topmed_data))],
binned_topmed_data['IQS'], align='center', width=0.4, label='TOPMed',
color='#2C5E90')
plt.xticks([x + 0.2 for x in range(len(binned_caapa_data))],
binned_caapa_data['MAF_Bin'], rotation=45)
plt.xlabel('Minor allele frequency bins')
plt.ylabel('Mean IQS')
plt.title('Mean IQS between Imputed and WGS Genotypes by MAF Bins (CAAPA vs
TOPMed)')
plt.legend()
# Add text annotations for the mean IQS values on the bars
for x, y in enumerate(binned_caapa_data['IQS']):
    plt.text(x, y, f'{y:.2f}', ha='center', va='bottom', color='#FF8A3B')
for x, y in enumerate(binned_topmed_data['IQS']):
    plt.text(x + 0.4, y, f'{y:.2f}', ha='center', va='bottom', color='#2C5E90')
```

```
plt.tight_layout()
plt.show()
```

Then the following command was used to run the Python script on the terminal:

```
python3 MAFvsIQS.py
```

c) Squared correlation (r^2):

```
import pandas as pd
import matplotlib.pyplot as plt
# Load concordance data for CAAPA from the text file
caapa_data = pd.read_csv('CAAPA0003.vcf_per_variant_results.txt', sep='\t')
# Load concordance data for TOPMed from the text file
topmed_data = pd.read_csv('TOPMed0003.vcf_per_variant_results.txt', sep='\t')
# Define the MAF bins
maf_bins = [0.01, 0.05, 0.25, 0.5]  # Customize the MAF bins as needed
# Bin the data based on WGS MAF values for CAAPA
caapa_data['MAF_Bin'] = pd.cut(caapa_data['WGS_MAF'], bins=maf_bins)
# Bin the data based on WGS MAF values for TOPMed
topmed_data['MAF_Bin'] = pd.cut(topmed_data['WGS_MAF'], bins=maf_bins)
# Group data by MAF bins and calculate the mean r2 in each bin for CAAPA
binned_caapa_data = caapa_data.groupby('MAF_Bin')['r2'].mean().reset_index()
# Group data by MAF bins and calculate the mean r2 in each bin for TOPMed
binned_topmed_data = topmed_data.groupby('MAF_Bin')['r2'].mean().reset_index()
# Create the bar plot comparing mean r2 between CAAPA and TOPMed
plt.figure(figsize=(10, 6))
plt.bar(range(len(binned_caapa_data)), binned_caapa_data['r2'], align='center',
width=0.4, label='CAAPA', color='#FF8A3B')
plt.bar([x + 0.4 for x in range(len(binned_topmed_data))], binned_topmed_data['r2'],
align='center', width=0.4, label='TOPMed', color='#2C5E90')
plt.xticks([x + 0.2 for x in range(len(binned_caapa_data))],
binned_caapa_data['MAF_Bin'], rotation=45)
plt.xlabel('Minor allele frequency bins')
plt.ylabel('Mean r2')
plt.title('Mean r2 between Imputed and WGS Genotypes by MAF Bins (CAAPA vs
TOPMed)')
plt.legend()
# Add text annotations for the mean r2 values on the bars
for x, y in enumerate(binned_caapa_data['r2']):
    plt.text(x, y, f'{y:.2f}', ha='center', va='bottom', color='black')
```

```
for x, y in enumerate(binned_topmed_data['r2']):
    plt.text(x + 0.4, y, f'{y:.2f}', ha='center', va='bottom', color='black')
plt.tight_layout()
plt.show()
```

Then the following command was used to run the Python script on the terminal:

```
python3 MAFvsr2.py
```

3.6.3.3.9 Creating a Venn diagram showing the overlap of SNPs between the WGS and imputed datasets

This was done using the following steps:

i. Harmonize variant id in TOPMed vcf.gz file to match CAAPA and WGS files:

```
bcftools annotate --rename-chrs CHR topmed_sorted.vcf.gz -Oz -o
corrected_TOPMed.vcf.gz
bcftools index corrected_TOPMed.vcf.gz
```

ii. Annotate the snp Ids in wgs vcf to match snp vcf:

```
bcftools annotate --set-id '%CHROM\:%POS\:%REF\:%ALT'
corrected_TOPMed.vcf.gz -Oz -o TOPMed.vcf.gz
bcftools index TOPMed.vcf.gz
```

iii. Identify Variants common in the three files:

```
bcftools isec -n=3 wgs2.final.vcf.gz caapa_sorted.vcf.gz TOPMed.vcf.gz -p
vcf_comparison3
wc -l vcf_comparison3/sites.txt
```

iv. Identify Variants common between each two of the three files:

```
bcftools isec -n=2 caapa_sorted.vcf.gz TOPMed.vcf.gz -p
vcf_comparison_caapa_vs_topmed
bcftools isec -n=2 wgs2.final.vcf.gz TOPMed.vcf.gz -p
vcf_comparison_topmed_vs_wgs && bcftools isec -n=2 wgs2.final.vcf.gz
caapa_sorted.vcf.gz -p vcf_comparison_caapa_vs_wgs
```

- For CAAPA versus WGS:

```
wc -l vcf_comparison_caapa_vs_wgs/sites.txt
```

- For TOPMed versus WGS:

```
wc -l vcf_comparison_topmed_vs_wgs/sites.txt
```

- For TOPMed versus CAAPA:

```
wc -l vcf_comparison_caapa_vs_topmed/sites.txt
```

v. Calculate unique variants in each file by firstly identifying the total number of variants:

- For WGS:

bcftools stats wgs2.final.vcf.gz> WGS_stat.txt

- For CAAPA:

bcftools stats caapa_sorted.vcf.gz> CAAPA_stat.txt

- For TOPMed:

bcftools stats TOPMed.vcf.gz> TOPMed_stat.txt

3.6.3.3.10 Calculating the average Non−reference Discordance Rate (NDR)

To calculate the average NDR across the 17 samples, bcftools stats was used following these steps:

i. For WGS vs TOPMed

bcftools stats --samples-file sample_order.txt wgs2.final.vcf.gz TOPMed.vcf.gz > TOPMED_WGS-IMP_comp.txt'

ii. For WGS vs CAAPA

bcftools stats --samples-file sample_order.txt wgs2.final.vcf.gz caapa_sorted.vcf.gz > CAAPA_WGS-IMP_comp.txt

iii. Extract the NDR table to CAAPA2_WGS-IMP_comp.txt and 'TOPMED2_WGS-IMP_comp.txt'

iv. Visualise the results using the following Python script:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
# Load the bcftools stats output for TOPMed into a DataFrame
topmed_data = pd.read_csv('TOPMED2_WGS-IMP_comp.txt', sep='\t')
# Load the bcftools stats output for CAAPA into a DataFrame
caapa_data = pd.read_csv('CAAPA2_WGS-IMP_comp.txt', sep='\t')
# Extract the NDR columns
topmed_ndr = topmed_data['non-reference discordance rate']
caapa_ndr = caapa_data['non-reference discordance rate']
# Create a DataFrame for plotting
plot_data = pd.DataFrame({'TOPMed': topmed_ndr, 'CAAPA': caapa_ndr})
# Create a violin plot with a box plot inside
plt.figure(figsize=(10, 6))
sns.violinplot(data=plot_data, inner='box')
plt.xlabel('Dataset')
plt.ylabel('Non-Reference Discordance Rate (NDR)')
```

```
plt.title('Distribution of Non-Reference Discordance Rate (NDR) between TOPMed
and CAAPA')
plt.tight_layout()
plt.show()
```

v. The following command was used to run the Python script on the terminal:
`python3 NDR5.py`

3.7 Concluding remarks and challenges

In recent years, Sudan has faced significant political unrest, marked by protests against the Al-Bashir regime in 2018, a military coup in October 2021, and an ongoing conflict between the Sudanese armed forces and the Rapid Support Forces militia in Khartoum state as of April 15, 2023. The COVID-19 pandemic and subsequent lockdown in 2020 have only added to the complexity of the situation. During our sample collection between July 2019 and March 2023, we encountered challenges due to the difficult circumstances, which unfortunately prevented us from obtaining the necessary sample size for our GWAS study. This was mainly caused by the loss of the second batch of samples in the war in Sudan. It is important to note that HIV was not tested in this study due to its extremely low rate of only 0.1% in Sudan (286).

CHAPTER 4 - Genetic basis of mycetoma susceptibility explored through familial studies

4.1 Overview

As outlined in the first two chapters, evidence suggests a genetic component for mycetoma; however, no studies have been conducted to estimate its size. This chapter primarily emphasises 46 carefully curated pedigrees of families with more than one case of mycetoma. The familial aggregation study first concentrated on two factors, namely the sibling recurrence risk (K_s) and the sibling recurrence risk ratio (λ_s). The former determined the likelihood of a disease recurring in siblings of an affected person, while the latter involved dividing the sibling recurrence risk (K_s) by the population prevalence of the disease (144). Both estimates were calculated using an adaptation of Olson and Cordell's formula described in the methods chapter. Another estimate quantified in this study was heritability (h^2), which measures the variability in an individual's likelihood of developing a disease that can be attributed to their genetic makeup (287). This quantity was estimated using three software programs, SOLAR, Mendel and ASSOC, implemented in the SAGE software package.

The second part of this study employed convenience sampling methodology, wherein 22 individuals hailing from four families (with a minimum of four mycetoma cases) were selected for WGS based on the data gathered from the 46 pedigrees mentioned earlier.

The chapter is structured into the following sections:

- Section one: Summarises the pedigrees' characteristics and presents the calculated estimates.
- Section two: Reports the WGS data results for each family.
- Section three: Demonstrates the resulting network and pathway analyses based on the WGS data findings.
- Section four: Emphasises the results' significance and addresses the limitations inherent to the familial-based approaches.

4.2 Familial aggregation analyses

In 2017, based on several studies reporting mycetoma clustered in families (9,55,56), I started an exploratory analysis of the records for 7296 mycetoma patients housed within the MRC surveillance database to investigate familial aggregation. This

investigation identified 13.5% of patients with a familial history of mycetoma, with this figure increasing to 15% for patients diagnosed with *M. mycetomatis* eumycetoma (Mycetoma Research Centre, 2017, unpublished data). Building on these preliminary insights, I initially focused on estimating the contribution of host genetics to mycetoma susceptibility using familial aggregation. This analysis identifies the clustering of diseases within families, laying the foundation for subsequent studies using WGS and GWAS (288).

4.2.1 Pedigrees statistics

Characteristics of the 46 pedigrees are summarised in **Table 4.1**. Fifteen per cent of individuals whose disease status was recorded had mycetoma (59.6% of males and 40.4% of females). Thus, the sex ratio of the cases in this study was male to female, 1.47:1.0, whereas the sex ratio for all individuals was 1.09:1 (male: female). The mean (SD) age for males with mycetoma (when the data was collected) was 32.81 (13.14), while affected females had a mean age of 32.69 (14.65). Consanguineous marriages were found in ten pedigrees with seventeen loops of kin marriages (fifteen between first cousins and two between second cousins), leading to the software's inability to define the number of generations for those pedigrees. An example of a loop in one of the ten pedigrees is demonstrated in **Figure 4.1**.

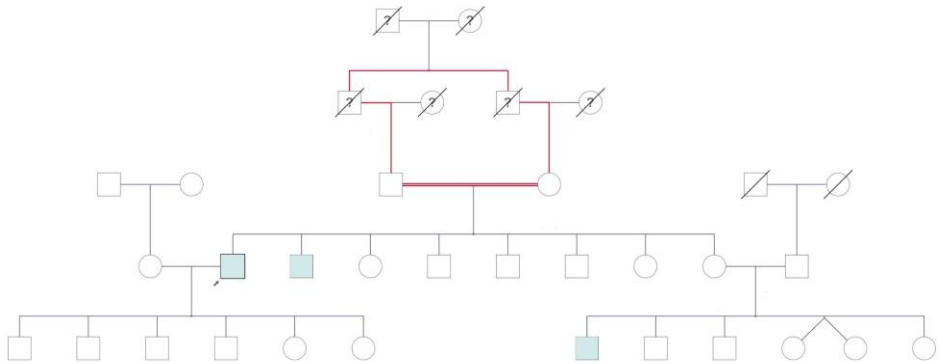


Figure 4.1 *One of the ten pedigrees with a loop (highlighted in red) that represents a consanguineous marriage. Round symbols indicate female; square, male; diamond, unspecified gender; slash through circle or square: dead female or male; blue-filled symbols, mycetoma; unfilled symbols, unaffected; question mark within the symbol, unknown affection status; the number immediately under the symbol, identity number of an individual within pedigree; double lines, consanguinity; arrow, proband.*

Twenty-four families had data collected over three generations, and twelve over four generations. The mean (SD) family size was 22.26 (21.03) individuals, ranging from 9 to 112 family members. Approximately 32% of individuals in our dataset were founders (did not have known ascendants), whilst 68% were non-founders with at least one known ascendant. No singletons – i.e., cases whose pedigrees contain only one individual – were included in our dataset.

Table 4.1 *Characteristics of members of mycetoma-affected families.*

Characteristic	**n**	**%**
Sex (n = 1024 individuals)		
Female	488	47.66
Male	536	52.34
Mycetoma (n = 1024 individuals)		
Affected	156	15.23
Unaffected	580	56.64
Unknown	288	28.12
Individuals (n = 1024 individuals)		
Founder	325	31.74
Non-founder	699	68.26
Singletons	0	0
Types of affected relative pair (n = 194 affected pairs)		
Full sibling	67	34.54
Half-sibling	5	2.57
Parent-offspring	19	9.79
Avuncular	34	17.53
Grandparent–grandchild	1	0.52
Cousin	68	35.05
Generations in pedigree (n = 46 families)		
Three	24	52.17
Four	12	26.09
Undefined (containing loops)	10	21.74

Avuncular: uncle/aunt to niece/nephew

Two hundred and ninety-three sibships were recorded, with a mean (SD) size of 2.39 (2.31), ranging from 1 to 10 siblings, and 67 affected full-sibling pairs. The dataset's remaining types of affected relative pairs included 35.05% cousins, 17.53% avuncular

pairs (i.e., uncle/aunt to niece/nephew), 9.79% parent–offspring pairs, 2.57% half-sibling pairs, and 0.52% grandparent–grandchild pairs.

4.2.2 Sibling recurrence risk

The sibling recurrence risk (K_S) in this dataset was estimated to have lower and upper bounds of 0.16 and 0.25 under single and complete ascertainment, respectively. Hence, 16%–25% of siblings of probands were themselves affected, a rate much higher than the estimated population risk of 0.87% from a previous study conducted in the same endemic area where most of this study pedigrees were collected (57). Consequently, the sibling recurrence risk ratio (λ_s) for mycetoma ranged from 18.4–28.7 i.e., the risk of a sibling of an affected individual being affected is at least 18 times higher than the risk for a member of the general population.

4.2.3 Heritability

Heritability (h^2) elucidates the degree to which differences in genetic factors contribute to the observed variations in disease susceptibility within a given population. The heritability of mycetoma was estimated with two commonly used genetic analysis packages, SOLAR and Mendel. Additionally, despite being relatively older in software development, the ASSOC program within the SAGE software package was utilised due to its robust methodology and established reputation for conducting heritability estimation in family-based genetic studies.

Mycetoma heritability estimated by SOLAR was 0.0, with a (nonsensical) standard error of 0.0. To verify this result, Mendel software was used and gave a similar result. Conversely, the ASSOC program gave a heritability estimation of 0.23 (SE 0.06, $p = 4.14\times10^{-4}$), i.e., about 23% of mycetoma susceptibility variance is accountable to genetic factors. The discrepancy between the utilised programs is most likely due to prior assumptions of no consanguinity that render them incapable of capturing specific genetic effects, leading to an underestimated or zero heritability estimate.

4.3 Whole genome sequencing study

For the next phase of my familial investigation, I enlisted the expertise of BGI Genomics to perform WGS on 22 individuals from four families with multiple cases of mycetoma. My objective was to determine if rare genetic variants were present in the genomes of individuals with mycetoma compared to their healthy family members.

After analysing the WGS data, I predicted how the identified single nucleotide variants affected protein secondary structure through amino acid substitutions. Additionally, I conducted network and pathway analyses using the candidate variants.

4.3.1 Per family analyses

The average number of variants detected across the four multicase families was 4,697,342, including SNPs, InDels, CNVs and SVs. The process for identifying candidate variants in each family is outlined in **Figure 4.2**. In short, I selected variants seen in the cases but not in their related controls. These variants were then filtered and prioritised based on their MAFs of less than 1%, pathogenicity classification (reported and/or predicted), evolutionary conservation scores, and identification in publicly available databases and literature data. The resulting candidate variants detected in the mycetoma cases are listed in **Table 4.8**. All candidate SNPs and InDels had a total read depth (number of reads that support the variant call) of >20, supporting the validity of the identified variants.

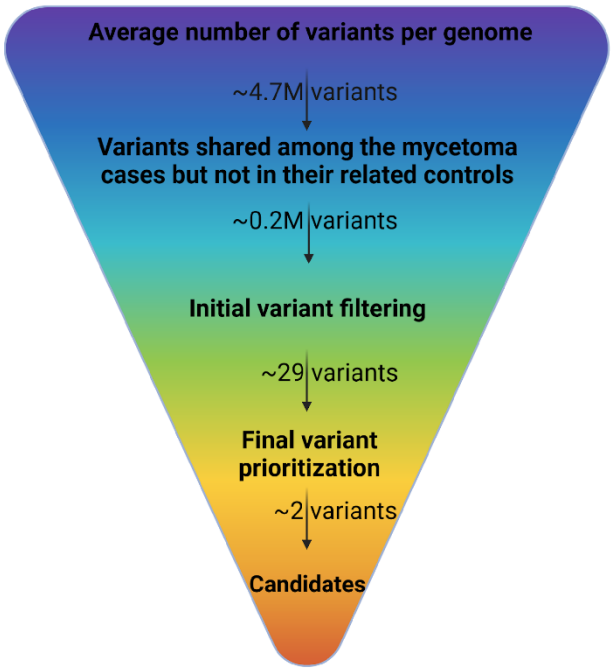


Figure 4.2 *Whole genome sequencing (WGS) data analysis and variant prioritisation scheme. The number of variants after each step represents the average across all four families.*

4.3.1.1 Family 1

Ten members of this family, spanning two generations, were affected by mycetoma (**Figure 4.3**). Among them were two brothers (001 and 011) and two maternal first

cousins (083 and 025), who all had a severe form that resulted in amputation of their lower limbs, and a third cousin (002) who had a mild lesion that was surgically removed. It is worth noting that due to insufficient public knowledge about mycetoma in the past, the grandparents' generation was marked with question marks as they may have contracted the infection unknowingly.

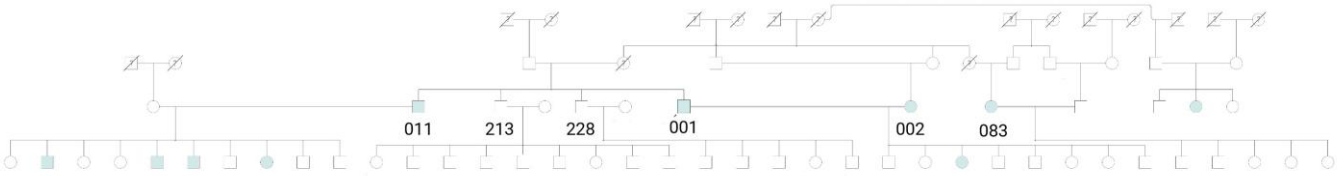


Figure 4.3 *Family 1 pedigree. Round symbols indicate female; square, male; diamond, unspecified gender; slash through circle or square: dead female or male; blue-filled symbols, mycetoma; unfilled symbols, unaffected; question mark within a symbol, unknown affection status; a number immediately under a symbol, identity number of an individual within pedigree; double lines, consanguinity; arrow, proband.*

The participants enlisted for the WGS investigation comprised the two affected siblings (001 and 011), two of their maternal first cousins (002 and 083), and two healthy control siblings (213 and 228), as illustrated in **Figure 4.4**.

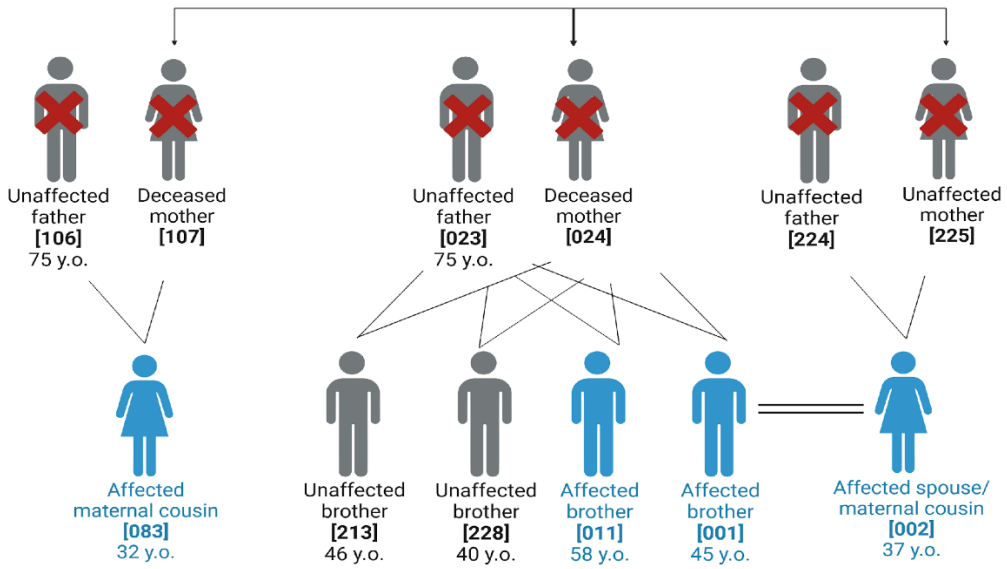


Figure 4.4 *Family 1 members recruited in the WGS study. The mycetoma cases (highlighted in blue) included two brothers, identified as 001 and 011, and two maternal cousins, 002 and 083. Two healthy brothers are designated as controls and identified as 213 and 228 (shown in dark grey)—Y.o., years old. The X mark is for individuals not recruited in the study due to death or unavailability.*

The number of variants after each step of the filtering and prioritisation is summarised in **Table 4.2**. It is worth mentioning that the inclusion of the two cousins in the analysis led to a significant reduction in the number of genetic variants to a total of 41,252 variants after the subsetting step. This was the lowest number among the four families, with Family 2 having 391,723 variants, Family 3 had 177,321 variants, and Family 4 had 198,591 variants.

Table 4.2 *The number of variants after the analysis of Family 1 genomes.*

Family 1	**SNPs**	**InDels**	**CNVs**	**SVs**
The average number of variants called in all samples	3,749,568	937,629	4,426	5,719
Number of variants after subsetting	30240	9829	600	583
Number of variants after initial filtering	15	9	1	0
Number of variants after final prioritisation	1	0	1	0

SNPs: single nucleotide polymorphism; InDels: insertion/deletions; CNVs: copy number variants; SVs: structural variants.

Although no variants were detected in all the affected individuals, an analysis revealed the presence of two heterozygote variants (CNV and SNP) in the affected siblings (001 and 011). The first variant was a 7.4-kilobase (-kb) duplication (chr8:6871802-6879300-DUP) that resulted in transcript amplification of the defensin alpha 3 (*DEFA3)* gene. This gene is a member of the alpha-defensin gene family, encoding human neutrophil peptides (HNPs) stored in the azurophilic granules of human neutrophils (289). The granules containing HNPs usually undergo limited secretion and are typically directed to fuse with phagolysosomes, where high concentrations of HNPs can directly kill phagocytosed microbes (290,291). This process involves defensins forming hexamers that create pores in the pathogen's cell membrane, resulting in permeabilisation and cell death (292,293). Alpha-defensins are well known to exhibit CNVs (294). A study conducted by Chen et al. in 2019 found that mice exhibiting high copy numbers of defensin alpha 1 (*DEFA1)* and *DEFA3* were more susceptible to severe sepsis and had higher mortality rates than those with low copy numbers and wild-type mice (295). This functional study followed a previous one in 2010 when the same team investigated the relationship between CNVs in the two genes and the susceptibility to severe sepsis in the Chinese Han population (296). The study found

an increased genomic copy number of *DEFA1/DEFA3* was associated with an elevated risk of severe sepsis in that population. The (chr8:6871802-6879300-DUP) CNV, identified in this study, may have a comparable effect on the immune response of the affected sibling.

The two affected brothers were heterozygote carriers of a missense variant in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene. The encoded protein functions as an anion channel that controls ion and water secretion and absorption in the apical plasma membrane of epithelial cells (297). According to the GTEx database, *CFTR* is mainly expressed in the pancreas, colon, minor salivary gland, small intestine terminal ileum, lung, and sun-exposed skin (lower leg) (298). Thousands of variants in *CFTR* are linked with cystic fibrosis (OMIM 219700), an autosomal recessive disease most commonly found in populations of Northern European ancestry (299). The two patients were free of symptoms of cystic fibrosis. The clinical significance of the identified loss of function (LOF) variant NM_000492.4:c.3485G>A (p.Arg1162Gln) according to ACMG criteria in Varsome (last accessed: 22nd May 2023) was likely pathogenic (detailed criteria in Table 4.8). The variant was predicted to be pathogenic by eighteen in silico pathogenicity engines, including SIFT, MutPred, FATHMM-MKL, FATHMM-XF, LIST-S2, M-CAP, MVP, DEOGEN2, and EIGEN PC. The variant had a very low allele frequency in the gnomAD genomes version 2.1.1 database of 0.0002004 (last accessed: 8th October 2023) in the African/African American dataset (300).

For the two other patients, the affected cousin 002 had a homozygous missense variant in the pleckstrin homology domain-containing family O member 2 (*PLEKHO2*) gene, the same candidate variant identified in Family 3. The affected siblings who shared the same variant in Family 3 (152 and 167) were her paternal first cousins. The pedigree diagram in **Figure 4.5** shows how 002 connects Family 1 and Family 3. In contrast, I did not identify any candidate variant in the other cousin (083).

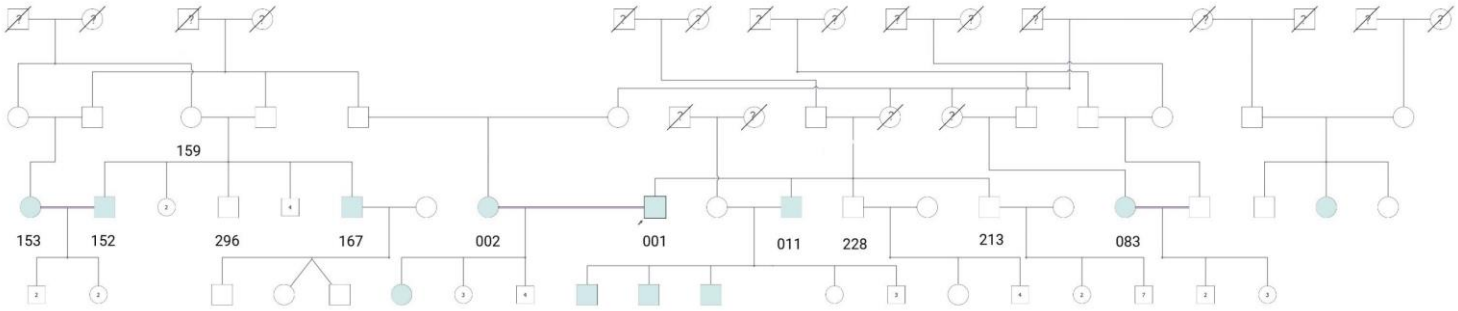


Figure 4.5 *A pedigree demonstrating how the affected cousin 002 connects Family 1 and Family 3. Round symbols indicate female; square, male; diamond, unspecified gender; slash through circle or square: dead female or male; blue-filled symbols, mycetoma; unfilled symbols, unaffected; question mark within a symbol, unknown affection status; a number immediately under a symbol, identity number of an individual within pedigree; double lines, consanguinity; a number within a symbol, number of offspring; arrow, proband.*

4.3.1.2 Family 2

The second family reported eleven confirmed cases and one suspected case of mycetoma. The recruited cases included two half-siblings and their maternal cousin. In the control group were the siblings' sister, one of their mothers, and the affected cousin's parents (**Figure 4.6**).

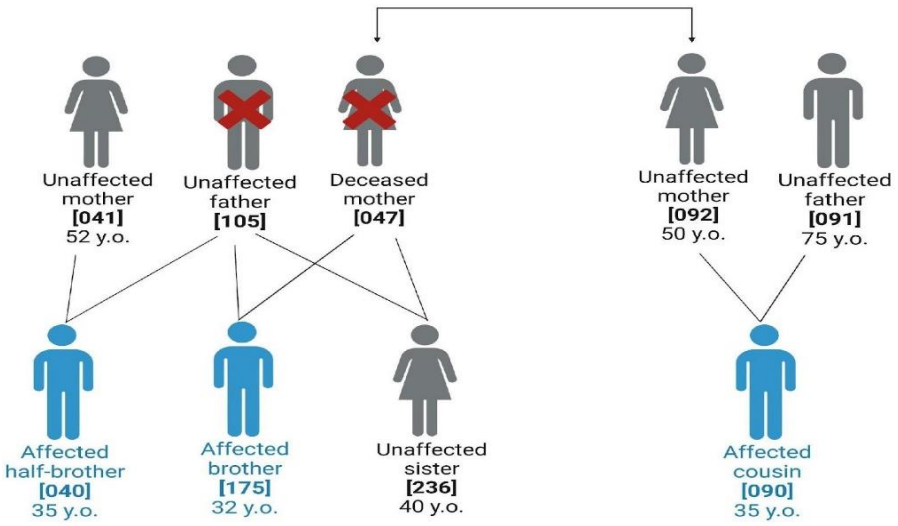


Figure 4.6 *Members of Family 2 recruited in the WGS study. The mycetoma cases (highlighted in blue) included two half-brothers, identified as 175 and 040, and one maternal cousin, 090. One healthy sister (236), mother of the half-brother (041) and parents of the cousin (091 and 092) are designated as controls (shown in dark grey)—Y.o., years old. The X mark is for individuals not recruited in the study due to death or unavailability.*

The steps taken to prioritise the candidate variant are shown in **Table 4.3**. This family had the highest number of variants after the subsetting step (391,723 variants). However, the subsequent filtering and prioritisation drastically reduced the number into two candidate variants, one InDel identified in all affected individuals and one SV identified in the affected cousin. The InDel variant was a frameshift deletion that spans from 77 bp upstream to 35 bp downstream of the start codon, leading to the loss of the start codon (start loss variant) in the caspase 8 (*CASP8)* gene. This gene encodes one of the cysteine proteases (caspases) that plays several roles, including initiation and execution of extrinsic apoptosis and suppression of a form of programmed necrosis known as necroptosis (301). The top five tissues where *CASP8* is expressed according to GTEx are cultured fibroblasts, adrenal gland, adipose subcutaneous, breast mammary tissue and small intestine terminal ileum. Mutations in *CASP8* can contribute to the development of caspase 8 deficiency (OMIM 607271), which is an autoimmune lymphoproliferative syndrome (ALPS)-related disorder (302). Patients suffering from this disorder are susceptible to infections due to impaired T-, B- and NK-cell activation, defective FAS-induced apoptosis, lymphadenopathy and splenomegaly (302,303).

Table 4.3 *The number of variants after the analysis of Family 2 genomes.*

Family 2	**SNPs**	**InDels**	**CNVs**	**SVs**
The average number of variants called in all samples	3,749,568	937,629	4,426	5,719
Number of variants after subsetting	271171	118867	13	1672
Number of variants after initial filtering	17	10	0	3
Number of variants after final prioritisation	0	1	-	1

SNPs: single nucleotide polymorphism; InDels: insertion/deletions; CNVs: copy number variants; SVs: structural variants.

The identified LOF variant (NM_001080125.2:c.-77_35del) was classified as likely pathogenic by Varsome (last accessed: 27th May 2023) according to the ACMG variant classification criteria (**Table 4.8**). GnomAD allele frequency for this variant was not available due to the failure of the reference dataset to pass quality control. The variant was homozygous in 175 and heterozygous in his half-sibling 040 and cousin

090. An Identity by Descent (IBD) analysis was conducted on the three affected individuals to confirm if the candidate variant came from a common ancestor. The results in **Table 4.4** showed that the half-siblings (040 and 175) shared both alleles (fully identical) at approximately 73% of the variants on the region where the candidate variant is located, with a Z2 value of 0.73, higher than the expected Z2 value of 0.5 for half-siblings (304). This suggests a recent shared ancestry. Additionally, they had a high PI-HAT value of 0.86, compared to the expected value of 0.25 for second-degree relatives, indicating a strong genetic resemblance. The PI-HAT value for first cousins is typically around 0.125; however, it was 0.5 in this family, indicating that the affected half-siblings share approximately 50% of variants on the analysed region with 090.

Table 4.4 *Proportion of loci shares to be zero alleles (Z0), one allele (Z1), or two alleles (Z=2) and PI_HAT values between affected individual pairs of Family 2.*

Individual ID1	Individual ID2	Z0	Z1	Z2	PI_HAT
M040	**M175**	0	0.27	0.73	0.86
M040	**M090**	0	1	0	0.5
M175	**M090**	0	1	0	0.5

According to the heatmap in **Figure 4.7**, the half-siblings 175 and 040 had a more significant percentage of shared alleles with two alleles (Z2=0.73) compared to those with only one shared allele (Z1=0.27). The first cousin's Z1 value is also noteworthy, considerably higher than the anticipated value for third-degree relatives (Z1=0.25). Moreover, there were no loci with zero allele share. These results signify that the related cases shared the variant due to a recent common ancestor.

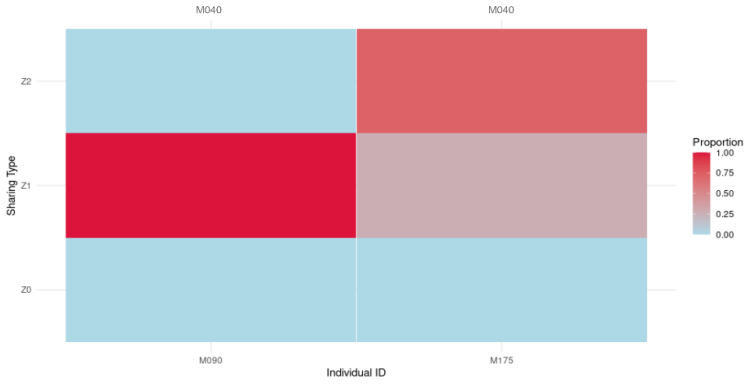


Figure 4.7 *Proportion of loci shares to be zero alleles (Z0), one allele (Z1), or two alleles (Z = 2) between the half-siblings (175 and 040) and their maternal cousin (090) in Family 2.*

Sanger sequencing confirmed the variant's zygosity in 090 and its absence in the controls 041, 091 and 236 (**Figure 4.8**). Both 040 and 175 Sanger chromatograms had high noise to validate the WGS candidate variant; however, the WGS variant had a total read depth of >20 in the half-siblings VCF files, supporting its validity. No protein secondary structure change was predicted for this variant since the utilised software only makes predictions for single nucleotide changes.

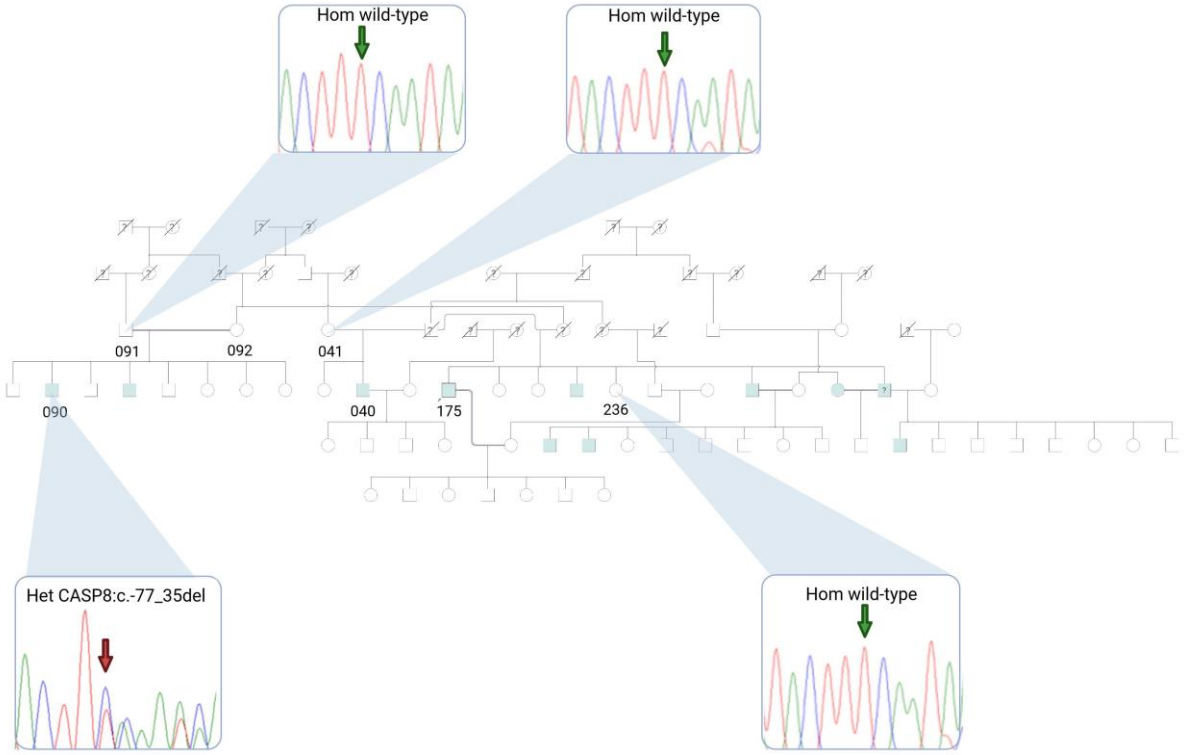


Figure 4.8 *Family 2 pedigree and Sanger sequencing results. Round symbols indicate female; square, male; diamond, unspecified gender; slash through circle or square: dead female or male; blue-filled symbols, mycetoma; unfilled symbols, unaffected; question mark within symbol, unknown affection status; number immediately under the symbol, identity number of an individual within pedigree; double lines, consanguinity; black arrow, proband, green arrow; control genotype, red arrow; case genotype. Hom: Homozygous; Het: Heterozygous.*

Further analysis of the genomes of the affected cousin 090 against his parents revealed a de novo structural variant (Chr12-27280789-27792522-DEL) resulting from a 511 kb deletion in chromosome 12p11.23 region. Based on Ensembl's VEP tool, the identified SV has a significant impact on four coding genes: aryl hydrocarbon receptor nuclear translocator-like 2 (*ARNTL2*), PTPRF interacting protein, binding protein 1 (*PPFIBP1*), single-pass membrane protein with coiled-coil domains 2 (*SMCO2*), and serine/threonine kinase 38 like (*STK38L*). The variation results in transcript ablation of

ARNTL2, *SMCO2*, and *STK38L*, as well as deletion from the start of the transcript to intron 2 of the *PPFIBP1* gene. This LOF variant had a 0.00009228 global allele frequency in gnomAD SVs version 2.1 with zero allele frequency in the African/African American dataset (305). Two of the four affected genes are known to play a role in immunity. The first one, *ARNTL2*, encodes a transcriptional activator which forms a core component of the circadian clock and suppresses the expression of IL-21 via binding to its promoter in non-obese diabetic (NOD) mice (306). The second gene, *STK38L,* also known as *NDR2,* encodes a protein that its activation is identified as a critical step in initiating T-cell receptor (TCR)-mediated activation of lymphocyte function-associated antigen-1 (LFA-1), an integrin molecule found on the surface of T-cells, and its activation plays a crucial role in T-cell adhesion and interaction with antigen-presenting cells (306). NDR2 was also found to regulate the TLR9-mediated innate immune response in murine macrophages (307). TLR9 is known for its recognition of DNA rich in unmethylated CpG motifs, and it is activated in response to fungal DNA, among other microbial DNA (308).

4.3.1.3 Family 3

This highly consanguineous family had four affected individuals, as shown in the family pedigree in **Figure 4.11**. For the WGS study, the cases included two affected brothers (152 and 167) and their double first cousin (the spouse of 152), while the controls were the affected siblings' mother and one of their healthy brothers (**Figure 4.9**). After the initial filtering step, this family had the highest number of variants, with 31 variants (**Table 4.5**).

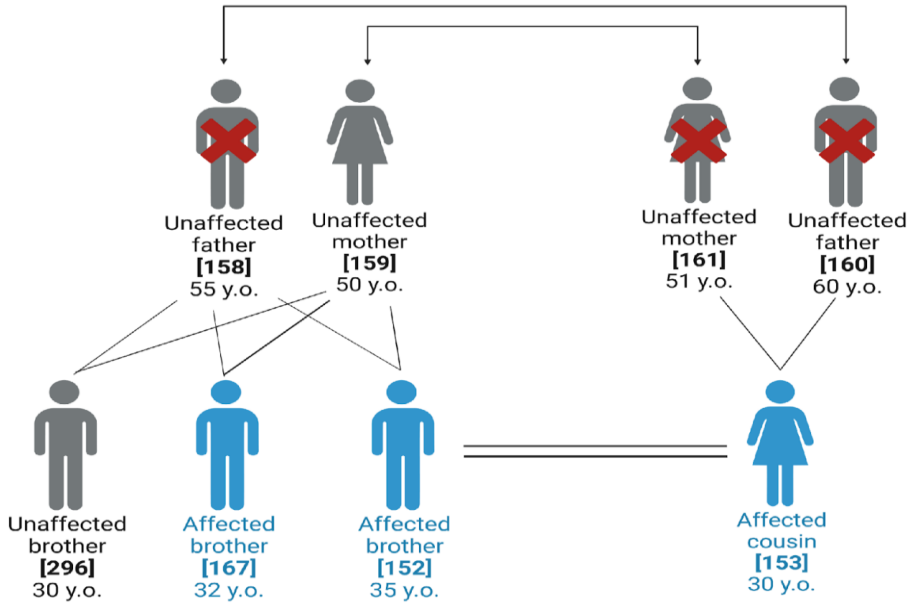


Figure 4.9 *Members of Family 3 recruited in the WGS study. The mycetoma cases (highlighted in blue) include two brothers, identified as 152 and 167, and one maternal cousin, 153. One healthy brother (296) and the siblings' mother (159) are designated as controls (shown in dark grey)—Y.o., years old. The X mark is for individuals not recruited in the study due to death or unavailability.*

Table 4.5 *The number of variants after the analysis of Family 3 genomes.*

Family 3	**SNPs**	**InDels**	**CNVs**	**SVs**
The average number of variants called in all samples	3,749,568	937,629	4,426	5,719
Number of variants after subsetting	136629	40395	154	143
Number of variants after initial filtering	11	9	4	7
Number of variants after final prioritisation	1	0	0	0

SNPs: single nucleotide polymorphism; InDels: insertion/deletions; CNVs: copy number variants; SVs: structural variants.

The prioritised candidate variant found in the genomes of the two affected brothers was a homozygous missense variant in the *PLEKHO2* gene. This gene is mainly expressed in the spleen, whole blood, Epstein-Barr virus (EBV)-transformed lymphocytes, lung, cultured fibroblasts and adipose subcutaneous, according to GTEx (309). The full extent of PLEKHO2 roles and mechanisms of action are still being elucidated, but this protein has been implicated in promoting macrophage survival,

differentiation, and maturation in mice (310). The identified variant NM_025201.5:c.379G>A (p.Asp127Asn) was predicted pathogenic by DANN, EIGEN PC, FATHMM-MKL, LRT and SIFT4G pathogenicity engines within Varsome platform (last accessed: 22nd May 2023). The variant was not present in gnomAD genomes but had a very low allele frequency of 0.0000358 in gnomAD exomes (last accessed: 8th October 2023). Notably, no homozygous individuals for the variant were previously reported in gnomAD exomes, nor was the variant found in the African/African American dataset (311). This variant was homozygous in the two affected brothers (152 and 167) and their paternal first cousin 002, while their mother and healthy brother were only carriers. The IBD analysis included the former three affected individuals (**Table 4.6**). The expected Z2 for full siblings is 0.25, while it should be zero for third-degree relatives such as first cousins. However, in this highly consanguineous family, both Z2 and PI-HAT values were above 0.9, indicating that the affected individuals share a very high proportion (over 90%) of their alleles, and these alleles are identical by descent in that particular region on chromosome 15 where the identified candidate is located.

Table 4.6 *Proportion of loci shares to be zero alleles (Z0), one allele (Z1), or two alleles (Z=2) and PI_HAT values between affected individual pairs of Family 2.*

Individual ID1	Individual ID2	Z0	Z1	Z2	PI_HAT
M152	**M167**	0	0.09	0.91	0.96
M152	**M002**	0	0.02	0.98	0.99
M167	**M002**	0	0.08	0.92	0.96

According to the heatmap in **Figure 4.10**, the affected individual 152 and his cousin 002 had a substantially high percentage of shared alleles with two alleles (Z2=0.98) compared to the sibling pair with a Z2 of 0.91. The single shared allele values (Z1) for all three individual pairs were almost negligible when compared to the expected values for full siblings (Z1=0.5) and third-degree relatives (Z1=0.25). These high Z2 and PI-HAT values of the analysed region on chromosome 15 suggest a very close genetic relationship and shared ancestry within that region.

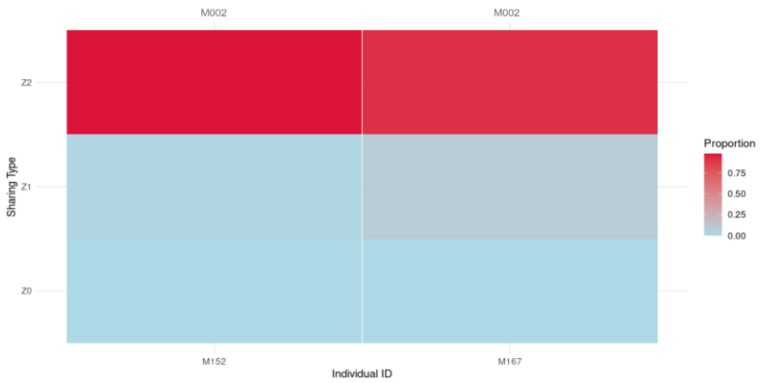


Figure 4.10 *Proportion of loci shares to be zero alleles (Z0), one allele (Z1), or two alleles (Z = 2) between the affected siblings in Family 3 and their paternal cousin.*

Sanger sequencing confirmed the presence of this variant in Family 3 members and their paternal cousin 002 (**Figure 4.11**). The confirmation of variant presence and zygosity for the affected cousin (153) could not be established as a sample for Sanger sequencing could not be obtained.

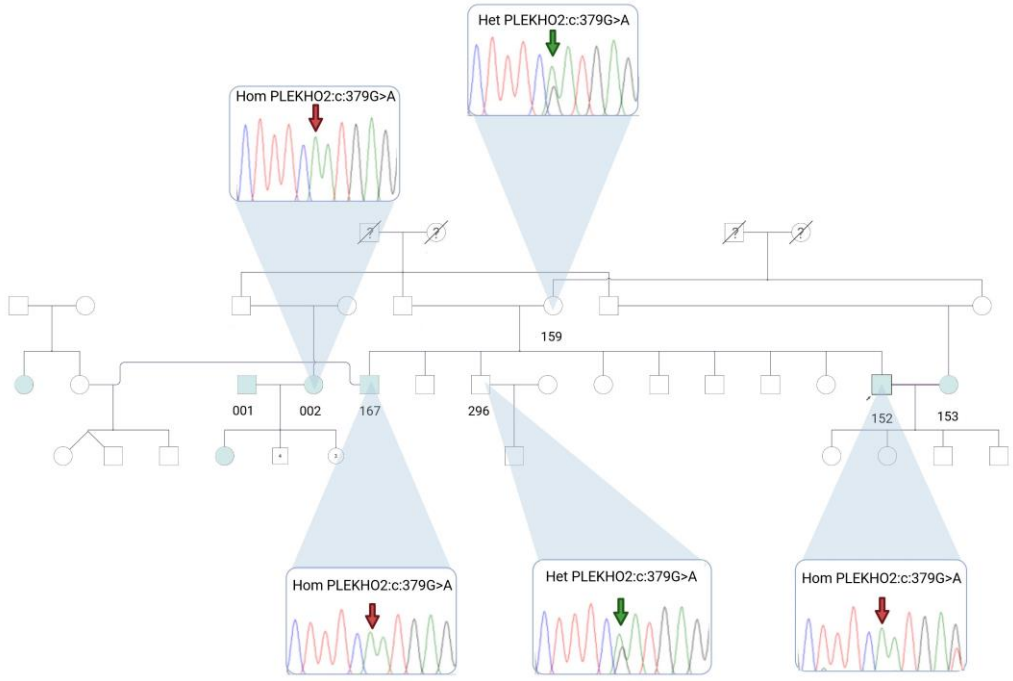


Figure 4.11 *Family 3 pedigree with patient 002 and their Sanger sequencing results. Round symbols indicate female; square, male; diamond, unspecified gender; slash through circle or square: dead female or male; blue-filled symbols, mycetoma; unfilled symbols, unaffected; question mark within symbol, unknown affection status; number immediately under the symbol, identity number of an individual within pedigree; double lines, consanguinity; black arrow, proband, green arrow; control genotype, red arrow; case genotype. Hom: Homozygous; Het: Heterozygous.*

The effect of the amino acid substitution due to the candidate variant (p.Asp127Asn) on the PLEKHO2 protein secondary structure is demonstrated in **Figure 4.12**. This substitution was predicted to be stabilising with ΔΔG of 0.159 kcal/mol accompanied by a decrease of molecule flexibility ($\Delta\Delta S_{Vib}$: -0.049 kcal.mol^{-1}.K^{-1}). These changes are due to additional interatomic interactions formed by the asparagine residue in the mutant protein, as shown in **Figure 4.12C**. This mutation introduces an extra steric clash (results when two non-bonded atoms overlap due to proximity) with valine (Val129) residue.

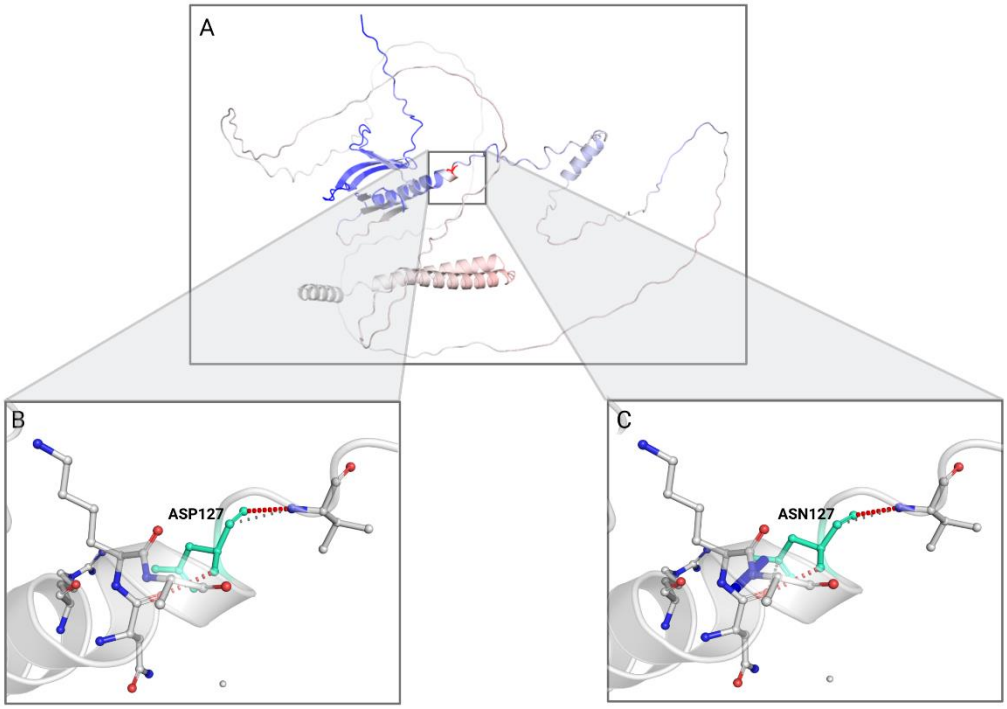


Figure 4.12 *Alterations in PLEKHO2 stability upon Asp127Asn mutation. (A) Visual representation of the chain in which the mutation occurs based on the vibrational entropy difference (ΔΔS) between wild-type and mutant structures of PLEKHO2. Amino acids are coloured according to ΔΔS, where blue represents a rigidification of the structure and red represents a gain in flexibility. (B) Interatomic interactions for wild-type PLEKHO2. (C) Interatomic interactions for mutant PLEKHO2. Both wild-type and mutant residues are coloured in light green and illustrated as sticks. A scale of colour definition for the interaction is provided by DynaMut software: red for hydrogen bonds; blue for halogen bonds; yellow for ionic interactions; purple for metal complex interactions; sky blue for aromatic contacts; green for hydrophobic contacts; pink for carbonyl contacts.*

4.3.1.4 Family 4

The last family recruited in the WGS study had four siblings with mycetoma. The participants in the WGS were two affected siblings, their father and one unaffected sister (**Figure 4.13**). This family WGS data analysis (**Table 4.7**) revealed a heterozygous missense variant in the rap guanine nucleotide exchange factor 2 (*RAPGEF2*) gene. This gene is expressed predominantly in EBV-transformed lymphocytes, minor salivary glands, thyroid, cultured fibroblasts, and adipose subcutaneous, according to the GTEx database. The encoded protein acts as a guanine nucleotide exchange factor (GEF) for a group of proteins called Rap GTPases (312–315). RAPGEF2 (also termed RA-GEF-1) helps activate Rap proteins by promoting the exchange of guanosine diphosphate (GDP) for guanosine triphosphate (GTP) (316). This activation of Rap proteins then triggers various downstream intracellular signalling events.

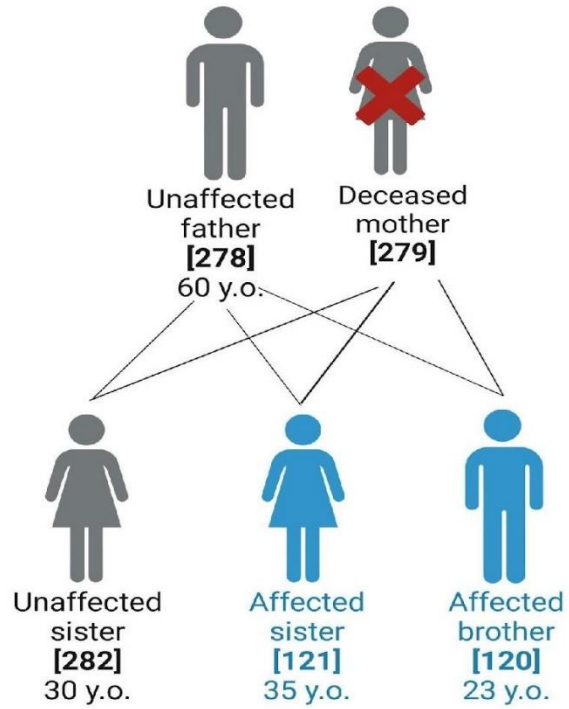


Figure 4.13 *Members of Family 4 recruited in the WGS study. The mycetoma cases (highlighted in blue) include two siblings, identified as 120 and 121. One healthy sister (282) and the siblings' father (278) are designated as controls (shown in dark grey)—Y.o., years old. The X mark is for individuals not recruited in the study due to death or unavailability.*

The identified missense variant NM_014247.3: c.700A>G (p.Ile234Val) was predicted pathogenic by seven engines, which were FATHMM-MKL, LRT, Mutation assessor,

EIGEN, EIGEN PC, LIST-S2, PrimateAI included in Varsome platform (last accessed: 21st May 2023). The identified variant was not found in gnomAD genomes and had a very low allele frequency of 0.00001606 in gnomAD exomes with no presence in the African/African American dataset (last accessed: 8th October 2023) (317). The variant was validated in the cases' DNA using Sanger sequencing, as demonstrated in **Figure 4.14**. The results obtained from Sanger sequencing analysis indicated that the heterozygous missense variant identified in the affected sibling is also present in the unaffected sister. This finding raises concerns about the potential risk of infection for the unaffected 30-year-old sister This is due to the fact that individuals between the ages of 20 and 40 are most vulnerable to the disease (70) in addition to the high sibling recurrence risk ratio (λ_s) for mycetoma found in the pedigrees data.

Table 4.7 *The number of variants after the analysis of Family 4 genomes.*

Family 4	**SNPs**	**InDels**	**CNVs**	**SVs**
The average number of variants called in all samples	3,749,568	937,629	4,426	5,719
Number of variants after subsetting	160141	37777	259	414
Number of variants after initial filtering	18	8	2	0
Number of variants after final prioritisation	1	0	1	-

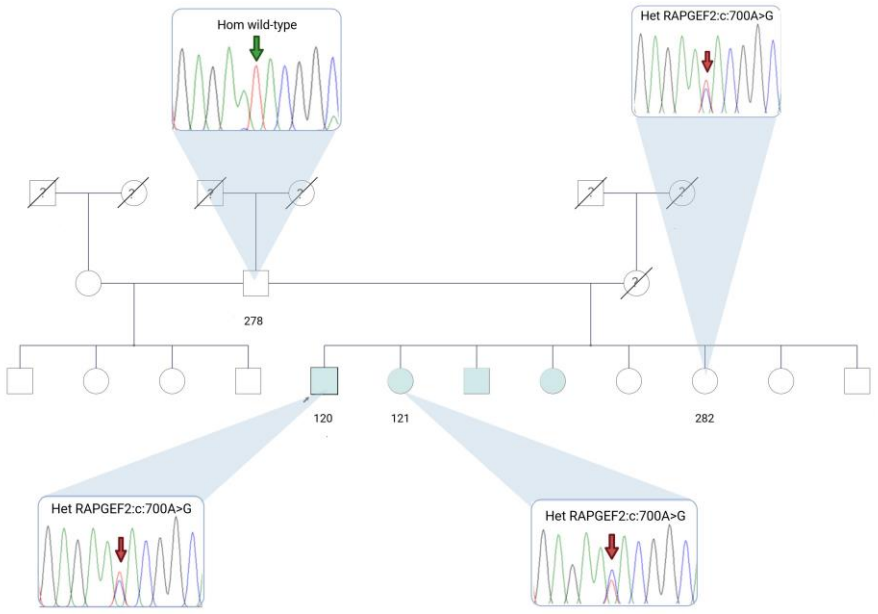


Figure 4.14 *Family 4 pedigree and Sanger sequencing results. Round symbols indicate female; square, male; diamond, unspecified gender; slash through circle or square: dead female or male; blue-filled symbols, mycetoma; unfilled symbols, unaffected; question mark within symbol, unknown affection status; number immediately under the symbol, identity number of an individual within pedigree; double lines, consanguinity; black arrow, proband, green arrow; control genotype, red arrow; case genotype. Hom: Homozygous; Het: Heterozygous.*

The variant (p.Ile234Val) had a destabilising effect (ΔΔG: -0.347 kcal/mol) on the RAPGEF2 secondary structure with an increase of molecule flexibility ($\Delta\Delta S_{Vib}$: 0.250 $kcal.mol^{-1}.K^{-1}$). The change in interatomic interactions in the residues' environment is displayed in **Figure 4.15**. A steric clash with Asp231 is predicted in the mutant RAPGEF2 with a loss of hydrophobic contacts with Val230 and ARG243 residues.

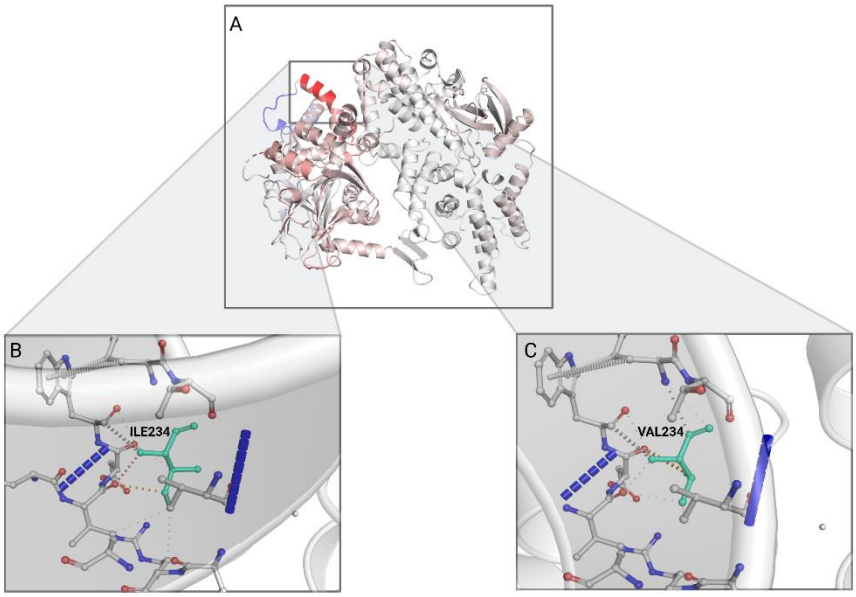


Figure 4.15 *Alterations in RAPGEF2 stability upon Ile234Val mutation. (A) Visual representation of the chain where the mutation occurs based on the vibrational entropy difference (ΔΔS) between wild-type and mutant structures of RAPGEF2. Amino acids are coloured according to ΔΔS, where blue represents a rigidification of the structure and red represents a gain in flexibility. (B) Interatomic interactions for wild-type RAPGEF2. (C) Interatomic interactions for mutant RAPGEF2. Both wild-type and mutant residues are coloured in light green and illustrated as sticks. A scale of colour definition for the interaction is provided by DynaMut software: red for hydrogen bonds; blue for halogen bonds; yellow for ionic interactions; purple for metal complex interactions; sky blue for aromatic contacts; green for hydrophobic contacts; pink for carbonyl contact.*

Table 4.8 *Whole genome sequencing variants classification.*

FID	Case ID	Gender	*Gene* (Transcript)	Variant	Coding impact	Zygosity	MAF	ACMG classification	ACMG criteria	Mode of Inheritance (gene)	Pathogenicity computational verdict	Read depth (DP)
1	001	M	*DEFA3* (NM_005217.4)	Duplication	Transcript amplification	Het	NA	Variant of uncertain significance	NA	AD/AR	NA	NA
			CFTR (NM_000492.4)	c.3485G>A (p.Arg1162Gln)	Missense	Het	0.0002004	Likely Pathogenic	PM2, PP1 & PP3	AD/AR	Pathogenic by 18 engines	43
	011	M	*DEFA3* (NM_005217.4)	7.5kb duplication	Transcript amplification	Het	NA	Variant of uncertain significance	NA	AD/AR	NA	NA
			CFTR (NM_000492.4)	c.3485G>A (p.Arg1162Gln)	Missense	Het	0.0002004	Likely Pathogenic	PM2, PP1 & PP3	AD/AR	Pathogenic by 18 engines	49
	002	F	*PLEKHO2* (NM_025201.5)	c.379G>A (p.Asp127Asn)	Missense	Hom	Not Found	No previous association with a monogenic disease	NA	AR	Pathogenic by five engines	24
2	175	M	*CASP8* (NM_001080125.2)	c.-77_35del	Start loss	Hom	NA	Likely Pathogenic	PVS1 & PP1	AR	NA	45
	040	M	*CASP8* (NM_001080125.2)	c.-77_35del	Start loss	Het	NA	Likely Pathogenic	PVS1 & PP1	AR	NA	27
	090	M	*CASP8* (NM_001080125.2)	c.-77_35del	Start loss	Het	NA	Likely Pathogenic	PVS1 & PP1	AR	NA	57
			ARNTL2 (NM_001248002.2) *, PPFIBP1*	511kb deletion	Transcript ablation	NA	0.00	Variant of uncertain significance	NA	AD/AR	NA	NA

			*(*NM_001198915.2*)* *, SMCO2* *(*NM_001145010.1*)* *, STK38L* *(*NM_015000.4*)*									
3	152	M	*PLEKHO2* (NM_025201.5)	c.379G>A (p.Asp127Asn)	Missense	Hom	Not Found	No previous association with a monogenic disease	NA	AR*	Pathogenic by five engines	30
	167	M	*PLEKHO2* (NM_025201.5)	c.379G>A (p.Asp127Asn)	Missense	Hom	Not Found	No previous association with a monogenic disease	NA	AR*	Pathogenic by five engines	47
4	120	M	*RAPGEF2* (NM_014247.3)	c.700A>G (p.Ile234Val)	Missense	Het	Not Found	No previous association with a monogenic disease	NA	AD*	Pathogenic by seven engines	39
	121	F	*RAPGEF2* (NM_014247.3)	c.700A>G (p.Ile234Val)	Missense	Het	Not Found	No previous association with a monogenic disease	NA	AD*	Pathogenic by seven engines	39

FID: Family ID; Hom: Homozygous; Het: Heterozygous MAF; minor allele frequency in African/African American according to gnomAD genomes browser; NA: not applicable; AD: autosomal dominant; AR: autosomal recessive; * predicted by DOMINO; PM2: gnomAD genomes homozygous allele count = 0; PP1: Co-segregation with the disease in multiple affected family members in a gene definitively known to cause the disease; PP3: Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.); PVS1: Null variant (start loss) in gene *CASP8* (this gene has 14 reported pathogenic LOF variants); Read depth (DP): number of reads that support the variant call.

4.3.2 Collective analyses of the whole genome sequencing data

The four families' candidate genes (*DEFA3, CASP8, PLEKHO2* and *RAPGEF2*) were used to perform network and pathway analyses, as described in the following two sections. The SV candidate identified in 090 was omitted because it was found in one individual. These collective analyses were conducted to explore the functional and contextual aspects of candidate genes within the broader biological landscape.

4.3.2.1 Network analysis

Network analysis revealed the interaction between the families' candidate genes and other predicted genes potentially contributing to mycetoma pathogenesis. **Figure 4.16** illustrates the biological interactions of these candidate genes, with physical interaction links comprising the most significant proportion (77.6%). Co-expression (8%), co-localisation (3.6%), genetic interactions (2.87%), pathway (1.88%), and shared protein domains (0.6%) were also observed. I excluded the predicted interactions comprising 5.4% of the network interactions to focus solely on the other experimentally confirmed ones. The size of the candidate genes' circles reflects their importance in the network and their potential role in susceptibility to mycetoma.

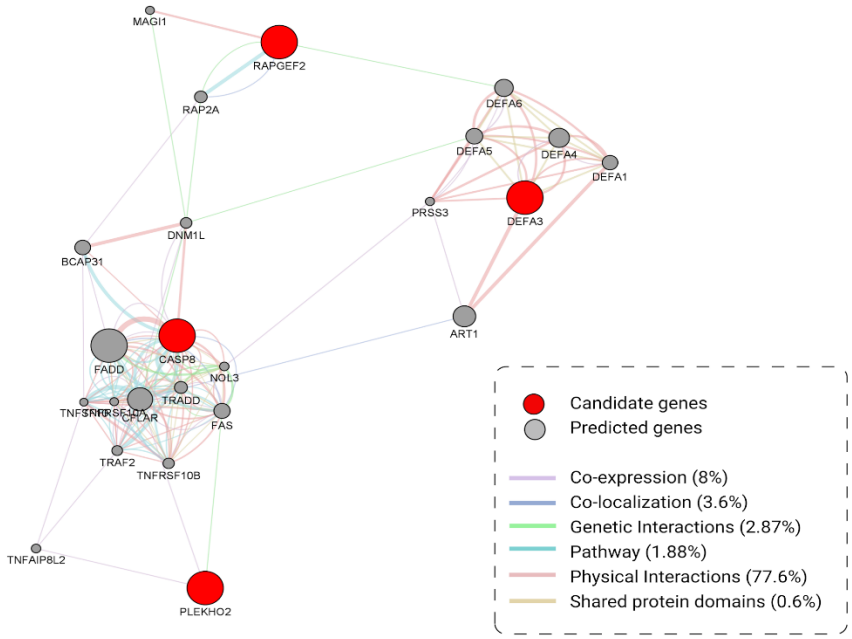


Figure 4.16 *Gene network analysis of the four identified candidate genes. The candidate genes are coloured in red, and the predicted genes are coloured in grey. The underlying molecular links were attributed to the co-expression associations, co-localization, genetic interactions, pathway links, physical interactions, and shared protein domains.*

Centiscape identified *CASP8* and its cluster of genes as critical network regulators, according to **Table 4.9**. This finding could have significant implications for future research on the role *CASP8* may play in mycetoma pathogenesis.

Table 4.9 *Centiscape node centralities. Top five genes ordered by degree centrality.*

Gene name	Degree	Betweenness
CASP8	37	44.9
FAS	35	10.3
CFLAR	34	93.7
FADD	33	10.2
TRADD	29	1.3

4.3.2.2 Pathway analysis

Pathway analysis was done to determine the biological pathways in which the mycetoma susceptibility genes belong and to understand the functional relevance of the genes in the respective pathways. The most significantly enriched pathways by the overrepresentation analysis were the innate immune system (p = 5.06E-04) and immune system (p = 0.006). The findings demonstrate that *CASP8* was identified in the top three overrepresented pathways, whereas *PLEKHO2* and *DEFA3* were implicated in immune-related pathways, as delineated in **Table 4.10**.

Table 4.10 *Top enriched pathways by overrepresentation analysis in Reactome.*

Pathway name	Submitted entities found	Entities *p*-value	Entities FDR
Innate Immune System	*CASP8*; *PLEKHO2;DEFA3*	5.06E-04*	0.01**
Immune System	*CASP8*; *PLEKHO2;DEFA3*	0.006*	0.02**
Signal Transduction	*CASP8*; *RAPGEF2*	0.07	0.07

FDR: false discovery rate; *$p \leq 0.05$; **FDR<0.05

An illustrative Voronoi diagram termed Reacfoam was constructed based on the pathway enrichment analysis (**Figure 4.17**). This diagram showed the top enriched pathways containing the candidate genes found in the four families.

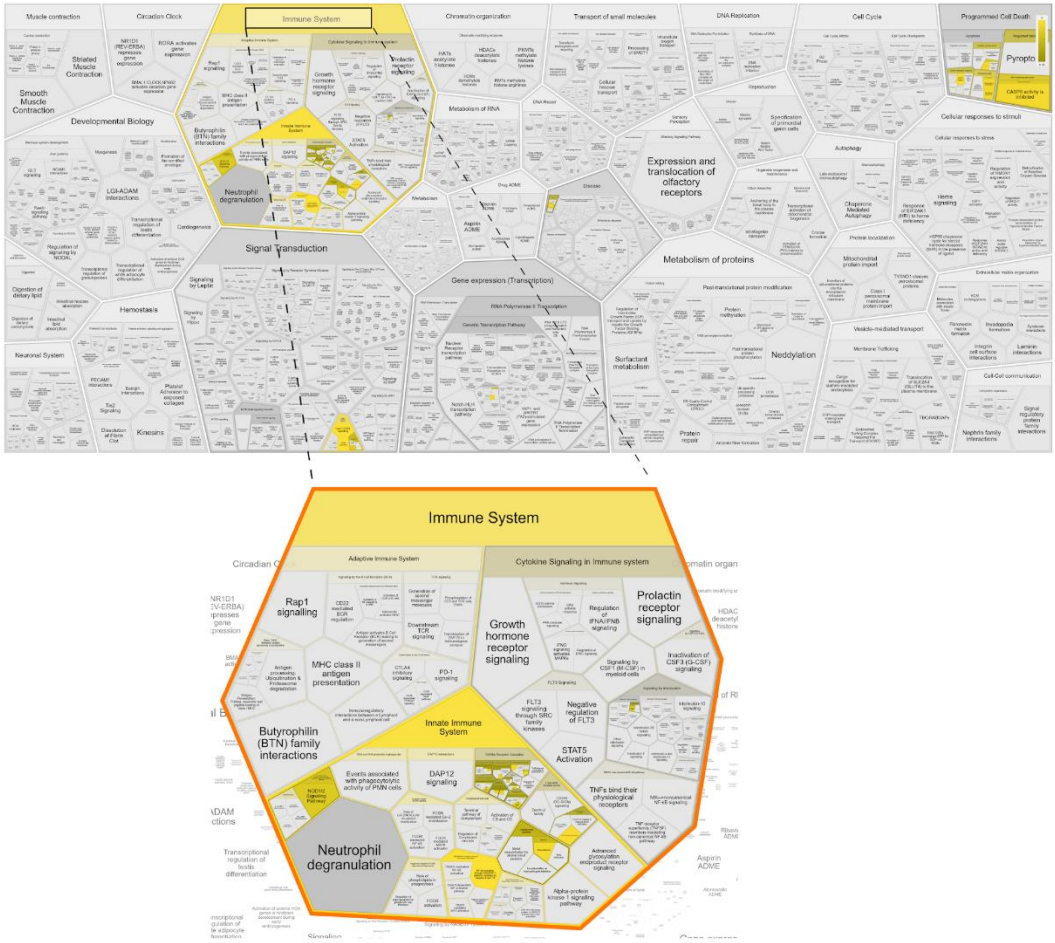


Figure 4.17 *The Reacfoam diagram displaying the top enriched pathways of the immune system, including the innate immune system, influenced by the four candidate genes. The yellow colour code denotes the overrepresentation of the pathways according to the input candidate genes. Light grey signifies pathways which are not significantly overrepresented.*

4.4 Discussion

Nearly all consanguineous marriages in this dataset were between first cousins (88%). Sudan is among the top nine countries with a high consanguinity rate of 40-49% (119). Across successive generations, the Sudanese population's consanguinity rate has remained consistently above 40%, indicating its deep-rooted presence (120,121). Most of the enrolled participants in this study come from poor rural communities where consanguineous marriages are prevalent. Consanguinity is a significant risk factor for human susceptibility to infectious diseases (125). Furthermore, research conducted in Sudan has revealed a strong association between consanguinity and genetic disorders, particularly those that exhibit autosomal recessive inheritance patterns (126–129,132).

A crucial area for improvement in utilising the three heritability estimation software programs, initially developed in high-income countries, is the inability to deal with the intricate pedigree structures commonly observed within our Sudanese consanguineous community. The primary challenge lies in the fact that these software programs were designed with an assumption of non-consanguinity, implying that the input pedigrees they handle are expected to be free of any loop structures. In other words, these programs are optimised to work within pedigrees that follow a more traditional family tree structure without any instances of intermarriage between close relatives. Consequently, due to consanguinity, these software programs may have encountered difficulties accurately estimating heritability when exposed to pedigrees containing loops. Adding to this limitation is that mycetoma, the trait under investigation in this study, is a binary trait, differing from the continuous traits for which these programs were originally developed. This divergence in trait type adds another layer of complexity to the heritability estimation process, as these programs are inherently more tailored to continuous trait analysis. Notably, the ASSOC program within the SAGE software package emerged as the most effective tool for estimating heritability within these complex pedigrees, generating results that were compatible with the increased risks observed in siblings of affected individuals.

Adopting familial aggregation approaches in genetic studies presents several limitations that warrant consideration. One primary constraint lies in the specificity of the obtained estimates within the studied population that renders extrapolation to other diverse populations or ethnic groups. Furthermore, the genetic analysis of pedigree data, which is fundamental to calculating these estimates, introduces challenges. This process necessitates accurate measurement of phenotypes, appropriate covariates, and constructing a kinship matrix. Traditionally, this matrix is crafted from meticulously assembled pedigrees that connect individuals with reported genealogical relationships. However, reliance on self-reported genealogy can lead to inaccuracies in relationship specification. Such errors might emerge due to factors like recording mistakes and variations in the cultural interpretation of biological kinship connections (e.g., it is common in the targeted communities in this study to call their cousins brothers). Moreover, the situation can be worsened when a collection of pedigrees is sourced from a common geographical region, potentially concealing hidden kinship ties between pedigrees that appear distinct (I had to merge several pedigrees in the

dataset after discovering distinct relatedness). The potential pedigree information errors were absent because at least two family members were usually present during the data collection process. This is because mycetoma patients are usually accompanied by another family member when visiting the MRC clinic. Additionally, during field visits to endemic areas, the patients could have the entire family coming for their support, as demonstrated in **Figure 4.18**, taken during a field mission to Sennar state.

Figure 4.18 *Pedigree information collection from a mycetoma patient accompanied by her family.*

The second part of this familial approach focused on families with a concentration of cases to bridge the gap between rare Mendelian disorders, often associated with consanguinity and complex disorders such as mycetoma. The average number of variants across the four families' genomes was ~4.6 million, consistent with the out-of-Africa model of human origin, with Africans having the highest number of variant sites per genome (~5 million variants) when compared to individuals of East Asian, European, or South Asian ancestry (~4.0–4.2 million variants) (318,319). The identified candidate variants across the four families varied in types, coding impact

and zygosity. However, network analysis showed how the underlying genes were biologically connected and which genes can have potential functional significance based on their centrality in the network. According to the pathway analysis results, the innate immune system pathway was the most significantly enriched (p = 5.06E-04). This aligns with the conclusions of the inaugural study on the genetic susceptibility to mycetoma conducted in 2007 by van de Sande and colleagues, which indicated that the initial response to mycetoma is facilitated by neutrophils (64). Several genes were identified as central, including *CASP8* and its closely associated genes involved in the extrinsic/death receptor pathway. Among the candidate genes, *CASP8* was the only one present in the top three overrepresented pathways: innate immune system, immune system, and signal transduction. It is known that caspase 8 is involved in regulating the signalling pathways of inflammasomes, which are multiprotein complexes that recognise and initiate inflammatory responses against pathogens as an essential part of innate immunity (320,321). Inflammasomes also play an important role in cytokine production, including IL-1β, either by activating caspase-1 or through a noncanonical caspase 8-dependent pathway associated with antifungal response (322). A study on inflammasome activation in mouse dendritic cells found that caspase 8 cleaves pro-IL-1β into its active form in response to β-glucans found in the cell walls of *Candida albicans* (323). IL-1β is an important pro-inflammatory cytokine linked to eumycetoma pathogenesis (33,324). This cytokine recruits neutrophils, promotes cell adhesion, and activates Th17 cells (325). The binding of IL-1β to its receptor leads to the activation of Nuclear factor kappa B (NF-κB) and subsequent expression of pro-inflammatory cytokines such as TNFα, CCL5 and IL-8, which acts as a chemoattractant for neutrophils (213,326–328). Polymorphisms in CCL5 and IL-8 have previously been associated with susceptibility to mycetoma (32,64). As such, further research is necessary to investigate the role of this pathway that includes inflammasomes, IL-1β, CCL5 and IL-8 in defence against mycetoma pathogens and whether the non-canonical caspase 8-dependent inflammasome system is activated in eumycetoma, as shown in species of *Candida*.

It should be noted that this WGS study had several limitations worth considering. First, while Mendelian disorders result from mutations in single genes and are relatively rare in the general population, concentrated family studies may reveal substantial genetic contributions within specific families, akin to Mendelian inheritance. However, these

mutations might not be as relevant to the outbred population where complex disorders prevail. Studying these rare mutations can provide insights into the disorder's biological mechanisms. Linkage disequilibrium (LD) mapping would complement such investigations by identifying common genetic variants associated with complex disorders in the broader population, thus connecting familial and population-level genetics. Second, functional studies are required to assess the pathogenicity of the identified candidate variants and determine the molecular basis of the disease. Third, several raw data files such as BAM and VCF files were required to validate the read depth of the CNV identified in Family 1 and the SV identified in 090 from Family 2. However, these files are stored at the IEND server and cannot be accessed due to the ongoing war; therefore, the variants' read depth was marked as NA. Lastly, despite the vital relevance of the identified CNV variant in Family 1, a contradictory study conducted on the α-defensin locus in 2504 samples from the 1000 Genomes (1KG) dataset reported that the African super-population differs substantially from the other super-populations (329). Approximately a third of African samples had genotypes with no DEFA3 copies, while the samples with the highest DEFA3 copy numbers were also African. All seven 1KG African populations had individuals with high DEFA3 copy numbers; however, none were East African. Thus, a functional investigation on the validity and impact of this variant in Family 1 is crucial. Furthermore, affirmative quantitative polymerase chain reaction (qPCR) analysis is crucial since this genetic variation is often overlooked due to the technical challenge of accurate copy number measurement.

In conclusion, this family-based study is the first to be conducted on mycetoma. With a sibling recurrence risk calculated at ~16% and a heritability of ~23% for mycetoma, these findings provide additional evidence of the size of genetic factors contributing to mycetoma development. Furthermore, I identified four candidate genes in the four multicase families primarily involved in the innate immune system and signal transduction pathways. These results could be a foundational framework for further research into the underlying biological mechanisms contributing to mycetoma susceptibility. The familial aggregation study highlights the importance of employing specialised methodologies and tailored software solutions to accommodate consanguineous populations' genetic complexities. While this study contributes insights into mycetoma susceptibility within our Sudanese context, its implications

extend beyond geographic boundaries. It requires further comprehensive investigations into familial aggregation and susceptibility in other regions endemic with mycetoma to validate the findings and expand the scope to diverse populations to unveil variations in genetic susceptibilities across different genetic backgrounds and environmental factors.

CHAPTER 5 - Unravelling the genetic landscape of mycetoma susceptibility through population-based genome-wide association studies

5.1 Overview

In this chapter, which is complementary to chapter four, I explore how genome-wide association studies (GWAS) can be utilised to identify common variants associated with mycetoma in addition to estimating the inbreeding in this dataset using the F_{ROH} inbreeding coefficient and SNP heritability. This population-based study aimed to explore the role of common variants with small effects and shed light on the intricate genetic landscape underlying susceptibility to mycetoma.

The chapter is divided into three sections:

- Section one: Summarises the study participants' characteristics.
- Section two: Demonstrates the resulting dataset after quality control.
- Section three: Details the GWAS data analyses performed.

5.2 Characteristics of the study participants

The data analysis included 474 subjects (309 cases and 165 controls). This was under the planned sample size due to the loss of the second batch of samples in the war in Sudan. Table 5.1 summarises the socio-demographic characteristics of participants. Both groups had higher percentages of males (over 70%). As expected, given the study's design, the cases were, on average, much younger than the controls (34 vs. 53 years). The enrolled cases belonged to 70 ethnic groups, while the controls represented 26 ethnic groups (see Appendices 9 and 10 for the complete list). Three ethnic groups (Kawahla, Jaalyeen and Hassania) were listed in the top five in both cases and controls. Due to the ongoing debate over the classification of Kordofanian languages as either part of the Niger-Congo (114) or Nilo-Saharan (330) language families, the four Nuba participants were designated as belonging to both groups (**Table 5.1**). Aljazeera, Sennar and White Nile were the top states with the highest number of participants in Aljazeera state (37% of cases and 56% of controls).

Table 5.1 *Basic characteristics of the study participants (n=309 cases and 165 controls).*

Characteristic	Cases, Number (%)	Controls, Number (%)
Sex		
Male	233 (75.4%)	115 (71%)
Female	76 (24.6%)	50 (32%)
Age		
Mean (SD)	34 (12.6)	53.3 (10.3)
Median (range)	32 (11-75)	52 (40-80)
Occupation		
	Farmer, 122 (40%)	**Farmer,** 72 (44%)
	Merchant, 48 (16%)	**Merchant,** 30 (18%)
	Housewife, 31 (10%)	**Housewife,** 30 (18%)
	University student, 19 (6%)	**Daily labourer,** 8 (4.8%)
	Unemployed, 19 (6%)	**Driver,** 4 (2.5%)
	Driver, 10 (3.2%)	**Employee,** 3 (1.8%)
	Student, 8 (2.6%)	**Construction worker,** 2 (1.2%)
	Daily labourer, 7 (2.2%)	**Teacher,** 2 (1.2%)
	Other, 43 (14%)	**Other,** 14 (8.5%)
Linguistic family		
Afro-Asiatic	237 (76.7%)	149 (90%)
Nilo-Saharan	58 (18.8%)	5 (3%)
Nilo-Saharan and Niger-Congo	4 (1.3%)	-
N/A	10 (3.2%)	11 (7%)
Self-reported ethnicity		
	Kawahla, 52 (17%)	**Kawahla,** 62 (40%)
	Bargo, 18 (5.8%)	**Lahawi,** 42 (27%)
	Gamoeya, 18 (5.8%)	**Jaalyeen,** 7 (5%)
	Baggara, 17 (5.5%)	**Nefedya,** 7 (5%)
	Hassania, 14 (4.5%)	**Hassania,** 3 (2%)
	Jaalyeen, 13 (4.2%)	**Hodor,** 3 (2%)
	Tama, 12 (3.8%)	**Rezegat,** 3 (2%)
	Shokria, 11 (3.6%)	**Rofaa,** 3 (2%)
	Other, 154 (49.8%)	**Other,** 24 (15%)
State		
	Aljazeera, 113 (36.6%)	**Aljazeera,** 93 (56.4%)
	Sennar, 45 (14.6%)	**Sennar,** 67 (40.6%)
	White Nile, 40 (13%)	**White Nile,** 2 (1.2%)
	North Kordofan, 33 (10.6%)	**Kassala,** 2 (1.2%)
	South Darfur, 14 (4.5%)	**North Kordofan,** 1 (0.6%)
	Other, 64 (20.7%)	

N/A: not available.

The case distribution based on location can be seen in **Figure 5.1**. By recruiting cases from the MRC, we formed a cohort comprising individuals from diverse ethnic backgrounds representing the 18 states of Sudan in addition to three cases from Chad.

Figure 5.1 *Mycetoma cases distribution.*

Based on the risk of mycetoma, occupations were classified into three categories: high, moderate, and low (**Figure 5.2**). This categorisation is subjective, as the risk level may vary depending on specific factors such as geographic location, work environment, and personal hygiene practices. However, a general categorisation was provided based on the potential risk associated with each occupation. Farming was, by far, the most frequent occupation in both groups (41% in cases and 44% in controls). Mycetoma patients frequently rely on farming as their sole means of income within their communities. Some individuals, particularly those who have undergone leg amputations, often become merchants (second occupation in rank among mycetoma cases) for convenience. A significant number of patients, around 6%, lost their jobs due to disability.

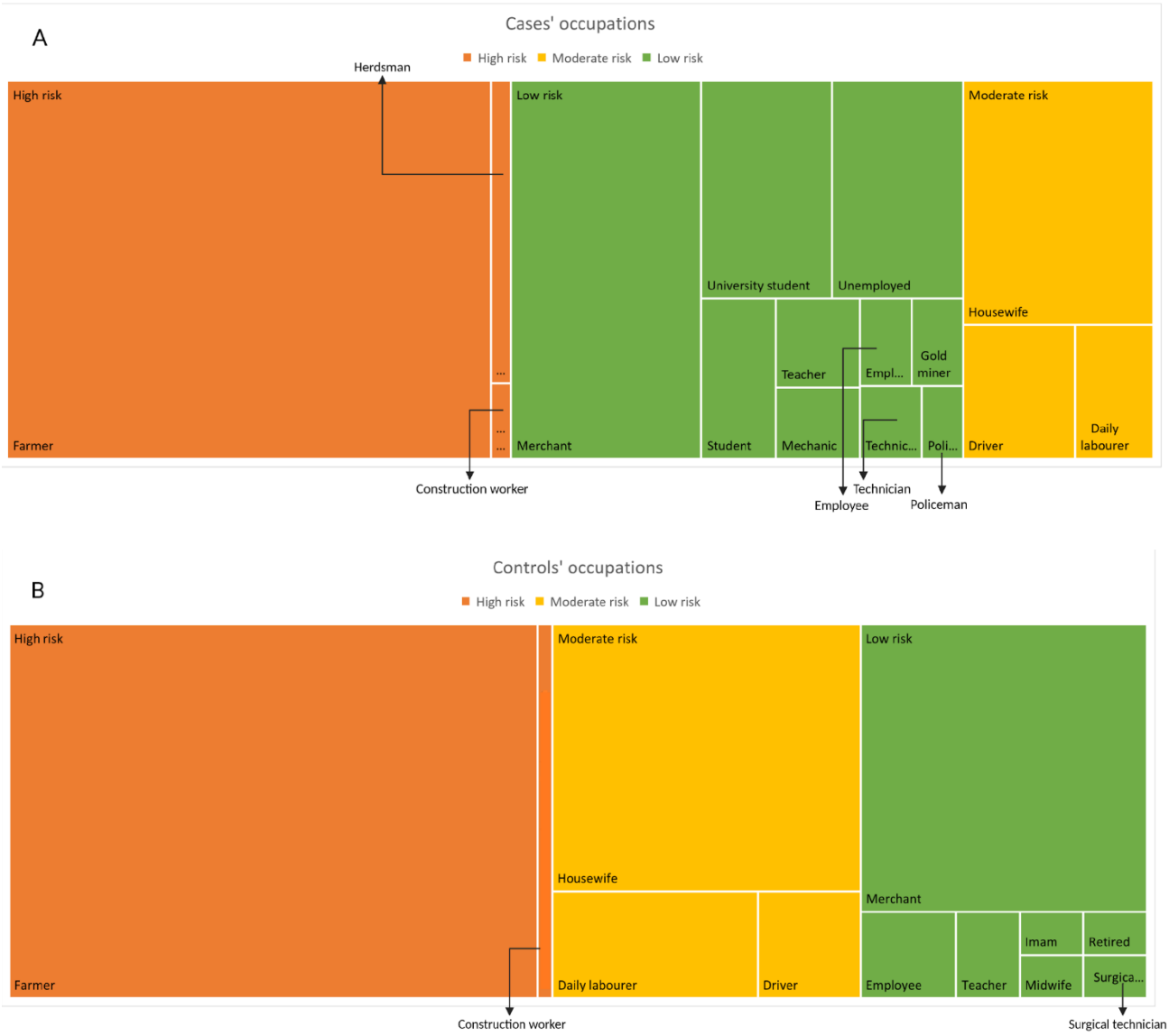


Figure 5.2 *Treemap demonstrating the study participants' occupations based on their frequency. (A) Cases' occupations distributed according to mycetoma risk. (B) Controls occupations distributed according to mycetoma risk.*

Table 5.2 showed further characteristics of the cases. As indicated in the table, the patients' average duration of mycetoma infection was ten years. Most lesions were located on the lower extremities, accounting for 81% of cases. The upper extremities were affected in 10% of cases, while 2.2% of patients had multiple lesions in different sites.

Of all the patients enrolled, 77% had no family history of mycetoma. Most patients with a family history were from families with multiple cases discovered during field visits to endemic areas. Those familial relationships are listed in **Table 5.2**.

Table 5.2 *The demographic features of mycetoma cases (n=309).*

Characteristic	**Number (%)**
Duration of illness before the date of survey (in years)	
Mean (SD)	10 (8)
Median (range)	7 (0.3-38)
Site of the lesion	
Upper extremities	31 (10%)
Lower extremities	250 (81%)
Head, neck, trunk, back and perineal area	18 (5.8%)
Multiple sites	7 (2.2%)
N/A	3 (1%)
Family history of mycetoma	
Parent	3 (1%)
Offspring	1 (0.3%)
Sibling(s)	24 (7.8%)
Avuncular	15 (4.9%)
Grandparent	2 (0.6%)
Cousin(s)	24 (7.8%)
No family history	238 (77%)
N/A	2 (0.6%)

Avuncular: uncle/aunt to niece/nephew; N/A: not available.

5.3 Findings of data quality control

Quality control included the identification and exclusion of individuals and SNPs with missing genotype proportion of more than 2%, individuals with conflicting sex information or outlying heterozygosity, and SNPs that were not in HWE or had low MAF of less than 0.05.

Summary statistics of the raw data showed that 473 (99.8%) individuals had a genotyping missingness rate of less than 0.02, while 95.6% of the SNPs had per-SNP missingness of less than 0.02 (**Figure 5.3**). Three affected individuals who had a heterozygosity rate deviating ±3 SD from the samples' heterozygosity rate mean were excluded. Eight individuals had discrepancies between self-reported sex and genetically inferred sex and were removed.

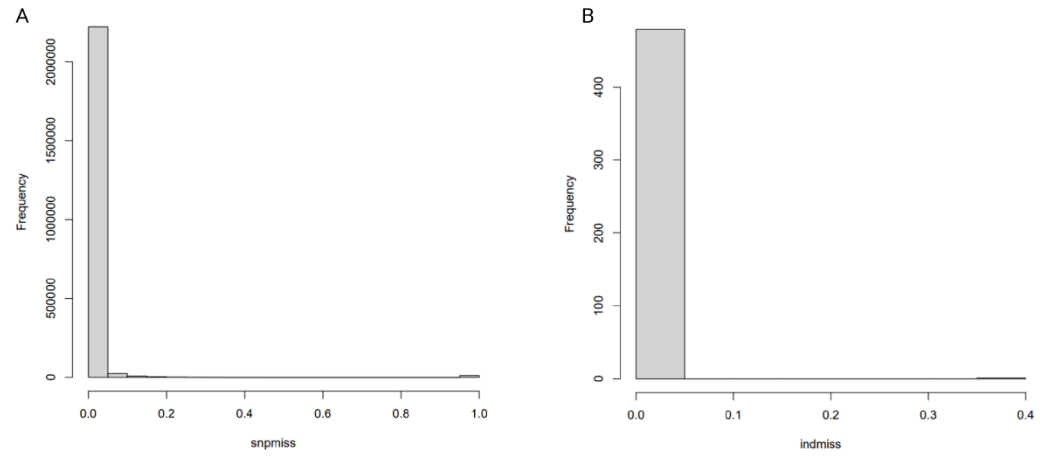


Figure 5.3 *Histograms of SNPs (A) and individuals (B) missingness.*

Thirteen individuals were excluded for having pi-hat=1 (duplicates). Distortion from HWE at p<1e−6 was observed for 698 (0.05%) SNPs among controls and for 378 (0.026%) SNPs among the cases at p<1e−10. A total of 1415704 SNPs remained after excluding low-frequency SNPs with MAF<0.05 (**Figure 5.4**).

These steps resulted in a dataset with 1,414,628 SNPs and 450 subjects (297 cases and 153 controls), upon which additional data quality filters were applied.

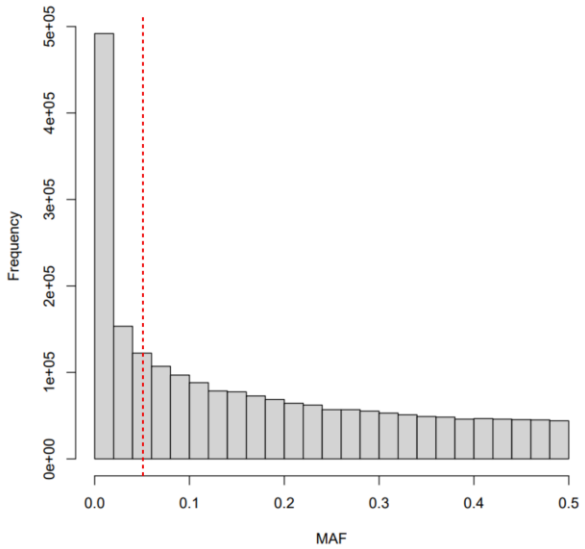


Figure 5.4 *SNP minor allele frequency (MAF). Low-frequency SNPs (MAF < 5%) were excluded.*

Multidimensional Scaling (MDS) was employed to detect population outliers by visualising genetic relationships between individuals. In MDS plots, individuals from the same population cluster together due to shared genetic profiles. Outliers, those

distant from their designated clusters, are identified as individuals with significant genetic differences from their assigned population. Detecting outliers is crucial for quality control in GWAS, as they can introduce noise and lead to spurious results. It also helps identify subpopulations within the dataset, which may have distinct genetic risk factors for the disease under study. Both visual inspection and statistical methods are used to identify and handle outliers effectively, ensuring the reliability of genetic associations in GWAS analyses (274).

The results of the MDS analysis indicated that all individuals within the mycetoma dataset belonged to the African superpopulation cluster of the 1000 Genomes (1KG) Project reference panel. This effectively eliminated the possibility of any population outliers, as shown in **Figure 5.5**. Consequently, none of the individuals were removed from the association analysis due to population stratification.

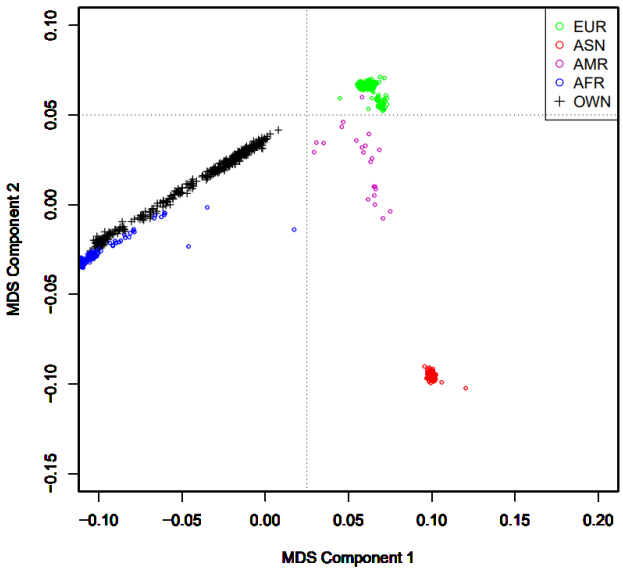


Figure 5.5 *Multidimensional scaling (MDS) plot of 1000 Genomes (1KG) reference panel against the mycetoma dataset. The black crosses (+ = "OWN") in the middle left part represent the first two MDS components of the individuals in the mycetoma dataset. Blue circles represent Africans (AFR); green circles represent Europeans (EUR); pink circles represent Admixed Americans (AMR); red circles represent Asians (ASN).*

5.4 Data analysis results

5.4.1 Estimating the genomic inbreeding coefficient (F_{ROH})

The F_{ROH} assessment determines an individual's genomic inbreeding coefficient by analysing the runs of homozygosity (ROH) in their genetic profile. The term ROH refers to genome segments where an individual has inherited identical alleles from both parents due to shared ancestry (identical by descent). By indicating the probability of inheriting two identical alleles from a recent common ancestor, the genomic inbreeding coefficient can reveal whether there has been inbreeding or shared ancestry within the population.

The ROH calculation and F_{ROH} estimation were performed using PLINK and detectRUNS R package. A total of 10294 runs of homozygosity (ROH) were identified, while the mean F_{ROH} in this dataset was 0.0317 (**Figure 5.6**).

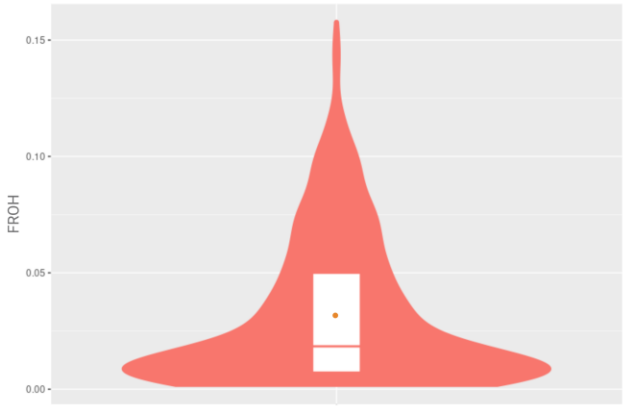


Figure 5.6 *Violin plot showing ROH-based inbreeding coefficient (F_{ROH}) calculated in the GWAS dataset. The orange dot represents the mean value (0.0317).*

5.4.2 Association analysis

The genome-wide association analysis was done using mixed linear model association analysis (MLMA) implemented in the genome-wide complex trait analysis (GCTA) software package and generalised mixed model (GMM) implemented in the scalable and accurate implementation of generalised mixed model (SAIGE) R package. The total number of SNPs tested was 1,414,628. The following sections will include the results of the two methods as well as a comparison between the produced results (section 5.4.2.3)

5.4.2.1 Association analysis using genome-wide complex trait analysis (GCTA)

I used a linear mixed model (LMM) approach for this association analysis, accounting for relatedness between individuals, as the dataset contained families. The Quantile-quantile (QQ)-plot in **Figure 5.7** is a visualisation technique commonly used in GWAS to assess the presence of systematic biases or inflation in the association test results. To determine the inflation of test statistics compared to the expected null distribution, the genomic inflation factor (lambda) was measured. Lambda values close to 1 indicate minimal inflation, suggesting that the observed *p*-values follow the expected distribution closely. The lambda value after the analysis was 0.999, suggesting minimal genomic inflation. This indicated that the GWAS analysis has effectively controlled for potential confounding factors and population stratification. Thus, it provided confidence in the validity and reliability of the GWAS results, as it indicated that the association test statistics align with the null hypothesis.

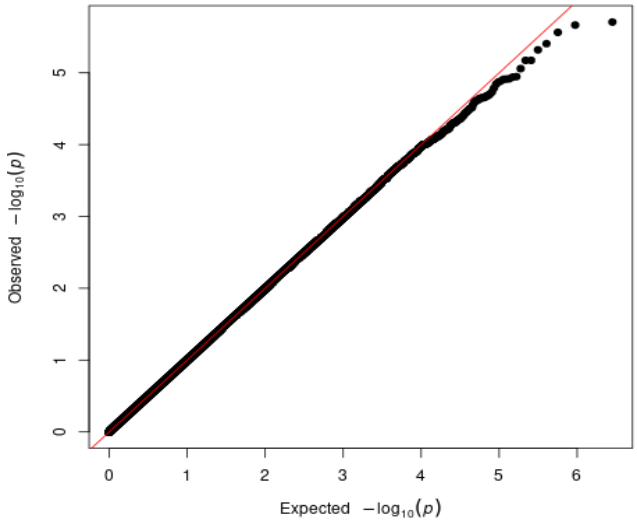


Figure 5.7 *Quantile-quantile (QQ) plot of -log$_{10}$ (p) from GCTA's GWAS analysis. The x-axis represents the expected -log$_{10}$ transformed p-values (based on the null distribution), and the y-axis represents the observed -log$_{10}$ transformed p-values.*

Although the QQ plot suggested that the overall set of test statistics was consistent with the global null hypothesis of no associations, several suggestive associations were identified, with eight SNPs surpassing the genome-wide threshold of $p < 1.0\times10^{-5}$, as demonstrated in Figure 5.8. The Manhattan plot in **Figure 5.8** graphically represents the association results by displaying SNPs' genomic locations (chromosomes) vs. their corresponding association significance (-log$_{10}$ transformed p-values).

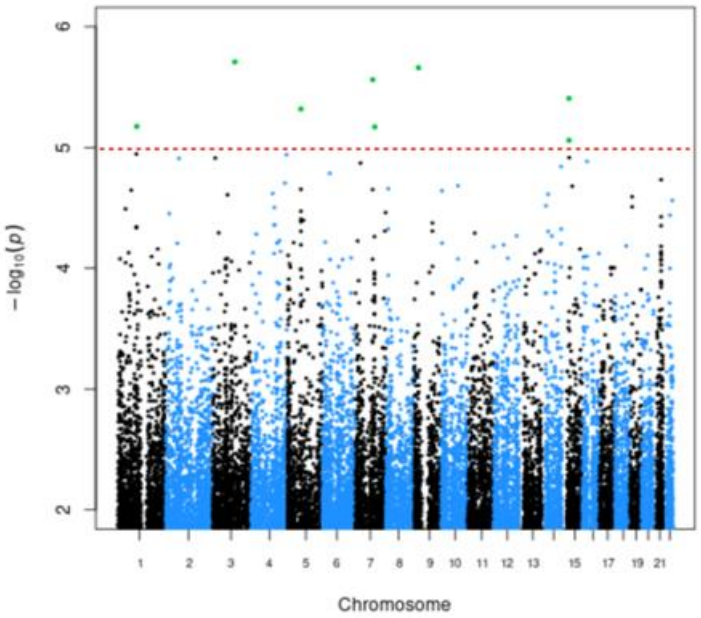


Figure 5.8 *Manhattan plot of GCTA association results. It displays chromosomes on the x-axis and the SNPs' corresponding association significance (-log$_{10}$ transformed p-values) on the y-axis. The top eight SNPs are depicted as green dots. The dashed red line denotes the genome-wide suggestive significance threshold of $p < 1.0\times10^{-5}$.*

The GCTA LMM revealed eight loci associated with mycetoma at suggestive levels of significance (p <1.0x10^{-5}). Two loci lie on chromosome 7, two on chromosome 15 and one on chromosomes 1, 3, 5 and 9 (**Table 5.3**). The highest scoring SNP was rs16861260 (p = 1.97E-06), an intergenic variant located 1.118 kb downstream to the gene upstream transcription factor family member 3 (*UCF3*), also known as *KIAA2018,* on chromosome 3. The associated SNP on chromosome 5, rs9291949 (p = 4.80E-06), is intronic in the adenylate kinase 6 (*AK6*) gene. On chromosome 7, the associated loci included two top SNPs; the first was rs10952971 (p = 2.73E-06), an intergenic variant located 16.1 kb near a pseudogene named RNA, U6 small nuclear 274 (*RNU6-274P*). The second locus included rs45529834 (p = 6.72E-06), an intronic variant mapped to four genes: cleavage and polyadenylation specific factor 4 (*CPSF4*), pentatricopeptide repeat domain 1 (*PTCD1*), ATP synthase membrane subunit F (*ATP5J2*) and *ATP5J2-PTCD1* readthrough (fusion gene). Chromosome 15 also included suggestive signals in two loci located on q14. The lead SNPs on 15q14 were rs10520048 (p = 3.93E-06) and rs319883 (p = 8.75E-06) intronic variants mapped to long non-coding RNAs (lncRNAs) genes.

Table 5.3 *Lead SNPs showing association with mycetoma under the linear mixed model of the GCTA software package.*

SNP	Chr	P (GRCh37)	Ref	Alt	MAF	Beta (SE)	p-value	Nearest gene	Global MAF	1KG (AFR)	GnomAD (AFR)	Cytoband
rs167677	1	96468117	G	A	0.18	-0.17 (0.04)	6.71E-06	*RP11-147C23.1**	0.34	0.86	0.83	p21.3
rs16861260	3	113366113	G	A	0.06	-0.3 (0.06)	1.97E-06	*UCF3*	0.09	0.11	0.11	q13.2
rs9291949	5	68648770	G	A	0.4	-0.14 (0.03)	4.80E-06	*AK6*	0.44	0.37	0.37	q13.2
rs10952971	7	89400148	C	A	0.23	-0.17 (0.04)	2.73E-06	*RNU6-274P*	0.27	0.01	0.08	q21.13
rs45529834	7	99049682	C	T	0.19	-0.18 (0.04)	6.72E-06	*PTCD1; ATP5J2-PTCD1; CPSF4; ATP5J2*	0.05	0.16	0.14	q22.1
rs2039461	9	20145988	T	C	0.27	-0.16 (0.03)	2.17E-06	*MLLT3*	0.3	0.1	0.17	p21.3
rs319883	15	35955716	A	G	0.28	-0.15 (0.03)	8.75E-06	*DPH6-DT**	0.29	0.6	0.8	q14
rs10520048	15	36345304	C	T	0.28	-0.15 (0.03)	3.93E-06	*RP11-184D12.1**	0.28	0.3	0.29	q14

Chr: chromosome; P: position in GRCh37 coordinate; Ref: reference allele; Alt: alternate allele; MAF: minor allele frequency; SE: standard error; 1KG: 1000 Genomes; *lincRNA: long intergenic non-coding RNA.

5.4.2.2 Association analysis using the scalable and accurate implementation of generalised mixed model (SAIGE)

Single-variant association analysis was done using SAIGE, which allows association testing in binary traits using a GMM that accounts for population structure and relatedness among individuals. This software package is known to have the most computationally efficient method when studying a binary trait with a small, correlated dataset (331).

Figure 5.9 demonstrates the QQ plot produced from this analysis. The lambda value was 1.032, suggesting that the association results are generally reliable and not substantially biased due to confounding or population structure.

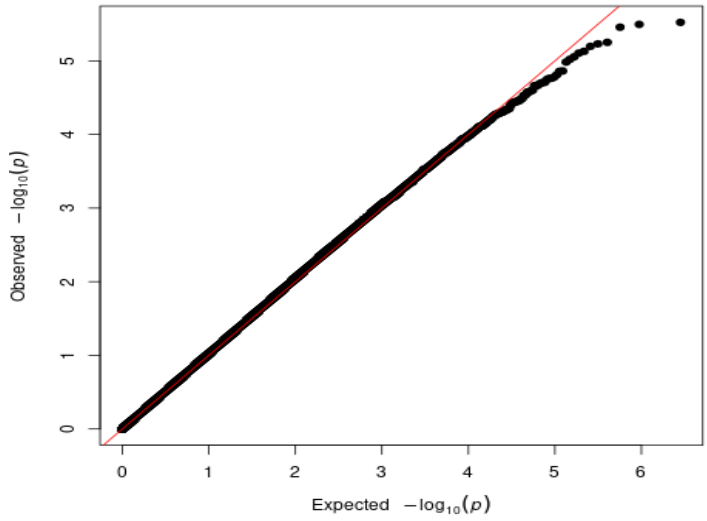


Figure 5.9 *Quantile-quantile (QQ) plot of -log$_{10}$ (p) from SAIGE's GWAS analysis. The x-axis represents the expected -log$_{10}$ transformed p-values (based on the null distribution), and the y-axis represents the observed -log$_{10}$ transformed p-values.*

Again, the QQ plot suggested that the overall set of test statistics was consistent with the global null hypothesis of no associations. In **Figure 5.10**, the Manhattan plot indicates that ten SNPs attained a suggestive significance level with a p-value of less than 1.0×10^{-5}. Three SNPs were found in both GCTA and SAIGE analyses, specifically rs16861260 on chromosome 3 and rs319883 and rs10520048 on chromosome 15 (**Table 5.4**). The minor allele A of the highest-scoring SNP, rs16861260, was identified by both GCTA and SAIGE and offered approximately 4.6-fold protection against mycetoma with an odds ratio (OR) of 0.22 (p = 1.97E-06 for GCTA and p = 6.35E-06 for SAIGE). Moreover, three loci on chromosome 15q14 were identified as having a protective effect against mycetoma. One of these loci

encompasses rs319885 (*p* = 7.43E-06) and rs319883 (*p* = 5.87E-06) mapped to an RNA gene known as DPH6 divergent transcript (*DPH6-DT*). In all three risk loci on 15q14, the minor alleles offered 2.3-fold increase in protection against mycetoma susceptibility (**Table 5.4**).

On the other hand, two risk loci on chromosome 2 were associated with increased odds of susceptibility to mycetoma. The first locus included rs1868403, mapped to the C1D nuclear receptor corepressor (*C1D*) gene with an OR of 2.68 (95% CI 2.23 to 3.13). The second associated SNP, rs4368333 (*p* = 3.20E-06), is intergenic about 28.4 kb upstream of the gene potassium voltage-gated channel modifier subfamily S member 3 (*KCNS3*). The A minor allele of this SNP was associated with a 2-fold increased susceptibility to mycetoma with an OR of 2.18 (95% CI 1.84 to 2.52). Another top SNP increasing the risk of mycetoma was found on chromosome 14q32.2. The associated SNP was rs4073738 (*p* = 3.49E-06; OR 2.26, 95% CI 1.9 to 2.61), an intronic variant in the gene brain-enriched guanylate kinase-associated (*BEGAIN*). Chromosome 7 had a risk locus different from those identified in GCTA analysis with rs4076072 (*p* = 5.61E-06) located in Sad1 and UNC84 domain containing 1 (*SUN1*) gene.

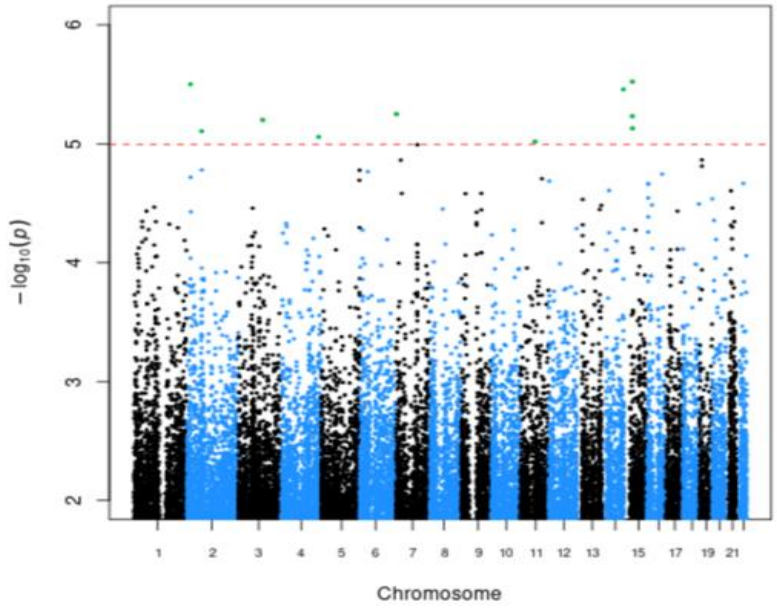


Figure 5.10 *Manhattan plot of SAIGE association results. It displays chromosomes on the x-axis and the SNPs' corresponding association significance (-log$_{10}$ transformed p-values) on the y-axis. The top ten SNPs are depicted as green dots. The dashed red line denotes genome-wide suggestive significance threshold of $p < 1 \times 10^{-5}$.*

Table 5.4 *Lead SNPs showing association with mycetoma under the generalised mixed model of SAIGE R package.*

SNP	Chr	P(GRCh37)	Ref	Alt	MAF	OR (95% CI)	p-value	Nearest gene	Global MAF	1KG (AFR)	GnomAD (AFR)	Cytoband
rs4368333	2	18030665	C	A	0.5	2.18 (1.84 – 2.52)	3.20E-06	*KCNS3*	0.37	0.42	0.43	p24.2
rs1868403	2	68310101	A	G	0.22	2.68 (2.23 – 3.13)	7.83E-06	*C1D*	0.22	0.22	0.24	p14
rs16861260	3	113366113	G	A	0.06	0.22 (-0.5 – 0.9)	6.35E-06	*UCF3*	0.09	0.11	0.11	q13.2
rs17062162	4	176770168	T	C	0.1	0.33 (-0.2 – 0.84)	8.81E-06	*GPM6A*	0.02	0.06	0.05	q34.2
rs4076072	7	892902	C	T	0.08	0.26 (-0.4 – 0.86)	5.61E-06	*SUN1*	0.17	0.08	0.08	p22.3
rs3019751	11	68923597	A	G	0.45	2.21 (1.85 – 2.57)	9.55E-06	*RP11-554A11.8**	0.5	0.56	0.52	q13.3
rs4073738	14	101015380	G	A	0.45	2.26 (1.9 – 2.61)	3.49E-06	*BEGAIN*	0.46	0.62	0.58	q32.2
rs319885	15	35955025	T	C	0.29	0.44 (0.08 – 0.81)	7.43E-06	*DPH6-DT**	0.29	0.86	0.8	q14
rs319883	15	35955716	C	T	0.29	0.44 (0.07 – 0.81)	5.87E-06	*DPH6-DT**	0.29	0.86	0.8	q14
rs10520048	15	36345304	C	T	0.28	0.44 (0.08 – 0.8)	3.00E-06	*RP11-184D12.1**	0.28	0.29	0.29	q14

Chr: chromosome; P: position in GRCh37 coordinate; Ref: reference allele; Alt: alternate allele; MAF: minor allele frequency; OR: odds ratio; CI: confidence interval; 1KG: 1000 Genomes; *lincRNA: long intergenic non-coding RNA.

5.4.2.3 Pairwise comparison of the association results

To evaluate the similarity of the results produced by GCTA and SAIGE, a pairwise comparison of -log$_{10}$(*p*-values) was done using R (**Figure 5.11**). The results of SAIGE and GCTA were concordant, and both software packages identified three suggestive signals associated with mycetoma, as detailed in section 5.4.2.2.

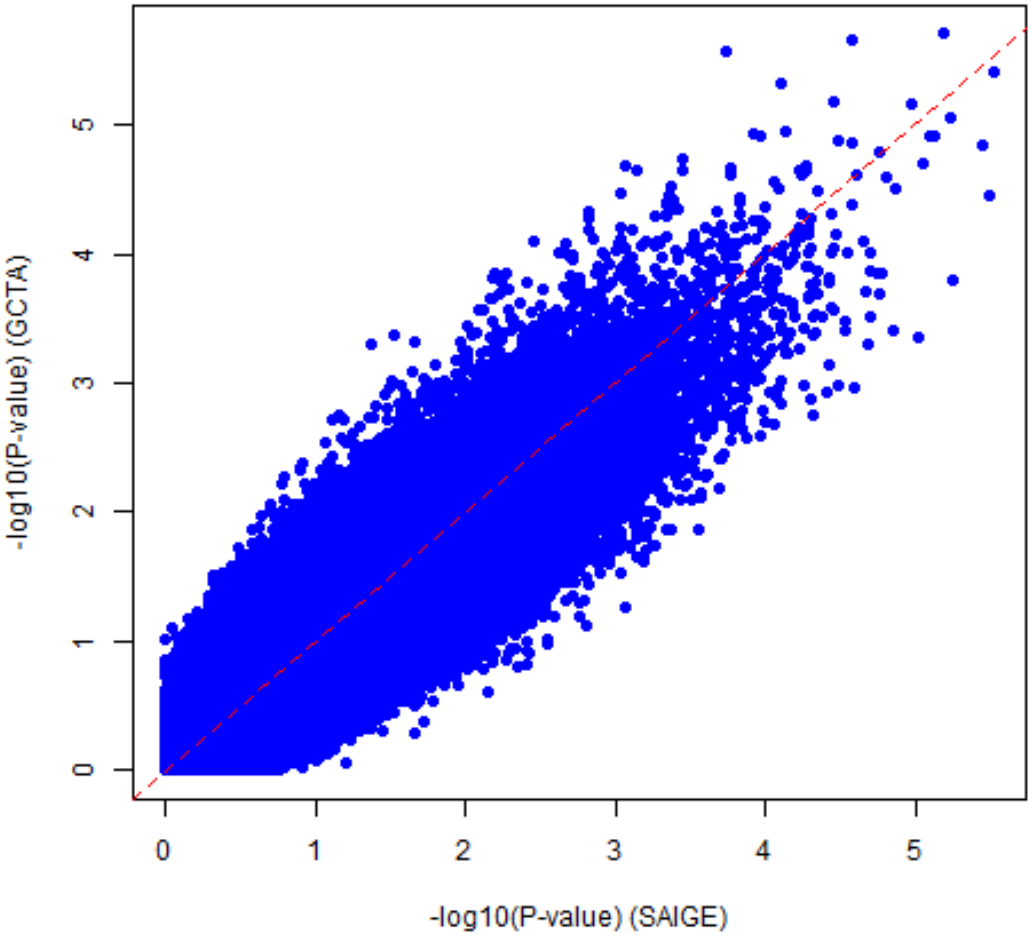


Figure 5.11 *Pairwise comparison plot of –log10(P-values) for the associations based on GMM using SAIGE on the x-axis versus associations based on MLMA using GCTA on the y-axis.*

5.4.3 Regional association

LocusZoom was used to visualise the genomic regions of the lead SNPs produced by the two association analysis models.

5.4.3.1 Lead SNPs from GCTA association analysis

The most promising suggestive signals (rs9291949, rs10952971, rs45529834 and rs167677) were further interrogated regarding their reliability given the LD pattern (**Figure 5.12**).

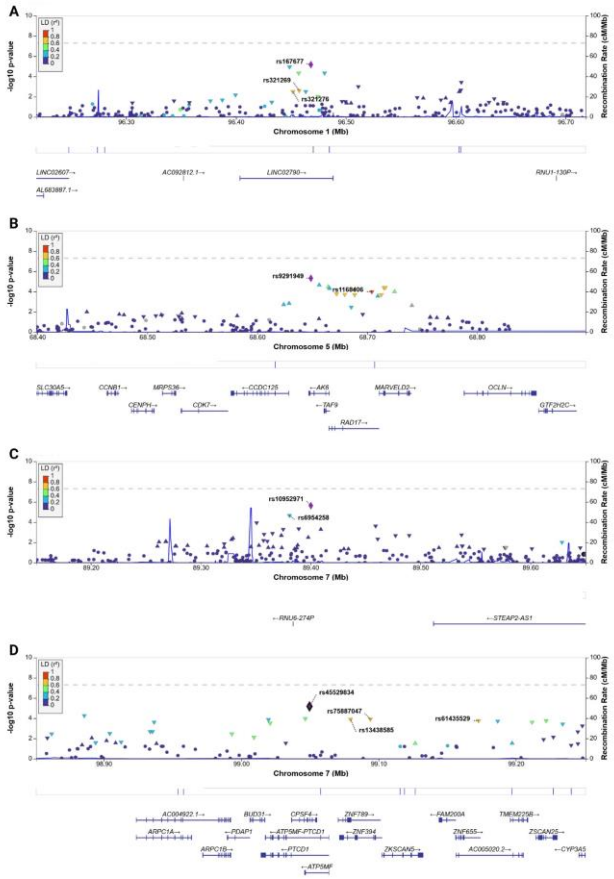


Figure 5.12 *Regional association plots for (A) rs167677, (B) rs9291949, (C) rs10952971, and (D) rs45529834. The x-axis displays the chromosomal position while the y-axis represents SNPs' corresponding association significance (-log$_{10}$ transformed p-values). The recombination rate is displayed as a continuous blue line. Individual SNPs are represented by a circle/triangle filled with a colour corresponding to the extent of LD with the lead SNP (a purple diamond) from dark blue (r^2 = 0) to red (r^2 = 1). Grey-filled circles/triangles refer to SNPs with no LD information. Symbol orientation indicates the effect sign; regular triangles indicate positive associations, whereas inverted triangles indicate negative associations. The dashed line denotes the genome-wide significance of $p < 5 \times 10^{-8}$. The lower section shows where the genes are located and their corresponding exons with the direction of transcription indicated by arrows.*

5.4.3.2 Lead SNPs from SAIGE association analysis

Figure 5.13 demonstrates the top SNPs from SAIGE association testing, which gave promising LD patterns. This includes rs4368333, rs1868403, rs4076072 and rs4073738. However, no significant LD pattern was present for rs17062162 and rs3019751.

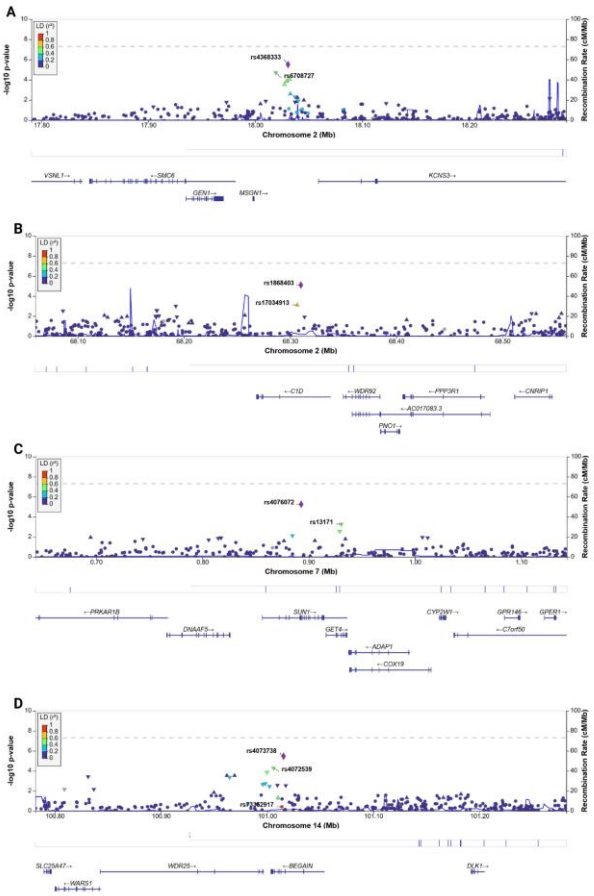


Figure 5.13 *Regional association plots for (A) rs4368333, (B) rs1868403, (C) rs4076072 and (D) rs4073738. The x-axis displays the chromosomal position while the y-axis represents SNPs' corresponding association significance (-log$_{10}$ transformed p-values). The recombination rate is displayed as a continuous blue line. Individual SNPs are represented by a circle/triangle filled with a colour corresponding to the extent of LD with the lead SNP (a purple diamond) from dark blue (r^2 = 0) to red (r^2 = 1). Grey-filled circles/triangles refer to SNPs with no LD information. Symbol orientation indicates the effect sign; regular triangles indicate positive associations, whereas inverted triangles indicate negative associations. The dashed line denotes genome-wide significance of $p < 5 \times 10^{-8}$. The lower section shows where the genes are located and their corresponding exons with the direction of transcription indicated by arrows.*

5.4.3.3 Common Lead SNPs

Lastly, the LD patterns for rs16861260, rs319883 and rs10520048 were plotted; however, rs319883 was the only one with a significant LD pattern. It had a perfect LD (r^2=1) with rs319885 (another top SNP according to SAIGE association testing), as depicted in **Figure 5.14**.

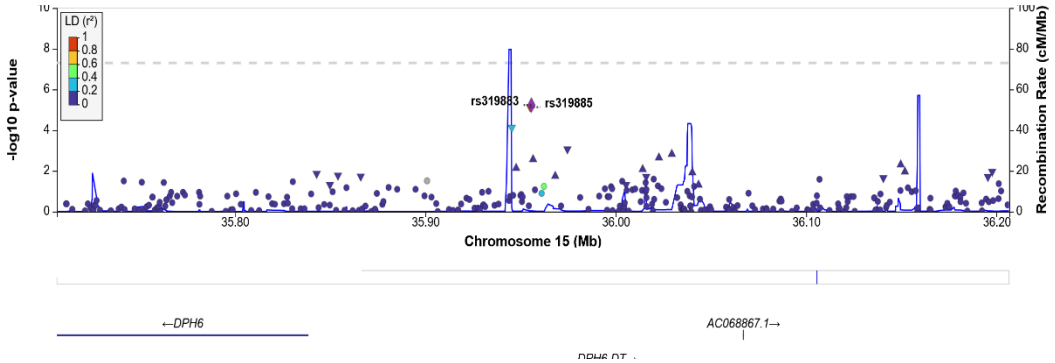


Figure 5.14 *Regional association plot for rs319883. The x-axis displays the chromosomal position while the y-axis represents SNPs' corresponding association significance (-log10 transformed p-values). The recombination rate is displayed as a continuous blue line. Individual SNPs are represented by a circle/triangle filled with a colour corresponding to the extent of LD with rs319883 (a purple diamond) from dark blue (r^2 = 0) to red (r^2 = 1). The other lead SNP, rs319885, is displayed as a red triangle behind rs319883. Grey-filled circles/triangles refer to SNPs with no LD information. Symbol orientation indicates the effect sign; regular triangles indicate positive associations, whereas inverted triangles indicate negative associations. The dashed line denotes genome-wide significance of $p < 5 \times 10^{-8}$. The lower section shows where the genes are located and their corresponding exons with the direction of transcription indicated by arrows.*

5.4.4 Annotation and Functional Analysis

The functional mapping and annotation of genome-wide association studies (FUMA) web application provided additional information to interpret the overall set of GWAS results. This was done on the GWAS summary statistics of GCTA and SAIGE association analyses.

5.4.4.1 GCTA lead SNPs

To assess the functional consequence of the input SNPs on genes, ANNOVAR enrichment test in FUMA was used. The 1KG Phase 3 dataset for Africans was chosen as a reference dataset to represent the background distribution of variants across the genome. By comparing the observed variants to this reference dataset, the enrichment test could identify functional categories significantly overrepresented or underrepresented in the mycetoma dataset, providing insights into the potential functional implications of the variants. The functional consequence of the SNPs on genes in GCTA's summary statistics is displayed in **Figure 5.15**. Most of the annotated SNPs were intronic (53.6%, p = 9E-04), followed by intergenic (22%, p = 1.38E-06), intronic noncoding-RNA (ncRNA) variants (12.6%, p = 0.74) and 3' untranslated regions (3' UTRs) variants (8%, p = 3.5E-06).

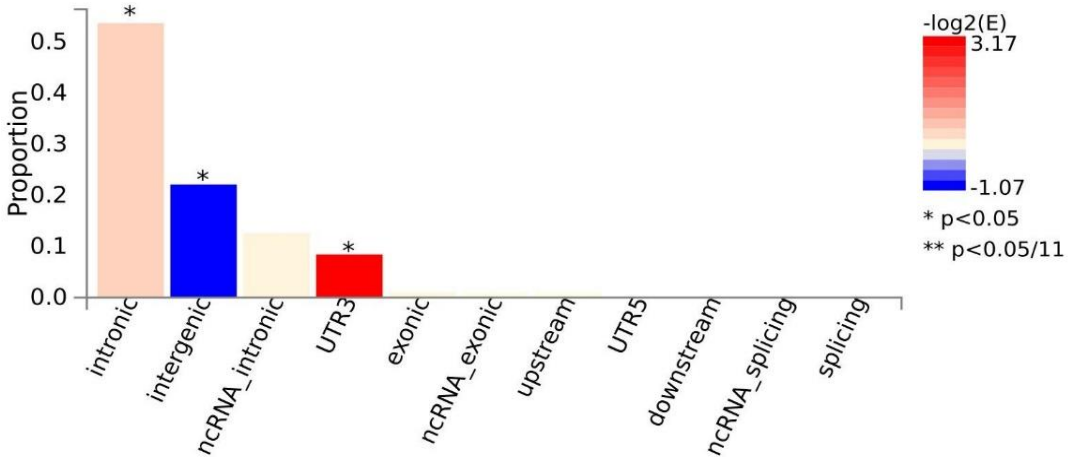


Figure 5.15 *Histogram displaying the proportion of SNPs (all SNPs in LD with GCTA's lead SNPs) with corresponding functional annotation assigned by ANNOVAR. Bars are coloured according to log2 (enrichment) relative to all SNPs in the 1000 Genomes Phase 3 dataset for Africans.*

Gene-based association analysis was done using the MAGMA tool implemented in FUMA. The input SNPs were mapped to 19150 protein-coding genes. **Figure 5.16** demonstrates the top six associated genes identified by MAGMA. The potassium inwardly-rectifying channel subfamily J member 6 (*KCNJ6*) gene on chromosome 21 was significantly associated with mycetoma (p = 3.36E-07).

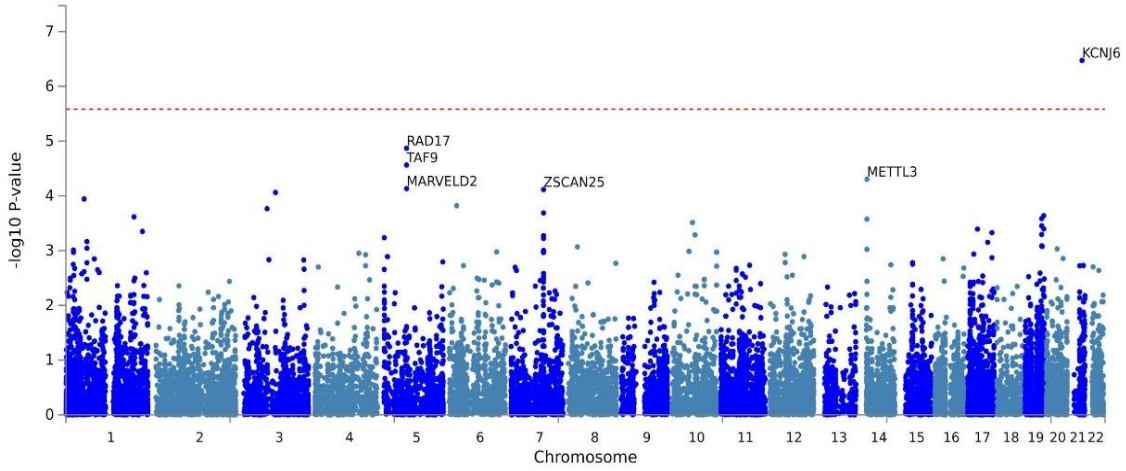


Figure 5.16 *Gene-based genome-wide association Manhattan plot of GCTA summary statistics, with the top 6 associated genes labelled. The y-axis denotes the –log$_{10}$(p-value) for association, while the x-axis is the physical position of the SNPs across the genome. The red dashed line indicates the genome-wide significance threshold (p = 0.05/19150 = 2.611E-6).*

FUMA also offers three key functional mapping strategies for interpreting GWAS findings. First, positional mapping helps identify nearby genes to GWAS-identified SNPs, aiding in the discovery of candidate genes. Second, Expression Quantitative Trait Loci (eQTL) mapping uncovers SNPs associated with gene expression changes, shedding light on the functional consequences of GWAS variants. Lastly, chromatin interaction mapping explores 3D organisation, revealing SNPs interacting with distal regulatory elements, thus unravelling long-range regulatory effects. These strategies collectively help bridge the gap between genetic associations and functional insights.

A total of 87 candidate genes were identified using FUMA functional mapping strategies, as listed in Appendix 11. These include 19, 13, and 81 genes identified by positional mapping, eQTL mapping, and chromatin interaction mapping, respectively (**Figure 5.17**). Mutual genes identified by the three mapping strategies were *USF3, PTCD1* and *ATP5J2*.

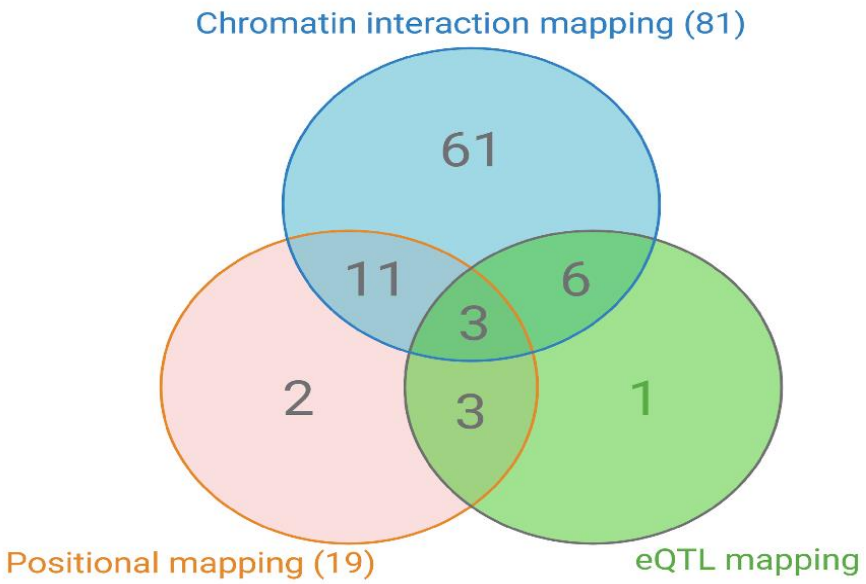


Figure 5.17 *The number of overlapping genes between the three functional mapping strategies (positional mapping, eQTL, and chromatin interactions) identified from GCTA's summary statistics. The intersecting areas of the circles represent the number of overlapping genes.*

Circos plots in FUMA display genomic and functional information in a circular layout. They are commonly used to visualise relationships and connections between various elements, such as genomic risk loci, genes, functional annotations, and GWAS results, in an informative way. The circos plot in **Figure 5.18A** showed multiple chromatin interactions between the genomic risk locus (chr1:96400001–96440000) and genes calponin 3 *(CNN3*), asparagine-linked glycosylation 14 homolog *(ALG14),* RWD domain-containing protein 3 *(RWDD3),* polypyrimidine tract-binding protein 2 *(PTBP2),* dihydropyrimidine dehydrogenase *(DPYD)* and sorting nexin 7 *(SNX7)*. Based on the RNA-seq data from GTEx version 8, a heatmap of gene expression showed lower expression levels of *CCN3* and *RWDD3* in whole blood (**Figure 5.19**).

The 3q13.2 locus was identified to have chromatin interactions with 19 genes (**Figure 5.18B**), including genes involved in immunoregulation such as B- and T-lymphocyte attenuator (*BTLA*), Cluster of Differentiation 200 (*CD200*) and its receptor CD200 receptor 1 (*CD200R1*), and germinal centre-associated signalling and motility (*GCSAM*). The heatmap of gene expression in **Figure 5.19** showed higher relative expression levels of

BTLA, GCSAM, and SID1 transmembrane family member 1 (*SIDT1*) genes in Epstein-Barr virus-transformed B lymphocytes. The gene *GCSAM* was also relatively upregulated in suprapubic and lower leg skin tissues. The lead SNP rs9291949 of chromosome 5 locus interacted with the 12 genes (**Figure 5.18C**) including TATA-box binding protein associated factor 9 (*TAF9*) which had a lower expression level in whole blood (**Figure 5.19**). The two loci on chromosome 7 had the highest share of interactions with 41 genes (**Figure 5.18D**), six identified by chromatin interaction and eQTL mappings. **Figure 5.18E** illustrates the correlation between the two risk loci on chromosome 15 and six genes. Among these genes was RAS guanyl-releasing protein 1 (*RASGRP1*), which is highly expressed in T-cells and, to a lesser extent, in B and natural killer cells. The encoded protein, RASGRP1, is critical in regulating the Ras signalling pathway, which is vital for immune responses. Research has shown that T-cell development and function rely on RASGRP1, as it connects the TCR to the Ras-mitogen-activated protein kinase signalling pathway (332,333). No interactions were identified in the chromosome 9 locus.

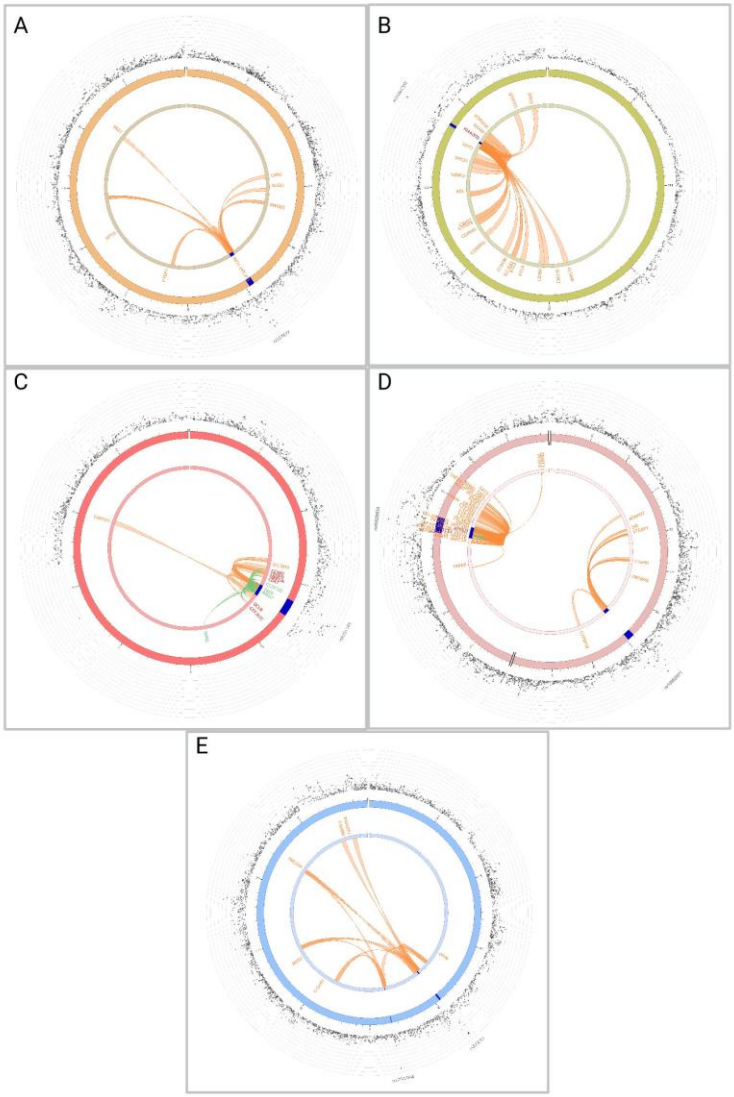


Figure 5.18 *Circos plot generated via FUMA for the top SNPs of GCTA. The outermost layer is the Manhattan plot, and the middle layer highlights genomic risk loci in blue, while the innermost layer highlights eQTLs and/or chromatin interactions. Only SNPs with $p < 0.05$ are displayed in the outer ring. SNPs in genomic risk loci are colour-coded as a function of their maximum r^2 to one of the independent significant SNPs in the locus. The rsID of the top SNPs is displayed in the outermost layer. For the innermost layer, if the gene is mapped only by chromatin interactions or only by eQTLs, it is coloured orange or green, respectively. It is coloured red when both map the gene. (A) Circos plot for chromosome 1 top SNP. (B) Circos plot for chromosome 3 top SNP. (C) Circos plot for chromosome 5 top SNP. (D) Circos plot for the two top SNPs on chromosome 7. (E) Circos plot for chromosome 15 top SNPs.*

The 87 mapped genes were used as the input for the second step of FUMA's analysis, GENE2FUNC, to identify enriched pathways or gene sets associated with mycetoma. This was done by utilising the Molecular Signatures Database (MSigDB) (281), which provides a comprehensive collection of curated gene sets representing diverse biological functions and pathways. This analysis revealed two gene sets in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways primarily related to drug and linoleic acid metabolism (**Table 5.5**). Furthermore, based on the gene ontology (GO) analysis, the molecular function significantly enriched was the Estrogen 16α-hydroxylase activity. Among nine gene sets reported in the GWAS catalogue, serum metabolite levels and sex hormones levels were the most relevant (complete list in Appendix 12).

Table 5.5 *The MSigDB significantly enriched gene sets from the GCTA association analysis dataset.*

Category	Gene set	genes	p-value	Adj. p-value
GO molecular function	Estrogen 16 alpha-hydroxylase activity	CYP3A5, CYP3A4, CYP3A43	2.96E-06	5.24E-3
KEGG	Drug metabolism - other enzymes	DPYD, CYP3A5, CYP3A4, CYP3A43	5.21E-5	9.69E-3
	Linoleic acid metabolism	CYP3A5, CYP3A4, CYP3A43	2.32E-04	2.16E-2
GWAS catalogue reported genes	Serum metabolite levels	DPYD, ARPC1A, ARPC1B, PDAP1, BUD31, CPSF4, ZNF789, ZKSCAN5, CYP3A5, CYP3A43, TRIM4	1.87E-7	1.66E-4
	Sex hormone levels	ZNF789, ZKSCAN5, CYP3A4	7.05E-6	5.20E-3

GO: Gene Ontology; Adj. P-value: Adjusted p-value after accounting for multiple testing.

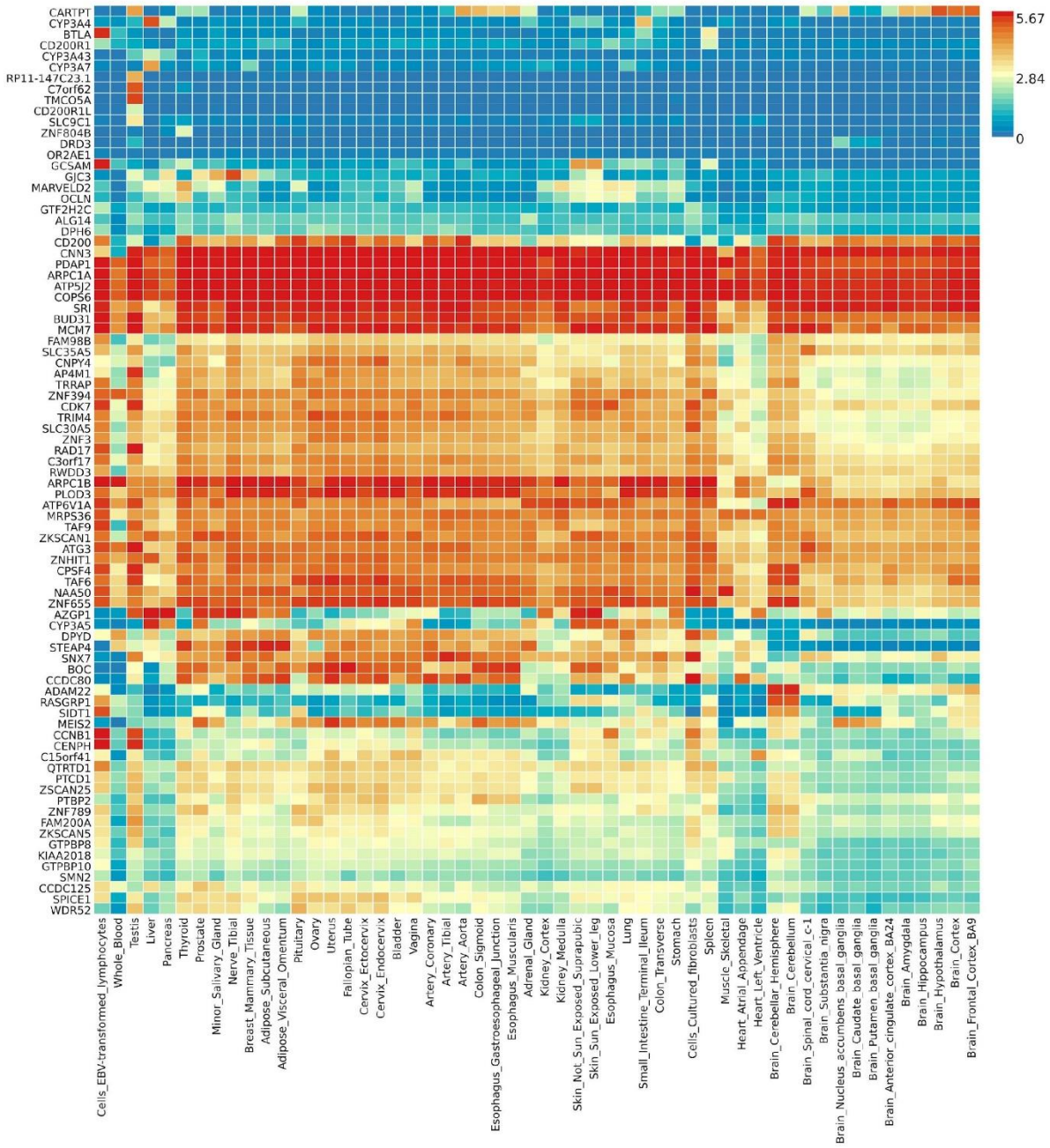


Figure 5.19 *Gene expression heatmap constructed with GTEx v8 (53 tissues) for GCTA's mapped genes. Genes (x-axis) and tissues (y-axis) are ordered by clusters with expression levels displayed as log2 transformed. The colour scale represents the expression levels ranging from red for high expression to blue for low expression.*

5.4.4.2 SAIGE lead SNPs

Figure 5.20 demonstrates the functional consequence of the SNPs on genes in the SAIGE's summary statistics. The annotated SNPs were 41.6% intergenic (*p* = 0.62), 36% intronic ncRNA (*p* = 1E-04), 11% 3'UTRs (*p* = 4E-04) and 8% intronic (*p* = 2E-04).

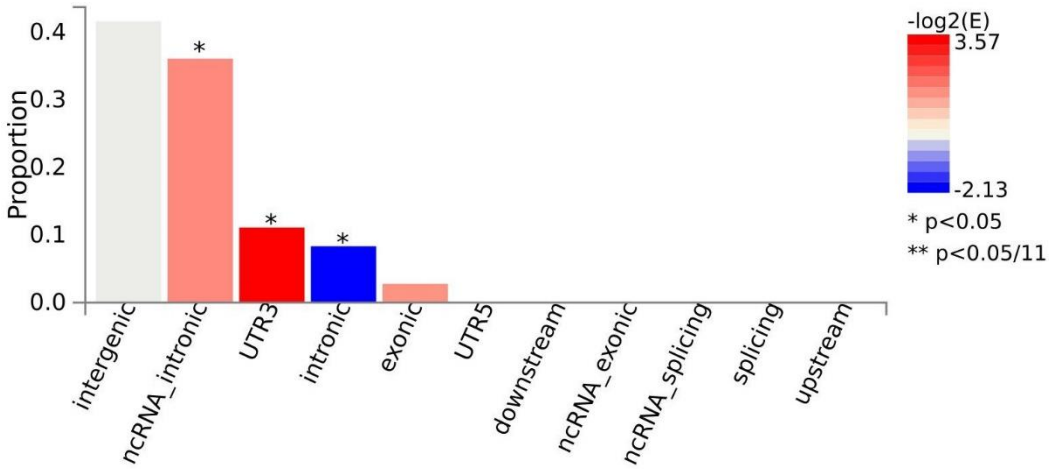


Figure 5.20 *Histogram displaying the proportion of SNPs (all SNPs in LD with SAIGE's lead SNPs) with corresponding functional annotation assigned by ANNOVAR. Bars are coloured according to log2 (enrichment) relative to all SNPs in the 1000 Genomes Phase 3 dataset for Africans (AFR).*

MAGMA gene-based association analysis was done for the input SNPs mapped to 19239 protein-coding genes. The analysis identified no individual gene that was significantly associated with mycetoma (*p* = 2.599E-6), although subthreshold (*p* < 8.6E-05) associations were observed with zinc finger and SCAN domain containing 25 (*ZSCAN25),* complement component 4 binding protein beta *(C4BPB),* cytochrome P450 family 3 subfamily A member 5 *(CYP3A5)* and forkhead box P1 *(FOXP1)* as displayed in **Figure 5.21**.

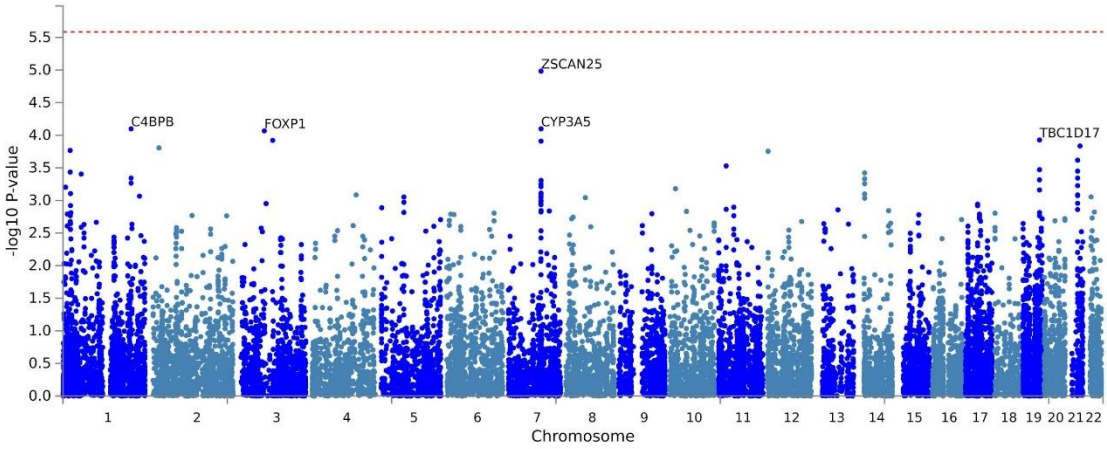


Figure 5.21 *Gene-based genome-wide association Manhattan plot of SAIGE summary statistics, with the top 5 associated genes labelled. The y-axis denotes the –log$_{10}$(p-value) for association, while the x-axis is the physical position of the SNPs across the genome. The red dashed line indicates the genome-wide significance threshold (p = 0.05/19239 = 2.599E-6).*

The functional mapping strategies (positional mapping, eQTL, and chromatin interactions) yielded 48 protein-coding genes (Appendix 13). These include four, nine, and forty-two genes identified by positional mapping, eQTL mapping, and chromatin interaction mapping, respectively (**Figure 5.22**). All three mapping strategies identified two genes, *UCF3* and *SUN1*, while three genes, namely, *C1D,* protein phosphatase 3 regulatory subunit B, Alpha *(PPP3R1)* and cytochrome C oxidase assembly factor COX19 *(COX19),* were predicted by two mapping strategies (eQTL and chromatin interactions).

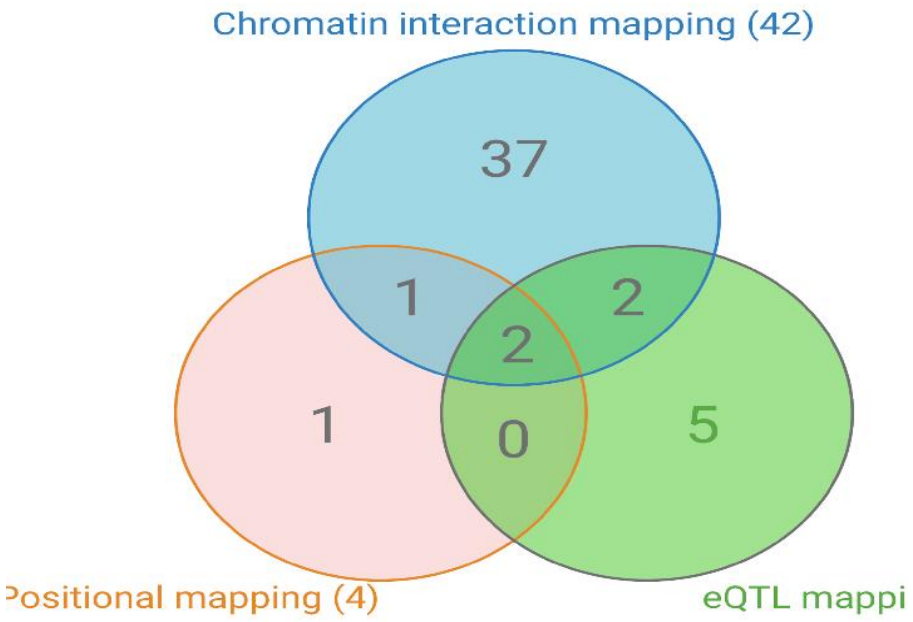


Figure 5.22 *The number of overlapping genes between the three functional mapping strategies (positional mapping, eQTL, and chromatin interactions) identified from SAIGE's summary statistics. The intersecting areas of the circles represent the number of overlapping genes.*

Circos plots of the loci on chromosomes 3 and 15, identified by both GCTA and SAIGE, are displayed in **Figure 5.23B** and **Figure 5.23D**, respectively. **Figure 5.23A** illustrates that only one of the two risk loci on chromosome 2, which harbours the top SNP rs1868403, had chromatin interactions with eight genes. Meanwhile, as evidenced in **Figure 5.23C**, the 11q13.3 locus interacted with five genes, whereas no interactions were found for the loci on chromosomes 4, 7, and 14.

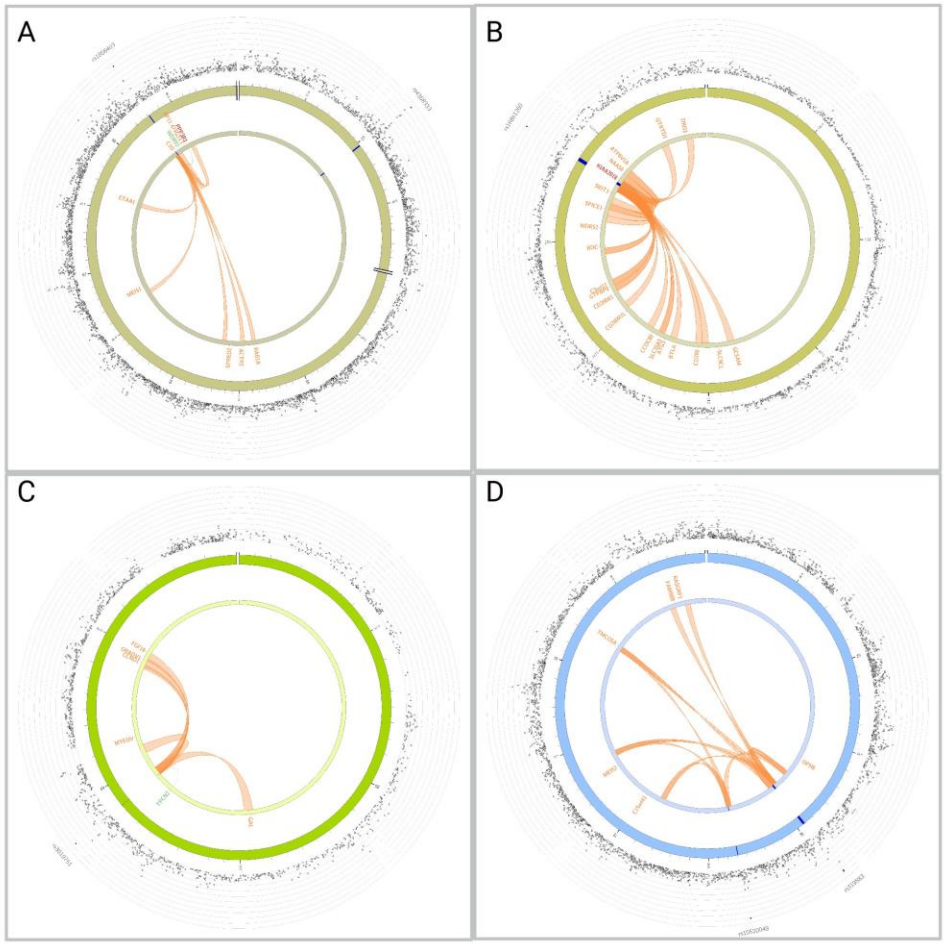


Figure 5.23 *Circos plot generated via FUMA for SAIGE's top SNPs. The outermost layer is the Manhattan plot, and the middle layer highlights genomic risk loci in blue, while the innermost layer highlights eQTLs and/or chromatin interactions. Only SNPs with p < 0.05 a are displayed in the outer ring. SNPs in genomic risk loci are colour-coded as a function of their maximum r^2 to one of the independent significant SNPs in the locus. The rsID of the top SNPs is displayed in the outermost layer. For the innermost layer, if the gene is mapped only by chromatin interactions or only by eQTLs, it is coloured orange or green, respectively. It is coloured red when both map the gene. (A) Circos plot for chromosome 2 top SNPs. (B) Circos plot for chromosome 3 top SNP. (C) Circos plot for chromosome 11 top SNP. (D) Circos plot for the two top SNPs on chromosome 15.*

Table 5.6 summarises the MSigDB gene sets significantly enriched in the mycetoma dataset produced from SAIGE association analysis. Based on the GO analysis, the biological process significantly enriched was the negative regulation of lymphocyte migration. Additionally, the immunologic signature that appeared significantly enriched

was the genes upregulated in bone marrow-derived macrophages from signal transducer and activator of transcription 6 (*STAT6*) knockout mice. Three microRNA (MIR) target gene sets were significantly enriched, as displayed in **Table 5.6**.

Table 5.6 *The MSigDB significantly enriched gene sets from the SAIGE association analysis dataset.*

Category	**Gene set**	**genes**	**p-value**	**Adj. p-value**
GO biological processes	Negative regulation of lymphocyte migration	GCSAM; CD200; CD200R1	4.95E-06	0.036
Immunologic signatures	Genes up-regulated in bone marrow-derived macrophages with STAT6	*FAM98B; RAB1A; ETAA1; C1D; QTRTD1; COX19*	3.57E-06	0.01
MicroRNA targets	CTCAAGA_MIR526B	*RAB1A; PPP3R1; CD200R1; DRD3*	2.75E-05	0.006
	TGTATGA_MIR4853P	*MEIS2; SPRED2; PPP3R1; GPM6A*	0.0005	0.048
	TAATGTG_MIR323	*CCND1; RAB1A; USF3; GPM6A*	0.0006	0.048

GO: Gene Ontology; Adj. p-value: Adjusted p-value after accounting for multiple testing.

The gene expression heatmap for the 48 mapped genes (**Figure 5.24**) showed similar results to GCTA's mapped genes with higher relative expression levels of *SIDT1, BTLA* and *GCSAM* in EBV-transformed B lymphocytes and relatively upregulated *GCSAM* in both skin tissue types (suprapubic and lower leg). Besides, two genes, cyclin D1 (*CCND1*) and sprouty-related EVH1 domain-containing protein 2 (*SPRED2*), were downregulated in whole blood.

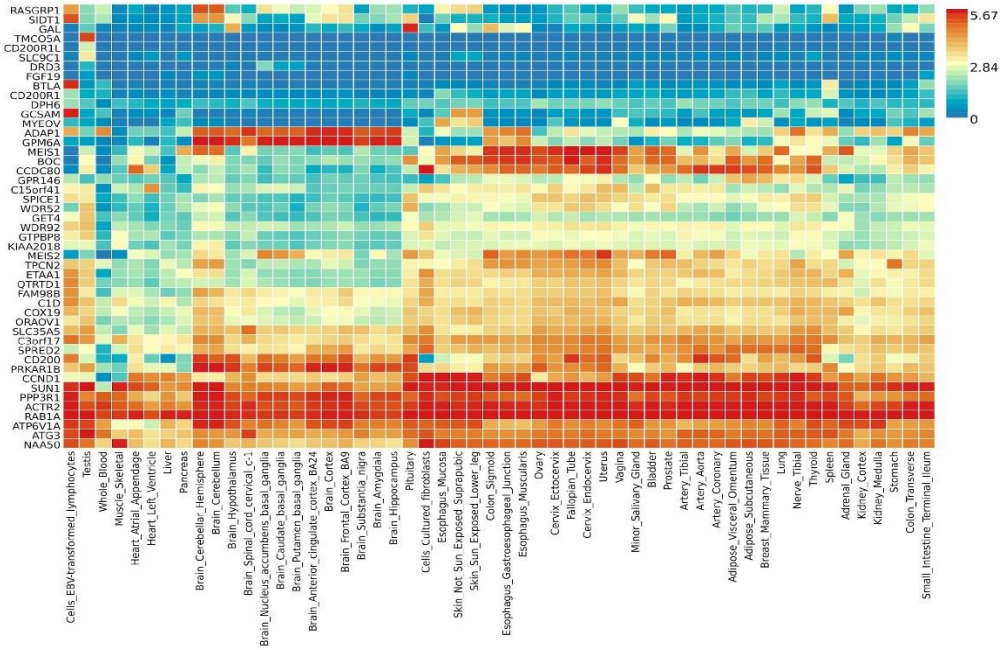


Figure 5.24 *Gene expression heatmap constructed with GTEx v8 (53 tissues) for SAIGE's mapped genes. Genes (x-axis) and tissues (y-axis) are ordered by clusters with expression levels displayed as log2 transformed. The colour scale represents the expression levels ranging from red for high expression to blue for low expression.*

5.4.5 SNP heritability

SNP heritability (h^2_{SNP}) was estimated to measure the contribution of the GWAS-assessed SNPs to the phenotypic variation of mycetoma. As displayed in **Table 5.7**, the estimated h^2_{SNP} for mycetoma was 0.78 (SE=0.13), indicating that around 80% of the phenotypic variation can be linked to the genetic variants examined.

Table 5.7 *The variance sources for mycetoma and the SNP heritability estimate.*

Variance Component	**Variance**	**SE**
V(G)	0.145693	0.027452
V(e)	0.041310	0.024477
Vp	0.187003	0.012554
h^2_{SNP} =V(G)/Vp	0.779095	0.131652

V(P): total phenotypic variance; V(G): genetic variance; V(e): environmental variance; h^2_{SNP}: SNP heritability; SE: standard error.

5.4.6 Imputation accuracy evaluation

I also conducted an analysis to assess the efficacy of two commonly employed imputation reference panels. These panels comprise reference genotypes that aid in inferring unobserved or missing genotypes in study sets. Specifically, this study centred on the Sudanese population and sought to establish whether the imputed genotypes (imputed using the two publicly available reference panels) accurately matched the available WGS data. A total of 17 samples, which had both high-coverage WGS and GWAS data, were included in the analysis. The samples' IDs are listed in **Table 5.8**. I imputed the genotyped data of the 17 samples remotely using two reference panels— the Trans-Omics for Precision Medicine (TOPMed; n = 97,256) hosted at the TOPMed Imputation Server (TIS) (282) and the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA; n = 883) hosted at the Michigan Imputation Server (MIS) (284).

5.4.6.1 Per sample comparison of imputed datasets

In the per-sample comparison results, CAAPA imputation yielded higher imputation accuracy parameters across the board, except for the squared correlation (r^2) metric (**Table 5.8**). The CAAPA imputation results also boasted a significantly lower average root mean squared error (RMSE) (0.20) than the TOPMed dataset (0.83).

Table 5.8 *Imputation accuracy scores per sample using CAAPA and TOPMed reference panels.*

Sample	**Concordance**		**r^2**		**precision**		**recall**		**RMSE**	
	CAAPA	TOPMed	CAAPA	TOPMED	CAAPA	TOPMED	CAAPA	TOPMED	CAAPA	TOPMED
M001	0.958	0.405	0.825	0.867	0.976	0.531	0.99	0.979	0.203	0.836
M002	0.959	0.419	0.828	0.859	0.977	0.552	0.99	0.979	0.2	0.827
M040	0.956	0.42	0.812	0.88	0.974	0.546	0.991	0.988	0.207	0.82
M041	0.959	0.43	0.824	0.888	0.976	0.554	0.99	0.988	0.204	0.81
M083	0.957	0.416	0.824	0.868	0.977	0.55	0.989	0.981	0.207	0.823

M090	0.956	0.37	0.794	0.879	0.973	0.483	0.99	0.977	0.215	0.853
M091	0.958	0.406	0.824	0.873	0.976	0.538	0.989	0.978	0.205	0.825
M120	0.957	0.4	0.821	0.888	0.975	0.525	0.99	0.984	0.205	0.825
M121	0.957	0.358	0.8	0.911	0.974	0.456	0.99	0.985	0.213	0.846
M159	0.959	0.412	0.824	0.876	0.976	0.544	0.99	0.987	0.201	0.824
M167	0.958	0.409	0.8	0.872	0.974	0.539	0.99	0.982	0.209	0.826
M213	0.961	0.399	0.834	0.871	0.978	0.531	0.99	0.981	0.198	0.831
M228	0.958	0.39	0.825	0.884	0.976	0.511	0.99	0.982	0.204	0.837
M236	0.96	0.321	0.833	0.924	0.978	0.385	0.989	0.985	0.198	0.868
M278	0.959	0.408	0.826	0.89	0.976	0.532	0.991	0.985	0.199	0.821
M282	0.958	0.391	0.814	0.889	0.975	0.509	0.99	0.986	0.206	0.835
M296	0.957	0.405	0.805	0.882	0.974	0.527	0.99	0.986	0.208	0.829

R^2: squared correlation; RMSE: Root Mean Squared Error.

5.4.6.2 Per variant comparison of imputed datasets

I classified the imputed variants by MAF and evaluated the imputation accuracy for common and rare variants to gain a more nuanced comprehension of the performance of each imputation panel. The reference panel CAAPA yielded the best and most consistent results in terms of concordance (≥95%) across all MAF bins, with a significant difference between CAAPA and TOPMed in imputation concordance for variants with MAF>5% (**Figure 5.25A**). Similar results were found for IQS (**Figure 5.25B**), which is known to produce a more reliable assessment.

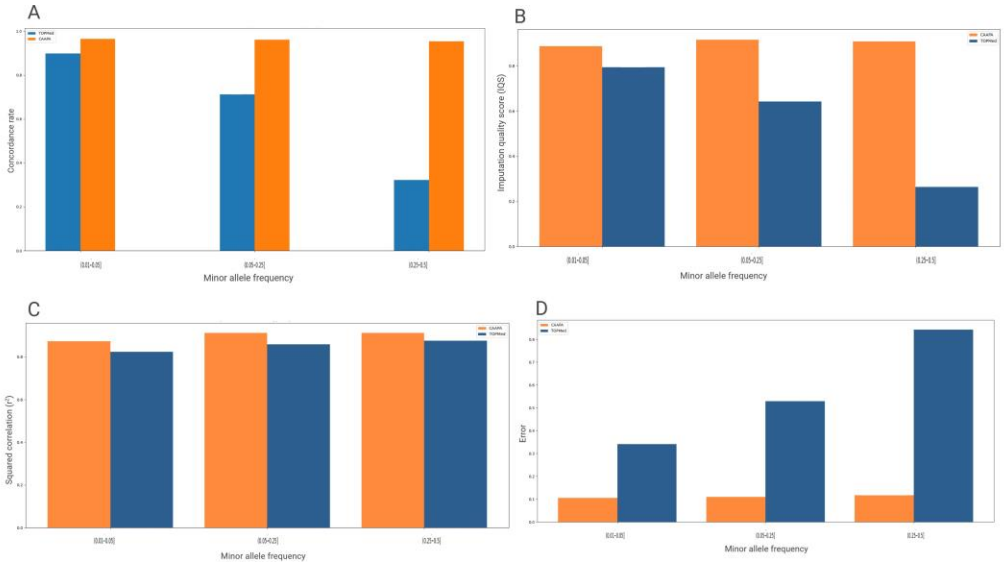


Figure 5.25 *Average imputation accuracy parameters across allele frequency bins using CAAPA and TOPMed imputation datasets. (A) The concordance rate, (B) The imputation quality score (IQS), (C) The squared correlation (r^2) and (D) error.*

To better assess the impact of MAF on imputation accuracy and error rates, I constructed a graph depicted in **Figure 5.26**. Each dot plotted on the graph corresponds to an imputed variant, with r2 and error rate being the plotted parameters. The markers on the graph formed clusters based on their MAF and followed a waterfall-like pattern. The graph revealed that rare-to-moderate-frequency SNPs (<0.05) displayed higher r2 and lower error rates, whereas variants with MAF>0.05 exhibited relatively lower r2 and higher error rates. Furthermore, the CAAPA imputed dataset contained a denser amount of SNPs, prompting examining the union and intersection of SNPs imputed by CAAPA and TOPMed, as explained in the following section.

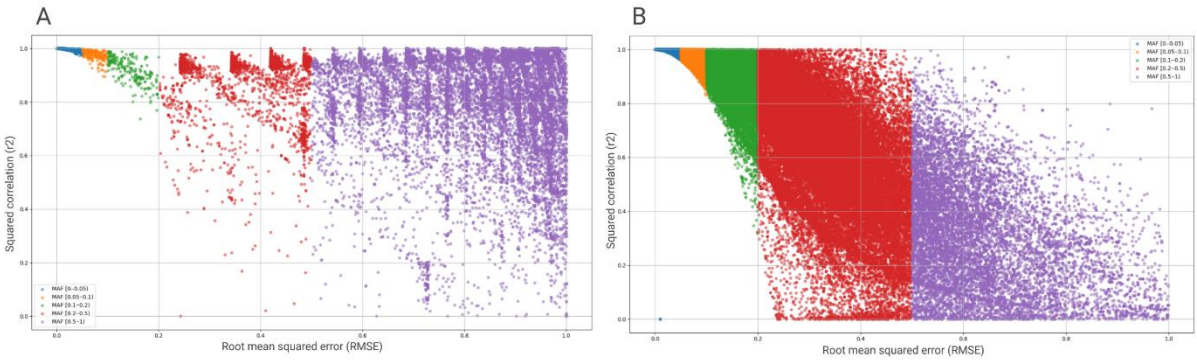


Figure 5.26 *Squared correlation (r2) plotted against root mean squared error (RMSE) for (A) TOPMed and (B) CAAPA imputation results.*

5.4.6.3 Union and intersection of the imputed and WGS datasets

I investigated the overlap between TOPMed, CAAPA, and the WGS dataset, as shown in **Figure 5.27** and **Table 5.9**. Although the total number of SNPs imputed by TOPMed (~9 million) was larger than the CAAPA (8.2 million), the latter had a substantially better overlap with the SNPs in the WGS data. CAAPA had the lowest homozygous alternative mismatch between WGSs and imputed data (**Table 5.9**).

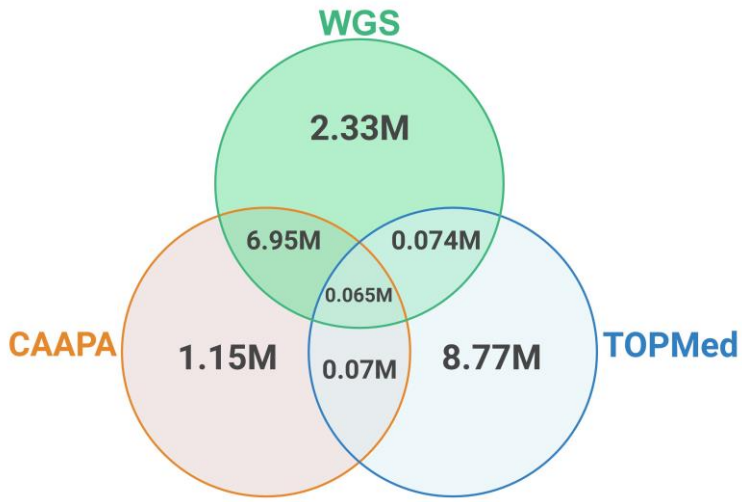


Figure 5.27 *Venn diagram showing the overlap of SNPs between the WGSs and datasets imputed using CAAPA and TOPMed panels.*

5.4.6.4 Estimating non-reference discordance rates (NDRs)

Next, I assessed the level of discordance of genotypes imputed by CAAPA and TOPMed by estimating the non-reference discordance rates (NDRs) against the WGS dataset for

the 17 participants. Overall, the CAAPA panel showed the highest concordance rates (i.e., lowest NDR = 5.7%) when compared to the TOPMed (NDR = 7.5%) panel (**Figure 5.28**). The NDR was almost constant in the CAAPA dataset, while the NDR showed substantial variation in the TOPMed dataset, ranging from 6.5 to 8.6%.

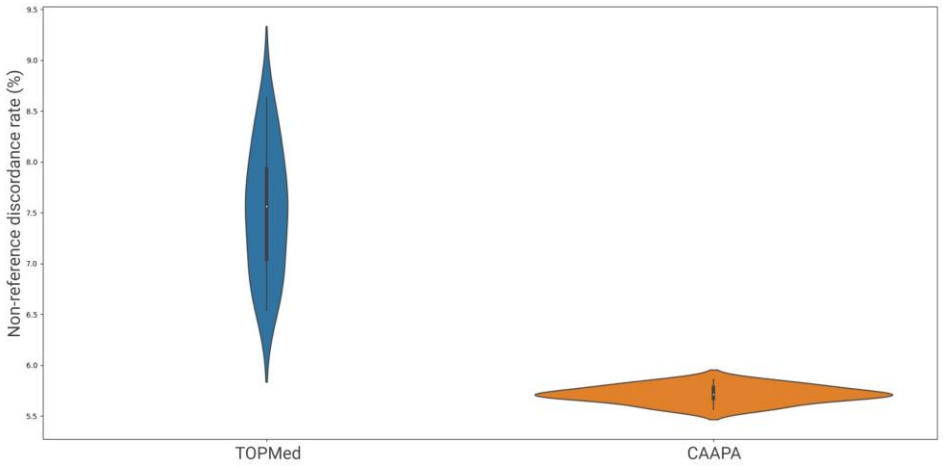


Figure 5.28 *Violin plot comparing the non-reference discordance rate (NDR) distribution between genotypes imputed using TOPMed vs. CAAPA in the 17 participants dataset.*

Taken together, the WGS-based comparisons showed that the CAAPA consistently outperformed TOPMed, making the former a more appealing choice for genotype imputation of Sudanese datasets.

Table 5.9 *Comparison of the whole-genome sequence data with the two imputed datasets*

Parameters compared	**CAAPA**	**TOPMed**
Total number of SNPs imputed	8,236,069	8,976,738
Number of sites common between imputed and sequence data	6,948,977	74,471
Number of sites unique to imputed data	1,152,207	8,767,382
Percentage of Ref/Ref mismatches	37.6%	27.5%
Percentage of Alt/Alt mismatches	5%	7.4%

5.5 Discussion

This chapter explored how genome-wide association studies (GWAS) can be utilised to identify common variants associated with mycetoma. As previously reported (56,74,334), Aljazeera state had the highest number of mycetoma cases. This state is recognised for

having the largest agriculture scheme in Sudan. The majority of its residents rely on arable farming or animal grazing, which increases the likelihood of exposure to mycetoma causative organisms. The Afro-Asiatic linguistic family was the most prominent among the cases and controls, 76.7% and 90%, respectively. This is because the three most affected states, namely Aljazeera, Sennar, and White Nile, are primarily inhabited by ethnic groups belonging to this family (104).

Both study groups were predominantly males, which could be attributed to their greater willingness to participate in research studies relative to females. The geographical distribution of cases on the map (**Figure 5.1**) closely corresponds to the projected distribution of eumycetoma cases, as per the MRC's records spanning over 2 decades, from 1991 to 2018 (334).

Utilising the largest genomic dataset at our disposal, I determined the inbreeding coefficient (F_{ROH}) in our Sudanese population. Though pinpointing specific recessive variants can present challenges, investigating the impacts of inbreeding can provide valuable insights into the overall influence of all recessive variants on phenotypes (335). Inbreeding leads to autozygosity, or the inheritance of an allele that is identical by descent, resulting in homozygosity. The mean F_{ROH} in the dataset was 0.0317, placing Sudan among the world's highest (336). The estimate indicates that the inbreeding in our Sudanese population has not changed over time, with a similar inbreeding coefficient of 0.032 in a study conducted in 1988 in the Khartoum population from Sudan (120). A recent study found the highest F_{ROH} estimates in Sudan's Rashaayda (0.077) and Kababish (0.038) populations among 14 ethnolinguistic groups in the Sahel/Savannah belt (330).

The study identified 15 mycetoma-associated SNPs at $p<1.0x10^{-5}$ utilising commonly used association testing software packages (GCTA and SAIGE). The results of the two software packages were largely concordant. Furthermore, three suggestive signals associated with mycetoma were in common between GCTA and SAIGE. The study revealed that minor alleles in four SNPs were associated with decreased susceptibility to mycetoma. Among these SNPs were rs16861260, located on chromosome 3 with an OR of 0.22 and a *p*-value of 7.83E-06, as well as rs319883, rs319885, and rs10520048 on

chromosome 15 with ORs of 0.44 and *p*-values of 5.87E-06, 7.43E-06, and 3.00E-06 respectively. In contrast, chromosome 2 harboured two risk loci that significantly increase the likelihood of developing mycetoma, each by over two-fold. These risk loci include the G minor allele of rs1868403, which has an OR of 2.68 (95% CI 2.23 to 3.13, p = 7.83E-06), as well as the A minor allele of rs4368333, which has an OR of 2.18 (95% CI 1.84 to 2.52, p = 3.20E-06). Through in silico functional analyses, the minor allele A of rs16861260 correlated with higher relative expression levels of *SIDT1*, *BTLA*, and *GCSAM* in EBV-transformed B lymphocytes. Furthermore, *GCSAM* was relatively upregulated in suprapubic and lower leg skin tissues. Additionally, gene set and pathway enrichment analyses indicated that the negative regulation of lymphocyte migration was significantly enriched (Adj. *p*-value = 0.036). Moreover, the immunologic signature that appeared significantly enriched was the genes upregulated in bone marrow-derived macrophages from *STAT6* knockout mice (Adj. *p*-value = 0.01).

Notably, out of the 15 lead SNPs, five were associated with lncRNA genes, suggesting a potential involvement in susceptibility or interactions between the host and fungus. Long non-coding RNA genes produce transcripts that exceed 200 nucleotides in length and do not undergo protein translation. Depending on the catalogue utilised, the total number of lncRNAs varies between 17,944 and 270,044 (337). These molecules regulate gene expression, chromatin structure, and cellular processes. They can function in gene silencing, scaffold protein complexes, contribute to immune function, and impact disease progression (337–340).

I compared whole-genome sequencing data from the familial study with imputed genotypes to evaluate the performance of the CAAPA and TOPMed reference panels. I examined various metrics, including concordance rates, correlation, precision, recall, imputation quality scores, and non-reference discordance rates. My analysis revealed that imputation accuracy varied across different minor allele frequency bins, with performance differences observed across different MAF ranges. The CAAPA consistently outperformed TOPMed, making it the superior option for genotype imputation in Sudanese datasets. This emphasises the importance of selecting appropriate reference

panels to achieve accurate genotype imputation and underscores the significance of including various African populations in frequently used reference panels.

This pioneering study marks the first-ever attempt to estimate the genetic heritability of mycetoma. The study’s findings revealed that a significant proportion of the disease's heritability was attributed to common genetic variants, which account for a staggering 78% of the total heritability. Nonetheless, it's essential to recognise that heritability alone does not pinpoint specific risk variants, warranting further investigations to fine-map the exact genetic variants involved.

Despite the potential underpowering of the GWAS, the results serve as a fruitful starting point for further exploration of gene-environment interactions. The fact that only half of the collected samples were genotyped does not diminish the value of the findings, which can provide valuable insights into the complex interplay between genetic and environmental factors. However, it is imperative to replicate these findings with larger sample sizes and give due consideration to the significant disparities in association patterns across various ethnolinguistic groups. This is a crucial factor to keep in mind for forthcoming genetic studies undertaken in the region.

CHAPTER 6 - Exploring genetic pathways to mycetoma susceptibility: a profound discussion

6.1 Overview

The research presented in this thesis is the first to systematically explore the genetic factors contributing to mycetoma susceptibility. The research aimed to use familial and population data to determine the extent to which genetic factors contribute to mycetoma aetiology and identify rare and common genetic variants associated with the disease. I investigated various models of mycetoma inheritance using hypothesis-free approaches. I used data from consanguineous families to start with monogenic models and progressed to more complex models with population-based association studies. Through my thesis findings, I aimed to enhance our comprehension of the underlying mechanisms of mycetoma pathogenesis. This knowledge has the potential to revolutionise how we identify high-risk populations and individuals, as well as discover new drug targets. Ultimately, this could reduce disease burdens and improve health outcomes for vulnerable communities.

6.2 Principle findings

6.2.1 Host genetic factors in mycetoma: a review of key contributors

In chapter two, I reviewed the previous research on genetic susceptibility to mycetoma to understand the knowledge gaps that needed to be filled. Additionally, I reviewed studies on genetic predisposition to other morphologically related fungal infections aiming to identify possible candidate genes associated with the neglected disease since the previous work was limited. Most previous studies on mycetoma host genetics adopted the candidate gene approach in which the researchers already had a predefined list of polymorphisms to test their association with mycetoma. The associated polymorphisms were located in 13 candidate genes belonging to four pathways: innate immunity, adaptive immunity, sex hormone biosynthesis and host enzymes. A gene-gene interaction network was constructed to gain insights into the interaction between these susceptibility genes and identify potential therapeutic targets. This analysis aimed to pinpoint central genes that could be targeted in the future. None of the candidate genes

was identified in this research's family-based or population-based studies. This is likely due to the small sample size in all the previous studies, which was at most 360 samples, and none was replicated. Small sample sizes in candidate gene approaches can hinder the statistical power and generalisability of results, potentially leading to spurious associations and limited applicability to the broader population (341). Moreover, without replication of findings in independent cohorts, the robustness and reliability of identified candidate genes come into question, impeding the establishment of causal relationships. These drawbacks were due to the disease's low worldwide incidence, limited research, close association with impoverished communities, and insufficient funding dedicated to its investigation. Despite the complex and multifactorial nature of mycetoma, the research on host genetics was limited to the Mycetoma Research Centre in Sudan and Erasmus Medical Centre in the Netherlands, with the most recent research conducted in 2015 (32,62,229). To advance genomic research in neglected tropical diseases like mycetoma, it is crucial to foster collaboration among researchers, healthcare institutions, and international organisations. This collaborative effort can improve patient outcomes and public health interventions.

6.2.2 Familial insights into the genetic basis of mycetoma susceptibility

In the fourth chapter, I conducted a thorough analysis using family-based data to determine the role of genetics in mycetoma predisposition. This involved examining 46 pedigrees with 1024 individuals and the genomes of 22 individuals from four multi-case families. Through the pedigree analysis, I estimated the heritability (h^2) of mycetoma and found that approximately 23% of the variation in the condition can be attributed to genetic factors. While this estimate offers significant insights into the susceptibility of mycetoma in Sudan, additional research is necessary to validate the findings in other regions where mycetoma is prevalent. This will broaden the scope of the investigation to different populations, revealing genetic susceptibilities that vary across diverse genetic backgrounds and environmental factors. This study's estimate echoes findings from a study on malaria, a well-studied infectious disease, in Kenyan children, where host genetics accounted for approximately 24% of the total variability in malaria severity (342). The study on malaria heritability, along with similar studies in other countries, paved the

way for further research that identified specific genetic loci associated with protection or susceptibility to malaria (343). This example highlights the potential for research on mycetoma to follow a similar path, exploring the genetic factors that influence susceptibility and severity of the disease. Furthermore, I used the pedigree data to calculate the sibling recurrence risk ratio (λ_s) for mycetoma. This involved comparing the risk of a sibling of an affected individual developing mycetoma to the risk in the general population. The results showed that the ratio ranged from 18.4 to 28.7, indicating that the risk of a sibling developing the condition is at least 18 times higher than that of the general population. The two estimates served as the foundation for a more advanced genetic investigation utilising WGS.

The WGS investigation centred around four families from Wad El Nimar village, each with a minimum of four cases of mycetoma. The genomes analysis identified four genes that could be potential candidates: *DEFA3* in Family 1, *CASP8* in Family 2, *PLEKHO2* in Family 3, and *RAPGEF2* in Family 4. The observation that the four families are from the same village, potentially connected at some ancestral level, raises intriguing questions about the genetic dynamics within the community. The discovery of four distinct candidate genes, especially in families 1 and 3, which share a connection through one individual 002, adds complexity to the genetic landscape. Despite the expectation of identifying a shared variant among the families, none were common. This divergence could be attributed to the high genetic diversity in the Sudanese population (average number of variants across the four families' genomes was ~4.6 million). The challenges of WGS could also be a possible explanation for this divergence, where rare variants with large effects might overshadow more common ones including low-impact non-coding or modifier variants. Modifier variants, in this context, refer to genetic elements that may not independently cause a condition but can influence its severity or manifestation. Further research is essential to unravel the intricate genetic interplay within these families and to uncover potentially overlooked common variants that could contribute to the understanding of mycetoma predisposition.

Though the genetic variations within the genes differed in terms of their type, coding impact, and zygosity, network analysis revealed their interconnectedness. The

overrepresentation analysis determined that all four genes were part of two main pathways: the immune system and signal transduction. Figure 6.1 illustrates how these candidate genes may play a role in defending against eumycetoma. In Family 1, the candidate gene was *DEFA3* from the myeloid α-defensins gene family. This gene family produces small, cysteine-rich peptides (made up of 29-30 amino acid residues) known as human neutrophil peptides (HNPs), which are crucial in innate immunity and host defence (289). Produced mainly by neutrophils and other immune cells (344), they exhibit potent antimicrobial properties, effectively targeting bacteria, fungi, and some viruses by disrupting cell membranes and interfering with microbial processes (345). These peptides play a pivotal role in the immune response, acting as rapid responders to infection or inflammation and aiding in the recruitment of immune cells (346). They can trigger macrophages to release TNF-α and IFN-γ and enhance bacterial phagocytosis (347). They can also increase the expression of TNF-α and IL-1β in monocytes activated with *Staphylococcus aureus* (348). During pulmonary inflammation, they upregulate IL-8 expression in airway epithelial cells, attracting and recruiting a greater number of neutrophils to the site of injury or infection (349). Beyond their direct antimicrobial functions, HNPs participate in wound healing and the regulation of angiogenesis (350–352). The mechanisms of α-defensins against fungal pathogens are poorly understood, and there is a lack of specific mechanistic data regarding their interactions with fungi (353). While α-defensins are primarily recognised for their antimicrobial activity against bacteria, their interactions with fungi and other pathogens may involve different mechanisms that remain to be elucidated. Based on these peptides' antimicrobial and immunomodulatory properties, I postulated the role of α-defensins, including DEFA3, in immunity against eumycetoma (**Figure 6.1C**). It is plausible that α-defensins play a part in the initial immune response against the fungal pathogens, aiding in their neutralisation. Their cationic nature may enable them to interact with fungal membranes, disrupting the integrity of the fungal cells. Additionally, α-defensins might contribute to the recruitment of immune cells and the modulation of the inflammatory response in eumycetoma. However, a deeper understanding of their precise mechanisms and effectiveness against eumycetoma causative agents is needed. Experimental studies and in vitro assays that explore the interactions between α-defensins and fungal pathogens are necessary to

shed light on the potential role of these peptides in antifungal immunity and their therapeutic implications for eumycetoma management.

The second candidate gene identified in Family 2 cases was *CASP8,* which encodes caspase-8, a crucial member of the caspase family of proteases that plays a multifaceted role in regulating immunity and immune responses. Its involvement in immunity primarily regulates programmed cell death, particularly apoptosis and pyroptosis (354). While its specific functions in fungal immunity are still being explored, several insights shed light on its contributions (355). Perhaps the most relevant to mycetoma is its role in inflammasome activation, an integral component of the innate immune response to fungal infections (356–358). Activation of the noncanonical inflammasome leads to the release of proinflammatory cytokines like IL-1β and IL-18, which are essential for host antifungal defences. This pathway has been suggested by previous studies on eumycetoma pathogenesis that linked elevated serum levels of pro-inflammatory cytokines such as IL-1β to surgical treatment, smaller lesion size and shorter disease duration (48,357). In this study, *CASP8* was identified to be central in network analysis and was the only candidate present in the top three overrepresented pathways (innate immune system, immune system, and signal transduction), indicating a promising role in immunity against mycetoma (**Figure 6.1B**). On the contrary, the third candidate gene, *PLEKHO2* (identified in Family 3), inhibits cell death pathways. The encoded protein possesses a pleckstrin homology (PH) domain that plays a role in intracellular signalling and cell survival (359). PLEKHO2 modulates cell death processes such as apoptosis and necroptosis by interacting with crucial signalling molecules like caspase-8 and receptor interaction protein kinase 1 (RIPK1) (310,360). In one study, the researchers observed upregulated expression of PLEKHO2 during macrophage differentiation and maturation (310). To elucidate the physiological role of PLEKHO2, they generated PLEKHO2-knockout mice, finding that they were viable, healthy, and fertile without apparent developmental abnormalities. However, PLEKHO2-deficient mice exhibited decreased macrophage populations in the spleen, peritoneal cavity, and blood. Additionally, bone marrow-derived macrophages (BMDMs) lacking PLEKHO2 displayed increased apoptotic cell death in the absence of macrophage-colony stimulating factor (M-CSF) despite unaffected BMDM differentiation and proliferation.

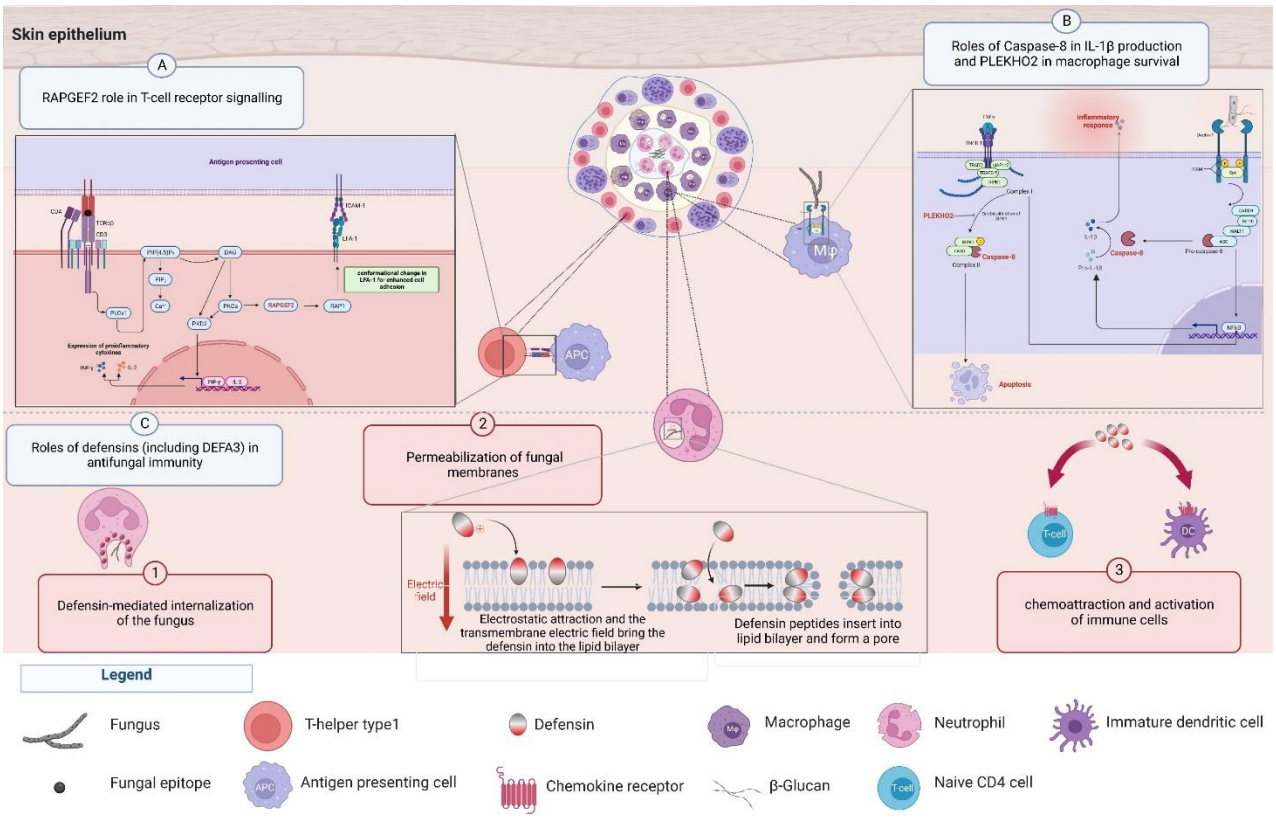


Figure 6.1 *Proposed roles of WGS candidate genes in immunity against eumycetoma.* ***(A)*** *RapGEF2 role in T-cell receptor (TCR) signalling. Upon TCR triggering, phospholipase C-γ (PLC-γ) is activated. This enzyme cleaves phosphatidylinositol-(4,5)-bisphosphate (PtdIns(4,5)P2) to produce two crucial second messenger molecules, polyunsaturated diacylglycerol DAG and inositol-(1,4,5)-trisphosphate (Ins(1,4,5)P3). DAG accumulation in the plasma membrane activates protein kinase C (PKC) and protein kinase D (PKD). PKCθ associates with RapGEF2 and phosphorylates it on Ser960, facilitating the activation of Rap1. When Rap1 is activated, LFA-1's adhesiveness to its binding partner ICAM-1 increases, resulting in an overall increase in adhesiveness of the T-cell to the antigen-presenting cell (APC).* ***(B)*** *Role of Caspase-8 mediated activation of IL-1β. Stimulation of dectin-1 with the fungal β-glucan leads to engagement of spleen tyrosine kinase (Syk) that triggers activation of nuclear factor-kappa B (NF-κB) and transcription of the gene encoding pro-IL-1β via the CARD9–Bcl10–MALT1 complex. The production of bioactive IL-1β is then done by the noncanonical inflammasome composed of CARD9, B cell lymphoma 10 (Bcl10), mucosa-associated lymphoid tissue 1 (MALT1) and caspase-8. Conversely, PLEKHO2 plays a pivotal role in macrophage survival. It acts as a crucial inhibitor of apoptosis, dependent on RIPK1 kinase activity, thus exerting significant control over cell death pathways, particularly in response to TNFα signalling.* ***(C)*** *Potential mechanisms by which α-*

defensins enhance host antifungal immunity via (1) facilitating fungal antigen uptake through defensin-mediated internalisation and (2) Fungal cell membranes permeabilisation through the fusion of granules that contain abundant HNPs to phagosomes. In this setting, fungi are exposed to a high concentration of HNPs, which insert into the lipid bilayer, leading to transmembrane pore formation. (3) In extracellular regions, HNPs could be chemotactic for immature dendritic cells and T-cells. Figure created using BioRender.com based on mechanistic data on defensins (291,361–363)*, caspase-8* (364–367)*, PLEKHO2* (310,360) *and RAPGEF2* (368–371)*.*

Furthermore, PLEKHO2 deficiency led to heightened activation of apoptosis-inducing caspases (caspase-3) and initiator caspases (caspase-8 and -9) in BMDMs. These findings provided genetic evidence supporting the role of PLEKHO2 in macrophage survival. In a later study, the same team identified PLEKHO2 as an inhibitor of apoptosis and necroptosis, and this inhibitory effect was dependent on the kinase activity of RIPK1(360). PLEKHO2 protected RIPK1 ubiquitination, inhibiting its activation, impeding complex II formation, and ultimately preventing cell death in response to TNFα (**Figure 6.1B**). Although the exact mechanism and clinical implication of the identified homozygous missense variant in Family 3 (NM_025201.5:c.379G>A) need to be investigated, PLEKHO2 represents an intriguing protein in the context of immune-related processes and potential therapeutic targets.

The final candidate identified in Family 4 was RAPGEF2, which is central to TCR signalling (**Figure 6.1A**). It acts as a guanine nucleotide exchange factor (GEF) for the small GTPase Rap1, a pivotal player in this signalling cascade (371). When TCR engagement occurs upon contact with APCs, RAPGEF2 initiates 'inside-out' signalling mechanisms. The primary function of RAPGEF2 is to activate integrins, especially the integrin LFA-1, through the activation of Rap1 (370). This activation leads to conformational changes in integrins, increasing their affinity for ligands on the APC surface. Consequently, RAPGEF2's actions enhance adhesion between T-cells and APCs, facilitating stable and long-lived interactions crucial for effective immune responses and T-cell activation. Mahgoub and his colleagues postulated in 1977 that genetic defects in cell-mediated immunity play a significant role in the pathogenesis of eumycetoma (29). Variations in this gene, such as the missense variant identified in Family 4 (NM_014247.3:c.700A>G), may have implications for eumycetoma by

influencing the immune response, the localisation of immune cells to affected tissues or the degree of inflammation associated with the disease. However, this requires further investigation to understand better the role of RAPGEF2 (and its subsequent gene variation) in the complex interactions between the immune system and the eumycetoma causative agents.

The prevalence of cystic fibrosis (CF) in Sudan is notably low, with only 35 reported cases (372). In that case series study, the consanguinity rate (%) among the patients' families was 71.4%. The genetic testing of the underlying *CFTR* gene variants was done only on three patients and identified two variants: c.3305G>A and c.1736A>G. The affected individuals displayed indications of pulmonary disease, with both *S. aureus* and *Pseudomonas aeruginosa* in their sputum (372). In the mycetoma WGS study, the affected brothers in Family 1 were carriers of a rare missense variant, c.3485G>A (p.Arg1162Gln) in the *CFTR* gene. According to the ACMG criteria, this variant was likely pathogenic, as depicted in Table 4.8. Furthermore, this variant had never been reported in Sudan, other Arab, or African countries (373,374), thereby suggesting a further investigation since there is a possibility that the occurrence of CF is higher in African populations than previously thought. However, past research might have primarily concentrated on identifying prevalent European mutations, such as deltaF508, which was detected at a lower frequency of 48% in Africa compared to the global average of 70–90% (375). Moreover, CF may be more prevalent than currently recognized, potentially being underdiagnosed due to its shared clinical presentation with other prevalent diseases. The manifestation of CF with symptoms like failure to thrive could lead to misdiagnosis, particularly as it might be initially perceived as malnutrition. Furthermore, the onset of *S. aureus* pneumonia as the primary lung infection in CF cases could add to the challenge of accurate diagnosis, especially considering the subsequent emergence of more problematic pathogens like *P. aeruginosa* and *Mycobacterium abscesses*. The prevalence of diarrheal diseases, especially in contexts of poor nutrition, raises the question of a potential heterozygous advantage in maintaining the high frequency of detrimental CF mutations. This advantage has been theorized to provide protection against diarrheal diseases, prevalent in Africa. While CF is commonly associated with individuals of European ancestry who may not be exposed to certain pathogens, there

may be a broader association with other infections, such as mycetoma, within the Sudanese population.

Consanguinity was a significant influencer in this family-based study (section 4.4). In societies where consanguinity is a common practice (as the one targeted in the WGS study) or where the population is small and isolated, various effects have been observed, such as a higher likelihood of monogenic disorders and the increased risk of complex diseases that involve recessive variants with intermediate or large effect sizes (335). In these isolated communities, limited gene flow and a smaller gene pool contribute to the preservation of genetic variations, including rare mutations that might be associated with increased susceptibility to infectious diseases. Understanding the impact of consanguinity on genetic variation and susceptibility to mycetoma in isolated communities such as the one in Wad El Nimar village is crucial for implementing targeted public health measures, genetic counselling, and tailored interventions to address the health challenges posed by this condition.

6.2.3 Decoding mycetoma susceptibility through population-based genome-wide association studies

Chapter 5 of the research primarily focused on identifying common genetic variations associated with mycetoma. However, the study's diverse cohort of individuals from 77 different ethnic groups, spread across the 18 states of Sudan, offered a unique opportunity to examine inbreeding within this population. This was the most extensive genomic dataset used to estimate the Sudanese population's genomic inbreeding coefficient (F_{ROH}). Pedigree-based estimates of consanguinity and homozygosity have limitations, notably in capturing distant generation close-kin marriages and underestimating cumulative inbreeding effects (124). Therefore, the genotyping data was employed to estimate the F_{ROH} by identifying uninterrupted runs of homozygosity (ROH). This approach provides a more comprehensive understanding of inbreeding effects by considering the uninterrupted stretches of homozygous regions in the genome, offering insights beyond the constraints of pedigree-based assessments (124). The mean F_{ROH} value in the dataset was one of the highest globally, standing at 0.0317 (336). This estimate indicated that the level of consanguinity in the Sudanese population has

remained consistent over time. According to a study conducted by Saha et al. in 1988, the population of Khartoum, Sudan, exhibited an inbreeding coefficient of 0.032 based on data on spousal relationships (120). The rate of consanguinity, which has consistently exceeded 40% from 1988 to 2023, provides substantial evidence for its persistence (120,121). This finding underscores the need for further research and inquiry into the potential genetic implications of such high rates of consanguineous marriages.

The GWAS was underpowered due to the fact that only half of the collected samples were genotyped while the other half were lost during the ongoing war in Sudan. A p-value cut-off of $<1.0x10^{-5}$ was used for the association testing to identify suggestive associations. This threshold was selected to avoid false negative results, particularly in African populations with weaker LD (157,165). Moreover, given that this is the first GWAS on mycetoma, the objective was to retain as many associated markers as possible for future replication studies. By utilising GCTA and SAIGE software packages, I pinpointed 15 SNPs associated with mycetoma. Notably, both software packages found three SNPs that may protect against mycetoma. Two of the SNPs, rs319883 and rs10520048, are intronic variants located in lncRNA genes (DPH6-DT and RP11-184D12.1, respectively) on chromosome 15. The third SNP, rs16861260, is an intergenic variant located 1.12 kb downstream from UCF3 on chromosome 3. The minor alleles of these SNPs were all associated with a lower susceptibility to mycetoma, with rs16861260 offering the highest level of protection (around 4.6-fold), with an odds ratio (OR) of 0.22. Additionally, the C minor alleles of the other two SNPs, rs319883 (p = 5.87E-06) and rs10520048 (p = 3.00E-06) on chromosome 15, provided roughly 2.3-fold protection against susceptibility to mycetoma, with odds ratios of 0.44 for both SNPs.

In silico functional mapping strategies (positional mapping, eQTL mapping, and chromatin interaction mapping) revealed that the minor allele A of rs16861260 was associated with higher relative expression levels of *SIDT1*, *BTLA*, and *GCSAM* in EBV-transformed B lymphocytes. Moreover, *GCSAM* exhibited relative upregulation in suprapubic and lower leg skin tissues. Among the three genes, the most promising candidate was *BTLA,* which encodes an essential co-signalling molecule that falls under the CD28 superfamily. It shares its structure and function with programmed cell death-1 (PD-1) and cytotoxic T

lymphocyte associated antigen-4 (CTLA-4) (376). BTLA is present in most lymphocytes, and its primary function is to induce immunosuppression by inhibiting the activation and proliferation of B and T cells (376). According to a previous study, BTLA+ αβ T cells have a central memory phenotype that helps fight *Mycobacterium tuberculosis* (Mtb) infection (377). These cells exhibit increased cell proliferation and cytokine secretion, indicating that BTLA expression in αβT cells plays a crucial role in protective immune memory against Mtb infection among patients with active pulmonary tuberculosis. Gene set and pathway enrichment analyses revealed significant enrichment in the negative regulation of lymphocyte migration (Adj. *p*-value = 0.036). This process is crucial for regulating immune responses, as controlled migration of lymphocytes is essential for effective immune surveillance and response to infection. The genes *GCSAM*, *CD200* and its receptor *CD200R* were associated with the negative regulation of lymphocyte migration. CD200 is a glycoprotein expressed on various cell types that serves as an immune checkpoint by interacting with its receptor, CD200R, to inhibit immune responses, particularly in myeloid cells (378). The CD200-CD200R interaction is crucial in maintaining immune homeostasis and preventing excessive inflammation (379). All the previously mentioned genes were associated with the highest-scoring SNP, rs16861260. Thus, suggesting a potential functional impact of this genetic variant on the expression of genes associated with lymphocyte function and migration regulation and providing insights into this genetic locus's immunological implications. Moreover, gene set enrichment analysis revealed a significantly enriched immunologic signature related to genes upregulated in BMDMs from *STAT6* knockout mice (Adj. *p*-value = 0.01). This refers to a set of genes that exhibit increased expression in macrophages derived from bone marrow cells where the *STAT6* gene has been knocked out (380). STAT6 is a transcription factor that plays a crucial role in the signalling pathway of IL-4 and IL-13 (381), members of the Th2 family of cytokines (382). The candidate genes belonging to this immunologic signature gene set were associated with the top SNPs, rs1868403 on chromosome 2, rs16861260 on chromosome 3, rs4076072 on chromosome 7 and rs319883 on chromosome 15.

I evaluated the performance and accuracy of two commonly used reference panels (CAAPA and TOPMed) for genotype imputation of our Sudanese GWAS dataset. By

comparing WGS data with imputed genotypes, I assessed various metrics, including concordance rates, R-squared values, precision, recall, and imputation quality scores. I investigated the impact of MAF bins on imputation accuracy and observed variations in performance across different MAF ranges. The comparison between CAAPA and TOPMed for genotype imputation revealed significant consequences associated with underrepresentation of Africans. Despite TOPMed's extensive dataset, which is considered the most comprehensive and diverse panel with 97,256 individuals (~22% of African ancestry (161)), it showed lower accuracy when applied to our Sudanese dataset. The consistent outperformance of CAAPA (n = 883) highlights the limitations of relying on broad panels that lack adequate representation from specific populations. This discrepancy in imputation performance emphasizes the critical need for inclusivity in reference panels. The superiority of CAAPA in our Sudanese context suggests that its inclusion of individuals of solely African ancestry significantly enhances its effectiveness. Therefore, it is urgent to refine reference panels to be more region-specific and inclusive, ensuring a more accurate representation of genetic variations in populations with distinct demographic histories, such as those in Northeast Africa. In a recent study, a dataset from various regions of Africa, including Kenya (East), Ghana and Burkina Faso (West), and South Africa (South), was utilised to evaluate distinct imputation reference panels (383). The outcomes indicated that TOPMed was the most efficient. Nonetheless, when considering our Northeastern population, CAAPA, comprising individuals of African heritage residing in the Americas, alongside some Atlantic African individuals (284), was a more practical alternative for genotype imputation. Albeit a relatively compact panel, it demonstrated to be a valuable selection.

This study also estimated that 78% of mycetoma's genetic heritability could be attributed to common genetic variants, marking a significant step in understanding the genetic basis of the disease. However, it's crucial to note that while heritability is high, the specific risk variants remain unidentified, necessitating further investigations for fine mapping.

6.3 Conclusions

i. The foundational base for this research rests on familial aggregation estimates, providing key insights into the genetic underpinnings of mycetoma. The heritability

(h^2) of mycetoma was estimated at approximately 23%, signifying a substantial genetic contribution to the condition. Moreover, the sibling recurrence risk ratio (λ_s) ranged from 18.4 to 28.7, indicating a significantly elevated risk for siblings of affected individuals compared to the general population. These estimates form the bedrock upon which subsequent genetic investigations were built.

ii. Whole-genome sequencing (WGS) was pivotal in identifying candidate genes associated with mycetoma susceptibility. Notable candidates such as *DEFA3, CASP8, PLEKHO2*, and *RAPGEF2* were uncovered, shedding light on potential players in the immune response against mycetoma. DEFA3 from the α-defensins gene family emerged as a promising candidate, with its antimicrobial properties and implications for innate immunity. *CASP8*, encoding caspase-8, was identified as central in network analysis and exhibited a promising role in immunity against mycetoma. PLEKHO2, known for its inhibitory effect on cell death pathways, demonstrated a unique role in macrophage survival, protecting against apoptosis. RAPGEF2, crucial in T-cell receptor signalling, indicated potential implications for the immune response against mycetoma, linking T-cell activation and effective immune responses to the disease. These candidate genes shall be validated and can collectively present intricate connections within immune pathways, providing a foundation for understanding the genetic basis of mycetoma susceptibility.

iii. The Genome-Wide Association Study (GWAS) uncovered promising genetic markers associated with mycetoma susceptibility. The identified 15 SNPs exhibited suggestive associations with three notable candidates (rs16861260, rs319883, and rs10520048). Utilising a less stringent *p*-value cut-off of $<1.0 \times 10^{-5}$ to identify suggestive associations, the analysis highlighted the potential protective effects of specific alleles, notably with rs16861260 offering a remarkable 4.6-fold decrease in susceptibility. In silico functional analyses delved into the mechanistic implications of these genetic markers, emphasising the significance of genes like *BTLA* in immune regulation.

The fact that only half of the samples were genotyped does not diminish the value of these findings. On the contrary, it emphasises the robustness of the observed associations and their potential to unravel critical aspects of mycetoma

susceptibility. With an impressive SNP heritability (h^2_{SNP}) of nearly 80%, these findings provide a strong foundation for future explorations, warranting a more extensive sample size to elucidate the intricate genetic factors contributing to mycetoma comprehensively.

iv. Exploration of the genetics of the Sudanese population revealed unique characteristics. The genomic inbreeding coefficient (F_{ROH}) was one of the highest globally, standing at 0.0317, indicating a consistent level of consanguinity over time. The study also emphasised the superiority of African ancestry reference panels, particularly CAAPA, over TOPMed for the Sudanese population in genotype imputation. The underperformance of TOPMed in our Sudanese dataset underscores the necessity of expanding and refining reference panels to include diverse populations, particularly those from regions historically underrepresented. This improvement will not only enhance the accuracy of downstream genetic studies but also ensure equitable benefits from genomic research for all populations, preventing biases and contributing to a more inclusive understanding of genetic factors across diverse human groups.

This research collectively advances our understanding of mycetoma's genetic basis, providing valuable insights for future investigations and potentially informing targeted interventions.

6.4 Limitations

The Lack of functional studies pose a notable limitation in assessing the potential pathogenicity of the identified WGS candidate variants and understanding their mechanistic roles in mycetoma pathogenesis. While the genetic information offers valuable insights, a comprehensive understanding of these variants' functional implications necessitates further investigations into their specific roles and impact on the disease.

The GWAS conducted in this research faced challenges related to sample size. The study was considered underpowered as only half of the initially collected samples were genotyped. Tragically, the other half of the samples was lost during the conflict in Sudan before they could be processed for genotyping. This loss underscores the

broader challenges associated with conducting research in regions affected by conflict, where external factors can significantly impact the execution and outcomes of scientific investigations.

Collecting data in rural Sudan through field missions faced logistical hurdles, especially during the rainy season from June to September. Roadblocks, which were likely worsened by adverse weather conditions, presented significant challenges to the research team's mobility. These obstructions not only affected the efficiency of data collection but also highlighted the practical difficulties of conducting scientific research in geographically challenging and politically unstable environments. Conducting research in low- and middle-income countries (LMICs) poses substantial challenges rooted in systemic issues. Limited investment in research infrastructure and insufficient funding for research activities present formidable obstacles. The global funding dynamics frequently reflect high-income countries (HICs) agendas, as seen in the application of tools developed in HICs for genetic studies in LMICs without adequate consideration of population structure differences. Neglected diseases, prevalent in LMICs, may struggle to secure funding due to perceived irrelevance to HICs, exacerbating health disparities. Moreover, the increased risk of political instability in some LMICs, like Sudan, adds an additional layer of complexity, impacting the continuity and safety of research initiatives. Researchers in these settings often find themselves competing with more immediate and essential health needs, such as access to clean water, sanitation, and vaccinations, which may take precedence in resource allocation. Collectively, these challenges underscore the critical need for global research initiatives that address the unique contexts and priorities of LMICs to foster equitable advancements in science and healthcare.

6.5 Future directions

Securing a more extensive set of samples from affected individuals, with a particular focus on the grandchildren generations, is critical for advancing the WGS familial study. The inclusion of additional samples is essential to robustly confirm the segregation of predisposing variants identified in the WGS familial study, providing a

more comprehensive understanding of the genetic landscape associated with mycetoma in familial contexts.

To enhance the validity of genetic findings, future directions should prioritize conducting functional studies to validate the pathogenicity of candidate variants and elucidate their roles in the molecular development of mycetoma. By delving into the functional aspects of these variants, we can bridge the gap between genetic associations and the mechanistic understanding of how these variants contribute to the disease's manifestation.

The WGS study highlights the possibility of conducting additional research on the CFTR gene among Sudanese communities. Analysing other Sudanese genome sequencing data for variations in the CFTR gene could reveal wider connections beyond cystic fibrosis. This method could provide valuable information on genetic factors that affect susceptibility or immunity to different infections, including mycetoma.

While the GWAS has unveiled potential genes linked to mycetoma susceptibility, it was hampered by limitations in statistical power and failed to achieve the intended sample size. A crucial next step involves completing the planned GWAS sample size to bolster statistical robustness and conclusively confirm the associations identified. This larger sample size is pivotal for ensuring the reliability and generalizability of the genetic associations uncovered by the study.

In culmination, this research has not only established the substantial heritability of mycetoma at approximately 23% but has laid the foundational base for further genetic investigations. Findings from whole-genome sequencing and genome-wide association studies have identified promising candidate genes and genetic markers, proposing unprecedented insights into the immune response against mycetoma. Moreover, the exploration of the Sudanese genetic population has unveiled unique characteristics and emphasized the importance of diverse reference panels for accurate genomic studies. Collectively, these findings not only advance our understanding of mycetoma susceptibility but also pave the way for targeted control and prevention strategies, offering hope for improved health outcomes in affected populations globally.

REFERENCES

1. Molyneux DH, Savioli L, Engels D. Neglected tropical diseases: progress towards addressing the chronic pandemic. *The Lancet*. [Online] Elsevier Ltd; 2017;389(10066): 312–325. Available from: doi:10.1016/S0140-6736(16)30171-4

2. World Health Organization. *Neglected tropical diseases*. [Online] Available from: https://www.who.int/health-topics/neglected-tropical-diseases#tab=tab_1

3. Khan Burki T. Starting from scratch—the unique neglect of mycetoma. *Lancet Infect Dis*. [Online] Elsevier Ltd; 2016;16(9): 100–112. Available from: doi:10.1016/S1473-3099(16)30283-3

4. Zaias N, Taplin D, Rebell G. Mycetoma. *Arch Dermatol.* 1969;99: 215–225.

5. Abbas M, Scolding PS, Yosif AA, EL Rahman RF, EL-Amin MO, Elbashir MK, et al. The disabling consequences of Mycetoma. *PLoS Neglected Tropical Diseases*. [Online] PLOS; 2018;12(12). Available from: doi:10.1371/journal.pntd.0007019 [Accessed: 4th September 2023]

6. Fahal AH. Mycetoma: A thorn in the flesh. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. [Online] 2004;98(1): 3–11. Available from: doi:10.1016/S0035-9203(03)00009-9

7. World Health Organization. *Mycetoma*. [Online] Available from: https://www.who.int/news-room/fact-sheets/detail/mycetoma

8. Hounsome N, Hassan R, Bakhiet SM, Deribe K, Bremner S, Fahal AH, et al. Role of socioeconomic factors in developing mycetoma: Results from a household survey in Sennar State, Sudan. *PLoS Neglected Tropical Diseases*. [Online] PLOS; 2022;16(10). Available from: doi:10.1371/journal.pntd.0010817 [Accessed: 1st September 2023]

9. Fahal A, Mahgoub ES, EL Hassan AM, Abdel-Rahman ME, Alshambaty Y, Hashim A, et al. A New Model for Management of Mycetoma in the Sudan. Vinetz JM (ed.) *PLoS Neglected Tropical Diseases*. [Online] Public Library of Science;

2014;8(10): e3271. Available from: doi:10.1371/journal.pntd.0003271

10. Reis CMS, Reis-Filho EG de M. Mycetomas: An epidemiological, etiological, clinical, laboratory and therapeutic review. *Anais Brasileiros de Dermatologia*. [Online] Sociedade Brasileira de Dermatologia; 2018;93(1): 8–18. Available from: doi:10.1590/abd1806-4841.20187075

11. van de Sande WWJ, Fahal AH, Goodfellow M, Mahgoub ES, Welsh O, Zijlstra EE. Merits and Pitfalls of Currently Used Diagnostic Tools in Mycetoma. *PLoS Neglected Tropical Diseases*. [Online] 2014;8(7). Available from: doi:10.1371/journal.pntd.0002918

12. Nenoff P, Van De Sande WWJ, Fahal AH, Reinel D, Schöfer H. *Eumycetoma and actinomycetoma - An update on causative agents, epidemiology, pathogenesis, diagnostics and therapy*. [Online] Journal of the European Academy of Dermatology and Venereology. 2015. p. 1873–1883. Available from: doi:10.1111/jdv.13008

13. Lichon V, Khachemoune A. *Mycetoma: a review*. [Online] American Journal of Clinical Dermatology. 2006. p. 315–321. Available from: doi:10.2165/00128071-200607050-00005

14. Welsh O, Al-Abdely HM, Salinas-Carmona MC, Fahal AH. Mycetoma Medical Therapy. *PLoS Neglected Tropical Diseases*. [Online] 2014;8(10). Available from: doi:10.1371/journal.pntd.0003218

15. Zijlstra EE, van de Sande WWJ, Welsh O, Mahgoub ES heikh, Goodfellow M, Fahal AH. Mycetoma: a unique neglected tropical disease. *The Lancet. Infectious diseases*. [Online] Elsevier Ltd; 2016;16(1): 100–112. Available from: doi:10.1016/S1473-3099(15)00359-X

16. Sampaio FMS, Galhardo MCG, Quintella LP, Souza PRC de, Coelho JMC de O, Valle ACF do, et al. Eumycetoma by Madurella mycetomatis with 30 years of evolution: therapeutic challenge. *Anais Brasileiros de Dermatologia*. [Online] 2013;88(6): 82–84. Available from: doi:10.1590/abd1806-4841.20132136

17. van de Sande WWJ. Global Burden of Human Mycetoma: A Systematic Review and Meta-analysis. Vinetz JM (ed.) *PLoS Neglected Tropical Diseases*. [Online] Public Library of Science; 2013;7(11): e2550. Available from: doi:10.1371/journal.pntd.0002550

18. Hassan R, Deribe K, Fahal AH, Newport M, Bakhiet S. Clinical epidemiological characteristics of mycetoma in Eastern Sennar locality, Sennar State, Sudan. *PLoS Neglected Tropical Diseases*. [Online] United States; 2021;15(12): e0009847. Available from: doi:10.1371/journal.pntd.0009847

19. Maiti PK, Ray A, Bandyopadhyay S. Epidemiological aspects of mycetoma from a retrospective study of 264 cases in West Bengal. *Tropical Medicine and International Health*. [Online] Trop Med Int Health; 2002;7(9): 788–792. Available from: doi:10.1046/j.1365-3156.2002.00915.x [Accessed: 1st September 2023]

20. Dubey N, Capoor MR, Hasan AS, Gupta A, Ramesh V, Sharma S, et al. Epidemiological profile and spectrum of neglected tropical disease eumycetoma from Delhi, North India. *Epidemiology and infection*. [Online] Cambridge University Press; 2019;147: e294. Available from: doi:10.1017/S0950268819001766 [Accessed: 1st September 2023]

21. López-Martínez R, Méndez-Tovar LJ, Bonifaz A, Arenas R, Mayorga J, Welsh O, et al. Update on the epidemiology of mycetoma in Mexico. A review of 3933 cases. *Gaceta medica de Mexico*. Mexico; 2013;149(5): 586–592.

22. Ahmed A, Adelmann D, Fahal A, Verbrugh H, Van Belkum A, De Hoog S. Environmental occurrence of Madurella mycetomatis, the major agent of human eumycetoma in Sudan. *Journal of Clinical Microbiology*. [Online] 2002;40(3): 1031–1036. Available from: doi:10.1128/JCM.40.3.1031-1036.2002

23. Samy AM, van de Sande WWJ, Fahal AH, Peterson AT. Mapping the potential risk of mycetoma infection in Sudan and South Sudan using ecological niche modeling. *PLoS neglected tropical diseases*. [Online] 2014;8(10): e3250. Available from: doi:10.1371/journal.pntd.0003250

24. de Hoog GS, Ahmed SA, Najafzadeh MJ, Sutton DA, Keisari MS, Fahal AH, et al. Phylogenetic Findings Suggest Possible New Habitat and Routes of Infection of Human Eumyctoma. Wanke B (ed.) *PLoS Neglected Tropical Diseases*. [Online] 2013;7(5): e2229. Available from: doi:10.1371/journal.pntd.0002229

25. Azrag RS, Bakhiet SM, Mhmoud NA, Almalik AM, Mohamed AH, Fahal AH. A possible role for ticks in the transmission of Madurella mycetomatis in a mycetoma-endemic village in Sudan. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. [Online] Oxford Academic; 2021;115(4): 364–374. Available from: doi:10.1093/trstmh/trab030 [Accessed: 2nd April 2023]

26. Fahal AH, Suliman SH, Hay R. Mycetoma: The Spectrum of Clinical Presentation. *Tropical Medicine and Infectious Disease*. [Online] Multidisciplinary Digital Publishing Institute (MDPI); 2018;3(3). Available from: doi:10.3390/TROPICALMED3030097 [Accessed: 1st January 2023]

27. EL Hassan, AM; Fahal, AH; EL Hag, IA; Khalil E. The pathology of mycetoma: light microscopic and ultrastructural features. *Sudan Med J.* 1994;(32): 23–45.

28. Fahal AH, El Toum EA, El Hassan AM, Mahgoub ES, Gumaa SA. The host tissue reaction to madurella mycetomatis: New classification. *Medical Mycology*. [Online] 1995;33(1): 15–17. Available from: doi:10.1080/02681219580000041

29. Mahgoub ES, Gumaa SA, El Hassan AM. Immunological status of mycetoma patients. *Bulletin de la Societe de pathologie exotique et de ses filiales*. [Online] 1977;70(1): 48–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/579331

30. Hassan AME, Fahal AH, Ahmed AO, Ismail A, Veress B. The immunopathology of actinomycetoma lesions caused by Streptomyces somaliensis. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. [Online] 2001;95(1): 89–92. Available from: doi:10.1016/S0035-9203(01)90346-3

31. Nasr A, Abushouk A, Hamza A, Siddig E, Fahal AH. Th-1, Th-2 Cytokines Profile among Madurella mycetomatis Eumycetoma Patients. *PLoS Neglected Tropical Diseases*. [Online] 2016;10(7): 1–11. Available from:

doi:10.1371/journal.pntd.0004862

32. Mhmoud NA, Fahal AH, van de Sande WWJ. The association between the interleukin-10 cytokine and CC chemokine ligand 5 polymorphisms and mycetoma granuloma formation. *Medical mycology*. [Online] 2015;53(3): 295–301. Available from: doi:10.3109/13693786.2012.745201

33. Abushouk A, Nasr A, Masuadi E, Allam G, Siddig EE, Fahal AH. The Role of Interleukin-1 cytokine family (IL-1β, IL-37) and interleukin-12 cytokine family (IL-12, IL-35) in eumycetoma infection pathogenesis. *PLoS Neglected Tropical Diseases*. [Online] 2019;13(4): e0007098. Available from: doi:10.1371/journal.pntd.0007098

34. Siddig EE, Mohammed Edris AM, Bakhiet SM, van de Sande WWJ, Fahal AH. Interleukin-17 and matrix metalloprotease-9 expression in the mycetoma granuloma. *PLoS Neglected Tropical Diseases*. [Online] 2019;13(7): 1–17. Available from: doi:10.1371/journal.pntd.0007351

35. Sumangala B, Shetty NSS, Kala B, Kavya M. Mycetoma Foot caused by Fusarium solani in a Young Boy from Karnataka, India. *International Journal of Current Microbiology and Applied Sciences*. [Online] 2016;5(7): 307–310. Available from: doi:10.20546/ijcmas.2016.507.032

36. Taylor PR, Leal SM, Sun Y, Pearlman E. Aspergillus and Fusarium Corneal Infections Are Regulated by Th17 Cells and IL-17–Producing Neutrophils . *The Journal of Immunology*. [Online] The American Association of Immunologists; 2014;192(7): 3319–3327. Available from: doi:10.4049/jimmunol.1302235

37. Wethered DB, Markey MA, Hay RJ, Mahgoub ES, Gumaa SA. Humoral immune responses to mycetoma organisms: Characterization of specific antibodies by the use of enzyme-linked immunosorbent assay and immunoblotting. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. [Online] 1988;82(6): 918–923. Available from: doi:10.1016/0035-9203(88)90042-9

38. van Hellemond JJ, Vonk AG, de Vogel C, Koelewijn R, Vaessen N, Fahal AH, et

al. Association of Eumycetoma and Schistosomiasis. Ozkan AT (ed.) *PLoS Neglected Tropical Diseases*. [Online] Public Library of Science; 2013;7(5): e2241. Available from: doi:10.1371/journal.pntd.0002241

39. Bakhiet SM, Fahal AH, Musa AM, Mohamed ESW, Omer RF, Ahmed ES, et al. A holistic approach to the mycetoma management. Franco-Paredes C (ed.) *PLoS Neglected Tropical Diseases*. [Online] Public Library of Science; 2018;12(5): e0006391. Available from: doi:10.1371/journal.pntd.0006391 [Accessed: 16th April 2019]

40. Fahal AH, Bakhiet SM. Mycetoma and the environment. *PLOS Neglected Tropical Diseases*. [Online] Public Library of Science; 2023;17(11): e0011736. Available from: https://doi.org/10.1371/journal.pntd.0011736

41. Fahal AH. Mycetoma: A thorn in the flesh. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. [Online] 2004;98(1): 3–11. Available from: doi:10.1016/S0035-9203(03)00009-9

42. Bonifaz A, Tirado-Sánchez A, Calderón L, Saúl A, Araiza J, Hernández M, et al. Mycetoma: Experience of 482 Cases in a Single Center in Mexico. *PLoS Neglected Tropical Diseases*. [Online] Public Library of Science; 2014;8(8): e3102. Available from: doi:10.1371/journal.pntd.0003102 [Accessed: 11th October 2023]

43. Hassan R, Deribe K, Simpson H, Bremner S, Elhadi O, Alnour M, et al. Individual Risk Factors of Mycetoma Occurrence in Eastern Sennar Locality, Sennar State, Sudan: A Case-Control Study. *Tropical Medicine and Infectious Disease*. [Online] Multidisciplinary Digital Publishing Institute (MDPI); 2022;7(8). Available from: doi:10.3390/tropicalmed7080174 [Accessed: 4th September 2023]

44. Ezaldeen EA, Fahal AH, Osman A. Mycetoma Herbal Treatment: The Mycetoma Research Centre, Sudan Experience. *PLoS Neglected Tropical Diseases*. [Online] Public Library of Science; 2013;7(8). Available from: doi:10.1371/journal.pntd.0002400

45. Elkheir LYM, Haroun R, Mohamed MA, Fahal AH. Madurella mycetomatis causing eumycetoma medical treatment: The challenges and prospects. Samy AM (ed.) *PLOS Neglected Tropical Diseases*. [Online] Public Library of Science; 2020;14(8): e0008307. Available from: doi:10.1371/journal.pntd.0008307

46. Blackwell JM, Jamieson SE, Burgner D. HLA and infectious diseases. *Clinical Microbiology Reviews*. [Online] American Society for Microbiology (ASM); 2009;22(2): 370–385. Available from: doi:10.1128/CMR.00048-08 [Accessed: 8th April 2023]

47. Newport MJ, Finan C. Genome-wide association studies and susceptibility to infectious diseases. *Briefings in Functional Genomics*. [Online] Oxford Academic; 2011;10(2): 98–107. Available from: doi:10.1093/bfgp/elq037 [Accessed: 28th September 2023]

48. Kwok AJ, Mentzer A, Knight JC. Host genetics and infectious disease: new tools, insights and translational opportunities. *Nature Reviews Genetics*. [Online] Nature Publishing Group; 2021;22(3): 137–153. Available from: doi:10.1038/s41576-020-00297-6 [Accessed: 2nd November 2021]

49. Kariuki SN, Williams TN. Human genetics and malaria resistance. *Human Genetics*. [Online] Springer; 2020;139(6–7): 801–811. Available from: doi:10.1007/s00439-020-02142-6 [Accessed: 13th October 2023]

50. McLaren PJ, Fellay J. HIV-1 and human genetic variation. *Nature Reviews Genetics*. [Online] Nature Publishing Group; 2021;22(10): 645–657. Available from: doi:10.1038/s41576-021-00378-0 [Accessed: 13th October 2023]

51. van de Sande WWJ, Maghoub ES, Fahal AH, Goodfellow M, Welsh O, Zijlstra E. The Mycetoma Knowledge Gap: Identification of Research Priorities. *PLoS Neglected Tropical Diseases*. [Online] 2014;8(3): 1–7. Available from: doi:10.1371/journal.pntd.0002667

52. Mclaren ML, Mahgoub ES, Georgakopoulos E. Preliminary investigation on the use of the enzyme linked immunosorbent assay (ELISA) in the serodiagnosis of

mycetoma. *Medical Mycology*. [Online] 1978;16(3): 225–228. Available from: doi:10.1080/00362177885380301

53. Taha A. A serological survey of antibodies to Streptomyces somaliensis and Actinomadura madurae in the Sudan using enzyme linked immunosorbent assay (ELISA). *Transactions of the Royal Society of Tropical Medicine and Hygiene*. [Online] 1983;77(1): 49–50. Available from: doi:10.1016/0035-9203(83)90011-1

54. Verbrugh H, Goedhart H, Ahmed AOA, van der Zee R, van de Sande WWJ, Janse D-J, et al. Translationally Controlled Tumor Protein from Madurella mycetomatis, a Marker for Tumorous Mycetoma Progression. *The Journal of Immunology*. [Online] American Association of Immunologists; 2014;177(3): 1997–2005. Available from: doi:10.4049/jimmunol.177.3.1997

55. Al Dawi AF, Mustafa MI, Fahal AH, EL Hassan AM, Bakhiet S, Magoub ES, et al. The association of HLA-DRB1 and HLA-DQB1 alleles and the occurrence of eumycetoma in Sudanese patients. *Khartoum Medical Journal*. 2013;6.

56. Fahal A, Mahgoub ES, Hassan AM EL, Abdel-Rahman ME. Mycetoma in the Sudan: An Update from the Mycetoma Research Centre, University of Khartoum, Sudan. Wanke B (ed.) *PLOS Neglected Tropical Diseases*. [Online] 2015;9(3): e0003679. Available from: doi:10.1371/journal.pntd.0003679

57. Hassan R, Deribe K, Fahal AH, Newport M, Bakhiet S. Clinical epidemiological characteristics of mycetoma in Eastern Sennar locality, Sennar State, Sudan. *PLoS Neglected Tropical Diseases*. [Online] PLOS; 2021;15(12). Available from: doi:10.1371/JOURNAL.PNTD.0009847 [Accessed: 30th December 2022]

58. Pérez-Blanco M, Hernández Valles R, García-Humbría L, Yegres F. Chromoblastomycosis in children and adolescents in the endemic area of the Falcón State, Venezuela. *Medical Mycology*. [Online] 2006;44(5): 467–471. Available from: doi:10.1080/13693780500543238

59. Tsuneto LT, Arce-Gomez B, Petzl-Erler ML, Queiroz-Telles F. HLA-a29 and genetic susceptibility to chromoblastomycosis. *Medical Mycology*. [Online]

Informa Healthcare; 1989;27(3): 181–185. Available from: doi:10.1080/02681218980000241

60. Wang X, Wang W, Lin Z, Wang X, Li T, Yu J, et al. CARD9 mutations linked to subcutaneous phaeohyphomycosis and T H17 cell deficiencies. *Journal of Allergy and Clinical Immunology*. [Online] Mosby Inc.; 2014;133(3). Available from: doi:10.1016/j.jaci.2013.09.033

61. Mhmoud NA, Fahal AH, van de Sande WWJ. The association between the interleukin-10 cytokine and CC chemokine ligand 5 polymorphisms and mycetoma granuloma formation. *Medical mycology*. [Online] 2015;53(3): 295–301. Available from: doi:10.3109/13693786.2012.745201

62. Verwer PEB, Notenboom CC, Eadie K, Fahal AH, Verbrugh HA, van de Sande WWJ. A Polymorphism in the Chitotriosidase Gene Associated with Risk of Mycetoma Due to Madurella mycetomatis Mycetoma–A Retrospective Study. *PLoS Neglected Tropical Diseases*. [Online] 2015;9(9): 1–11. Available from: doi:10.1371/journal.pntd.0004061

63. Ahmed SA bdalla, de Hoog GS, Stevens DA, Fahal AH, van de Sande WWJ. Polymorphisms in catechol-O-methyltransferase and cytochrome p450 subfamily 19 genes predispose towards Madurella mycetomatis-induced mycetoma susceptibility. *Medical mycology*. [Online] 2015;53(3): 295–301. Available from: doi:10.3109/13693781003636680

64. van de Sande WWJ, Fahal A, Verbrugh H, van Belkum A. Polymorphisms in Genes Involved in Innate Immunity Predispose Toward Mycetoma Susceptibility. *The Journal of Immunology*. [Online] 2007;179(5): 3065–3074. Available from: doi:10.4049/jimmunol.179.5.3065

65. Geneugelijk K, Kloezen W, Fahal AH, van de Sande WWJ. Active Matrix Metalloprotease-9 Is Associated with the Collagen Capsule Surrounding the Madurella mycetomatis Grain in Mycetoma. *PLoS Neglected Tropical Diseases*. [Online] 2014;8(3): 1–7. Available from: doi:10.1371/journal.pntd.0002754

66. Stechmiller JK. *Understanding the role of nutrition and wound healing*. [Online] Nutrition in Clinical Practice. Nutr Clin Pract; 2010. p. 61–68. Available from: doi:10.1177/0884533609358997 [Accessed: 4th September 2023]

67. Prasad AS. *Zinc in human health: Effect of zinc on immune cells*. [Online] Molecular Medicine. The Feinstein Institute for Medical Research; 2008. p. 353–357. Available from: doi:10.2119/2008-00033.Prasad [Accessed: 4th September 2023]

68. Huang Z, Liu Y, Qi G, Brand D, Zheng SG. *Role of vitamin A in the immune system*. [Online] Journal of Clinical Medicine. Multidisciplinary Digital Publishing Institute (MDPI); 2018. Available from: doi:10.3390/jcm7090258 [Accessed: 4th September 2023]

69. Hao X, Cognetti M, Burch-Smith R, Mejia EO, Mirkin G. *Mycetoma: Development of Diagnosis and Treatment*. [Online] Journal of Fungi. Multidisciplinary Digital Publishing Institute (MDPI); 2022. Available from: doi:10.3390/jof8070743 [Accessed: 12th September 2023]

70. Ahmed AOA, Van Leeuwen W, Fahal A, Van De Sande W, Verbrugh H, Van Belkum A. Mycetoma caused by Madurella mycetomatis: A neglected infectious burden. *Lancet Infectious Diseases*. [Online] 2004;4(9): 566–574. Available from: doi:10.1016/S1473-3099(04)01131-4

71. Ahmed A, Adelmann D, Fahal A, Verbrugh H, Van Belkum A, De Hoog S. Environmental occurrence of Madurella mycetomatis, the major agent of human eumycetoma in Sudan. *Journal of Clinical Microbiology*. [Online] 2002;40(3): 1031–1036. Available from: doi:10.1128/JCM.40.3.1031-1036.2002

72. Hashizume H, Taga S, Sakata MK, Taha MHM, Siddig EE, Minamoto T, et al. Detection of multiple mycetoma pathogens using fungal metabarcoding analysis of soil DNA in an endemic area of Sudan. *PLoS Neglected Tropical Diseases*. [Online] Public Library of Science; 2022;16(3): e0010274. Available from: doi:10.1371/journal.pntd.0010274 [Accessed: 4th September 2023]

73. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*. [Online] John Wiley & Sons, Ltd; 2012;21(8): 2045–2050. Available from: doi:10.1111/j.1365-294X.2012.05470.x [Accessed: 4th September 2023]

74. Hassan R, Simpson H, Cano J, Bakhiet S, Ganawa E, Argaw D, et al. Modelling the spatial distribution of mycetoma in Sudan. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. [Online] England; 2021;115(10): 1144–1152. Available from: doi:10.1093/trstmh/trab076

75. Ahmed AA, van de Sande W, Fahal AH. Mycetoma laboratory diagnosis: Review article. *PLoS Neglected Tropical Diseases*. [Online] Public Library of Science; 2017;11(8). Available from: doi:10.1371/journal.pntd.0005638

76. van de Sande W, Fahal A, Ahmed SA, Serrano JA, Bonifaz A, Zijlstra E. Closing the mycetoma knowledge gap. *Medical Mycology*. [Online] 2017; 1–12. Available from: doi:10.1093/mmy/myx061

77. Queiroz-Telles F, Fahal AH, Falci DR, Caceres DH, Chiller T, Pasqualotto AC. Neglected endemic mycoses. *The Lancet Infectious Diseases*. [Online] Elsevier Ltd; 2017;17(11): e367–e377. Available from: doi:10.1016/S1473-3099(17)30306-7

78. Dupont B, Datry A, Poirée S, Canestri A, Boucheneb S, Fourniols E. Role of a NSAID in the apparent cure of a fungal mycetoma. *Journal de Mycologie Medicale*. [Online] Elsevier Masson SAS; 2016;26(2): 86–93. Available from: doi:10.1016/j.mycmed.2016.03.003

79. Watanabe S, Tsubouchi I, Okubo A. Efficacy and safety of fosravuconazole L-lysine ethanolate, a novel oral triazole antifungal agent, for the treatment of onychomycosis: A multicenter, double-blind, randomized phase III study. *Journal of Dermatology*. [Online] Wiley-Blackwell; 2018;45(10): 1151–1159. Available from: doi:10.1111/1346-8138.14607 [Accessed: 8th April 2023]

80. DNDi. *World's first clinical trial for devastating fungal disease mycetoma shows*

efficacy of new, promising treatment. [Online] Available from: https://dndi.org/press-releases/2023/worlds-first-clinical-trial-for-mycetoma-shows-efficacy-new-promising-treatment/

81. Indo-Africa. *Top 10 Largest African Countries – Indo African Chamber Of Commerce & Industry*. [Online] Available from: https://www.indoafrican.org/top-10-largest-african-countries/ [Accessed: 5th September 2023]

82. World Atlas. *Sudan Maps & Facts*. [Online] Available from: https://www.worldatlas.com/maps/sudan [Accessed: 5th September 2023]

83. Fanack. *Water Resources in Sudan*. [Online] Available from: https://water.fanack.com/sudan/water-resources-in-sudan/ [Accessed: 5th September 2023]

84. Fanack. *Geography of Sudan*. [Online] Available from: https://fanack.com/sudan/geography-of-sudan/ [Accessed: 5th September 2023]

85. *States of Sudan | Mappr*. [Online] Available from: https://www.mappr.co/counties/states-of-sudan/ [Accessed: 26th July 2023]

86. World Popoluation Review. *Sudan Population 2023*. [Online] Available from: https://worldpopulationreview.com/countries/sudan-population [Accessed: 5th September 2023]

87. Kulneff C. A comparative study of urban and rural dairy management systems in Sudan. SLU, Dept. of Anatomy, Physiology and Biochemistry; 2006;

88. Elsammani MO, Salih AAA. *Nomads' Settlement in Sudan: Experiences, Lessons and Future action*. [Online] 2006 [Accessed: 6th September 2023]. Available from: www.sd.undp.org [Accessed: 6th September 2023]

89. Crowther N, Okamura K, Raja C, Rinnert D, Spencer E. Inequalities in Public Services in the Sudan. 2014;

90. UNICEF Sudan. *SUDAN EDUCATION: GLOBAL THEMATIC REPORT JANUARY – DECEMBER 2018*. 2019.

91. Ismail M. Regional disparities in the distribution of Sudan's health resources. *Eastern Mediterranean Health Journal*. [Online] 2020;26(9): 1105–1114. Available from: doi:10.26719/emhj.20.056 [Accessed: 6th September 2023]

92. Mustafa ME, Elshakh MM, Ebaidalla EM. Does foreign aid promote economic growth in Sudan? Evidence from ardl bounds testing analysis. *Journal of Economic Cooperation and Development*. 2019;40(3): 115–140.

93. Abbass Ali SM. Post-secession Sudan and South Sudan: A Comparative Study of Economic Performance, Export Diversification, and Institutions. *Journal of Asian and African Studies*. [Online] SAGE Publications Ltd; 2022; Available from: doi:10.1177/00219096221076106/ASSET/IMAGES/LARGE/10.1177_00219096221076106-FIG19.JPEG [Accessed: 11th September 2023]

94. World Bank. *World Development Indicators*. [Online] Databank. Available from: https://databank.worldbank.org/source/world-development-indicators [Accessed: 11th September 2023]

95. Ebaidalla EM. Assessing the potential and challenges for trilateral trade integration among Egypt, Sudan, and Ethiopia. *Development Studies Research*. [Online] Routledge; 2023;10(1). Available from: doi:10.1080/21665095.2023.2220563 [Accessed: 11th September 2023]

96. Fanack. *Economy of Sudan*. [Online] Available from: https://fanack.com/sudan/economy-of-sudan/ [Accessed: 11th September 2023]

97. The Economist Intelligence Unit. *Sudan Economy, Politics and GDP Growth*. [Online] 2023 [Accessed: 11th September 2023]. Available from: https://country.eiu.com/sudan [Accessed: 11th September 2023]

98. Hublin JJ, Ben-Ncer A, Bailey SE, Freidline SE, Neubauer S, Skinner MM, et al. New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature*. [Online] Nature; 2017;546(7657): 289–292. Available from: doi:10.1038/nature22336 [Accessed: 9th September 2023]

99. Wohns AW, Wong Y, Jeffery B, Akbari A, Mallick S, Pinhasi R, et al. A unified

genealogy of modern and ancient genomes. *Science*. [Online] American Association for the Advancement of Science; 2022;375(6583). Available from: doi:10.1126/science.abi8264 [Accessed: 4th May 2023]

100. Campbell MC, Hirbo JB, Townsend JP, Tishkoff SA. *The peopling of the African continent and the diaspora into the new world*. [Online] Current Opinion in Genetics and Development. Curr Opin Genet Dev; 2014. p. 120–132. Available from: doi:10.1016/j.gde.2014.09.003 [Accessed: 9th September 2023]

101. Elhassan N, Gebremeskel EI, Elnour MA, Isabirye D, Okello J, Hussien A, et al. The episode of genetic drift defining the migration of humans out of africa is derived from a large east african population size. *PLoS ONE*. [Online] Public Library of Science; 2014;9(5): e97674. Available from: doi:10.1371/journal.pone.0097674 [Accessed: 9th September 2023]

102. Ibrahim ME. *Genetic diversity of the Sudanese: Insights on origin and implications for health*. [Online] Human Molecular Genetics. Oxford University Press; 2021. p. R37–R41. Available from: doi:10.1093/hmg/ddab028 [Accessed: 9th September 2023]

103. Eberhard D, Simons GF, Fennig CD. *Ethnologue: Languages of the World*. [Online]. Twenty-six. Dallas, Texas: SIL International; 2023. Available from: https://www.ethnologue.com/ [Accessed: 7th September 2023]

104. Wolff HE. *Afro-Asiatic languages*. [Online] Encyclopedia Britannica. Available from: https://www.britannica.com/topic/Afro-Asiatic-languages [Accessed: 7th September 2023]

105. Bendor-Samuel JT. *Niger-Congo languages*. [Online] Encyclopedia Britannica. Available from: https://www.britannica.com/topic/Niger-Congo-languages [Accessed: 7th September 2023]

106. Dimmendaal, Gerrit J. and Goodman MF. *Nilo-Saharan languages*. [Online] Encyclopedia Britannica. Available from: https://www.britannica.com/topic/Nilo-Saharan-languages [Accessed: 7th September 2023]

107. Köhler ORA, Traill A. *Khoisan languages*. [Online] Encyclopedia Britannica. Available from: https://www.britannica.com/topic/Khoisan-languages [Accessed: 7th September 2023]

108. Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, et al. The genetic prehistory of southern Africa. *Nature Communications*. [Online] 2012;3(1): 1143. Available from: doi:10.1038/ncomms2140

109. WorldAtlas. *Ethnic Groups Of Sudan*. [Online] Available from: https://www.worldatlas.com/articles/the-ethnic-groups-in-sudan.html [Accessed: 7th September 2023]

110. Fan S, Spence JP, Feng Y, Hansen MEB, Terhorst J, Beltrame MH, et al. Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell*. [Online] Elsevier B.V.; 2023;186(5): 923-939.e14. Available from: doi:10.1016/j.cell.2023.01.042

111. Fan S, Kelly DE, Beltrame MH, Hansen MEB, Mallick S, Ranciaro A, et al. African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biology*. [Online] BioMed Central Ltd.; 2019;20(1): 1–14. Available from: doi:10.1186/s13059-019-1679-2 [Accessed: 10th September 2023]

112. Choudhury A, Aron S, Botigué LR, Sengupta D, Botha G, Bensellak T, et al. High-depth African genomes inform human migration and health. *Nature*. [Online] Nature Publishing Group; 2020;586(7831): 741–748. Available from: doi:10.1038/s41586-020-2859-7 [Accessed: 10th September 2023]

113. Ibrahim ME. *Genetic diversity of the Sudanese: Insights on origin and implications for health*. [Online] Human Molecular Genetics. Oxford Academic; 2021. p. R37–R41. Available from: doi:10.1093/hmg/ddab028 [Accessed: 2nd May 2023]

114. Bird N, Ormond L, Awah P, Caldwell EF, Connell B, Elamin M, et al. Dense sampling of ethnic groups within African countries reveals fine-scale genetic structure and extensive historical admixture. *Science Advances*. [Online]

American Association for the Advancement of Science; 2023;9(13). Available from: doi:10.1126/sciadv.abq2616 [Accessed: 7th September 2023]

115. Krings M, Salem AEH, Bauer K, Geisert H, Malek AK, Chaix L, et al. mtDNA analysis of Nile River Valley populations: A genetic corridor or a barrier to migration? *American Journal of Human Genetics*. [Online] Am J Hum Genet; 1999;64(4): 1166–1176. Available from: doi:10.1086/302314 [Accessed: 8th September 2023]

116. Hollfelder N, Schlebusch CM, Günther T, Babiker H, Hassan HY, Jakobsson M. Northeast African genomic variation shaped by the continuity of indigenous groups and Eurasian migrations. *PLoS Genetics*. [Online] Public Library of Science; 2017;13(8): e1006976. Available from: doi:10.1371/journal.pgen.1006976 [Accessed: 12th September 2023]

117. Braukämper U. Notes on the origin of Baggara Arab culture with special reference to the Shuwa. *Arabs and Arabic in the Lake Chad region*. 1994. p. 13–46.

118. Krafft C, Assaad R, Cortes-mendosa A, Honzay I, Krafft C. *The Structure of the Labor Force and Employment in Sudan*. 2023.

119. *Inbreeding by Country 2023*. [Online] Available from: https://worldpopulationreview.com/country-rankings/inbreeding-by-country [Accessed: 2nd August 2023]

120. Saha N, Sheikh FS. Inbreeding Levels in Khartoum. *Journal of Biosocial Science*. [Online] Cambridge University Press; 1988;20(3): 333–336. Available from: doi:10.1017/S0021932000006660 [Accessed: 19th August 2023]

121. Elhadi YAM, Alrawa SS, Alfadul ESAA, Mahgoub EAA, El-Osta A, Belal SA, et al. Consanguinity and willingness to perform premarital genetic screening in Sudan. *European Journal of Human Genetics*. [Online] Nature Publishing Group; 2023; 1–8. Available from: doi:10.1038/s41431-023-01438-1 [Accessed: 7th August 2023]

122. *Sudan - United States Department of State*. [Online] Available from:

https://www.state.gov/reports/2022-report-on-international-religious-freedom/sudan/ [Accessed: 20th August 2023]

123. Bittles AH, Hamamy HA. Endogamy and consanguineous marriage in Arab populations. *Genetic Disorders Among Arab Populations*. [Online] Springer Heidelberg Dordrecht London; 2010. p. 85–108. Available from: doi:10.1007/978-3-642-05080-0_4 [Accessed: 21st August 2023]

124. Bittles AH, Black ML. Consanguinity, human evolution, and complex diseases. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] National Academy of Sciences; 2010;107(SUPPL. 1): 1779–1786. Available from: doi:10.1073/pnas.0906079106 [Accessed: 20th August 2023]

125. Lyons EJ, Frodsham AJ, Zhang L, Hill AVS, Amos W. Consanguinity and susceptibility to infectious diseases in humans. *Biology Letters*. [Online] The Royal Society; 2009;5(4): 574–576. Available from: doi:10.1098/rsbl.2009.0133 [Accessed: 11th September 2023]

126. Mohammed AO, Attalla B, Bashir FMK, Ahmed FE, El Hassan AM, Ibnauf G, et al. Relationship of the sickle cell gene to the ethnic and geographic groups populating the Sudan. *Community Genetics*. [Online] S. Karger AG; 2006;9(2): 113–120. Available from: doi:10.1159/000091489 [Accessed: 11th September 2023]

127. Daak AA, Elsamani E, Ali EH, Mohamed FA, Abdel-Rahman ME, Elderdery AY, et al. Sickle cell disease in western Sudan: Genetic epidemiology and predictors of knowledge attitude and practices. *Tropical Medicine and International Health*. [Online] John Wiley & Sons, Ltd; 2016;21(5): 642–653. Available from: doi:10.1111/tmi.12689 [Accessed: 18th February 2023]

128. Elsayed LEO, Mohammed IN, Hamed AAA, Elseed MA, Johnson A, Mairey M, et al. Hereditary spastic paraplegias: Identification of a novel SPG57 variant affecting TFG oligomerization and description of HSP subtypes in Sudan. *European Journal of Human Genetics*. [Online] Nature Publishing Group; 2016. p.

100–110. Available from: doi:10.1038/ejhg.2016.108

129. Yahia A, Hamed AAA, Mohamed IN, Elseed MA, Salih MA, El-sadig SM, et al. Clinical phenotyping and genetic diagnosis of a large cohort of Sudanese families with hereditary spinocerebellar degenerations. *European Journal of Human Genetics*. [Online] Nature Publishing Group; 2023; 1–13. Available from: doi:10.1038/s41431-023-01344-6 [Accessed: 11th September 2023]

130. Yahia A, Ayed I Ben, Hamed AA, Mohammed IN, Elseed MA, Bakhiet AM, et al. Genetic diagnosis in Sudanese and Tunisian families with syndromic intellectual disability through exome sequencing. *Annals of Human Genetics*. [Online] Ann Hum Genet; 2022;86(4): 181–194. Available from: doi:10.1111/ahg.12460 [Accessed: 11th September 2023]

131. Bakhit Y, Ibrahim MO, Tesson C, Elhassan AA, Ahmed MA, Alebeed MA, et al. Intrafamilial and interfamilial heterogeneity of PINK1-associated Parkinson's disease in Sudan. *Parkinsonism and Related Disorders*. [Online] Parkinsonism Relat Disord; 2023;111. Available from: doi:10.1016/j.parkreldis.2023.105401 [Accessed: 11th September 2023]

132. Ellaithi M, Kamel A, Saber O. Consanguinity and Disorders of Sexual Developments in the Sudan. *Sudan Journal of Medical Sciences*. 2012;6: 267–269.

133. Saha N, Hamad RE, Mohamed S. Inbreeding effects on reproductive outcome in a sudanese population. *Human Heredity*. [Online] S. Karger AG; 1990;40(4): 208–212. Available from: doi:10.1159/000153932

134. Allison AC. Protection afforded by sickle-cell trait against subtertian malarial infection. *British Medical Journal*. [Online] BMJ Publishing Group; 1954;1(4857): 290–294. Available from: doi:10.1136/bmj.1.4857.290 [Accessed: 3rd May 2023]

135. Aidoo M, Terlouw DJ, Kolczak MS, McElroy PD, Ter Kuile FO, Kariuki S, et al. Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet*. [Online] Lancet; 2002;359(9314): 1311–1312. Available from:

doi:10.1016/S0140-6736(02)08273-9 [Accessed: 3rd May 2023]

136. Miller EN, Fadl M, Mohamed HS, Elzein A, Jamieson SE, Cordell HJ, et al. Y chromosome lineage- and village-specific genes on chromosomes 1p22 and 6q27 control visceral leishmaniasis in Sudan. *PLoS Genetics*. [Online] PLOS; 2007;3(5): 679–688. Available from: doi:10.1371/journal.pgen.0030071 [Accessed: 12th September 2023]

137. International Organization for Migration Displacement Tracking Matrix. *Weekly Displacement Snapshot (02)*. [Online] DTM Sudan. Available from: https://dtm.iom.int/reports/dtm-sudan-weekly-displacement-snapshot-02?close=true [Accessed: 11th September 2023]

138. United Nations Office for the Coordination of Humanitarian Affairs. *Sudan Humanitarian Update*. [Online] OCHA. Available from: https://reports.unocha.org/en/country/sudan/ [Accessed: 11th September 2023]

139. Comstock GW. Tuberculosis in twins: a re-analysis of the Prophit survey. *The American review of respiratory disease*. [Online] Am Rev Respir Dis; 1978;117(4): 621–624. Available from: doi:10.1164/ARRD.1978.117.4.621 [Accessed: 10th January 2023]

140. CHAKRAVARTTI M. A twin study on leprosy. *Topics in human genetics*. [Online] Stuttgart; 1973; 1–123. Available from: https://cir.nii.ac.jp/crid/1571980076160727552.bib?lang=en [Accessed: 10th January 2023]

141. Lin TM, Chen CJ, Wu MM, Yang CS, Chen JS, Lin CC, et al. Hepatitis B virus markers in Chinese twins. *Anticancer research*. Greece; 1989;9(3): 737–741.

142. Sørensen TIA, Nielsen GG, Andersen PK, Teasdale TW. Genetic and Environmental Influences on Premature Death in Adult Adoptees. *New England Journal of Medicine*. [Online] N Engl J Med; 1988;318(12): 727–732. Available from: doi:10.1056/nejm198803243181202 [Accessed: 10th January 2023]

143. Tenesa A, Haley CS. The heritability of human disease: Estimation, uses and

abuses. *Nature Reviews Genetics*. [Online] Nature Publishing Group; 2013;14(2): 139–149. Available from: doi:10.1038/nrg3377

144. Olson JM, Cordell HJ. Ascertainment bias in the estimation of sibling genetic risk parameters. *Genetic Epidemiology*. [Online] 2000;18(3): 217–235. Available from: doi:10.1002/(SICI)1098-2272(200003)18:3<217::AID-GEPI3>3.0.CO;2-8

145. Davey G, GebreHanna E, Adeyemo A, Rotimi C, Newport M, Desta K. Podoconiosis: a tropical model for gene–environment interactions? *Transactions of the Royal Society of Tropical Medicine and Hygiene*. [Online] 2007;101(1): 91–96. Available from: doi:10.1016/j.trstmh.2006.05.002 [Accessed: 27th November 2018]

146. Tekola Ayele F, Adeyemo A, Finan C, Hailu E, Sinnott P, Burlinson ND, et al. HLA Class II Locus and Susceptibility to Podoconiosis. *New England Journal of Medicine*. [Online] Massachusetts Medical Society ; 2012;366(13): 1200–1208. Available from: doi:10.1056/NEJMoa1108448 [Accessed: 27th November 2018]

147. Gebresilase T, Finan C, Suveges D, Tessema TS, Aseffa A, Davey G, et al. Replication of HLA class II locus association with susceptibility to podoconiosis in three Ethiopian ethnic groups. *Scientific Reports 2021 11:1*. [Online] Nature Publishing Group; 2021;11(1): 1–11. Available from: doi:10.1038/s41598-021-81836-x [Accessed: 12th January 2023]

148. Abel L, El-baghdadi J, Bousfiha AA, Casanova J, Schurr E, Abel L, et al. Human genetics of tuberculosis : a long and winding road. *Phisiological Transactions of The Royal Scoeity B*. [Online] 2015;369(figure 1): 20130428. Available from: http://dx.doi.org/10.1098/rstb.2013.0428

149. Boisson-Dupuis S, Baghdadi JE, Parvaneh N, Bousfiha A, Bustamante J, Feinberg J, et al. Il-12rβ1 deficiency in two of fifty children with severe tuberculosis from IRN, MAR, and TUR. *PLoS ONE*. [Online] PLOS; 2011;6(4). Available from: doi:10.1371/journal.pone.0018524 [Accessed: 1st August 2023]

150. Cooper AM, Solache A, Khader SA. *Interleukin-12 and tuberculosis: an old story*

revisited. [Online] Current Opinion in Immunology. Elsevier Current Trends; 2007. p. 441–447. Available from: doi:10.1016/j.coi.2007.07.004

151. Leiva-torres GA, Nebesio N, Vidal SM. *Innate antiviral immunity*. [Online] Bulletin of Russian State Medical University. 2020. Available from: doi:10.24075/brsmu.2020-05

152. Chapman SJ, Hill A V. Human genetic susceptibility to infectious disease. *Nat Rev Genet*. [Online] Nature Publishing Group; 2012;13(3): 175–188. Available from: doi:10.1038/nrg3114

153. Burgner D, Jamieson SE, Blackwell JM. Genetic susceptibility to infectious diseases: big is beautiful, but will bigger be even better? *Lancet Infectious Diseases*. [Online] 2006;6(10): 653–663. Available from: doi:10.1016/S1473-3099(06)70601-6

154. Lucek K, Willi Y. Drivers of linkage disequilibrium across a species' geographic range. *PLOS Genetics*. [Online] Public Library of Science; 2021;17(3): e1009477. Available from: doi:10.1371/JOURNAL.PGEN.1009477 [Accessed: 16th January 2023]

155. Belmont JW, Boudreau A, Leal SM, Hardenbol P, Pasternak S, Wheeler DA, et al. A haplotype map of the human genome. *Nature*. [Online] England; 2005;437(7063): 1299–1320. Available from: doi:10.1038/nature04226

156. Manolio TA. Genomewide Association Studies and Assessment of the Risk of Disease. *New England Journal of Medicine*. [Online] N Engl J Med; 2010;363(2): 166–176. Available from: doi:10.1056/nejmra0905980 [Accessed: 18th January 2023]

157. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. [Online] Europe PMC Funders; 2007;449(7164): 851. Available from: doi:10.1038/NATURE06258 [Accessed: 18th January 2023]

158. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method

for genome-wide association studies by imputation of genotypes. *Nature Genetics*. [Online] 2007;39(7): 906–913. Available from: doi:10.1038/ng2088

159. Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. A map of human genome variation from population-scale sequencing. *Nature*. [Online] 2010;467(7319): 1061–1073. Available from: doi:10.1038/nature09534

160. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*. [Online] Nat Genet; 2016;48(10): 1279–1283. Available from: doi:10.1038/ng.3643 [Accessed: 18th January 2023]

161. Kowalski MH, Qian H, Hou Z, Rosen JD, Tapia AL, Shan Y, et al. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genetics*. [Online] PLoS Genet; 2019;15(12). Available from: doi:10.1371/journal.pgen.1008500 [Accessed: 18th January 2023]

162. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. *Human Molecular Genetics*. [Online] Oxford University Press; 2015;24(R1): R111–R119. Available from: doi:10.1093/hmg/ddv260 [Accessed: 18th January 2023]

163. Teo YY, Small KS, Kwiatkowski DP. Methodological challenges of genome-wide association analysis in Africa. *Nature reviews. Genetics*. [Online] Europe PMC Funders; 2010;11(2): 149. Available from: doi:10.1038/NRG2731 [Accessed: 18th January 2023]

164. *H3Africa array annotations*. [Online] Available from: https://chipinfo.h3abionet.org/ [Accessed: 18th January 2023]

165. Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, Clark TG, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature Genetics*. [Online] Nature Publishing Group; 2009;41(6): 657–665. Available from: doi:10.1038/ng.388 [Accessed: 4th April 2023]

166. Band G, Le QS, Clarke GM, Kivinen K, Hubbart C, Jeffreys AE, et al. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nature Communications*. [Online] Nature Publishing Group; 2019;10(1): 5732. Available from: doi:10.1038/s41467-019-13480-z [Accessed: 18th January 2023]

167. Sazonovs A, Barrett JC. Rare-variant studies to complement genome-wide association studies. *Annual Review of Genomics and Human Genetics*. [Online] 2018;19: 97–112. Available from: doi:10.1146/annurev-genom-083117-021641

168. Yin R, Kwoh CK, Zheng J. Whole genome sequencing analysis. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. [Online] Elsevier; 2018. p. 176–183. Available from: doi:10.1016/B978-0-12-809633-8.20095-2

169. Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. Exome Sequencing: Current and Future Perspectives. *G3 (Bethesda, Md.)*. [Online] 2015;5(August): g3.115.018564-. Available from: doi:10.1534/g3.115.018564 [Accessed: 10th February 2018]

170. 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*. [Online] Nature Publishing Group; 2015;17(5): 405–424. Available from: doi:10.1038/gim.2015.30 [Accessed: 6th April 2023]

171. Jiao H, Kulyté A, Näslund E, Thorell A, Gerdhem P, Kere J, et al. Whole-exome sequencing suggests LAMB3 as a susceptibility gene for morbid obesity. *Diabetes*. [Online] 2016;65(10): 2981–2989. Available from: doi:10.2337/db16-0522

172. Hixson JE, Jun G, Shimmin LC, Wang Y, Yu G, Mao C, et al. Whole Exome Sequencing to Identify Genetic Variants Associated with Raised Atherosclerotic Lesions in Young Persons. *Scientific Reports*. [Online] Springer US; 2017;7(1): 1–

9. Available from: doi:10.1038/s41598-017-04433-x

173. Petersen BS, Fredrich B, Hoeppner MP, Ellinghaus D, Franke A. Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genetics*. [Online] BMC Genetics; 2017;18(1): 1–13. Available from: doi:10.1186/s12863-017-0479-5

174. Lionakis MS, Netea MG, Holland SM. Mendelian Genetics of Human Susceptibility to Fungal Infection. *Cold Spring Harbor Perspectives in Medicine*. [Online] 2014;4(6): a019638–a019638. Available from: doi:10.1101/cshperspect.a019638

175. Maskarinec SA, Johnson MD, Perfect JR. Genetic Susceptibility to Fungal Infections: What is in the Genes? *Current Clinical Microbiology Reports*. [Online] 2016;3(2): 81–91. Available from: doi:10.1007/s40588-016-0037-3

176. Arnadottir GA, Norddahl GL, Gudmundsdottir S, Agustsdottir AB, Sigurdsson S, Jensson BO, et al. A homozygous loss-of-function mutation leading to CYBC1 deficiency causes chronic granulomatous disease. *Nature Communications*. [Online] Nature Publishing Group; 2018;9(1). Available from: doi:10.1038/s41467-018-06964-x [Accessed: 19th January 2023]

177. Wang Z, Gerstein M, Snyder M. *RNA-Seq: A revolutionary tool for transcriptomics*. [Online] Nature Reviews Genetics. Nature Publishing Group; 2009. p. 57–63. Available from: doi:10.1038/nrg2484 [Accessed: 21st January 2023]

178. Yang WE, Woods CW, Tsalik EL. Host-based diagnostics for detection and prognosis of infectious diseases. In: Sails A, Tang Y-WBT-M in M (eds.) *Methods in Microbiology*. [Online] Academic Press; 2015. p. 465–500. Available from: doi:10.1016/bs.mim.2015.06.001

179. Westermann AJ, Gorski SA, Vogel J. *Dual RNA-seq of pathogen and host*. [Online] Nature Reviews Microbiology. Nature Publishing Group; 2012. p. 618–630. Available from: doi:10.1038/nrmicro2852 [Accessed: 19th January 2023]

180. Montoya DJ, Andrade P, Silva BJA, Teles RMB, Ma F, Bryson B, et al. Dual RNA-Seq of Human Leprosy Lesions Identifies Bacterial Determinants Linked to Host Immune Response. *Cell reports*. [Online] United States; 2019;26(13): 3574-3585.e3. Available from: doi:10.1016/j.celrep.2019.02.109

181. Khan MM, Ernst O, Manes NP, Oyler BL, Fraser IDC, Goodlett DR, et al. *Multi-omics strategies uncover host-pathogen interactions*. [Online] ACS Infectious Diseases. American Chemical Society; 2019. p. 493–505. Available from: doi:10.1021/acsinfecdis.9b00080 [Accessed: 22nd January 2023]

182. Eisfeld AJ, Halfmann PJ, Wendler JP, Kyle JE, Burnum-Johnson KE, Peralta Z, et al. Multi-platform 'Omics Analysis of Human Ebola Virus Disease Pathogenesis. *Cell host & microbe*. [Online] Cell Host Microbe; 2017;22(6): 817-829.e8. Available from: doi:10.1016/J.CHOM.2017.10.011 [Accessed: 22nd January 2023]

183. Mboowa G, Sserwadda I, Amujal M, Namatovu N. *Human Genomic Loci Important in Common Infectious Diseases: Role of High-Throughput Sequencing and Genome-Wide Association Studies*. [Online] Canadian Journal of Infectious Diseases and Medical Microbiology. Hindawi Limited; 2018. Available from: doi:10.1155/2018/1875217

184. Prieto-Granada CN, Lobo AZC, Mihm MC. Skin Infections. In: Kradin RLBT-DP of ID (Second E (ed.) *Diagnostic Pathology of Infectious Disease*. [Online] Elsevier; 2018. p. 542–647. Available from: doi:10.1016/B978-0-323-44585-6.00020-5

185. Krzyściak PM, Pindycka-Piaszczyńska M, Piaszczyński M. Chromoblastomycosis. *Postepy dermatologii i alergologii*. [Online] 2014;31(5): 310–321. Available from: doi:10.5114/pdia.2014.40949 [Accessed: 6th November 2019]

186. Turiansky GW, Benson PM, Sperling LC, Sau P, Salkin IF, McGinnis MR, et al. Phialophora verrucosa: A new cause of mycetoma. *Journal of the American Academy of Dermatology*. [Online] 1995;32(2 PART 2): 311–315. Available from: doi:10.1016/0190-9622(95)90393-3

187. Lionakis MS, Levitz SM. Host Control of Fungal Infections: Lessons from Basic Studies and Human Cohorts. *Annual Review of Immunology*. [Online] 2018;36(1). Available from: doi:10.1146/annurev-immunol-042617-053318

188. Gross O, Gewies A, Finger K, Schäfer M, Sparwasser T, Peschel C, et al. Card9 controls a non-TLR signalling pathway for innate anti-fungal immunity. *Nature*. [Online] 2006;442(7103): 651–656. Available from: doi:10.1038/nature04926

189. Glocker E-OO, Hennigs A, Nabavi M, Schäffer AA, Woellner C, Salzer U, et al. A homozygous CARD9 mutation in a family with susceptibility to fungal infections. *New England Journal of Medicine*. [Online] Massachussetts Medical Society; 2009;361(18): 1727–1735. Available from: doi:10.1056/NEJMoa0810719

190. Drewniak A, Gazendam RP, Tool ATJ, van Houdt M, Jansen MH, van Hamme JL, et al. Invasive fungal infection and impaired neutrophil killing in human CARD9 deficiency. *Blood*. [Online] 2013;121(13): 2385–2392. Available from: doi:10.1182/blood-2012-08-450551

191. Queiroz-Telles F, Mercier T, Maertens J, Sola CBS, Bonfim C, Lortholary O, et al. Successful Allogenic Stem Cell Transplantation in Patients with Inherited CARD9 Deficiency. *Journal of Clinical Immunology*. [Online] Springer New York LLC; 2019;39(5): 462–469. Available from: doi:10.1007/s10875-019-00662-z

192. Puel A, Cypowyj S, Bustamante J, Wright JF, Liu L, Lim HK, et al. Chronic mucocutaneous candidiasis in humans with inborn errors of interleukin-17 immunity. *Science*. [Online] 2011;332(6025): 65–68. Available from: doi:10.1126/science.1200439 [Accessed: 31st March 2018]

193. van de Veerdonk FL, Plantinga TS, Hoischen A, Smeekens SP, Joosten LAB, Gilissen C, et al. STAT1 Mutations in Autosomal Dominant Chronic Mucocutaneous Candidiasis. *New England Journal of Medicine*. [Online] 2011;365(1): 54–61. Available from: doi:10.1056/NEJMoa1100102 [Accessed: 31st March 2018]

194. Anderson MS, Venanzi ES, Klein L, Chen Z, Berzins SP, Turley SJ, et al.

Projection of an immunological self shadow within the thymus by the aire protein. *Science*. [Online] 2002;298(5597): 1395–1401. Available from: doi:10.1126/science.1075958 [Accessed: 1st April 2018]

195. Puel A, Döffinger R, Natividad A, Chrabieh M, Barcenas-Morales G, Picard C, et al. Autoantibodies against IL-17A, IL-17F, and IL-22 in patients with chronic mucocutaneous candidiasis and autoimmune polyendocrine syndrome type I. *The Journal of Experimental Medicine*. [Online] 2010;207(2): 291–297. Available from: doi:10.1084/jem.20091983 [Accessed: 1st April 2018]

196. Maziarz EK, Perfect JR. *Fungal Infections, An Issue of Infectious Disease Clinics of North America*. [Online] Infectious Disease Clinics of North America. 2016. 179–206 p. Available from: doi:10.1016/j.idc.2015.10.006 [Accessed: 5th April 2018]

197. Kumar V, Cheng S-C, Johnson MD, Smeekens SP, Wojtowicz A, Giamarellos-Bourboulis E, et al. Immunochip SNP array identifies novel genetic variants conferring susceptibility to candidaemia. *Nature Communications*. [Online] Nature Publishing Group; 2014;5: 4675. Available from: doi:10.1038/ncomms5675 [Accessed: 5th April 2018]

198. Cunha C, Aversa F, Romani L, Carvalho A. Human Genetic Susceptibility to Invasive Aspergillosis. *PLoS Pathogens*. [Online] 2013;9(8): e1003434. Available from: doi:10.1371/journal.ppat.1003434 [Accessed: 5th April 2018]

199. Latgé JP. The pathobiology of Aspergillus fumigatus. *Trends in Microbiology*. [Online] 2001;9(8): 382–389. Available from: doi:10.1016/S0966-842X(01)02104-7

200. Kesh S, Mensah NY, Peterlongo P, Jaffe D, Hsu K, VAN DEN Brink M, et al. TLR1 and TLR6 polymorphisms are associated with susceptibility to invasive aspergillosis after allogeneic stem cell transplantation. *Annals of the New York Academy of Sciences*. [Online] 2005;1062: 95–103. Available from: doi:10.1196/annals.1358.012

201. Carvalho A, Pasqualotto AC, Pitzurra L, Romani L, Denning DW, Rodrigues F. Polymorphisms in Toll-Like Receptor Genes and Susceptibility to Pulmonary Aspergillosis. *The Journal of Infectious Diseases*. [Online] 2008;197(4): 618–621. Available from: doi:10.1086/526500 [Accessed: 18th April 2018]

202. Carvalho A, De Luca A, Bozza S, Cunha C, D'Angelo C, Moretti S, et al. TLR3 essentially promotes protective class I-restricted memory CD8 T-cell responses to Aspergillus fumigatus in hematopoietic transplanted patients. *Blood*. [Online] American Society of Hematology; 2012;119(4): 967–977. Available from: doi:10.1182/blood-2011-06-362582 [Accessed: 18th April 2018]

203. Loetscher M, Gerber B, Loetscher P, Jones S a, Piali L, Clark-Lewis I, et al. Chemokine receptor specific for IP10 and mig: structure, function, and expression in activated T-lymphocytes. *The Journal of experimental medicine*. [Online] 1996;184(3): 963–969. Available from: doi:10.1084/jem.184.3.963

204. Mezger M, Steffens M, Beyer M, Manger C, Eberle J, Toliat MR, et al. Polymorphisms in the chemokine (C-X-C motif) ligand 10 are associated with invasive aspergillosis after allogeneic stem-cell transplantation and influence CXCL10 epression in monocyte-derived dendritic cells. *Blood*. [Online] 2008;111(2): 534–536. Available from: doi:10.1182/blood-2007-05-090928

205. Lambourne J, Agranoff D, Herbrecht R, Troke PF, Buchbinder A, Willis F, et al. Association of Mannose-Binding Lectin Deficiency with Acute Invasive Aspergillosis in Immunocompromised Patients. *Clinical Infectious Diseases*. [Online] 2009;49(10): 1486–1491. Available from: doi:10.1086/644619

206. Cunha C, Aversa F, Lacerda JF, Busca A, Kurzai O, Grube M, et al. Genetic PTX3 Deficiency and Aspergillosis in Stem-Cell Transplantation. *New England Journal of Medicine*. [Online] 2014;370(5): 421–432. Available from: doi:10.1056/NEJMoa1211161

207. Zaas AK, Liao G, Chien JW, Weinberg C, Shore D, Giles SS, et al. Plasminogen alleles influence susceptibility to invasive aspergillosis. Flint J (ed.) *PLoS Genetics*. [Online] Public Library of Science; 2008;4(6): e1000101. Available from:

doi:10.1371/journal.pgen.1000101 [Accessed: 17th April 2018]

208. Ali RS, Newport MJ, Bakhiet SM, Ibrahim ME, Fahal AH. *Host genetic susceptibility to mycetoma*. [Online] PLoS Neglected Tropical Diseases. Public Library of Science; 2020. p. 1–14. Available from: doi:10.1371/journal.pntd.0008053

209. *CR1 - Complement receptor type 1 precursor - Homo sapiens (Human). Uniprot*. [Online] Available from: https://www.uniprot.org/uniprot/P17927

210. Hull J, Thomson A, Kwiatkowski D. Association of respiratory syncytial virus bronchiolitis with theinterleukin 8 gene region in UK families. *Thorax*. [Online] BMJ Publishing Group; 2000;55(12): 1023. Available from: doi:10.1136/THORAX.55.12.1023 [Accessed: 8th April 2023]

211. Pluskota E, Stenina OI, Krukovets I, Szpak D, Topol EJ, Plow EF. Mechanism and effect of thrombospondin-4 polymorphisms on neutrophil function. *Blood*. [Online] Blood; 2005;106(12): 3970–3978. Available from: doi:10.1182/BLOOD-2005-03-1292 [Accessed: 8th April 2023]

212. Matheson MC, Ellis JA, Raven J, Walters EH, Abramson MJ. Association of IL8, CXCR2 and TNF-α polymorphisms and airway disease. *Journal of Human Genetics 2006 51:3*. [Online] Nature Publishing Group; 2006;51(3): 196–203. Available from: doi:10.1007/s10038-005-0344-7 [Accessed: 8th April 2023]

213. Hammond ME, Lapointe GR, Feucht PH, Hilt S, Gallegos CA, Gordon CA, et al. IL-8 induces neutrophil chemotaxis predominantly via type I IL-8 receptors. *The Journal of Immunology*. [Online] United States; 1995;155(3): 1428–1433. Available from: doi:10.4049/jimmunol.155.3.1428

214. Mercereau-Puijalon O, May J, Mordmüller B, Perkins DJ, Alpers M, Weinberg JB, et al. Nitric Oxide Synthase 2 Lambaréné (G-954C), Increased Nitric Oxide Production, and Protection against Malaria. *The Journal of Infectious Diseases*. [Online] 2002;184(3): 330–336. Available from: doi:10.1086/322037

215. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al.

The genetic structure and history of Africans and African Americans. *Science*. [Online] Science; 2009;324(5930): 1035–1044. Available from: doi:10.1126/science.1172257 [Accessed: 2nd May 2023]

216. Mellars P. *Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model*. [Online] Proceedings of the National Academy of Sciences of the United States of America. Proc Natl Acad Sci U S A; 2006. p. 9381–9386. Available from: doi:10.1073/pnas.0510792103 [Accessed: 3rd May 2023]

217. Underhill PA, Passarino G, Lin AA, Shen P, Mirazón Lahr M, Foley RA, et al. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Annals of Human Genetics*. [Online] John Wiley & Sons, Ltd; 2001;65(1): 43–62. Available from: doi:10.1046/j.1469-1809.2001.6510043.x [Accessed: 4th May 2023]

218. Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, et al. Hierarchical patterns of global human Y-chromosome diversity. *Molecular Biology and Evolution*. [Online] Oxford Academic; 2001;18(7): 1189–1203. Available from: doi:10.1093/oxfordjournals.molbev.a003906 [Accessed: 4th May 2023]

219. Mangalam AK, Taneja V, David CS. HLA Class II Molecules Influence Susceptibility versus Protection in Inflammatory Diseases by Determining the Cytokine Profile. *The Journal of Immunology*. [Online] 2013;190(2): 513–519. Available from: doi:10.4049/jimmunol.1201891

220. Dubaniewicz A, Lewko B, Moszkowska G, Zamorska B, Stepinski J. Molecular subtypes of the HLA-DR antigens in pulmonary tuberculosis. *International Journal of Infectious Diseases*. [Online] 2000;4(3): 129–133. Available from: doi:10.1016/S1201-9712(00)90073-0

221. Carrillo-Meléndez Hilda, Ortega-Hernández Esteban, Julio Granados, Sara Arroyo, Rodrigo Barquera RA. Role of HLA-DR Alleles to Increase Genetic Susceptibility to Onychomycosis in Nail Psoriasis. *Skin Appendage Disord*.

[Online] 2016;2: 22–25. Available from: doi:10.1159/000446444

222. Mhmoud N, Fahal A, Van De Sande WWJ. Association of IL-10 and CCL5 single nucleotide polymorphisms with tuberculosis in the Sudanese population. *Tropical Medicine and International Health*. [Online] 2013;18(9): 1119–1127. Available from: doi:10.1111/tmi.12141

223. Marques RE, Guabiraba R, Russo RC, Teixeira MM. Targeting CCL5 in inflammation. *Expert opinion on therapeutic targets*. [Online] England; 2013;17(12): 1439–1460. Available from: doi:10.1517/14728222.2013.837886

224. Antachopoulos C, Roilides E. *Cytokines and fungal infections*. [Online] British Journal of Haematology. 2005. p. 583–596. Available from: doi:10.1111/j.1365-2141.2005.05498.x

225. Ahmed AOA, Leeuwen W Van, Fahal A, Sande W Van De, Verbrugh H, Belkum A Van. Mycetoma caused by Madurella mycetomatis : a neglected infectious burden. *Review Literature And Arts Of The Americas*. 2004;4(September): 566–574.

226. Verwer PEB, Notenboom CC, Eadie K, Fahal AH, Verbrugh HA, van de Sande WWJ. A Polymorphism in the Chitotriosidase Gene Associated with Risk of Mycetoma Due to Madurella mycetomatis Mycetoma–A Retrospective Study. *PLoS Neglected Tropical Diseases*. [Online] 2015;9(9): 1–11. Available from: doi:10.1371/journal.pntd.0004061

227. Boot RG, Renkema GH, Verhock M, Strijland A, Bliek J, De Meulemeester TMAMO, et al. The human chitotriosidase gene - Nature of inherited enzyme deficiency. *Journal of Biological Chemistry*. [Online] American Society for Biochemistry and Molecular Biology; 1998;273(40): 25680–25685. Available from: doi:10.1074/jbc.273.40.25680

228. Seibold MA, Donnelly S, Solon M, Innes A, Woodruff PG, Boot RG, et al. Chitotriosidase is the primary active chitinase in the human lung and is modulated by genotype and smoking habit. *Journal of Allergy and Clinical Immunology*.

[Online] NIH Public Access; 2008;122(5): 944-950.e3. Available from: doi:10.1016/j.jaci.2008.08.023

229. Ahmed SA bdalla, de Hoog GS, Stevens DA, Fahal AH, van de Sande WWJ. Polymorphisms in catechol-O-methyltransferase and cytochrome p450 subfamily 19 genes predispose towards Madurella mycetomatis-induced mycetoma susceptibility. *Medical mycology*. [Online] 2015;53(3): 295–301. Available from: doi:10.3109/13693781003636680

230. Montojo J, Zuberi K, Rodriguez H, Kazi F, Wright G, Donaldson SL, et al. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics*. [Online] Oxford University Press; 2010;26(22): 2927. Available from: doi:10.1093/BIOINFORMATICS/BTQ562 [Accessed: 17th March 2023]

231. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*. [Online] 2003;13(11): 2498–2504. Available from: doi:10.1101/gr.1239303

232. Vera-Cabrera L, Ortiz-Lopez R, Elizondo-Gonzalez R, Perez-Maya AA, Ocampo-Candiani J. Complete genome sequence of Nocardia brasiliensis HUJEG-1. *Journal of Bacteriology*. [Online] 2012;194(10): 2761–2762. Available from: doi:10.1128/JB.00210-12 [Accessed: 28th May 2018]

233. Kirby R, Sangal V, Tucker NP, Zakrzewska-Czerwińska J, Wierzbicka K, Herron PR, et al. Draft genome sequence of the human pathogen Streptomyces somaliensis, a significant cause of actinomycetoma. *Journal of Bacteriology*. [Online] 2012;194(13): 3544–3545. Available from: doi:10.1128/JB.00534-12 [Accessed: 28th May 2018]

234. Vera-Cabrera L, Ortiz-Lopez R, Elizondo-Gonzalez R, Campos-Rivera MP, Gallardo-Rocha A, Molina-Torres CA, et al. Draft Genome Sequence of Actinomadura madurae LIID-AJ290, Isolated from a Human Mycetoma Case. *Genome Announcements*. [Online] American Society for Microbiology (ASM); 2014;2(2): e00201-14-e00201-14. Available from: doi:10.1128/genomeA.00201-

14 [Accessed: 28th May 2018]

235. Shaw JJ, Spakowicz DJ, Dalal RS, Davis JH, Lehr NA, Dunican BF, et al. Biosynthesis and genomic analysis of medium-chain hydrocarbon production by the endophytic fungal isolate Nigrograna mackinnonii E5202H. *Applied Microbiology and Biotechnology*. [Online] 2015;99(8): 3715–3728. Available from: doi:10.1007/s00253-014-6206-5 [Accessed: 28th May 2018]

236. Smit S, Derks MFL, Bervoets S, Fahal A, van Leeuwen W, van Belkum A, et al. Genome Sequence of *Madurella mycetomatis* mm55, Isolated from a Human Mycetoma Case in Sudan. *Genome Announcements*. [Online] American Society for Microbiology (ASM); 2016;4(3): e00418-16. Available from: doi:10.1128/genomeA.00418-16 [Accessed: 28th May 2018]

237. Belkum A Van, Fahal A, Sande WWJ Van De. Mycetoma Caused by Madurella mycetomatis: A Completely Neglected Medico-social Dilemma. *Hot Topics in Infection and Immunity in Children IX*. [Online] 2013; 179–189. Available from: doi:10.1007/978-1-4614-4726-9

238. Sagoo GS, Little J, Higgins JPT. Systematic Reviews of Genetic Association Studies. *PLoS Medicine*. [Online] Public Library of Science; 2009;6(3): e1000028. Available from: doi:10.1371/journal.pmed.1000028 [Accessed: 2nd April 2019]

239. Gorroochurn P, Hodge SE, Heiman GA, Durner M, Greenberg DA. Non-replication of association studies: 'Pseudo-failures' to replicate? *Genetics in Medicine*. [Online] 2007;9(6): 325–331. Available from: doi:10.1097/GIM.0b013e3180676d79 [Accessed: 1st April 2019]

240. Hassan R, Deribe K, Simpson H, Bremner S, Elhadi O, Alnour M, et al. Individual Risk Factors of Mycetoma Occurrence in Eastern Sennar Locality, Sennar State, Sudan: A Case-Control Study. *Tropical Medicine and Infectious Disease*. [Online] Multidisciplinary Digital Publishing Institute (MDPI); 2022;7(8). Available from: doi:10.3390/tropicalmed7080174 [Accessed: 27th July 2023]

241. Williams-Blangero S, Blangero J. Collection of pedigree data for genetic analysis

in isolate populations. *Human Biology*. [Online] Wayne State University Press; 2006;78(1): 89–101. Available from: doi:10.1353/hub.2006.0023 [Accessed: 28th January 2023]

242. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The human phenotype ontology in 2017. *Nucleic Acids Research*. [Online] 2017;45(D1): D865–D876. Available from: doi:10.1093/nar/gkw1039

243. *Home | GAS Power Calculator*. [Online] Available from: http://csg.sph.umich.edu/abecasis/gas_power_calculator/index.html [Accessed: 28th July 2023]

244. *DNA Genotek - DNA saliva collection - Research - Oragene OG-600*. [Online] Available from: https://www.dnagenotek.com/ROW/products/collection-human/oragene-dna/600-series/OG-600.html

245. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3-new capabilities and interfaces. *Nucleic Acids Research*. [Online] Oxford University Press; 2012;40(15): e115. Available from: doi:10.1093/nar/gks596 [Accessed: 19th February 2023]

246. *SNPCheck*. [Online] Available from: https://genetools.org/SNPCheck/snpcheck.htm [Accessed: 19th February 2023]

247. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. *Genome Research*. [Online] Cold Spring Harbor Laboratory Press; 2002;12(6): 996–1006. Available from: doi:10.1101/gr.229102 [Accessed: 19th February 2023]

248. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC bioinformatics*. [Online] BMC Bioinformatics; 2012;13: 134. Available from: doi:10.1186/1471-2105-13-134 [Accessed: 19th February 2023]

249. *H3Africa array annotations*. [Online] Available from: https://chipinfo.h3abionet.org/

250. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al.

Modernizing reference genome assemblies. *PLoS Biology*. [Online] PLoS Biol; 2011;9(7). Available from: doi:10.1371/journal.pbio.1001091 [Accessed: 27th July 2023]

251. S.A.G.E. *Statistical Analysis for Genetic Epidemiology, Release 6.3*. [Online] Department of Epidemiology & Biostatistics, Case Western University. 2012. Available from: http://darwin.cwru.edu/sage/

252. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics*. [Online] Am J Hum Genet; 1998;62(5): 1198–1211. Available from: doi:10.1086/301844 [Accessed: 6th February 2023]

253. Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM. Mendel: The Swiss army knife of genetic analysis programs. *Bioinformatics*. [Online] Oxford Academic; 2013;29(12): 1568–1570. Available from: doi:10.1093/bioinformatics/btt187 [Accessed: 7th March 2023]

254. Elston RC, Gray-McGuire C. A review of the 'Statistical Analysis for Genetic Epidemiology' (S.A.G.E.) software package. *Human genomics*. [Online] BMC; 2004;1(6): 456–459. Available from: doi:10.1186/1479-7364-1-6-456 [Accessed: 19th August 2023]

255. R Core Team. *R: A Language and Environment for Statistical Computing*. [Online] R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: http://www.r-project.org

256. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. [Online] Nature Publishing Group; 2020;581(7809): 434–443. Available from: doi:10.1038/s41586-020-2308-7 [Accessed: 9th February 2023]

257. Bomba L, Walter K, Soranzo N. *The impact of rare and low-frequency genetic variants in common disease*. [Online] Genome Biology. BioMed Central; 2017. p. 1–17. Available from: doi:10.1186/s13059-017-1212-4 [Accessed: 29th August

2023]

258. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] National Academy of Sciences; 2014;111(4): E455–E464. Available from: doi:10.1073/pnas.1322563111 [Accessed: 29th August 2023]

259. 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*. [Online] NIH Public Access; 2015;17(5): 405–424. Available from: doi:10.1038/gim.2015.30 [Accessed: 9th February 2023]

260. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biology*. [Online] BioMed Central Ltd.; 2016;17(1): 1–14. Available from: doi:10.1186/s13059-016-0974-4 [Accessed: 10th February 2023]

261. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. *Science*. [Online] American Association for the Advancement of Science; 2015;347(6220). Available from: doi:10.1126/science.1260419 [Accessed: 10th February 2023]

262. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. *The Genotype-Tissue Expression (GTEx) project*. [Online] Nature Genetics. Nature Publishing Group; 2013. p. 580–585. Available from: doi:10.1038/ng.2653 [Accessed: 10th February 2023]

263. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*. [Online] Oxford Academic; 2022;50(D1): D687–D692. Available from: doi:10.1093/nar/gkab1028 [Accessed: 10th February 2023]

264. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, et al. VarSome: the human genomic variant search engine. *Bioinformatics*. [Online] Oxford Academic; 2019;35(11): 1978–1980. Available from: doi:10.1093/bioinformatics/bty897 [Accessed: 10th February 2023]

265. Quinodoz M, Royer-Bertrand B, Cisarova K, Di Gioia SA, Superti-Furga A, Rivolta C. DOMINO: Using Machine Learning to Predict Genes Associated with Dominant Disorders. *American Journal of Human Genetics*. [Online] Elsevier; 2017;101(4): 623–629. Available from: doi:10.1016/j.ajhg.2017.09.001 [Accessed: 29th August 2023]

266. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen — The Clinical Genome Resource. *New England Journal of Medicine*. [Online] Massachusetts Medical Society; 2015;372(23): 2235–2242. Available from: doi:10.1056/nejmsr1406261 [Accessed: 10th February 2023]

267. Agapite J, Albou LP, Aleksander S, Argasinska J, Arnaboldi V, Attrill H, et al. Alliance of Genome Resources Portal: Unified model organism research platform. *Nucleic Acids Research*. [Online] Oxford Academic; 2020;48(D1): D650–D658. Available from: doi:10.1093/nar/gkz813 [Accessed: 10th February 2023]

268. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: Leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research*. [Online] Oxford University Press; 2019;47(D1): D1038–D1043. Available from: doi:10.1093/nar/gky1151

269. Desvignes JP, Bartoli M, Delague V, Krahn M, Miltgen M, Béroud C, et al. VarAFT: A variant annotation and filtration system for human next generation sequencing data. *Nucleic Acids Research*. [Online] Oxford University Press; 2018;46(W1): W545–W553. Available from: doi:10.1093/nar/gky471

270. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*. [Online] Oxford University Press; 2016;33(7): 1870–1874. Available from: doi:10.1093/MOLBEV/MSW054

271. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*. [Online] BioMed Central Ltd.; 2015;4(1): 7. Available from: doi:10.1186/s13742-015-0047-8 [Accessed: 11th February 2023]

272. Rodrigues CHM, Pires DEV, Ascher DB. DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Research*. [Online] Oxford University Press; 2018;46(W1): W350–W355. Available from: doi:10.1093/nar/gky300 [Accessed: 16th October 2023]

273. Scardoni G, Tosadori G, Faizan M, Spoto F, Fabbri F, Laudanna C. Biological network analysis with CentiScaPe: centralities and experimental dataset integration. *F1000Research*. [Online] 2014;3(0): 139. Available from: doi:10.12688/f1000research.4477.1

274. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*. [Online] John Wiley & Sons, Ltd; 2018;27(2): e1608. Available from: doi:10.1002/mpr.1608 [Accessed: 11th February 2023]

275. Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A map of human genome variation from population-scale sequencing. *Nature*. [Online] Nature Publishing Group; 2010;467(7319): 1061–1073. Available from: doi:10.1038/nature09534 [Accessed: 12th February 2023]

276. Biscarini F, Cozzi P, Gaspa G, Marras G. *detectRUNS: Detect Runs of Homozygosity and Runs of Heterozygosity in Diploid Genomes. R package version 0.9.6.* [Online] 2019. Available from: https://cran.r-project.org/web/packages/detectRUNS/vignettes/detectRUNS.vignette.html [Accessed: 18th July 2023]

277. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*. [Online] Elsevier; 2011;88(1): 76–82. Available from: doi:10.1016/j.ajhg.2010.11.011 [Accessed:

13th February 2023]

278. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*. [Online] Nat Genet; 2018;50(9): 1335–1341. Available from: doi:10.1038/s41588-018-0184-y [Accessed: 29th June 2023]

279. Boughton AP, Welch RP, Flickinger M, Vandehaar P, Taliun D, Abecasis GR, et al. LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics*. [Online] Oxford Academic; 2021;37(18): 3017–3018. Available from: doi:10.1093/bioinformatics/btab186 [Accessed: 13th February 2023]

280. Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*. [Online] 2017;8(1): 1826. Available from: doi:10.1038/s41467-017-01261-5

281. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Computational Biology*. [Online] PLOS; 2015;11(4): 1004219. Available from: doi:10.1371/journal.pcbi.1004219 [Accessed: 14th July 2023]

282. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. [Online] Nature; 2021;590(7845): 290–299. Available from: doi:10.1038/S41586-021-03205-Y [Accessed: 14th November 2023]

283. Schurz H, Müller SJ, Van Helden PD, Tromp G, Hoal EG, Kinnear CJ, et al. Evaluating the accuracy of imputation methods in a five-way admixed population. *Frontiers in Genetics*. [Online] Frontiers Media S.A.; 2019;10(feburay): 427294. Available from: doi:10.3389/fgene.2019.00034

284. Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, et al. A continuum of admixture in the Western Hemisphere revealed by the African

Diaspora genome. *Nature Communications*. [Online] Nature Publishing Group; 2016;7. Available from: doi:10.1038/NCOMMS12522 [Accessed: 28th August 2023]

285. *GitHub - TorkamaniLab/imputation_accuracy_calculator*. [Online] Available from: https://github.com/TorkamaniLab/imputation_accuracy_calculator [Accessed: 18th October 2023]

286. UNAIDS. *HIV Rates by Country 2023*. [Online] Worldpopulationreview.Com. p. 1–7. Available from: https://worldpopulationreview.com/country-rankings/hiv-rates-by-country [Accessed: 18th October 2023]

287. Mayhew AJ, Meyre D. Assessing the Heritability of Complex Traits in Humans: Methodological Challenges and Opportunities. *Current Genomics*. [Online] 2017;18(4): 332–340. Available from: doi:10.2174/1389202918666170307161450

288. Matthews AG, Finkelstein DM, Betensky RA. Analysis of familial aggregation studies with complex ascertainment schemes. *Statistics in medicine*. [Online] NIH Public Access; 2008;27(24): 5076. Available from: doi:10.1002/SIM.3327 [Accessed: 24th September 2023]

289. Selsted ME, Harwig SSL, Ganz T, Schilling JW, Lehrer RI. Primary structures of three human neutrophil defensins. *Journal of Clinical Investigation*. [Online] American Society for Clinical Investigation; 1985;76(4): 1436–1439. Available from: doi:10.1172/JCI112121

290. Faurschou M, Sørensen OE, Johnsen AH, Askaa J, Borregaard N. Defensin-rich granules of human neutrophils: Characterization of secretory properties. *Biochimica et Biophysica Acta - Molecular Cell Research*. [Online] Elsevier; 2002;1591(1–3): 29–35. Available from: doi:10.1016/S0167-4889(02)00243-4

291. Xu D, Lu W. *Defensins: A Double-Edged Sword in Host Immunity*. [Online] Frontiers in Immunology. Frontiers Media S.A.; 2020. p. 538653. Available from: doi:10.3389/fimmu.2020.00764

292. Kagan BL, Selsted ME, Ganz T, Lehrer RI. Antimicrobial defensin peptides form

voltage-dependent ion-permeable channels in planar lipid bilayer membranes. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] National Academy of Sciences; 1990;87(1): 210–214. Available from: doi:10.1073/pnas.87.1.210 [Accessed: 29th September 2023]

293. Wimley WC, Selsted ME, White SH. Interactions between human defensins and lipid bilayers: Evidence for formation of multimeric pores. *Protein Science*. [Online] Wiley-Blackwell; 1994;3(9): 1362–1373. Available from: doi:10.1002/pro.5560030902 [Accessed: 29th September 2023]

294. Aldred PMR, Hollox EJ, Armour JAL. Copy number polymorphism and expression level variation of the human α-defensin genes DEFA1 and DEFA3. *Human Molecular Genetics*. [Online] Oxford Academic; 2005;14(14): 2045–2052. Available from: doi:10.1093/hmg/ddi209 [Accessed: 29th September 2023]

295. Chen QX, Yang Y, Hou JC, Shu Q, Yin YX, Fu WT, et al. Increased gene copy number of DEFA1/DEFA3 worsens sepsis by inducing endothelial pyroptosis. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] National Academy of Sciences; 2019;116(8): 3161–3170. Available from: doi:10.1073/pnas.1812947116 [Accessed: 28th September 2023]

296. Chen Q, Hakimi M, Wu S, Jin Y, Cheng B, Wang H, et al. Increased genomic copy number of DEFA1/DEFA3 is associated with susceptibility to severe sepsis in Chinese han population. *Anesthesiology*. [Online] 2010;112(6): 1428–1434. Available from: doi:10.1097/ALN.0b013e3181d968eb

297. Bear CE, Li C, Kartner N, Bridges RJ, Jensen TJ, Ramjeesingh M, et al. Purification and functional reconstitution of the cystic fibrosis transmembrane conductance regulator (CFTR). *Cell*. [Online] Elsevier; 1992;68(4): 809–818. Available from: doi:10.1016/0092-8674(92)90155-6 [Accessed: 14th May 2023]

298. GTEx Project. *GTEx Portal*. [Online] Available from: https://gtexportal.org/home/gene/CFTR [Accessed: 29th September 2023]

299. Zielenski J, Tsui LC. *Cystic fibrosis: Genotypic and phenotypic variations*. [Online]

Annual Review of Genetics. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA; 1995. p. 777–807. Available from: doi:10.1146/annurev.ge.29.120195.004021 [Accessed: 14th May 2023]

300. *7-117267592-G-A | gnomAD v2.1.1 | gnomAD*. [Online] Available from: https://gnomad.broadinstitute.org/variant/7-117267592-G-A?dataset=gnomad_r2_1 [Accessed: 8th October 2023]

301. Tummers B, Green DR. *Caspase-8: regulating life and death*. [Online] Immunological Reviews. Blackwell Publishing Ltd; 2017. p. 76–89. Available from: doi:10.1111/imr.12541

302. Madkaikar M, Mhatre S, Gupta M, Ghosh K. *Advances in autoimmune lymphoproliferative syndromes*. [Online] European Journal of Haematology. John Wiley & Sons, Ltd; 2011. p. 1–9. Available from: doi:10.1111/j.1600-0609.2011.01617.x [Accessed: 27th May 2023]

303. Chun HJ, Zheng L, Ahmad M, Wang J, Speirs CK, Siegel RM, et al. Pleiotropic defects in lymphocyte activation caused by caspase-8 mutations lead to human immunodeficiency. *Nature 2002 419:6905*. [Online] Nature Publishing Group; 2002;419(6905): 395–399. Available from: doi:10.1038/nature01063 [Accessed: 27th May 2023]

304. Blay N, Casas E, Galvan-Femenia I, Graffelman J, De Cid R, Vavouri T. Assessment of kinship detection using RNA-seq data. *Nucleic Acids Research*. [Online] Oxford University Press; 2019;47(21): E136. Available from: doi:10.1093/nar/gkz776 [Accessed: 5th October 2023]

305. *DEL_12_126655 | gnomAD SVs v2.1 | gnomAD*. [Online] Available from: https://gnomad.broadinstitute.org/variant/DEL_12_126655?dataset=gnomad_sv_r2_1 [Accessed: 6th October 2023]

306. Lebailly B, He C, Rogner UC. Linking the circadian rhythm gene Arntl2 to interleukin 21 expression in type 1 diabetes. *Diabetes*. [Online] Diabetes; 2014;63(6): 2148–2157. Available from: doi:10.2337/db13-1702 [Accessed: 6th

October 2023]

307. Wen M, Ma X, Cheng H, Jiang W, Xu X, Zhang Y, et al. Stk38 protein kinase preferentially inhibits TLR9-activated inflammatory responses by promoting MEKK2 ubiquitination in macrophages. *Nature Communications*. [Online] Nature Publishing Group; 2015;6(1): 1–11. Available from: doi:10.1038/ncomms8167 [Accessed: 6th October 2023]

308. Ramirez-Ortiz ZG, Specht CA, Wang JP, Lee CK, Bartholomeu DC, Gazzinelli RT, et al. Toll-like receptor 9-dependent immune activation by unmethylated CpG motifs in Aspergillus fumigatus DNA. *Infection and Immunity*. [Online] American Society for Microbiology (ASM); 2008;76(5): 2123–2129. Available from: doi:10.1128/IAI.00047-08 [Accessed: 6th October 2023]

309. Project Gte. *GTEx Portal*. [Online] Available from: https://gtexportal.org/home/gene/PLEKHO2 [Accessed: 30th September 2023]

310. Zhang P, Zhou C, Lu C, Li W, Li W, Jing B, et al. PLEKHO2 is essential for M-CSF-dependent macrophage survival. *Cellular Signalling*. [Online] Pergamon; 2017;37: 115–122. Available from: doi:10.1016/j.cellsig.2017.06.006

311. *15-65152192-G-A | gnomAD v2.1.1 | gnomAD*. [Online] Available from: https://gnomad.broadinstitute.org/variant/15-65152192-G-A?dataset=gnomad_r2_1

312. Ohtsuka T, Hata Y, Ide N, Yasuda T, Inoue E, Inoue T, et al. nRap GEP: A novel neural GDP/GTP exchange protein for Rap1 small G protein that interacts with synaptic scaffolding molecule (S-SCAM). *Biochemical and Biophysical Research Communications*. [Online] Academic Press Inc.; 1999;265(1): 38–44. Available from: doi:10.1006/bbrc.1999.1619

313. De Rooij J, Boenink NM, Van Triest M, Cool RH, Wittinghofer A, Bos JL. PDZ-GEF1, a guanine nucleotide exchange factor specific for Rap1 and Rap2. *Journal of Biological Chemistry*. [Online] 1999;274(53): 38125–38130. Available from: doi:10.1074/jbc.274.53.38125

314. Liao Y, Kariya KI, Hu CD, Shibatohge M, Goshima M, Okada T, et al. RA-GEF, a novel Rap1A guanine nucleotide exchange factor containing a Ras/Rap1A-associating domain, is conserved between nematode and humans. *Journal of Biological Chemistry*. [Online] 1999;274(53): 37815–37820. Available from: doi:10.1074/jbc.274.53.37815

315. Pham N, Cheglakov I, Koch CA, De Hoog CL, Moran MF, Rotin D. The guanine nucleotide exchange factor CNrasGEF activates Ras in response to cAMP and cGMP. *Current Biology*. [Online] Current Biology Ltd; 2000;10(9): 555–558. Available from: doi:10.1016/S0960-9822(00)00473-5

316. Liao Y, Satoh T, Gao X, Jin TG, Hu CD, Kataoka T. RA-GEF-1, a Guanine Nucleotide Exchange Factor for Rap1, is Activated by Translocation Induced by Association with Rap1·GTP and Enhances Rap1-dependent B-Raf Activation. *Journal of Biological Chemistry*. [Online] Elsevier; 2001;276(30): 28478–28483. Available from: doi:10.1074/jbc.M101737200

317. *4-160251043-A-G | gnomAD v2.1.1 | gnomAD*. [Online] Available from: https://gnomad.broadinstitute.org/variant/4-160251043-A-G?dataset=gnomad_r2_1 [Accessed: 26th November 2023]

318. Rotimi CN, Bentley AR, Doumatey AP, Chen G, Shriner D, Adeyemo A. *The genomic landscape of African populations in health and disease*. [Online] Human Molecular Genetics. Oxford University Press; 2017. p. R225–R236. Available from: doi:10.1093/hmg/ddx253 [Accessed: 27th September 2023]

319. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. *Nature*. [Online] 2015;526(7571): 68–74. Available from: doi:10.1038/nature15393

320. Martinon F, Burns K, Tschopp J. The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proIL-beta. *Molecular cell*. [Online] Mol Cell; 2002;10(2): 417–426. Available from: doi:10.1016/S1097-2765(02)00599-3 [Accessed: 15th May 2023]

321. Orning P, Lien E. *Multiple roles of caspase-8 in cell death, inflammation, and innate immunity*. [Online] Journal of Leukocyte Biology. John Wiley & Sons, Ltd; 2021. p. 121–141. Available from: doi:10.1002/JLB.3MR0420-305R [Accessed: 15th May 2023]

322. Netea MG, Van De Veerdonk FL, Van Der Meer JWM, Dinarello CA, Joosten LAB. Inflammasome-independent regulation of IL-1-family cytokines. *Annual Review of Immunology*. [Online] Annual Reviews; 2015;33: 49–77. Available from: doi:10.1146/annurev-immunol-032414-112306 [Accessed: 9th October 2023]

323. Ganesan S, Rathinam VAK, Bossaller L, Army K, Kaiser WJ, Mocarski ES, et al. Caspase-8 modulates dectin-1 and complement receptor 3-driven IL-1β production in response to β-glucans and the fungal pathogen, Candida albicans. *Journal of immunology (Baltimore, Md. : 1950)*. [Online] J Immunol; 2014;193(5): 2519–2530. Available from: doi:10.4049/JIMMUNOL.1400276 [Accessed: 15th August 2022]

324. Nasr A, Abushouk A, Hamza A, Siddig E, Fahal AH. Th-1, Th-2 Cytokines Profile among Madurella mycetomatis Eumycetoma Patients. *PLoS Neglected Tropical Diseases*. [Online] 2016;10(7): 1–11. Available from: doi:10.1371/journal.pntd.0004862

325. Sims JE, Smith DE. *The IL-1 family: Regulators of immunity*. [Online] Nature Reviews Immunology. Nat Rev Immunol; 2010. p. 89–102. Available from: doi:10.1038/nri2691 [Accessed: 10th October 2023]

326. Ninomiya-Tsuji J, Kishimoto K, Hiyama A, Inoue JI, Cao Z, Matsumoto K. The kinase TAK1 can activate the NIK-IκB as well as the MAP kinase cascade in the IL-1 signalling pathway. *Nature*. [Online] Nature Publishing Group; 1999;398(6724): 252–256. Available from: doi:10.1038/18465 [Accessed: 10th October 2023]

327. Saperstein S, Chen L, Oakes D, Pryhuber G, Finkelstein J. IL-1 augments TNF - Mediated inflammatory responses from lung epithelial cells. *Journal of Interferon and Cytokine Research*. [Online] J Interferon Cytokine Res; 2009;29(5): 273–284.

Available from: doi:10.1089/jir.2008.0076 [Accessed: 10th October 2023]

328. Lebovic DI, Chao VA, Martini J-F, Taylor RN. IL-1β Induction of RANTES (Regulated upon Activation, Normal T Cell Expressed and Secreted) Chemokine Gene Expression in Endometriotic Stromal Cells Depends on a Nuclear Factor-κB Site in the Proximal Promoter. *The Journal of Clinical Endocrinology & Metabolism*. [Online] J Clin Endocrinol Metab; 2001;86(10): 4759–4764. Available from: doi:10.1210/jcem.86.10.7890 [Accessed: 10th October 2023]

329. Hughes T, Hansson L, Akkouh I, Hajdarevic R, Bringsli JS, Torsvik A, et al. Runaway multi-allelic copy number variation at the α-defensin locus in African and Asian populations. *Scientific Reports*. [Online] Nature Publishing Group; 2020;10(1): 1–8. Available from: doi:10.1038/s41598-020-65675-w [Accessed: 10th October 2023]

330. Fortes-Lima C, Tříska P, Čížková M, Podgorná E, Diallo MY, Schlebusch CM, et al. Demographic and Selection Histories of Populations Across the Sahel/Savannah Belt. *Molecular Biology and Evolution*. [Online] 2022;39(10). Available from: doi:10.1093/molbev/msac209 [Accessed: 19th September 2023]

331. Gurinovich A, Li M, Leshchyk A, Bae H, Song Z, Arbeev KG, et al. Evaluation of GENESIS, SAIGE, REGENIE and fastGWA-GLMM for genome-wide association studies of binary traits in correlated data. *Frontiers in Genetics*. [Online] Frontiers Media S.A.; 2022;13: 897210. Available from: doi:10.3389/fgene.2022.897210

332. Mao H, Yang W, Latour S, Yang J, Winter S, Zheng J, et al. RASGRP1 mutation in autoimmune lymphoproliferative syndrome-like disease. *Journal of Allergy and Clinical Immunology*. [Online] Mosby; 2018;142(2): 595-604.e16. Available from: doi:10.1016/j.jaci.2017.10.026

333. Ksionda O, Limnander A, Roose JP. *RasGRP Ras guanine nucleotide exchange factors in cancer*. [Online] Frontiers in Biology. Higher Education Press Limited Company; 2013. p. 508–532. Available from: doi:10.1007/s11515-013-1276-9 [Accessed: 22nd September 2023]

334. Hassan R, Cano J, Fronterre C, Bakhiet S, Fahal A, Deribe K, et al. Estimating the burden of mycetoma in Sudan for the period 1991–2018 using a model-based geostatistical approach. *PLOS Neglected Tropical Diseases*. [Online] Public Library of Science; 2022;16(10): e0010795. Available from: doi:10.1371/JOURNAL.PNTD.0010795 [Accessed: 18th February 2023]

335. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, et al. Runs of Homozygosity in European Populations. *American Journal of Human Genetics*. [Online] Elsevier; 2008;83(3): 359–372. Available from: doi:10.1016/j.ajhg.2008.08.007 [Accessed: 17th September 2023]

336. Clark DW, Okada Y, Moore KHS, Mason D, Pirastu N, Gandin I, et al. Associations of autozygosity with a broad range of human phenotypes. *Nature Communications*. [Online] Nature Publishing Group; 2019;10(1): 1–17. Available from: doi:10.1038/s41467-019-12283-6 [Accessed: 19th September 2023]

337. Ponting CP, Haerty W. *Genome-Wide Analysis of Human Long Noncoding RNAs: A Provocative Review*. [Online] Annual Review of Genomics and Human Genetics. Annual Reviews; 2022. p. 153–172. Available from: doi:10.1146/annurev-genom-112921-123710 [Accessed: 21st September 2023]

338. Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen LL, et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology*. [Online] Nature Publishing Group; 2023;24(6): 430–447. Available from: doi:10.1038/s41580-022-00566-8 [Accessed: 21st September 2023]

339. Shirahama S, Miki A, Kaburaki T, Akimitsu N. *Long Non-coding RNAs Involved in Pathogenic Infection*. [Online] Frontiers in Genetics. Frontiers Media S.A.; 2020. Available from: doi:10.3389/fgene.2020.00454 [Accessed: 21st September 2023]

340. Kwon GJ, Henao-Mejia J, Williams A. *Editorial: Long Non-coding RNAs and Immunity*. [Online] Frontiers in Immunology. Frontiers Media S.A.; 2019. p. 2378. Available from: doi:10.3389/fimmu.2019.02378 [Accessed: 21st September 2023]

341. Ioannidis JPA. Why most published research findings are false. *Getting to Good: Research Integrity in the Biomedical Sciences*. [Online] Public Library of Science; 2018. p. 2–8. Available from: doi:10.1371/journal.pmed.0020124 [Accessed: 25th July 2023]

342. Mackinnon MJ, Mwangi TW, Snow RW, Marsh K, Williams TN. Heritability of malaria in Africa. *PLoS Medicine*. [Online] Public Library of Science; 2005;2(12): 1253–1259. Available from: doi:10.1371/journal.pmed.0020340 [Accessed: 21st November 2023]

343. Kariuki SN, Williams TN. *Human genetics and malaria resistance*. [Online] Human Genetics. Springer; 2020. p. 801–811. Available from: doi:10.1007/s00439-020-02142-6 [Accessed: 22nd November 2023]

344. Agerberth B, Charo J, Werr J, Olsson B, Idali F, Lindbom L, et al. The human antimicrobial and chemotactic peptides LL-37 and α-defensins are expressed by specific lymphocyte and monocyte populations. *Blood*. [Online] United States; 2000;96(9): 3086–3093. Available from: doi:10.1182/blood.v96.9.3086

345. Raj PA, Dentino AR. Current status of defensins and their role in innate and adaptive immunity. *FEMS Microbiology Letters*. [Online] Oxford Academic; 2002;206(1): 9–18. Available from: doi:10.1111/j.1574-6968.2002.tb10979.x [Accessed: 7th November 2023]

346. Yang D, Biragyn A, Kwak LW, Oppenheim JJ. *Mammalian defensins in immunity: More than just microbicidal*. [Online] Trends in Immunology. Trends Immunol; 2002. p. 291–296. Available from: doi:10.1016/S1471-4906(02)02246-9 [Accessed: 7th November 2023]

347. Soehnlein O, Kai-Larsen Y, Frithiof R, Sorensen OE, Kenne E, Scharffetter-Kochanek K, et al. Neutrophil primary granule proteins HBP and HNP1-3 boost bacterial phagocytosis by human and murine macrophages. *Journal of Clinical Investigation*. [Online] American Society for Clinical Investigation; 2008;118(10): 3491–3502. Available from: doi:10.1172/JCI35740 [Accessed: 9th November 2023]

348. Chaly Y V, Paleolog EM, Kolesnikova TS, Tikhonov II, Petratchenko E V, Voitenok NN. Human neutrophil α-defensin modulates cytokine production in human monocytes and adhesion molecule expression in endothelial cells. *European Cytokine Network*. [Online] 2000;11(2): 257–266. Available from: https://www.jle.com/fr/revues/ecn/e-docs/neutrophil_a_defensin_human_neutrophil_peptide_modulates_cytoline_production_in_human_monocytes_and_adhesion_molecule_expression_in_endothelial_cells._90370/article.phtml?tab=texte [Accessed: 9th November 2023]

349. Van Wetering S, Mannesse-Lazeroms SPG, Van Sterkenburg MAJA, Daha MR, Dijkman JH, Hiemstra PS. Effect of defensins on interleukin-8 synthesis in airway epithelial cells. *American Journal of Physiology - Lung Cellular and Molecular Physiology*. [Online] American Physiological Society Bethesda, MD; 1997;272(5 16-5). Available from: doi:10.1152/ajplung.1997.272.5.l888 [Accessed: 9th November 2023]

350. Aarbiou J, Verhoosel RM, Van Wetering S, De Boer WI, Van Krieken JH, Litvinov S V., et al. Neutrophil Defensins Enhance Lung Epithelial Wound Closure and Mucin Gene Expression in Vitro. *American Journal of Respiratory Cell and Molecular Biology*. [Online] American Thoracic Society; 2004;30(2): 193–201. Available from: doi:10.1165/rcmb.2002-0267OC [Accessed: 7th November 2023]

351. Gursoy UK, Könönen E, Luukkonen N, Uitto V-J. Human Neutrophil Defensins and Their Effect on Epithelial Cells. *Journal of Periodontology*. [Online] John Wiley & Sons, Ltd; 2013;84(1): 126–133. Available from: doi:10.1902/jop.2012.120017 [Accessed: 7th November 2023]

352. Chavakis T, Cines DB, Rhee J-S, Liang OD, Schubert U, Hammes H-P, et al. Regulation of neovascularization by human neutrophil peptides (α-defensins): a link between inflammation and angiogenesis. *The FASEB Journal*. [Online] Wiley; 2004;18(11): 1306–1308. Available from: doi:10.1096/fj.03-1009fje

353. Ordonez SR, Veldhuizen EJA, van Eijk M, Haagsman HP. Role of soluble innate effector molecules in pulmonary defense against fungal pathogens. *Frontiers in*

Microbiology. [Online] Frontiers Media S.A.; 2017;8(OCT): 299957. Available from: doi:10.3389/fmicb.2017.02098

354. Orning P, Lien E. Multiple roles of caspase-8 in cell death, inflammation, and innate immunity. *Journal of Leukocyte Biology*. [Online] John Wiley & Sons, Ltd; 2021;109(1): 121–141. Available from: doi:10.1002/JLB.3MR0420-305R [Accessed: 8th November 2023]

355. Camilli G, Blagojevic M, Naglik JR, Richardson JP. Programmed Cell Death: Central Player in Fungal Infections. *Trends in Cell Biology*. [Online] Elsevier; 2021;31(3): 179–196. Available from: doi:10.1016/j.tcb.2020.11.005 [Accessed: 8th November 2023]

356. Dupaul-Chicoine J, Saleh M. *A new path to IL-1β production controlled by caspase-8*. [Online] Nature Immunology. Nature Publishing Group; 2012. p. 211–212. Available from: doi:10.1038/ni.2241 [Accessed: 9th October 2023]

357. Ganesan S, Rathinam VAK, Bossaller L, Army K, Kaiser WJ, Mocarski ES, et al. Caspase-8 Modulates Dectin-1 and Complement Receptor 3–Driven IL-1β Production in Response to β-Glucans and the Fungal Pathogen, Candida albicans . *The Journal of Immunology*. [Online] American Association of Immunologists; 2014;193(5): 2519–2530. Available from: doi:10.4049/jimmunol.1400276 [Accessed: 15th May 2023]

358. Ketelut-Carneiro N, Ghosh S, Levitz SM, Fitzgerald KA, Da Silva JS. A Dectin-1-Caspase-8 Pathway Licenses Canonical Caspase-1 Inflammasome Activation and Interleukin-1β Release in Response to a Pathogenic Fungus. *Journal of Infectious Diseases*. [Online] Oxford Academic; 2018;217(2): 329–339. Available from: doi:10.1093/infdis/jix568 [Accessed: 8th November 2023]

359. Haslam RJ, Koide HB, Hemmings BA. *Pleckstrin domain homology*. [Online] Nature. Nature Publishing Group; 1993. p. 309–310. Available from: doi:10.1038/363309b0 [Accessed: 10th November 2023]

360. Zhou C, Zhang X, Yang C, He Y, Zhang L. PLEKHO2 inhibits TNFα-induced cell

death by suppressing RIPK1 activation. *Cell Death and Disease*. [Online] Nature Publishing Group; 2021;12(8): 1–8. Available from: doi:10.1038/s41419-021-04001-2 [Accessed: 18th May 2023]

361. Yang D, Biragyn A, Kwak LW, Oppenheim JJ. *Mammalian defensins in immunity: More than just microbicidal*. [Online] Trends in Immunology. Elsevier; 2002. p. 291–296. Available from: doi:10.1016/S1471-4906(02)02246-9 [Accessed: 7th November 2023]

362. Rodríguez-García M, Oliva H, Climent N, Escribese MM, García F, Moran TM, et al. Impact of α-defensins1-3 on the maturation and differentiation of human monocyte-derived DCs. Concentration-dependent opposite dual effects. *Clinical Immunology*. [Online] Academic Press; 2009;131(3): 374–384. Available from: doi:10.1016/j.clim.2009.01.012

363. Gao X, Ding J, Liao C, Xu J, Liu X, Lu W. *Defensins: The natural peptide antibiotic*. [Online] Advanced Drug Delivery Reviews. Elsevier; 2021. p. 114008. Available from: doi:10.1016/j.addr.2021.114008

364. Tang J, Lin G, Langdon WY, Tao L, Zhang J. Regulation of C-type lectin receptor-mediated antifungal immunity. *Frontiers in Immunology*. [Online] Frontiers Media S.A.; 2018;9(FEB): 326162. Available from: doi:10.3389/fimmu.2018.00123

365. Hardison SE, Brown GD. C-type lectin receptors orchestrate antifungal immunity. *Nature Immunology*. [Online] Nature Publishing Group; 2012;13(9): 817–822. Available from: doi:10.1038/ni.2369 [Accessed: 9th November 2023]

366. Camilli G, Tabouret G, Quintin J. The Complexity of Fungal β-Glucan in Health and Disease: Effects on the Mononuclear Phagocyte System. *Frontiers in Immunology*. [Online] Frontiers Media S.A.; 2018;9(APR): 343771. Available from: doi:10.3389/fimmu.2018.00673

367. Gurung P, Kanneganti TD. *Novel Roles for Caspase-8 in IL-1β and Inflammasome Regulation*. [Online] American Journal of Pathology. Elsevier Inc.; 2015. p. 17–25. Available from: doi:10.1016/j.ajpath.2014.08.025 [Accessed: 9th

November 2023]

368. Michalczyk I, Sikorski AF, Kotula L, Junghans RP, Dubielecka PM. The emerging role of protein kinase Cθ in cytoskeletal signaling. *Journal of Leukocyte Biology*. [Online] Oxford Academic; 2012;93(3): 319–327. Available from: doi:10.1189/jlb.0812371 [Accessed: 9th November 2023]

369. Letschka T, Kollmann V, Pfeifhofer-Obermair C, Lutz-Nicoladoni C, Obermair GJ, Fresser F, et al. PKC-{theta} selectively controls the adhesion-stimulating molecule Rap1. *Blood*. [Online] Blood; 2008;112(12): 4617–4627. Available from: doi:10.1182/blood-2007-11-121111 [Accessed: 9th November 2023]

370. Navarro MN, Cantrell DA. Serine-threonine kinases in TCR signaling. *Nature Immunology*. [Online] Europe PMC Funders; 2014;15(9): 808–814. Available from: doi:10.1038/ni.2941 [Accessed: 9th November 2023]

371. Huse M. The T-cell-receptor signalling network. *Journal of Cell Science*. [Online] The Company of Biologists; 2009;122(9): 1269–1273. Available from: doi:10.1242/jcs.042762 [Accessed: 9th November 2023]

372. Ibrahim SA, Fadl Elmola MA, Karrar ZA, Arabi AME, Abdullah MA, Ali SK, et al. Cystic fibrosis in Sudanese children: First report of 35 cases. *Sudanese journal of paediatrics*. [Online] Sudan Association of Paediatricians; 2014;14(1): 39–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27493388%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4949914 [Accessed: 31st October 2023]

373. Al-Sadeq D, Abunada T, Dalloul R, Fahad S, Taleb S, Aljassim K, et al. Spectrum of mutations of cystic fibrosis in the 22 Arab countries: A systematic review. *Respirology*. [Online] John Wiley & Sons, Ltd; 2019;24(2): 127–136. Available from: doi:10.1111/resp.13437 [Accessed: 31st October 2023]

374. Abubakar Bobbo K, Ahmad U, Chau DM, Nordin N, Abdullah S. A comprehensive review of cystic fibrosis in Africa and Asia. *Saudi Journal of Biological Sciences*. [Online] Elsevier; 2023;30(7): 103685. Available from:

doi:10.1016/j.sjbs.2023.103685

375. Stewart C, Pepper MS. Cystic fibrosis on the African continent. *Genetics in Medicine*. [Online] Nature Publishing Group; 2016;18(7): 653–662. Available from: doi:10.1038/gim.2015.157 [Accessed: 31st October 2023]

376. Ning Z, Liu K, Xiong H. *Roles of BTLA in Immunity and Immune Disorders*. [Online] Frontiers in Immunology. 2021. Available from: doi:10.3389/fimmu.2021.654960

377. Zeng JC, Lin DZ, Yi LL, Liu G Bin, Zhang H, Wang WD, et al. BTLA exhibits immune memory for αβ T cells in patients with active pulmonary tuberculosis. *American Journal of Translational Research*. United States; 2014;6(5): 494–506.

378. Kotwica-Mojzych K, Jodłowska-Jędrych B, Mojzych M. *Cd200:Cd200r interactions and their importance in immunoregulation*. [Online] International Journal of Molecular Sciences. Multidisciplinary Digital Publishing Institute (MDPI); 2021. p. 1–21. Available from: doi:10.3390/ijms22041602 [Accessed: 13th November 2023]

379. Mukhopadhyay S, Plüddemann A, Hoe JC, Williams KJ, Varin A, Makepeace K, et al. Immune Inhibitory Ligand CD200 Induction by TLRs and NLRs limits macrophage activation to protect the host from meningococcal septicemia. *Cell Host and Microbe*. [Online] 2010;8(3): 236–247. Available from: doi:10.1016/j.chom.2010.08.005

380. Szanto A, Balint BL, Nagy ZS, Barta E, Dezso B, Pap A, et al. STAT6 transcription factor is a facilitator of the nuclear receptor PPARγ-regulated gene expression in macrophages and dendritic cells. *Immunity*. [Online] Immunity; 2010;33(5): 699–712. Available from: doi:10.1016/j.immuni.2010.11.009 [Accessed: 13th November 2023]

381. Goenka S, Kaplan MH. *Transcriptional regulation by STAT6*. [Online] Immunologic Research. NIH Public Access; 2011. p. 87–96. Available from: doi:10.1007/s12026-011-8205-2 [Accessed: 13th November 2023]

382. Iwaszko M, Biały S, Bogunia-Kubik K. *Significance of interleukin (IL)-4 and IL-13 in inflammatory arthritis*. [Online] Cells. Multidisciplinary Digital Publishing Institute (MDPI); 2021. Available from: doi:10.3390/cells10113000 [Accessed: 13th November 2023]

383. Sengupta D, Botha G, Meintjes A, Mbiyavanga M, Hazelhurst S, Mulder N, et al. Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations. *Cell Genomics*. [Online] Elsevier; 2023;3(6): 100332. Available from: doi:10.1016/j.xgen.2023.100332

Appendix 1 - Published outputs.

These include:

i. A review article on host genetic susceptibility to mycetoma.
ii. An issue brief presented at the NIHR NTD Phase One Annual Unit Meeting in Kigali in 2022.

PLOS NEGLECTED TROPICAL DISEASES

REVIEW

Host genetic susceptibility to mycetoma

Rayan S. Ali [1,2]*, Melanie J. Newport[2], Sahar Mubarak Bakhiet[1,3], Muntaser E. Ibrahim[3], Ahmed Hassan Fahal[1]

1 The Mycetoma Research Centre, University of Khartoum, Khartoum, Sudan, **2** Brighton and Sussex Centre for Global Health Research, Brighton and Sussex Medical School, Brighton, United Kingdom, **3** Institute of Endemic Diseases, University of Khartoum, Khartoum, Sudan

* r.ali@bsms.ac.uk

OPEN ACCESS

Citation: Ali RS, Newport MJ, Bakhiet SM, Ibrahim ME, Fahal AH (2020) Host genetic susceptibility to mycetoma. PLoS Negl Trop Dis 14(4): e0008053. https://doi.org/10.1371/journal.pntd.0008053

Editor: Abdallah M. Samy, Faculty of Science, Ain Shams University (ASU), EGYPT

Published: April 30, 2020

Funding: This article was produced as part of the NIHR Global Health Research Unit for NTDs project funded by the National Institute for Health Research, https://www.nihr.ac.uk/ (grant number 16/136/29). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Mycetoma is one of the badly neglected tropical diseases, characterised by subcutaneous painless swelling, multiple sinuses, and discharge containing aggregates of the infecting organism known as grains. Risk factors conferring susceptibility to mycetoma include environmental factors and pathogen factors such as virulence and the infecting dose, in addition to host factors such as immunological and genetic predisposition. Epidemiological evidence suggests that host genetic factors may regulate susceptibility to mycetoma and other fungal infections, but they are likely to be complex genetic traits in which multiple genes interact with each other and environmental factors, as well as the pathogen, to cause disease. This paper reviews what is known about genetic predisposition to fungal infections that might be relevant to mycetoma, as well as all studies carried out to explore host genetic susceptibility to mycetoma. Most studies were investigating polymorphisms in candidate genes related to the host immune response. A total of 13 genes had allelic variants found to be associated with mycetoma, and these genes lie in different pathways and systems such as innate and adaptive immune systems, sex hormone biosynthesis, and some genes coding for host enzymes. None of these studies have been replicated. Advances in genomic science and the supporting technology have paved the way for large-scale genome-wide association and next generation sequencing (NGS) studies, underpinning a new strategy to systematically interrogate the genome for variants associated with mycetoma. Dissecting the contribution of host genetic variation to susceptibility to mycetoma will enable the identification of pathways that are potential targets for new treatments for mycetoma and will also enhance the ability to stratify 'at-risk' individuals, allowing the possibility of developing preventive and personalised clinical care strategies in the future.

Background

Mycetoma is a badly neglected tropical disease. It is a chronic granulomatous infectious disease characterised by painless subcutaneous swelling associated with multiple sinuses and discharge that contain aggregates of the infecting organism known as grains [1,2]. The disease is classified according to its causative organisms into actinomycetoma, which is caused by actinomycetes bacteria, and eumycetoma, which is caused by fungi [3]. The suspected route of infection is through traumatic inoculation of environmental microorganisms into the

subcutaneous tissue [4,5]. Other mechanisms of transmission (for example, inoculation via insect bites) have not been excluded. Mycetoma has a worldwide distribution, but it is endemic in tropical and subtropical regions in what is known as the 'mycetoma belt' between the latitudes of 15˚S and 30˚N. This belt includes Sudan, Somalia, Senegal, Yemen, India, Mexico, Venezuela, Columbia, and Argentina [6,7]. Sudan seems to be the most highly endemic country for mycetoma worldwide [8]. Mycetoma is seen more frequently amongst impoverished communities in remote rural areas [9]. The majority of mycetoma patients are of low socioeconomic status with little health education. Hence, most of the patients present late with advanced disease, massive deformity, disability, and high morbidity [10,11].

Mycetoma was only recently recognised as a neglected tropical disease in May 2016 by the World Health Organization (WHO) [8,12], and much of the basic information on mycetoma is lacking, including the true incidence, prevalence, and burden of disease; the route of infection; and which risk factors predispose individuals to disease susceptibility. Risk factors that could predispose individuals to mycetoma include environmental factors such as climatic conditions and pathogen factors such as virulence and the infecting dose, in addition to host factors such as immunological status, genetic predisposition, nutritional status, immunosuppression from HIV, coinfections, and use of drugs such as antibiotics or steroids.

The host immune response towards mycetoma-causative organisms has been studied on a limited scale, targeting only a few of the approximately 70 different causative microorganisms of mycetoma. Innate immune responses are a prominent factor in mycetoma because the role of neutrophils in the early defence against mycetoma was demonstrated in previous studies that reported the presence of large numbers of neutrophils in the mycetoma lesion [13,14]. There are 3 host tissue reactions to mycetoma, which include neutrophil adherence and degranulation, resulting in grain disintegration; replacement of neutrophils with macrophages to engulf grain and neutrophil debris; and formation of epithelioid granuloma [14]. Cell-mediated immunity is also required for immunity in mycetoma, with T lymphocytes playing a central role. T helper (Th) type 1 lymphocyte responses provide protective immunity against mycetoma, whilst progression of the disease is linked to Th2 immune response, as previously demonstrated by the significantly higher levels of Th2 cytokines (interleukin [IL]-4, IL-5, IL-6, and IL-10) in mycetoma patients [15–17]. A recent study also pointed a possible role of IL-35 and IL-37 in the pathogenesis of *Madurella mycetomatis*-induced eumycetoma [18]. In 1977, Mahgoub and colleagues suggested that genetic defects in underlying cell-mediated immunity [19] may play a role in the pathogenesis of eumycetoma [16]. Coinfection with schistosomiasis was found to be associated with susceptibility to mycetoma in a small case–control study of 84 subjects in Sudan [16]. The authors hypothesised that the prolonged Th2 response, induced by schistosomiasis, predisposed individuals to develop mycetoma [16]. The humoral arm of the acquired immunity against mycetoma was evaluated by Wethered and his colleagues using an enzyme-linked immunosorbent assay (ELISA) to measure immunoglobulin levels in serum collected from cases and controls [20]. High levels of immunoglobulin M (IgM) were seen in most patients with mycetoma due to *M. mycetomatis*, whereas low levels of specific IgG were detected in some patients. Additionally, IgM, IgG, and complement factors were identified on the surface of the grains taken from actinomycetoma lesions caused by *Streptomyces somaliensis* [15].

In this review, genetic predisposition to fungal infections resembling mycetoma was reviewed to give clues about similar genetic variations that predispose individuals to mycetoma. Additionally, all studies that were previously done to explore host genetic susceptibility to mycetoma were also included.

Methods

An extensive literature review was conducted using the electronic databases PubMed/MED-LINE and Google Scholar. The search included combinations of the following key words: 'mycetoma', 'polymorphisms', and 'genetic susceptibility' for reviewing articles regarding susceptibility to mycetoma. For genetic susceptibility to fungal infections, the key words 'fungal', 'infection', 'mycoses', and 'genetic susceptibility' were used as search terms. All articles identified through the 2 electronic databases were analysed to ensure they were within the scope of this review. Only papers written in English were included, and year of publication was not restricted.

Host genetic susceptibility to mycetoma and other fungal infections

Whilst there is a clear link to low socioeconomic development, only a proportion of people exposed to mycetoma-causing organisms develop clinical disease, despite a shared environment and the ubiquitous environmental presence of the pathogens in endemic areas. This observation raises the hypothesis that host genetic factors have a role in determining the out-come of infection.

Three previous studies used ELISA to demonstrate the presence of antibodies against mycetoma-causative agents in the sera of unaffected residents in endemic areas, indicating exposure to the pathogens without development of disease [21–23].

Other evidence suggesting a genetic predisposition to mycetoma includes familial clustering of mycetoma patients, first observed by Al Dawi and her colleagues in 2013 in a hospital-based study at the Mycetoma Research Centre (MRC) that included 53 eumycetoma patients and 31 healthy controls [24]. This study showed that more than 62% of eumycetoma patients had at least one family member affected by the disease. A community-based study in 2014 also suggested a role for genetic inheritance, with more than half of patients (52%) having a family member with the disease [10]. In 2015, a comprehensive study from the MRC reported a family history of mycetoma in 12% of 6,792 patients seen during the period 1991–2014 [25]. No studies have been undertaken to date to estimate heritability, investigate the disease's mode of inheritance, or confirm the genetic component of susceptibility towards mycetoma based on the previously observed familial clustering. It is important to note that families share their environment as well as their genes, but in communities affected by mycetoma, families with multiple cases live in close proximity to families who have no cases, suggesting gene–environment interactions are indeed important.

Finally, there is a good evidence in other fungal infections that host genetic factors deter-mine susceptibility to disease [26]. Examples include phaeohyphomycosis, chromoblastomycosis, candidiasis, and aspergillosis, in which a complex range of clinical phenotypes is also observed.

In summary, there is some evidence suggesting host genetic factors as regulators of susceptibility to mycetoma and other fungal infections, but they are likely to be complex genetic traits in which multiple genes interact with each other and environmental factors, as well as the pathogen, to cause disease. It is therefore challenging to identify the genes involved. One approach that has successfully identified genes associated with infectious diseases is the characterisation of rare monogenic disorders that predispose individuals to infection. Whilst rare, investigation of such families offers insights into critical immune response pathways that can be further investigated at the population level. Examples here include mutations in genes within the interferon (IFN)-gamma pathway that predispose individuals to mycobacterial infection [27] or genes encoding terminal complement pathway components that predispose

individuals to meningococcal infection [28]. This approach has also been applied to fungal infections.

Monogenic disorders and fungal infections

Single-gene defects that lead to primary immunodeficiencies (PIDs) have helped to under-stand better the immunological defects that increase susceptibility to fungal infections. An example of an autosomal recessive PID is human caspase recruitment domain-containing protein 9 (CARD9) deficiency, which is caused by biallelic mutations in the gene *CARD9* [29]. *CARD9* encodes for an adaptor protein downstream from C-type lectin receptors (CLRs), which are pathogen recognition receptors (PRRs), like the Toll-like receptors (TLRs), that have a major role in fungal recognition [30]. In humans, CLRs (for example, dectin-1, dectin-2, and the macrophage inducible Ca 2+-dependent lectin receptor MINCLE) that can recognise fungal pathogen-associated molecular patterns (PAMPs) and their adaptor molecule CARD9 have a critical role in antifungal host defence, and several studies highlighted an enhanced fungus-specific infection susceptibility as a consequence of mutations or deletions in CLRs or *CARD9* [30,31]. For example, autosomal recessive mutations in *CARD9* cause significant defects in the Th17 response because of decreased proportions of circulating $IL17^+$ T lymphocytes [32]; diminished production of IL1β and IL6, which are essential for priming Th17 cell differentiation; and impaired neutrophil killing [33]. Interestingly, 2 compound heterozygous mutations and 1 homozygous frameshift mutation in *CARD9* were also found in patients with phaeohyphomycosis, which is a subcutaneous infection similar to mycetoma [34]. Phaeohyphomycosis is caused by several microorganisms, including *Phialophora verrucose*, which is one of the mycetoma-causative organisms [35]. The identified mutations did not affect *CARD9* expression but resulted in lack of the wild-type mature CARD9 protein, which consequently decreased TH17 cell proportions and cytokine production and impaired patients' immune responses. Recently, Queiroz-Telles and colleagues successfully used hematopoietic stem cell transplantation (HSCT) to treat 2 unrelated patients from Brazil and Morocco with deep/invasive dermatophytosis caused by CARD9 deficiency [36]. Both patients stopped antifungal therapy after achieving complete clinical remission, suggesting that the pathogenesis of fungal infections in these patients was mainly due to the disruption of leukocyte-mediated CARD9 immunity. *P. verrucosa* causes another disease that resembles mycetoma known as chromoblastomycosis (CBM). Familial inheritance to CBM was suggested by Perez-Blanco and colleagues, with a 3.5 times higher risk of developing CBM amongst members of common ancestry in an endemic state in Venezuela [37]. In Brazil, Tsuneto and col-leagues found in a study involving 32 CBM patients and 77 healthy matched controls that the human leukocyte antigen (HLA)-A29 was significantly associated with susceptibility to the dis-ease (*P* value = 0.03) [38]. The relative risk of individuals with HLA-A29 antigen was estimated to be 10-fold higher than those lacking the antigen.

Another known PID is chronic granulomatous disease (CGD), which results from defects in genes encoding the NADPH oxidase subunits (*CYBA* and *CYBB*, encoding the cytochrome B-245 alpha and beta chains, respectively; and *NCF-1*, *NCF-2*, and *NCF-4*, encoding neutrophil cytosolic factors 1, 2 and 4), which together are critical for the generation of superoxide within phagocytes [39]. In 65% of CGD cases, an X-linked recessive pattern of inheritance is found because of mutations in *CYBB* encoding subunit $gp91^{phox}$. The remaining 35% of cases are autosomal recessive resulting from mutations in *CYBA*-, *NCF-1*–, *NCF-2*–, and *NCF-4*–encoding subunits $p22^{phox}$, $p47^{phox}$, $p67^{phox}$, and $p40^{phox}$, respectively. The mutations in the genes encoding NADPH oxidase subunits include deletions, frameshifts, and nonsense and missense mutations [40]. This defect in CGD phagocytes increases susceptibility to fungal infections

such as aspergillosis (in one-third of all CGD cases) [39] and candidiasis [41]. The important function of neutrophils in immunity against mycetoma has been demonstrated [1,42], suggest-ing that investigating similar mutations that potentially lead to impaired oxygen-dependent fungicidal activity in eumycetoma may be fruitful.

Chronic mucocutaneous candidiasis (CMC) is a monogenetic immunodeficiency of cell-mediated immunity that develops as a consequence of defects in IL-17 and IL-22 immunity required for defence against fungal infections [43]. CMC is transmitted with autosomal dominant or recessive inheritance patterns and affects the skin, nails, and mucous membranes of affected patients [41]. The autosomal dominant pattern of CMC is mainly caused by gain-of-function missense mutations in the CC-domain of *STAT1* (Signal Transducer And Activator Of Transcription 1), which encodes a critical transcription factor downstream from IFN-⟨/® and IFN-© signalling. These mutations led to defective Th1 and Th17 lymphocyte responses and reduced IL-17, IL-22, and IFN-© production, explaining the increased susceptibility to fungal infection [44]. In addition, mutations in *IL17RA* and *IL17F* have been associated with autosomal dominant CMC, leading to impaired IL-17 immunity [44]. Autosomal recessive CMC results from a rare condition, autoimmune polyendocrinopathy candidiasis ectodermal dystrophy (APECED), which in turn results from mutations in the autoimmune regulator (*AIRE*) gene [39]. *AIRE* encodes a transcriptional regulator expressed by medullary thymic epithelial cells to regulate the expression of peripheral tissue-specific self-antigens and promote central tolerance via deletion of self-reactive T cells [45]. Loss-of-function mutations in *AIRE* lead to the production of neutralising autoantibodies against important cytokines with anti-fungal properties such as IL-17E, IL-17F, and IL-22 [46].

Since deficiency in cell-mediated immunity in mycetoma patients was previously suggested by Mahgoub and colleagues [19], the presence of similar mutations in eumycetoma should be studied. The importance of IL-17 immunity against *Aspergillus fumigatus* and *Fusarium oxysporum*, which are both causative agents of mycetoma [47,48], was demonstrated by Taylor and colleagues in murine models of fungal keratitis [49]. Both IL-17–producing neutrophils and Th17 cells conferred protective immunity against fungal hyphae and regulated the severity of corneal disease. In humans, Siddig and colleagues demonstrated the expression of IL-17A in the granuloma of different mycetoma-causative agents, reaffirming the importance of Th17 immunity against mycetoma infection [50].

Polygenic disorders and fungal infections

Although very informative, single-gene mutations leading to fungal infections are relatively rare in the general population, in which complex inheritance of traits is more likely. Recent advances in genome interrogation technologies have enabled researchers to detect genetic variations that predispose susceptibility to fungal infections and the interplay between innate and adaptive immune cells in addition to other effector cells that together constitute the host immune response to fungal invasion [41,51]. For example, genetic susceptibility to candidiasis was investigated using genome-wide association studies (GWASs), which revealed a 19.4-fold increased risk in individuals with 2 or more single-nucleotide polymorphisms (SNPs) in *LFA-3* (lymphocyte function-associated antigen 3), *LCE4A* (late cornified envelope 4A), and *TAGAP* (T cell activation RhoGTPase activating protein) loci [52].

Genetic susceptibility to aspergillosis is also polygenic [53]. *Aspergillus* spp. are seldom pathogenic in immunocompetent hosts because of innate immune responses mediated mainly through alveolar macrophages and neutrophils [54]. Most cases of aspergillosis occur in the context of down-regulation of the immune system by immunosuppressive therapies. Susceptibility to different forms of aspergillosis, including invasive allergic (IA) and chronic

noninvasive syndromes, has been shown to be associated with polymorphic variants in *TLR1*, *TLR6*, *TLR4*, *TLR9*, and *TLR3* [55–57]. Variants in the chemokine (C-X-C motif) ligand 10 (*CXCL10*), an inflammatory mediator that stimulates the directional migration of Th1 in addition to increasing T cell adhesion to endothelium [58], have been associated with invasive *A. fumigatus* infection [59]. In this study, the presence of a haplotype in *CXCL10* (rs1554013 [+11101 **C**/T], rs3921 [+1642 C/**G**], and rs4257674 [−1101 **A**/G]) was associated with the inability of immature dendritic cells (iDCs) to express *CXCL10* in patients with IA after allogeneic HSCT, which led to an increased risk of developing IA. Susceptibility to IA has also been linked to deficiencies in PRRs with opsonic activity such as mannose-binding lectin (*MBL*) and long pentraxin 3 (*PTX3*), which mediates fungal uptake and killing by phagocytes [30,60,61]. Another molecule with an opsonic activity that has been associated with susceptibility to IA in allogeneic HSCT patients is plasminogen (*PLG*) [62]. Using a novel method of candidate gene polymorphisms identification, Zaas and colleagues used computational genetic mapping analysis of suppressed murine model survival data to eventually identify an SNP in human *PLG* that increased the risk for IA in HSCT recipients.

Genetic susceptibility to mycetoma

Most of the studies done on host susceptibility to mycetoma focused on polymorphisms in candidate genes related to the host immune response.

Polymorphisms in innate immunity genes. Because of the significance of innate immunity in host defence, genetic variations in the genes of this system could have a major impact on immune responses to mycetoma-causative organisms. Thus, the first evidence of host genetic susceptibility to mycetoma was shown in 2007 by van de Sande and her colleagues when they reported associations between polymorphisms in genes involved in innate immunity and susceptibility to mycetoma [42]. Studies initially focused on neutrophil function given its importance in early defence against mycetoma [13,14]. Studies were undertaken on 11 SNPs in 8 genes: complement receptor 1 (*CR1*), *MBL*, tumour necrosis factor 〈 (*TNF*〈), macrophage chemoattractant protein-1 (*MCP-1*), IL 8 (*CXCL8*), C-X-C motif chemokine receptor 2 (*CXCR2*), nitric oxide synthase 2 (*NOS2*), and thrombospondin-4 (*TSP-4*). Five SNPs in *CR1*, *CXCL8*, its receptor *CXCR2*, *TSP-4*, and *NOS2* had significant differences in allele distribution between 125 Sudanese mycetoma patients and 140 matched controls. Furthermore, an SNP in *NOS2* was associated with lesion size. The SI1 and McC^{a} alleles of CR1 SNPs rs17047661 and rs17047660, respectively, had higher distribution in mycetoma patients than in matched endemic controls, and the authors speculated that these polymorphisms could result in conformational changes in the receptor that eventually might lead to a defect in the efficacy of *M. mycetomatis* phagocytosis. CXCL8 and TSP-4 are chemoattractants for neutrophils. The alleles −251A allele in *CXCL8* and the 29929C allele in *TSP4*, in addition to the +785C allele in *CXCR2* that encodes for the CXCL8 receptor, were associated with mycetoma in this study. The final polymorphism for which there were differences in allele frequency between mycetoma patients and healthy controls was found in *NOS2*, encoding nitric oxide synthase, which produces nitric oxide (NO), which mediates tumoricidal and bactericidal actions [63]. The G954C allele results in a 7-fold higher NOS activity compared to the wild type, which may explain why it was found more amongst the healthy endemic controls [64]. Higher production of CXCL8 and lower NO production in mycetoma patients were both highlighted as risk factors for developing mycetoma owing to impaired wound healing, which was previously demonstrated in mice [65,66].

The absence of significant differences in allele frequencies between cases and endemic matched controls in polymorphisms in the other 6 genes might be due to investigating a

limited number of variants in addition to the small number of enrolled subjects, resulting in a lack of statistical power. Thus, exploring other variants in these genes in addition to other genes involved in neutrophil function in a larger cohort could reveal additional variants with functional impact in susceptibility to mycetoma.

Polymorphisms in acquired immunity genes. Several HLA alleles (both class I and class II) have been associated with susceptibility or resistance to the development of infectious diseases [67–69]. Based on this, Al Dawi and colleagues analysed *HLA-DRB1* and *HLA-DQB1* allele frequencies amongst confirmed eumycetoma patients compared with matched controls [24]. Of the *HLA-DRB1* alleles, the HLA-DRB1*13 allele showed significant association with eumycetoma infection ($P = 0.044$, odds ratio [OR] = 2.629). Interestingly, the HLA-DRB1*02 allele had a high frequency in the control group (9.8%, $P = 0.047$) whilst it was absent in the eumycetoma patients, suggesting a protective role against eumycetoma. For *HLA-DQB1* alleles, the HLA-DQB1*5 allele showed a significantly higher frequency in mycetoma patients than in healthy controls ($P = 0.029$, OR = 3.471), indicating a possible association between this allele and the development of clinical mycetoma. This study included only 84 subjects (53 eumycetoma patients and 31 healthy matched controls) who were admitted to the MRC. Therefore, further studies are required on larger sample sizes and amongst communities living in mycetoma-endemic areas.

A number of studies have identified association between variants in cytokine genes involved in acquired immune responses and mycetoma. Owing to the important roles of C-C chemokine ligand 5 (CCL5) in attracting T cells, dendritic cells, eosinophils, and natural killer (NK) cells [70] and IL-10, which is one of the Th 2 cytokines [71], Mhmoud and colleagues studied the role of 3 functional SNPs in *CCL5* and 2 SNPs in *IL-10* in mycetoma granuloma formation [17]. Allele frequencies for 2 SNPs in *CCL5* (rs2280788 and rs280789) and 1 SNP in *IL-10* (rs280789) were significantly different between 149 mycetoma patients and 206 matched controls and were linked to elevated serum levels of both CCL5 and IL-10 in mycetoma patients. The 2 SNP alleles in *CCL5* that were significantly associated with mycetoma had opposite functional effects: the G-allele at SNP rs2280788 results in a higher *CCL5* transcription, whilst the C-allele in the other SNP, rs280789, is associated with diminished transcription of *CCL5*. Consequently, the exact role of CCL5 in mycetoma granuloma formation remains to be elucidated.

Polymorphisms in host enzyme genes. Two studies involving host enzymes that are part of the immune system have been reported. The first, reported in 2014 by Geneugelijk and colleagues [72], investigated the role of matrix metalloproteinases-2 (MMP-2), matrix metallo-proteinases-9 (MMP-9), and tissue inhibitor of metalloproteinases (TIMP) in the formation of fungal grains in *M. mycetomatis* eumycetoma patients. These enzymes are thought to be involved in the deposition of collagen around the grains to encapsulate them within the granulomas [72]. It is believed that excessive collagen formation may contribute to treatment failure in mycetoma patients. The study included 3 parts: (1) detection of MMP-2 and MMP-9 locally around fungal grains using immunohistochemical staining of 8 tissue sections taken from *M. mycetomatis* eumycetoma patients, (2) measuring absolute serum levels of MMP-2 and MMP-9 in the sera of 36 *M. mycetomatis*-infected patients and 36 healthy controls using ELISA, and (3) genotyping functional SNPs in the promoter regions of *MMP-2* (rs243865), *MMP-9* (rs3918242), and *TIMP-1* (rs4898) using genomic DNA from 125 Sudanese *M. mycetomatis* eumycetoma patients and 103 healthy endemic controls. Active MMP-9 was only found in the sera of 36% of eumycetoma patients, yet that cannot be attributed solely to mycetoma infection because no data about coinfection were recorded during the sample collection. No genetic differences were found between patients and healthy controls for *MMP-2* and *MMP-9*; however, there was a higher frequency of the T allele of an SNP in *TIMP-1* (372T/C) in 77 male

eumycetoma patients compared to 44 healthy male controls. This T allele is associated with a lower TIMP-1 protein expression in inflammatory bowel disease, and the authors suggested that it could explain the previously reported male predominance in mycetoma [73].

In 2015, a second study investigated the presence of chitin in the cell wall of *M. mycetoma-tis*. Chitin triggers the host innate immune response and the release of chitin-degrading enzymes such as chitotriosidase (*CHIT1*) and acidic mammalian chitinase (AMCase, also known as *CHIA*) [74]. Four polymorphisms in *CHIT1* and *CHIA* were studied in 112 eumycetoma patients and 103 matched controls that were from the same cohort used for the metalloproteinases study. Though both *CHIA* and *CHIT1* were expressed in mycetoma lesions, a

24-bp insertion in *CHIT1* resulting in impaired enzyme activity [75,76] was found to be significantly associated with eumycetoma. Because both chitotriosidase and AMCase were proven to exist in mycetoma lesions, further investigation to detect polymorphisms affecting the activity of these enzymes is still needed with a larger sample size.

Polymorphisms in sex hormone biosynthesis genes. A male predominance of mycetoma has been described in several studies [7,9,11,72], suggesting a possible role of hormonal influence in mycetoma susceptibility. Based on this observation, a study was carried out in 2015 on 125 eumycetoma patients and 103 matched controls (the same metalloproteinases study group) to investigate changes in hormonal levels that could result from polymorphisms within genes encoding enzymes essential for sex hormone biosynthesis [77]. The selected polymorphisms were rs4680 in Catechol-O-methyltransferase (*COMT*), rs1056836 in Cytochrome p450 subfamily 1B1 (*CYP1B1*), rs743572 in Cytochrome p450 subfamily 17 (*CYP17*), rs700518 in cytochrome p450 subfamily 19 (*CYP19*), and rs6203 in hydroxysteroid dehydrogenase 3B (*HSD3B*). Only alleles of the rs4680 SNP in *COMT*, which is quite common in African populations according to the Genome Aggregation Database (gnomAD) [78] and has been linked to neuropsychological disorders, and rs700518 in *CYP19* had significant differences in distribution amongst mycetoma patients compared to healthy endemic controls.

In summary, a total of 13 genes (Table 1) have allelic variants found to be associated with mycetoma, and these genes lie in different pathways and systems. All the studies mentioned above were focused on a predefined set of polymorphisms in candidate genes, selected for their possible role in immunity to mycetoma in relatively small sample sizes. No systematic genome-wide studies have been reported to date.

Concluding remarks and future perspectives

The studies described above suggest there is a role for the host's underlying genetic profile in determining susceptibility to mycetoma. However, the studies are generally small and take a candidate gene approach, and none of them have been replicated. Advances in genomic science and the supporting technology have paved the way for large-scale genome-wide association and next generation sequencing (NGS) studies that can underpin a new strategy to systematically interrogate the genome for variants associated with mycetoma. GWASs are used to detect common variants, whilst NGS technologies such as whole-exome sequencing (WES), in which the exons across the whole genomes of cases and controls are sequenced, are used to identify the rare variants with functional impact on disease pathogenesis [79]. Since several genomes for the most common causative organisms, including *Nocardia brasiliensis* [80], *S. somaliensis* [81], *Actinomadura madurae* [82], *Nigrograna mackinnonii* [83], and *M. mycetomatis* [84] are now publicly available, parallel analysis of pathogen and host genomes could give valuable insights into host–pathogen interactions in mycetoma [1] and would eventually aid in devising better strategies for risk assessment, treatment, and prognosis for mycetoma patients.

Table 1. A summary of genes with allelic variants that are significantly associated with mycetoma.

Gene	SNP	Rs-Number	*P* Value	Reference
CCL5	−28C/G	rs2280788	<0.0001	[17]
	1648044C>T[♭]	rs280789	<0.0001	[17]
CHIT1	24-bp insertion	rs3831317	0.004	[74]
COMT	19951271G>A[♭]	rs4680	0.006	[77]
CR1	207782889A>G[♭]	rs17047661	0.039	[42]
	207782856A>G[♭]	rs17047660	0.001	[42]
CXCL8	−251T/A	rs4073	0.008	[42]
CXCR2	219000310C>T[♭]	rs2230054	0.037	[42]
CYP19	51529112T>C[♭]	rs700518	0.004	[77]
HLA-DRB1	HLA-DRB1[♭]13	N/A	0.044	[24]
	HLADRB1[♭]02	N/A	0.047	[24]
HLA-DQB1	HLA-DQB1[♭]5	N/A	0.029	[24]
IL-10	−819T/C	rs1800871	0.0005	[17]
NOS2	−954G>C[♭]	rs1800482	0.0006	[42]
TIMP-1	372T/C	rs4898	0.0004 (males); 0.53 (females)	[72]
TSP4	79361265G>C[♭]	rs1866389	0.030	[42]

[♭] GRCh37 human reference genome.

Abbreviations: *CCL5*, C-C chemokine ligand 5; *CHIT1*, chitotriosidase; *COMT*, catechol-O-methyltransferase; *CR1*, complement receptor 1; *CXCL8*, C-X-C chemokine ligand 8; *CXCR2*, C-X-C motif chemokine receptor 2; *CYP19*, cytochrome p450 family 19; *HLA*, human leukocyte antigen; *IL-10*, interleukin 10; *NOS2*, nitric oxide synthase 2; N/A, not applicable; *TIMP-1*, tissue inhibitor of metalloproteinases 1; *TSP4*, thrombospondin-4.

https://doi.org/10.1371/journal.pntd.0008053.t001

It is essential to include large cohorts to avoid the multiple challenges of genetic association studies, including population stratification, genetic heterogeneity, and insufficient replication

Key learning points

1. Current evidence suggests a role for host genetic factors in regulating susceptibility to mycetoma.
2. Mycetoma is likely to be a complex genetic trait in which multiple genes interact with each other and environmental factors, as well as the pathogen, to cause the disease.
3. Reviewing genetic susceptibility to other morphologically related fungal infections can help to identify candidate genes involved in susceptibility to mycetoma.
4. A total of 13 genes had allelic variants found to be associated with mycetoma.
5. All the reviewed studies on genetic susceptibility to mycetoma took a candidate gene approach, and none have been replicated.
6. There is a need for a new strategy to systematically interrogate the genome for variants associated with mycetoma.

Top 5 papers

1. van de Sande WWJ, Fahal A, Verbrugh H, van Belkum A. Polymorphisms in Genes Involved in Innate Immunity Predispose Toward Mycetoma Susceptibility. J Immunol. 2007;179: 3065–3074.

2. Verwer PEB, Notenboom CC, Eadie K, Fahal AH, Verbrugh HA, van de Sande WWJ. A Polymorphism in the Chitotriosidase Gene Associated with Risk of Mycetoma Due to Madurella mycetomatis Mycetoma–A Retrospective Study. PLoS Negl Trop Dis. 2015;9: e0004061.

3. van de Sande WW, Fahal A, Tavakol M, van Belkum A. Polymorphisms in cate-chol-O-methyltransferase and cytochrome p450 subfamily 19 genes predispose towards Madurella mycetomatis- induced mycetoma susceptibility. Med Mycol 2010; 48: 959–968.

4. Mhmoud NA, Fahal AH, van de Sande WWJ. The association between the inter-leukin-10 cytokine and CC chemokine ligand 5 polymorphisms and mycetoma granuloma formation. Med Mycol. 2013;(5): 527–33.

5. Al Dawi AF, Mustafa MI, Fahal AH, EL Hassan AM, Bakhiet S, Magoub ES, et al. The association of HLA-DRB1 and HLA-DQB1 alleles and the occurrence of eumycetoma in Sudanese patients. Khartoum Med J. 2013;6: 923–929.

power [85,86]. Additionally, genotype–phenotype correlation analyses are mandatory to correlate the relative contribution of genetic findings to the specific clinical outcomes.

Dissecting the contribution that host genetic variation has on susceptibility to mycetoma will enable the identification of pathways that are potential targets for new treatments for mycetoma and will also enhance our ability to stratify 'at-risk' individuals, allowing the potential to develop preventive and personalised clinical care strategies in the future.

References

1. van Belkum A, Fahal A, van de Sande WW. Mycetoma caused by Madurella mycetomatis: a completely neglected medico-social dilemma. Adv Exp Med Biol. 2013; 764: 179–89. https://doi.org/10.1007/978-1-4614-4726-9_15 PMID: 23654067
2. Lichon V, Khachemoune A. Mycetoma: a review. American Journal of Clinical Dermatology. 2006; 315– 321. https://doi.org/10.2165/00128071-200607050-00005 PMID: 17007542
3. Nenoff P, van de Sande WW, Fahal AH, Reinel D, Schofer H. Eumycetoma and actinomycetoma—An update on causative agents, epidemiology, pathogenesis, diagnostics and therapy. J Eur Acad Derma-tol Venereol. 2015; 29(10): 1873–83. https://doi.org/10.1111/jdv.13008 PMID: 25726758
4. de Hoog GS, Ahmed SA, Najafzadeh MJ, Sutton DA, Keisari MS, Fahal AH, et al. Phylogenetic Findings Suggest Possible New Habitat and Routes of Infection of Human Eumyctoma. PLoS Negl Trop Dis. 2013; 7(5): e2229. https://doi.org/10.1371/journal.pntd.0002229 PMID: 23696914
5. Samy AM, van de Sande WWJ, Fahal AH, Peterson AT. Mapping the potential risk of mycetoma infection in Sudan and South Sudan using ecological niche modeling. PLoS Negl Trop Dis. 2014; 8: e3250. https://doi.org/10.1371/journal.pntd.0003250 PMID: 25330098
6. WHO. Mycetoma. World Health Organization [Internet]. 2016 [cited 2019 Mar 21]. Available from: https://www.who.int/buruli/mycetoma/en/

7. Fahal AH. Mycetoma: A thorn in the flesh. Trans R Soc Trop Med Hyg. 2004; 98: 3–11. https://doi.org/ 10.1016/s0035-9203(03)00009-9 PMID: 14702833

8. Burki TK. Starting from scratch-the unique neglect of mycetoma. Lancet Infect Dis. 2016; 16(9): 1011– 1012. https://doi.org/10.1016/S1473-3099(16)30283-3 PMID: 27684346

9. Queiroz-Telles F, Fahal AH, Falci DR, Caceres DH, Chiller T, Pasqualotto AC. Neglected endemic mycoses. Lancet Infect Dis. 2017 Nov; 17(11): e367–e377. https://doi.org/10.1016/S1473-3099(17) 30306-7 PMID: 28774696

10. Fahal A, Mahgoub ES, EL Hassan AM, Abdel-Rahman ME, Alshambaty Y, Hashim A, et al. A New Model for Management of Mycetoma in the Sudan. PLoS Negl Trop Dis. 2014; 8(10): e3271. https://doi. org/10.1371/journal.pntd.0003271 PMID: 25356640

11. Zijlstra EE, van de Sande WWJ, Welsh O, Mahgoub ES heikh, Goodfellow M, Fahal AH. Mycetoma: a unique neglected tropical disease. Lancet Infect Dis. 2016; 16: 100–112. https://doi.org/10.1016/ S1473-3099(15)00359-X PMID: 26738840

12. Ahmed SA, Van De Sande WWJ, Desnos-Ollivier M, Fahal AH, Mhmoud NA, De Hoog GS. Application of isothermal amplification techniques for identification of madurella mycetomatis, the prevalent agent of human mycetoma. J Clin Microbiol. 2015; 53: 3280–3285. https://doi.org/10.1128/JCM.01544-15 PMID: 26246484

13. EL Hassan AM, Fahal AH, EL Hag IA, Khalil EAG. The pathology of mycetoma: light microscopic and ultrastructural features. Sudan Med J. 1994;23–45.

14. Fahal AH, el Toum EA, el Hassan AM, Mahgoub ES, Gumaa SA. The host tissue reaction to Madurella mycetomatis: new classification. J Med Vet Mycol 1995; 33: 15–17. PMID: 7650573

15. El Hassan AM, Fahal AH, Ahmed AO, Ismail A, Veress B. The immunopathology of actinomycetoma lesions caused by Streptomyces somaliensis. Trans R Soc Trop Med Hyg. 2001; 95(1): 89–92. https:// doi.org/10.1016/s0035-9203(01)90346-3 PMID: 11280076

16. Nasr A, Abushouk A, Hamza Aet al. Th-1, Th-2 Cytokines profile among Madurella mycetomatis eumycetoma patients. PLoS Negl Trop Dis. 2016; 10: e0004862. https://doi.org/10.1371/journal.pntd. 0004862 PMID: 27434108

17. Mhmoud NA, Fahal AH, van de Sande WWJ. The association between the interleukin-10 cytokine and CC chemokine ligand 5 polymorphisms and mycetoma granuloma formation. Med Mycol. 2013;(5): 527–33. https://doi.org/10.3109/13693786.2012.745201 PMID: 23210681

18. Abushouk A, Nasr A, Masuadi E, Allam G, Siddig EE, Fahal AH. The Role of Interleukin-1 cytokine fam-ily (IL-1β, IL-37) and interleukin-12 cytokine family (IL-12, IL-35) in eumycetoma infection pathogenesis. PLoS Negl Trop Dis. 2019; 13: e0007098. https://doi.org/10.1371/journal.pntd.0007098 PMID: 30946748

19. Mahgoub ES, Gumaa SA, El Hassan AM. Immunological status of mycetoma patients. Bull Soc Pathol Exot Filiales. 1977; 70(1): 48–54. PMID: 579331

20. Wethered DB, Markey MA, Hay RJ, Mahgoub ES, Gumaa SA. Humoral immune responses to mycetoma organisms: characterization of specific antibodies by the use of enzyme-linked immunosorbent assay and immunoblotting. Trans R Soc Trop Med Hyg. 1988; 82: 918–23. https://doi.org/10.1016/ 0035-9203(88)90042-9 PMID: 2476875

21. Mclaren ML, Mahgoub ES, Georgakopoulos E. Preliminary investigation on the use of the enzyme linked immunosorbent assay (ELISA) in the serodiagnosis of mycetoma. Med Mycol. 1978; 16: 225– 228.

22. Taha A. A serological survey of antibodies to Streptomyces somaliensis and Actinomadura madurae in the Sudan using enzyme linked immunosorbent assay (ELISA). Trans R Soc Trop Med Hyg. 1983; 77: 49–50. https://doi.org/10.1016/0035-9203(83)90011-1 PMID: 6679364

23. van de Sande WW, Janse DJ, Hira V, et al. Translationally controlled tumor protein from Madurella mycetomatis, a marker for tumorous mycetoma progression. J Immunol 2006; 177: 1997–2005. https://doi.org/10.4049/jimmunol.177.3.1997 PMID: 16849514

24. Al Dawi AF, Mustafa MI, Fahal AH, EL Hassan AM, Bakhiet S, Magoub ES, et al. The association of HLA-DRB1 and HLA-DQB1 alleles and the occurrence of eumycetoma in Sudanese patients. Khartoum Med J. 2013; 6: 923–929.

25. Fahal A, Mahgoub el S, El Hassan AMet al. Mycetoma in the Sudan: an update from the Mycetoma Research Centre, University of Khartoum, Sudan. PLoS Negl Trop Dis. 2015; 9: e0003679. https://doi. org/10.1371/journal.pntd.0003679 PMID: 25816316

26. Gow NAR, Netea MG. Medical mycology and fungal immunology: new research perspectives addressing a major world health challenge. Philos Trans R Soc B Biol Sci. 2016; 371: 20150462.

27. Newport MJ, Huxley CM, Huston S, Hawrylowicz CM, Oostra BA, Williamson R, et al. A Mutation in the Interferon-γ – Receptor Gene and Susceptibility to Mycobacterial Infection. N Engl J Med. 1996; 335: 1941–1949. https://doi.org/10.1056/NEJM199612263352602 PMID: 8960473

28. Figueroa JE, Densen P. Infectious diseases associated with complement deficiencies. Clin Microbiol Rev. 1991; 4: 359–95. https://doi.org/10.1128/cmr.4.3.359 PMID: 1889047

29. Drummond RA, Lionakis MS. Mechanistic Insights into the Role of C-Type Lectin Receptor/CARD9 Signaling in Human Antifungal Immunity. Front Cell Infect Microbiol. 2016; 6:39. https://doi.org/10.3389/ fcimb.2016.00039 PMID: 27092298

30. Lionakis MS, Levitz SM. Host Control of Fungal Infections: Lessons from Basic Studies and Human Cohorts. Annu Rev Immunol. 2018; 36: 157–191. Epub 2017 Dec 13. https://doi.org/10.1146/annurev-immunol-042617-053318 PMID: 29237128

31. Gross O, Gewies A, Finger K, Schafer M, Sparwasser T, Peschel C, et al. Card9 coïtrols a non-TLR signalling pathway for innate antifungal immunity. Nature. 2006; 442: 651–656. https://doi.org/10.1038/ nature04926 PMID: 16862125

32. Glocker E-O, Hennigs A, Nabavi M, Schaffer AA, Woellner C, Salzer U, et al. A Homozygöus CARD9 Mutation in a Family with Susceptibility to Fungal Infections. N Engl J Med. 2009; 361: 1727–1735. https://doi.org/10.1056/NEJMoa0810719 PMID: 19864672

33. Drewniak A, Gazendam RP, Tool ATJ, van Houdt M, Jansen MH, van Hamme JL, et al. Invasive fungal infection and impaired neutrophil killing in human CARD9 deficiency. Blood. 2013; 121: 2385–2392. https://doi.org/10.1182/blood-2012-08-450551 PMID: 23335372

34. Wang X, Wang W, Lin Z, Wang X, Li T, Yu J, et al. CARD9 mutations linked to subcutaneous phaeohyphomycosis and T H17 cell deficiencies. J Allergy Clin Immunol. 2014; 133(3): 905–908. https://doi.org/ 10.1016/j.jaci.2013.09.033 PMID: 24231284

35. Turiansky GW, Benson PM, Sperling LC, Sau P, Salkin IF, McGinnis MR, et al. Phialophora verrucosa: A new cause of mycetoma. J Am Acad Dermatol. 1995; 32: 311–315. https://doi.org/10.1016/0190-9622(95)90393-3 PMID: 7829731

36. Queiroz-Telles F, Mercier T, Maertens J, Sola CBS, Bonfim C, Lortholary O, et al. Successful Allogenic Stem Cell Transplantation in Patients with Inherited CARD9 Deficiency. J Clin Immunol. Springer New York LLC; 2019; 39: 462–469. https://doi.org/10.1007/s10875-019-00662-z PMID: 31222666

37. Perez-Blanco M, Hernandez Valles R, Garcıa-Humḿbrıa L, Yegres F. Chŕomoblastomycosis iń children´ and adolescents in the endemic area of the Falcon State, Venezuela. Med Mycol. 2006; 44: 467–471. ´ https://doi.org/10.1080/13693780500543238 PMID: 16882614

38. Tsuneto LT, Arce-Gomez B, Petzl-Erler ML, Queiroz-Telles F. HLA-A29 and genetic susceptibility to chromoblastomycosis. J Med Vet Mycol 1989; 27: 181–185. https://doi.org/10.1080/ 02681218980000241 PMID: 2778577

39. Lionakis MS, Netea MG, Holland SM. Mendelian genetics of human susceptibility to fungal infection. Cold Spring Harb Perspect Med. 2014; 4(6): a019638. https://doi.org/10.1101/cshperspect.a019638 PMID: 24890837

40. Holland SM. Chronic Granulomatous Disease. Hematol Oncol Clin North Am. 2013; 27: 89–99. https://doi.org/10.1016/j.hoc.2012.11.002 PMID: 23351990

41. Maskarinec SA, Johnson MD, Perfect JR. Genetic Susceptibility to Fungal Infections: What is in the Genes? Curr Clin Microbiol Reports. 2016; 3: 81–91.

42. van de Sande WWJ, Fahal A, Verbrugh H, van Belkum A. Polymorphisms in Genes Involved in Innate Immunity Predispose Toward Mycetoma Susceptibility. J Immunol. 2007; 179: 3065–3074. https://doi. org/10.4049/jimmunol.179.5.3065 PMID: 17709521

43. Puel A, Cypowyj S, Bustamante J, Wright JF, Liu L, Lim HK, et al. Chronic mucocutaneous candidiasis in humans with inborn errors of interleukin-17 immunity. Science. 2011; 332(6025): 65–68. https://doi. org/10.1126/science.1200439 PMID: 21350122

44. van de Veerdonk FL, Plantinga TS, Hoischen A, Smeekens SP, Joosten LAB, Gilissen C, et al. STAT1 Mutations in Autosomal Dominant Chronic Mucocutaneous Candidiasis. N Engl J Med. 2011; 365: 54–61. https://doi.org/10.1056/NEJMoa1100102 PMID: 21714643

45. Anderson MS, Venanzi ES, Klein L, Chen Z, Berzins SP, Turley SJ, et al. Projection of an immunological self shadow within the thymus by the aire protein. Science. 2002; 298: 1395–1401. https://doi.org/ 10.1126/science.1075958 PMID: 12376594

46. Puel A, Doffinger R, Natividad A, Chrabieh M, Barcenas-M̈orales G, Picard C, et al. Autoantibodies against IL-17A, IL-17F, and IL-22 in patients with chronic mucocutaneous candidiasis and autoimmune polyendocrine syndrome type I. J Exp Med. 2010; 207: 291–297. https://doi.org/10.1084/jem.20091983 PMID: 20123958

47. van de Sande WWJ. Global Burden of Human Mycetoma: A Systematic Review and Meta-analysis. PLoS Negl Trop Dis. 2013; 7: e2550https://doi.org/10.1371/journal.pntd.0002550 PMID: 24244780

48. Sumangala B, Sahana Shetty NS, Kala B, Kavya M. Mycetoma Foot caused by Fusarium solani in a Young Boy from Karnataka, India. Int. J. Curr. Microbiol. App. Sci. 2016; 5: 307–310.

49. Taylor PR, Leal SM, Sun Y, Pearlman E. Aspergillus and Fusarium Corneal Infections Are Regulated by Th17 Cells and IL-17–Producing Neutrophils. J Immunol. 2014; 192: 3319–3327. https://doi.org/10. 4049/jimmunol.1302235 PMID: 24591369

50. Siddig EE, Mohammed Edris AM, Bakhiet SM, van de Sande WWJ, Fahal AH. Interleukin-17 and matrix metalloprotease-9 expression in the mycetoma granuloma. PLoS Negl Trop Dis. 2019; 13: e0007351. https://doi.org/10.1371/journal.pntd.0007351 PMID: 31295246

51. Maziarz EK, Perfect JR. Cryptococcosis. Infect Dis Clin North Am. 2016; 30: 179–206. https://doi.org/ 10.1016/j.idc.2015.10.006 PMID: 26897067

52. Kumar V, Cheng S-C, Johnson MD, Smeekens SP, Wojtowicz A, Giamarellos-Bourboulis E, et al. Immunochip SNP array identifies novel genetic variants conferring susceptibility to candidaemia. Nat Commun. 2014; 5: 4675. https://doi.org/10.1038/ncomms5675 PMID: 25197941

53. Cunha C, Aversa F, Romani L, Carvalho A. Human Genetic Susceptibility to Invasive Aspergillosis. PLoS Pathog. 2013; 9: e1003434. https://doi.org/10.1371/journal.ppat.1003434 PMID: 23950708

54. Latge JP. The pathobiology of Aspergillus fumigatus. Trends Microbiol. 2001; 9: 382–389. https://doi. org/10.1016/s0966-842x(01)02104-7 PMID: 11514221

55. Kesh S, Mensah NY, Peterlongo P, Jaffe D, Hsu K, VAN DEN Brink M, et al. TLR1 and TLR6 polymorphisms are associated with susceptibility to invasive aspergillosis after allogeneic stem cell transplanta-tion. Ann N Y Acad Sci. 2005; 1062: 95–103. https://doi.org/10.1196/annals.1358.012 PMID: 16461792

56. Carvalho A, Pasqualotto AC, Pitzurra L, Romani L, Denning DW, Rodrigues F. Polymorphisms in Toll-Like Receptor Genes and Susceptibility to Pulmonary Aspergillosis. J Infect Dis. 2008; 197: 618–621. https://doi.org/10.1086/526500 PMID: 18275280

57. Carvalho A, De Luca A, Bozza S, Cunha C, D'Angelo C, Moretti S, et al. TLR3 essentially promotes pro-tective class I-restricted memory CD8 T-cell responses to Aspergillus fumigatus in hematopoietic trans-planted patients. Blood. American Society of Hematology; 2012; 119: 967–977.

58. Loetscher M, Gerber B, Loetscher P, Jones S a, Piali L, Clark-Lewis I, et al. Chemokine receptor specific for IP10 and mig: structure, function, and expression in activated T-lymphocytes. J Exp Med. 1996; 184: 963–9. https://doi.org/10.1084/jem.184.3.963 PMID: 9064356

59. Mezger M, Steffens M, Beyer M, Manger C, Eberle J, Toliat MR, et al. Polymorphisms in the chemokine (C-X-C motif) ligand 10 are associated with invasive aspergillosis after allogeneic stem-cell transplanta-tion and influence CXCL10 epression in monocyte-derived dendritic cells. Blood. 2008; 111: 534–536. https://doi.org/10.1182/blood-2007-05-090928 PMID: 17957030

60. Lambourne J, Agranoff D, Herbrecht R, Troke PF, Buchbinder A, Willis F, et al. Association of Man-nose-Binding Lectin Deficiency with Acute Invasive Aspergillosis in Immunocompromised Patients. Clin Infect Dis. 2009; 49: 1486–1491. https://doi.org/10.1086/644619 PMID: 19827955

61. Cunha C, Aversa F, Lacerda JF, Busca A, Kurzai O, Grube M, et al. Genetic PTX3 Deficiency and Aspergillosis in Stem-Cell Transplantation. N Engl J Med. 2014; 370: 421–432. https://doi.org/10.1056/ NEJMoa1211161 PMID: 24476432

62. Zaas AK, Liao G, Chien JW, Weinberg C, Shore D, Giles SS, et al. Plasminogen alleles influence susceptibility to invasive aspergillosis. PLoS Genet. 2008; 4: e1000101. https://doi.org/10.1371/journal. pgen.1000101 PMID: 18566672

63. Forstermann U, Sessa WC. Nitric oxide synthases: Regulation and function. Eur Heart J. Oxford Uni-versity Press; 2012; 33: 829–37, 837a-837d. https://doi.org/10.1093/eurheartj/ehr304 PMID: 21890489

64. Mercereau-Puijalon O, May J, Mordmüller B, Perkins DJ, Alpers M, Weinberg JB, et al. Nitric Oxide Synthase 2 Lambarene (G-954C), Increased Nitric Oxide Production, and Protection against Malaria. J Infect Dis. 2002; 184: 330–336.

65. Dovi J V, He L-K, DiPietro LA. Accelerated wound closure in neutrophil-depleted mice. J Leukoc Biol. 2003; 73: 448–55. https://doi.org/10.1189/jlb.0802406 PMID: 12660219

66. Stenger S, Donhauser N, Thüring H, Rollinghoff M, Bogdan C. Reactivation of latent leishmaniasis by inhibition of inducible nitric oxide synthase. J Exp Med. 1996; 183: 1501–14. https://doi.org/10.1084/ jem.183.4.1501 PMID: 8666908

67. Mangalam AK, Taneja V, David CS. HLA Class II Molecules Influence Susceptibility versus Protection in Inflammatory Diseases by Determining the Cytokine Profile. J Immunol. 2013; 190: 513–519. https:// doi.org/10.4049/jimmunol.1201891 PMID: 23293357

68. Dubaniewicz A, Lewko B, Moszkowska G, Zamorska B, Stepinski J. Molecular subtypes of the HLA-DR antigens in pulmonary tuberculosis. Int J Infect Dis. 2000; 4: 129–133. https://doi.org/10.1016/s1201-9712(00)90073-0 PMID: 11179915

69. Hilda Carrillo-Melendez, Esteban Ortega-Hernandéz, Granados Julio, Arroyo Saźra, Rodrigo Barquera RA. Role of HLA-DR Alleles to Increase Genetic Susceptibility to Onychomycosis in Nail Psoriasis. Ski Appendage Disord. 2016; 2: 22–25.

70. Marques RE, Guabiraba R, Russo RC, Teixeira MM. Targeting CCL5 in inflammation. Expert Opin Ther Targets. England; 2013; 17: 1439–1460. https://doi.org/10.1517/14728222.2013.837886 PMID: 24090198

71. Antachopoulos C, Roilides E. Cytokines and fungal infections. Br J Haematol. 2005; 129(5): 583–596. https://doi.org/10.1111/j.1365-2141.2005.05498.x PMID: 15916680

72. Geneugelijk K, Kloezen W, Fahal AH, van de Sande WWJ. Active Matrix Metalloprotease-9 Is Associated with the Collagen Capsule Surrounding the Madurella mycetomatis Grain in Mycetoma. PLoS Negl Trop Dis. 2014; 8: e2754. https://doi.org/10.1371/journal.pntd.0002754 PMID: 24675764

73. Ahmed AO, van Leeuwen W, Fahal A, van de Sande W, Verbrugh H, van Belkum A. Mycetoma caused by Madurella mycetomatis: a neglected infectious burden. Lancet Infect Dis. 2004; 4: 566–74. https:// doi.org/10.1016/S1473-3099(04)01131-4 PMID: 15336224

74. Verwer PEB, Notenboom CC, Eadie K, Fahal AH, Verbrugh HA, van de Sande WWJ. A Polymorphism in the Chitotriosidase Gene Associated with Risk of Mycetoma Due to Madurella mycetomatis Myce-toma–A Retrospective Study. PLoS Negl Trop Dis. 2015; 9: e0004061. https://doi.org/10.1371/journal. pntd.0004061 PMID: 26332238

75. Boot RG, Renkema GH, Verhock M, Strijland A, Bliek J, De Meulemeester TMAMO, et al. The human chitotriosidase gene—Nature of inherited enzyme deficiency. J Biol Chem. 1998; 273: 25680–25685. https://doi.org/10.1074/jbc.273.40.25680 PMID: 9748235

76. Seibold MA, Donnelly S, Solon M, Innes A, Woodruff PG, Boot RG, et al. Chitotriosidase is the primary active chitinase in the human lung and is modulated by genotype and smoking habit. J Allergy Clin Immunol. 2008; 122: 944–950.e3. https://doi.org/10.1016/j.jaci.2008.08.023 PMID: 18845328

77. van de Sande WW, Fahal A, Tavakol M, van Belkum A. Polymorphisms in catechol-O-methyltransferase and cytochrome p450 subfamily 19 genes predispose towards Madurella mycetomatis- induced mycetoma susceptibility. Med Mycol 2010; 48: 959–968. https://doi.org/10.3109/13693781003636680 PMID: 20184498

78. Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536: 285–91. https://doi.org/10.1038/nature19057 PMID: 27535533

79. Newport M. The Contribution of Host Genetics to TB Disease. Curr Respir Med Rev. 2013; 9: 156–162.

80. Vera-Cabrera L, Ortiz-Lopez R, Elizondo-Gonzalez R, Perez-Maya AA, Ocampo-Candiani J. Complete genome sequence of Nocardia brasiliensis HUJEG-1. J Bacteriol. 2012; 194: 2761–2762. https://doi. org/10.1128/JB.00210-12 PMID: 22535940

81. Kirby R, Sangal V, Tucker NP, Zakrzewska-Czerwińska J, Wierzbicka K, Herron PR, et al. Draft genome sequence of the human pathogen Streptomyces somaliensis, a significant cause of actinomycetoma. J Bacteriol. 2012; 194: 3544–3545. https://doi.org/10.1128/JB.00534-12 PMID: 22689234

82. Vera-Cabrera L, Ortiz-Lopez R, Elizondo-Gonzalez R, Campos-Rivera MP, Gallardo-Rocha A, Molina-Torres CA, et al. Draft Genome Sequence of Actinomadura madurae LIID-AJ290, Isolated from a Human Mycetoma Case. Genome Announc. American Society for Microbiology (ASM); 2014; 2: e00201–14. https://doi.org/10.1128/genomeA.00201-14 PMID: 24675856

83. Shaw JJ, Spakowicz DJ, Dalal RS, Davis JH, Lehr NA, Dunican BF, et al. Biosynthesis and genomic analysis of medium-chain hydrocarbon production by the endophytic fungal isolate Nigrograna mackinnonii E5202H. Appl Microbiol Biotechnol. 2015; 99: 3715–3728. https://doi.org/10.1007/s00253-014-6206-5 PMID: 25672844

84. Smit S, Derks MFL, Bervoets S, Fahal A, van Leeuwen W, van Belkum A, et al. Genome Sequence of Madurella mycetomatis mm55, Isolated from a Human Mycetoma Case in Sudan. Genome Announc. 2016; 4: e00418–16. https://doi.org/10.1128/genomeA.00418-16 PMID: 27231361

85. Sagoo GS, Little J, Higgins JPT. Systematic Reviews of Genetic Association Studies. PLoS Med. 2009; 6: e1000028.

86. Gorroochurn P, Hodge SE, Heiman GA, Durner M, Greenberg DA. Non-replication of association stud-ies: "Pseudo-failures" to replicate? Genet Med. 2007; 9: 325–331. https://doi.org/10.1097/gim. 0b013e3180676d79 PMID: 17575498

Mycetoma

Is it just the germs?

Determining the role of host genes in mycetoma

Sudan

brighton and sussex medical school

Summary of brief

The infectious disease mycetoma is a devastating, hard-to-treat neglected tropical disease. Not everyone who is exposed to the microbes (germs) that cause mycetoma gets disease. Understanding why this is so could lead to better treatment and prevention policies for mycetoma. This research aimed to discover whether an individual's genetic make-up determines their susceptibility to mycetoma and if so, which genes are important.

Key Messages

- A systematic literature review confirmed that host genes are likely to influence the development (or not) of mycetoma following infection with one of the many microbes that can cause it.
- Genetic differences between family members with and without mycetoma were identified in genes belonging to immune response pathways relevant for controlling infectious diseases that warrant further investigation.
- Dissecting the contribution that host genetic variation makes to susceptibility to mycetoma will enable the identification of pathways that are potential targets for new treatments for mycetoma.
- It will also enhance our ability to stratify "at-risk" individuals allowing the development of preventive and personalised clinical care strategies in the future.

Background

Mycetoma is a chronic infectious neglected tropical disease (NTD) that has extensive health and socioeconomic impact on patients and communities. Without early detection and treatment it destroys skin, underlying tissue and bone, leading to deformity and disability, often resulting in amputation – the lower limb is most commonly affected.

It is thought the infection is introduced by penetrating thorn injuries. Not everyone exposed to the microbes that cause mycetoma develops disease suggesting host defences determine susceptibility to mycetoma. There is evidence that genetic factors influence these defences since cases cluster in some families but not others, despite a shared environment.

However, genetic studies to date have been small and findings have not been reproduced. We therefore adopted a systematic approach to investigate the genetic basis of mycetoma in Sudan, a country with one of the highest rates of mycetoma in the world. We identified families with more than one case of mycetoma and compared the DNA sequences of family members with and without mycetoma.

Further studies are ongoing, investigating the role of these genetic variants in susceptibility to mycetoma in the wider population.

Key findings

- Our systematic review supported the hypothesis that host genetic factors influence the development (or not) of mycetoma following infection with one of the many microbes that can cause it.
- Familial cases are common in areas of Sudan where mycetoma is endemic.
- DNA sequencing studies in members from multi-case families identified genetic variants that were present in immune response genes in affected but not unaffected family members.

Priority actions / recommendations

- Our research forms the basis for further large-scale studies that aim to corroborate the role of host genes in predisposing individuals to the development of mycetoma.
- Further research to identify the key molecular pathways involved in the pathogenesis of mycetoma is necessary to develop the new diagnostic tools and treatments required to bring this debilitating disease under control, with a longer-term goal of elimination.
- With further evaluative research, genetic variants that predispose to mycetoma could be used to screen vulnerable communities and enable efficient prevention interventions or at least early disease detection and management.

Without early detection and treatment mycetoma destroys skin, underlying tissue and bone

Host genes are likely to influence the development (or not) of mycetoma following infection with one of the many microbes that can cause i

Mycetoma

Sudan

Further reading:

- Ali RS, Newport MJ, Bakhiet SM, Ibrahim ME, Fahal AH. *Host genetic susceptibility to mycetoma*. PLoS Negl Trop Dis. 2020;14(4):e0008053

Authors:

Rayan S. Ali[1,2], Sahar Mubarak Bakhiet[1,3], Muntaser E. Ibrahim[3], Ahmed Hassan Fahal[1], Melanie J. Newport[2]

Institutions:

1The Mycetoma Research Centre, University of Khartoum, Khartoum, Sudan

2Brighton and Sussex Centre for Global Health Research, Brighton and Sussex Medical School, Brighton, UK

3Institute of Endemic Diseases, University of Khartoum, Khartoum, Sudan

This research was funded by the National Institute for Health Research (NIHR) Global Health Research Unit on Neglected Tropical Diseases at BSMS using UK aid from the UK Government to support global health research. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK government.

FUNDED BY NIHR | National Institute for Health Research

UKaid from the British people

Appendix 2 - Pedigree collection datasheet.

Collection of pedigree information

Ethnic group:

Location:	Household:	Date:

Table_Apx 1 Datasheet for pedigrees.

ID	Name	Sex	Birthdate	Birthplace	Father's Name	Father Alive	Father's Birthplace	Father's Residence	Mother's Name	Mother Alive	Mother's Birthplace	Mother's Residence

Appendix 3 - Genome-wide association studies collection datasheet

Table_Apx 2 Datasheet for cases.

GWAS_ID	MRC_ID	Name	Sex	Age	Ethnic group	Age Of Onset	Occupation	Family history (Yes/No)	Relation (fam_hist)	Locality	State	Tel	Eumycetoma Site

Table_Apx 3 Datasheet for controls.

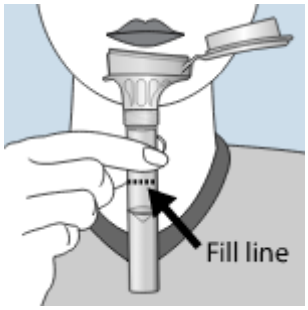
GWAS_ID	Name	Sex	Age	Ethnic group	Occupation	Family history (Yes/No)	Relation (fam_hist)	Locality	State	Tel

Appendix 4 - OrageneTM (OG-600) protocol for DNA collection from a saliva sample

(Source: https://www.dnagenotek.com/ROW/support/collection-instructions/oragene-dna/OG-500andOG-600.html)

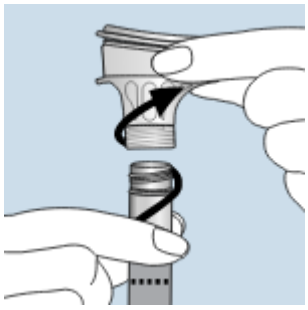
- Do not eat, drink, smoke, or chew gum for 30 minutes before giving your saliva sample.
- Do not remove the plastic film from the funnel lid.

Step 1: Spit into the funnel until the amount of liquid saliva (not bubbles) reaches the fill line shown below.

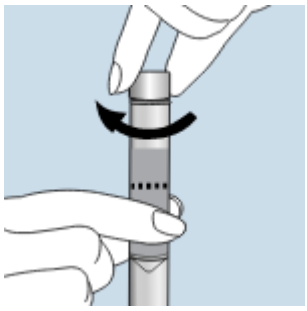


Step 2: Hold the tube upright with one hand. Close the funnel lid with the other hand (as shown) by firmly pushing the lid until you hear a loud click. The liquid in the lid will be released into the tube to mix with the saliva. Make sure that the lid is closed tightly.

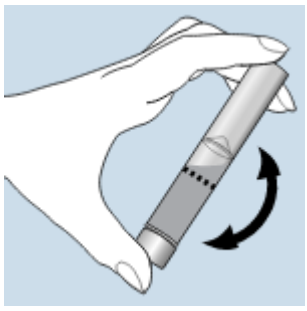
Step 3: Hold the tube upright. Unscrew the funnel from the tube.



Step 4: Use the small cap to close the tube tightly.



Step 5: Shake the capped tube for 5 seconds. Discard or recycle the funnel.



Appendix 5 - OrageneTM (OG-600) protocol for DNA isolation from a saliva sample

(Source: https://www.dnagenotek.com/ROW/products/collection-human/oragene-dna/600-series/OG-600.html)

Laboratory protocol for manual purification of DNA from 0.5 mL of sample:

1. Mix the sample in the DNA Genotek kit by inversion and gentle shaking for a few seconds.
2. Incubate the sample at 50°C in a water incubator for a minimum of 1 hour or in an air incubator for a minimum of 2 hours.
3. Transfer 500 µL of the mixed sample to a 1.5 mL microcentrifuge tube.
4. For 500 µL of the sample, add 20 µL (1/25th volume) of PT-L2P to the microcentrifuge tube and mix by vertexing for a few seconds.
5. Incubate on ice for 10 minutes.
6. Centrifuge at room temperature for 5 minutes at 15,000 × g.
7. Carefully transfer the clear supernatant with a pipette tip into a fresh microcentrifuge tube. Discard the pellet containing impurities.
8. To 500 µL of supernatant, add 600 µL of room temperature 95% to 100% ethanol. Mix gently by inversion 10 times.
9. Allow the sample to stand at room temperature for 10 minutes to allow the DNA to fully precipitate.
10. Place the tube in the microcentrifuge in a known orientation. Centrifuge at room temperature for 2 minutes at 15,000 × g.
11. Carefully remove the supernatant with a pipette tip and discard it. Take care to avoid disturbing the DNA pellet.
12. Carefully add 250 µL of 70% ethanol. Let stand at room temperature for 1 minute. Completely remove the ethanol without disturbing the pellet.
13. Add 100 µL of TE solution to dissolve the DNA pellet. Vortex for at least 5 seconds.
14. To ensure complete rehydration of the DNA (pellet and smear) incubate at room temperature overnight followed by vertexing or at 50°C for 1 hour with occasional vertexing.
15. Store DNA samples in aliquots at -20°C for long-term storage.

Appendix 6 - Agarose gel electrophoresis

This was adapted and modified from Protocols.io (https://www.protocols.io/view/agarose-gel-electrophoresis-e6nvw9pdwgmk/v3)

To prepare 0.8% agarose gel to check genomic DNA integrity:

1. Measure 0.8 g of agarose
2. Mix agarose powder with 100 ml of 1X Tris-borate-EDTA (TBE) buffer.
3. Heat till the mixture becomes transparent.
4. Add 4 µl of ethidium bromide when the temperature of the mixture reaches ~ 55° C.
5. Pour the agarose into the gel tray with the well comb in place.
6. After polymerisation of the gel, remove the comb and additional parts. Any bubbles can be pushed towards the sides/edges of the gel with a pipette tip.
7. Let the newly poured gel sit at room temperature for 20-30 mins, until it has completely solidified.
8. Once solidified, place the agarose gel into the gel box (electrophoresis unit).
9. Add loading buffer to the DNA samples. The function of loading buffer is twofold. Firstly, it adds a visible dye which aids in gel loading and allows for an estimation of DNA migration distance. Secondly, it has a high glycerol content which enhances the density of the DNA sample, preventing it from diffusing in the buffer and causing it to settle at the bottom of the gel well.
10. Fill gel box with 1X TBE buffer until the gel is covered.
11. Carefully load 5 µl of the samples into subsequent wells of the gel.
12. Start the device and run the gel at Voltage 300, Current 50 for 45 minutes.
13. Turn OFF power, disconnect the electrodes from the power source, and then carefully remove the gel from the gel box.
14. Visualize the DNA integrity under UV light.

Appendix 7 - Polymerase chain reaction protocol

A 20 µl PCR reaction volume was prepared using Maxime™ PCR PreMix (iNtRON Biotechnology, South Korea) with ratios following the manufacturer protocol demonstrated in **Table_Apx 4**.

Table_Apx 4 *PCR reaction volume.*

Reagent	**1x (in µl)**
H2O	17
10 µM of forward primer	1
10 µM of reverse primer	1
Template DNA	1

We amplified DNA using touch-down PCR settings detailed in **Table_Apx 5** below.

Table_Apx 5 *Touchdown PCR program.*

Cycles	**Phase**	**Time**	**Temperature (°C)**
	Initial denaturation	10 minutes	95
10x	Denaturation	30 seconds	95
	Annealing*	30 seconds	65-55
	Extension	1 minute	72
30x	Denaturation	30 seconds	95
	Annealing	30 seconds	55
	Extension	1 minute	72
	Final extension	10 minutes	72

*Annealing temperature (Ta) is reduced by 1.0°C each cycle.

Appendix 8 - Functional mapping and annotation of genome-wide association studies (FUMA) parameters

Table_Apx 6 The parameters used to perform FUMA functional annotation.

Parameter	Value
Lead SNP P-value	1.00E-05
P-value cutoff	0.05
R2 threshold for independent significant SNPs	0.6
Second R2 threshold	0.1
Reference panel population	1KG/Phase3 African
Non-GWAS tagged SNPs	yes
Minor allele frequency	0
Maximum distance between LD blocks to merge into a locus	250kb
Gene mapping	Positional; eQTL; chromatin interactions
Gene type	protein_coding
Distance from gene start/stop site to map SNPs	10kb
MHS region excluded from annotation	yes
MAGMA gene expression analysis repository	GTEx v.8

Appendix 9 - Ethnic groups of mycetoma cases recruited in the GWAS

Table_Apx 7 Ethnic groups of mycetoma cases recruited.

Ethnicity	n	%	Linguistic family	Geographic group
Agalyen	3	0.97	Afro-Asiatic	Central
Ahamda	3	0.97	Afro-Asiatic	Central
Arab Awlad Got	1	0.32	Afro-Asiatic	Western
Araki	1	0.32	Afro-Asiatic	Central
Aringa	2	0.65	Nilo-Saharan	Western
Awamra	2	0.65	Afro-Asiatic	Central
Baggara	17	5.5	Afro-Asiatic	Western
Balanga	1	0.32	Afro-Asiatic	Western
Bani Fadul	1	0.32	Afro-Asiatic	Western
Bani Garar	2	0.65	Afro-Asiatic	Western
Bani Tameem	1	0.32	Afro-Asiatic	Western
Bardo	1	0.32	Nilo-Saharan	Western
Bargo	18	5.83	Nilo-Saharan	Western
Barno	1	0.32	Afro-Asiatic	Western
Barti	1	0.32	Nilo-Saharan	Western
Batahen	5	1.62	Afro-Asiatic	Central
Bederia	7	2.27	Afro-Asiatic	Central and west
Beja	7	2.27	Afro-Asiatic	Eastern
Berged	1	0.32	Nilo-Saharan	Western
Bilala	1	0.32	Nilo-Saharan	Western
Bozaa	3	0.97	Afro-Asiatic	Western
Dar Hamed	5	1.62	Afro-Asiatic	Western
Fadnya	1	0.32	Afro-Asiatic	Central
Falata	4	1.29	Afro-Asiatic	Central
Fong	1	0.32	Nilo-Saharan	Central
Gamoeya	18	5.83	Afro-Asiatic	Central
Garabi	1	0.32	Nilo-Saharan	Western
Gemer	1	0.32	Nilo-Saharan	Western
Gorane	1	0.32	Nilo-Saharan	Western
Halawi	2	0.65	Afro-Asiatic	Central

Halfa	1	0.32	Nilo-Saharan	Northern
Hamar	7	2.27	Afro-Asiatic	Western
Hassania	14	4.53	Afro-Asiatic	Central
Hawari	1	0.32	Afro-Asiatic	Western
Hawsa	7	2.27	Afro-Asiatic	Western
Hosonati	1	0.32	Afro-Asiatic	Central
Hotia	1	0.32	Afro-Asiatic	Western
Jaalyeen	13	4.21	Afro-Asiatic	Northern
Kababesh	3	0.97	Afro-Asiatic	Western
Kawahla	52	16.8	Afro-Asiatic	Central
Khawalda	1	0.32	Afro-Asiatic	Central
Kurtan	1	0.32	Afro-Asiatic	Southern
Maganen	1	0.32	Afro-Asiatic	Western
Magarba	1	0.32	Afro-Asiatic	Central
Magdiya	1	0.32	Afro-Asiatic	Southern
Maghata	1	0.32	Afro-Asiatic	Central
Mahas	1	0.32	Afro-Asiatic	Northern
Manaser	3	0.97	Afro-Asiatic	Northern
Mararet	6	1.94	Nilo-Saharan	Western
Mema	1	0.32	Nilo-Saharan	Western
Mosalamya	4	1.29	Afro-Asiatic	Central
Nefedya	1	0.32	Afro-Asiatic	Central
Nuba	4	1.29	Nilo-Saharan and Niger-Congo	Western
Rekabya	2	0.65	Afro-Asiatic	Northern
Rezegat	4	1.29	Afro-Asiatic	Western
Robatab	1	0.32	Afro-Asiatic	Northern
Rofaa	8	2.59	Afro-Asiatic	Central
Rotamat	1	0.32	Afro-Asiatic	Central
Sadarna	1	0.32	Afro-Asiatic	Central
Shaigia	3	0.97	Afro-Asiatic	Northern
Shaila	1	0.32	Afro-Asiatic	Northern
Shanabla	2	0.65	Afro-Asiatic	Western
Shankhab	4	1.29	Afro-Asiatic	Southern
Shargawi	1	0.32	Afro-Asiatic	Northern

Shokria	11	3.56	Afro-Asiatic	Central
Tama	12	3.88	Nilo-Saharan	Western
Tongor	1	0.32	Nilo-Saharan	Western
Zaghawa	9	2.91	Nilo-Saharan	Western
NA	9	2.91		

Appendix 10 - Ethnic groups of control subjects recruited in the GWAS

Table_Apx 8 Ethnic groups of the recruited controls.

Ethnicity	n	%	Linguistic family	Geographic group
Abo Hashim	1	0.65	Afro-Asiatic	Central
Awaida	1	0.65	Afro-Asiatic	Central
Awamra	1	0.65	Afro-Asiatic	Central
Baggara	2	1.29	Afro-Asiatic	Western
Bederia	2	1.29	Afro-Asiatic	Central and west
Beja	2	1.29	Afro-Asiatic	Eastern
Esenawi	1	0.65	Nilo-Saharan	Northern
Fur	1	0.65	Nilo-Saharan	Western
Gamoeya	2	1.29	Afro-Asiatic	Central
Gemer	1	0.65	Nilo-Saharan	Western
Hadiya	1	0.65	Afro-Asiatic	Eastern
Hassania	3	1.94	Afro-Asiatic	Central
Hodor	3	1.94	Afro-Asiatic	Northern
Jaalyeen	7	4.52	Afro-Asiatic	Northern
Kawahla	62	40	Afro-Asiatic	Central
Lahawi	42	27.1	Afro-Asiatic	Eastern and Central
Magarba	2	1.29	Afro-Asiatic	Central
Magdiya	2	1.29	Afro-Asiatic	Southern
Mohamadi	1	0.65	Afro-Asiatic	Central
Nefedya	7	4.52	Afro-Asiatic	Central
Rezegat	3	1.94	Afro-Asiatic	Western
Rofaa	3	1.94	Afro-Asiatic	Central
Sadarna	1	0.65	Afro-Asiatic	Central
Shokria	1	0.65	Afro-Asiatic	Central
Tama	1	0.65	Nilo-Saharan	Western
Zaghawa	1	0.65	Nilo-Saharan	Western
NA	11	6.67		

Appendix 11 - Mapped genes from GCTA's summary statistics

Table_Apx 9 The eighty-seven mycetoma-associated candidate genes identified by positional mapping, eQTL mapping and chromatin interaction mapping strategies implemented in FUMA.

Symbol	Chr	start	end	Top SNP(s)
CNN3	1	95362507	95392834	rs167677
ALG14	1	95439963	95538501	rs167677
RWDD3	1	95699711	95712781	rs167677
RP11-147C23.1	1	96403457	96488436	rs167677
PTBP2	1	97187221	97289294	rs167677
DPYD	1	97543299	98386605	rs167677
SNX7	1	99127236	99226056	rs167677
GCSAM	3	1.12E+08	1.12E+08	rs16861260
SLC9C1	3	1.12E+08	1.12E+08	rs16861260
CD200	3	1.12E+08	1.12E+08	rs16861260
BTLA	3	1.12E+08	1.12E+08	rs16861260
ATG3	3	1.12E+08	1.12E+08	rs16861260
SLC35A5	3	1.12E+08	1.12E+08	rs16861260
CCDC80	3	1.12E+08	1.12E+08	rs16861260
CD200R1L	3	1.13E+08	1.13E+08	rs16861260
CD200R1	3	1.13E+08	1.13E+08	rs16861260
GTPBP8	3	1.13E+08	1.13E+08	rs16861260
C3orf17	3	1.13E+08	1.13E+08	rs16861260
BOC	3	1.13E+08	1.13E+08	rs16861260
WDR52	3	1.13E+08	1.13E+08	rs16861260
SPICE1	3	1.13E+08	1.13E+08	rs16861260
SIDT1	3	1.13E+08	1.13E+08	rs16861260
KIAA2018	3	1.13E+08	1.13E+08	rs16861260
NAA50	3	1.13E+08	1.13E+08	rs16861260
ATP6V1A	3	1.13E+08	1.14E+08	rs16861260
QTRTD1	3	1.14E+08	1.14E+08	rs16861260
DRD3	3	1.14E+08	1.14E+08	rs16861260
SLC30A5	5	68389473	68426896	rs9291949
CCNB1	5	68462837	68474072	rs9291949
CENPH	5	68485375	68506184	rs9291949

MRPS36	5	68513587	68525956	rs9291949
CDK7	5	68530668	68573250	rs9291949
CCDC125	5	68576002	68628636	rs9291949
TAF9	5	68646811	68665840	rs9291949
RAD17	5	68665120	68710628	rs9291949
MARVELD2	5	68710939	68740157	rs9291949
OCLN	5	68788119	68853931	rs9291949
GTF2H2C	5	68856035	68890550	rs9291949
SMN2	5	69345350	69374349	rs9291949
CARTPT	5	71014990	71016875	rs9291949
ADAM22	7	87563458	87832204	rs10952971
SRI	7	87834433	87856308	rs10952971
STEAP4	7	87905744	87936206	rs10952971
ZNF804B	7	88388682	88966346	rs10952971
C7orf62	7	88423420	88425035	rs10952971
GTPBP10	7	89964537	90020769	rs10952971
TRRAP	7	98475556	98610866	rs45529834
ARPC1A	7	98923521	98985787	rs45529834
ARPC1B	7	98971872	98992424	rs45529834
PDAP1	7	98989671	99006452	rs45529834
BUD31	7	99006264	99017239	rs45529834
PTCD1	7	99014362	99063787	rs45529834
ATP5J2-PTCD1	7	99017372	99063820	rs45529834
CPSF4	7	99036545	99054994	rs45529834
AC073063.1	7	99040552	99040747	rs45529834
ATP5J2	7	99046098	99063954	rs45529834
ZNF789	7	99070464	99101273	rs45529834
ZNF394	7	99084142	99097947	rs45529834
ZKSCAN5	7	99102274	99132323	rs45529834
FAM200A	7	99143931	99156159	rs45529834
ZNF655	7	99156029	99174076	rs45529834
GS1-259H13.10	7	99156291	99205433	rs45529834
ZSCAN25	7	99214569	99230030	rs45529834
CYP3A5	7	99245817	99277621	rs45529834
CYP3A7	7	99302660	99332819	rs45529834

CYP3A4	7	99354604	99381888	rs45529834
CYP3A43	7	99425636	99463718	rs45529834
OR2AE1	7	99473610	99474680	rs45529834
TRIM4	7	99474581	99517223	rs45529834
GJC3	7	99520892	99527243	rs45529834
AZGP1	7	99564343	99573780	rs45529834
ZKSCAN1	7	99613204	99639312	rs45529834
ZNF3	7	99661656	99680171	rs45529834
COPS6	7	99686577	99689823	rs45529834
MCM7	7	99690351	99699563	rs45529834
AP4M1	7	99699172	99707968	rs45529834
TAF6	7	99704693	99717464	rs45529834
CNPY4	7	99717236	99723134	rs45529834
GATS	7	99798283	99869855	rs45529834
PLOD3	7	1.01E+08	1.01E+08	rs45529834
ZNHIT1	7	1.01E+08	1.01E+08	rs45529834
DPH6	15	35509546	35838394	rs319883; rs10520048
C15orf41	15	36871812	37102449	rs319883; rs10520048
MEIS2	15	37181406	37393504	rs319883; rs10520048
TMCO5A	15	38214140	38259925	rs319883; rs10520048
FAM98B	15	38746328	38779911	rs319883
RASGRP1	15	38780304	38857776	rs319883

Appendix 12 - Genes reported in GWAS catalogue.

Table_Apx 10 GWAS catalogue previously reported genes.

Gene set	**genes**	***p*-value**	**Adjusted-*p***
Facial emotion recognition (sad faces)	ARPC1A, ARPC1B, PDAP1, BUD31, CPSF4, ZNF789, ZNF394, ZKSCAN5	2.61E-15	1.16E-11
Tacrolimus trough concentration in kidney transplant patients	CPSF4, ZNF789, ZKSCAN5, CYP3A4, CYP3A43, TRIM4	6.56E-15	1.45E-11
Ticagrelor levels in individuals with acute coronary syndromes treated with ticagrelor	ARPC1A, CYP3A5, CYP3A4, CYP3A43, ZKSCAN1	3.26E-11	4.81E-8
Dehydroepiandrosterone sulphate levels	ARPC1A, ZKSCAN5, CYP3A43, TRIM4	7.58E-8	8.38E-5
Serum metabolite levels	DPYD, ARPC1A, ARPC1B, PDAP1, BUD31, CPSF4, ZNF789, ZKSCAN5, CYP3A5, CYP3A43, TRIM4	1.87E-7	1.66E-4
Sex hormone levels	ZNF789, ZKSCAN5, CYP3A4	7.05E-6	5.20E-3
Palmitic acid (16:0) levels	CNN3, ALG14, RWDD3	2.37E-5	1.31E-2
Stearic acid (18:0) levels	CNN3, ALG14, RWDD3	2.37E-5	1.31E-2
Restless legs syndrome	STEAP4, ZNF804B, DPH6, MEIS2	3.62E-5	1.78E-2

Appendix 13 - Mapped genes from SAIGE's summary statistics

Table_Apx 11 Forty-eight mycetoma-associated candidate genes identified by positional mapping, eQTL mapping and chromatin interaction mapping strategies implemented in FUMA.

Symbol	Chr	Start	End	Top SNP(s)
RAB1A	2	65297835	65357240	rs1868403
ACTR2	2	65454887	65498387	rs1868403
SPRED2	2	65537985	65659771	rs1868403
MEIS1	2	66660584	66801001	rs1868403
ETAA1	2	67624451	67637677	rs1868403
C1D	2	68268262	68338080	rs1868403
WDR92	2	68350068	68384692	rs1868403
RP11-474G23.1	2	68358370	68488362	rs1868403
PPP3R1	2	68405989	68483369	rs1868403
GCSAM	3	1.12E+08	1.12E+08	rs16861260
SLC9C1	3	1.12E+08	1.12E+08	rs16861260
CD200	3	1.12E+08	1.12E+08	rs16861260
BTLA	3	1.12E+08	1.12E+08	rs16861260
ATG3	3	1.12E+08	1.12E+08	rs16861260
SLC35A5	3	1.12E+08	1.12E+08	rs16861260
CCDC80	3	1.12E+08	1.12E+08	rs16861260
CD200R1L	3	1.13E+08	1.13E+08	rs16861260
CD200R1	3	1.13E+08	1.13E+08	rs16861260
GTPBP8	3	1.13E+08	1.13E+08	rs16861260
C3orf17	3	1.13E+08	1.13E+08	rs16861260
BOC	3	1.13E+08	1.13E+08	rs16861260
WDR52	3	1.13E+08	1.13E+08	rs16861260
SPICE1	3	1.13E+08	1.13E+08	rs16861260
SIDT1	3	1.13E+08	1.13E+08	rs16861260
KIAA2018	3	1.13E+08	1.13E+08	rs16861260
NAA50	3	1.13E+08	1.13E+08	rs16861260
ATP6V1A	3	1.13E+08	1.14E+08	rs16861260
QTRTD1	3	1.14E+08	1.14E+08	rs16861260
DRD3	3	1.14E+08	1.14E+08	rs16861260

GPM6A	4	1.77E+08	1.77E+08	rs17062162
PRKAR1B	7	588834	767287	rs4076072
SUN1	7	855528	936072	rs4076072
GET4	7	916189	936073	rs4076072
ADAP1	7	937540	995043	rs4076072
COX19	7	938415	1015235	rs4076072
GPR146	7	1084212	1098897	rs4076072
GAL	11	68451247	68458643	rs3019751
TPCN2	11	68816365	68858072	rs3019751
MYEOV	11	69061605	69182494	rs3019751
CCND1	11	69455855	69469242	rs3019751
ORAOV1	11	69467844	69490184	rs3019751
FGF19	11	69513000	69519410	rs3019751
DPH6	15	35509546	35838394	rs319883; rs10520048
C15orf41	15	36871812	37102449	rs319883; rs10520048
MEIS2	15	37181406	37393504	rs319883; rs10520048
TMCO5A	15	38214140	38259925	rs319883; rs10520048
FAM98B	15	38746328	38779911	rs319883
RASGRP1	15	38780304	38857776	rs319883

Copyright permission for the published review article

This paper is published in PLOS NTDs which indicate that journal allows reuse of publications:

‘Reuse of PLOS Article Content

PLOS applies the Creative Commons Attribution 4.0 International (CC BY) license to articles and other works we publish. If you submit your paper for publication by PLOS, you agree to have the CC BY license applied to your work. Under this Open Access license, you as the author agree that anyone can reuse your article content in whole or part for any purpose, for free, even for commercial purposes. These permitted uses include but are not limited to self-archiving by authors of submitted, accepted and published versions of their papers in institutional repositories. Anyone may copy, redistribute, reuse, or modify the content as long as the author and original source are properly cited. This facilitates freedom in reuse and also ensures that PLOS content can be mined without barriers for the needs of research.’

Link: https://journals.plos.org/plosone/s/licenses-and-copyright