

UNRAVELING THE GENETIC LANDSCAPE OF MACROCEPHALY: A BIOINFORMATICS APPROACH TO CAUSAL VARIANTS & BIOLOGICAL PATHWAYS

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Science in Bioinformatics and Data Science

BY

K.HAFSA FATHIMA (41738004)



**DEPARTMENT OF BIOINFORMATICS
SCHOOL OF BIO AND CHEMICAL ENGINEERING**

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY (DEEMED TO BE UNIVERSITY)

Accredited with Grade "A++" by NAAC | Category-1 University by UGC |

Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai- 600119

APRIL-2024



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A++" by NAACI Category-1 University by UGC Approved by AICTE
JeppiaarNagar, Rajiv Gandhi Salai, Chennai-600119

www.sathyabama.ac.in

CERTIFICATE

This is to certify that the dissertation titled **"UNRAVELING THE GENETIC LANDSCAPE OF MACROCEPHALY: A BIOINFORMATICS APPROACH TO CAUSAL VARIANTS & BIOLOGICAL PATHWAYS"** was carried out by **MS. HAFSA FATHIMA.K (Reg No.41738004)**, Department of Bioinformatics and Data Science, Sathyabama Institute of Science and Technology, Chennai in partial fulfillment of the requirements of the award of the **Degree of Bachelor of Science in Bioinformatics and Data Science 2021-2024**, is a bonafide record of the project work done under my guidance and supervision. This thesis or part, therefore, has not been submitted elsewhere for any other degree or diploma.

Date: 27/04/2024

Place: CHENNAI

A handwritten signature in blue ink, appearing to read 'Daniel Alex Anand', is written over a horizontal line.

Dr. Daniel Alex Anand

ACKNOWLEDGEMENTS

I am delighted to express my heartfelt acknowledgement to the **BOARD OF MANAGEMENT OF SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement and support throughout the duration of my project. Their guidance has been invaluable in ensuring the successful completion of this project, and I am sincerely grateful for their contribution.

I convey my thanks to **Dr. Rekha Chakravarthy, Dean, School of Science and Humanities** for providing me necessary support and details at the right time during the progressive reviews.

I sincerely thank **Dr. H. Jemmy Christy, Head of the Department, Dept. of Bioinformatics**, for her unwavering support and timely provision of necessary details during the progressive reviews. Her support has been instrumental in helping me to overcome obstacles and achieve success in this project.

I am deeply indebted to my **Project Guide, Dr. Daniel Alex Anand**, for his invaluable guidance, suggestions, and constant encouragement throughout the project. His expertise and insights have been instrumental in shaping my work and ensuring its successful completion.

I also wish to extend my gratitude to my family for their support and assistance in various ways, which have contributed to the successful completion of this project.

HAFSA FATHIMA.K



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

**Accredited with Grade "A++" by NAACI Category-1 University by UGC Approved by AICTE
Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai- 600119**

www.sathyabama.ac.in

DEPARTMENT OF BIOINFORMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **HAFSA FATHIMA.K (41738004)**, who carried out the final year project entitled "**UNRAVELING THE GENETIC LANDSCAPE OF MACROCEPHALY: A BIOINFORMATICS APPROACH TO CAUSAL VARIANTS & BIOLOGICAL PATHWAYS**" under our supervision from January 2024 to April 2024.

A handwritten signature in blue ink, appearing to read 'Daniel Alex Anand', with a long horizontal stroke extending to the right.

Dr. DANIEL ALEX ANAND

Internal Guide

A handwritten signature in blue ink, appearing to read 'H. Jemmy Christy', with a long horizontal stroke extending to the right.

Dr. H. JEMMY CHRISTY

Head of the Department

Submitted for Viva Voce Examination held on 27/04/2024

A handwritten signature in blue ink, appearing to read 'Internal Examiner', with a long horizontal stroke extending to the right.

Internal Examiner

A handwritten signature in blue ink, appearing to read 'External Examiner', with a long horizontal stroke extending to the right.

External Examiner

DECLARATION

I, **HAFSA FATHIMA.K**, hereby declare that the Project Report entitled **UNRAVELING THE GENETIC LANDSCAPE OF MACROCEPHALY: A BIOINFORMATICS APPROACH TO CAUSAL VARIANTS & BIOLOGICAL PATHWAYS** done by me under the guidance of **Dr. Alex Anand** at Sathyabama Institute of Science and Technology, Jeppiaar Nagar, Rajiv Gandhi Road, Chennai-600119, Tamil Nadu, India, is submitted in partial fulfillment of the requirements for the award of **Bachelor of Science degree in Bioinformatics and Data Science**.

DATE: 27/04/2024

PLACE: CHENNAI

K. Hafsa Fathima

SIGNATURE OF THE CANDIDATE

ABSTRACT

Macrocephaly, characterised by an abnormally enlarged head circumference, presents a complex and heterogeneous phenotype with diverse underlying genetic etiologies. This thesis employs a bioinformatics approach to unravel the genetic landscape of Macrocephaly, exploring causal variants and biological pathways implicated in its development. Understanding the genetic landscape of macrocephaly is crucial for elucidating its molecular mechanisms, facilitating accurate diagnosis, and guiding therapeutic interventions. This study proposes a bioinformatics-driven approach to unravel the causal variants and biological pathways associated with macrocephaly. Leveraging advanced computational tools, large-scale genomic datasets, and integrative analysis techniques, we aim to identify and prioritise candidate variants implicated in macrocephaly across diverse populations. Through meticulous data analysis using Python libraries like Pandas, Seaborn, and Matplotlib for distribution analysis, and databases such as DisGeNET for gene-disease associations and pathway enrichment, the research prioritises candidate genes and elucidates biological pathways linked to macrocephaly. The study also explores pathways that may impact the condition and the role of genes like PTEN, providing insights into gene-disease relationships. It also explores comparative and correlation analysis. Furthermore, integrative analyses using pathway enrichment techniques and functional genomics are utilised to identify the underlying molecular pathways causing this illness. Genetic variants play a pivotal role in the diverse spectrum of human phenotypes and diseases, offering insights into molecular mechanisms and potential therapeutic targets. The thesis successfully leverages genome-wide association studies (GWAS) and whole-exome sequencing (WES) to deepen the understanding of macrocephaly's genetic underpinnings. Key pathways such as Ras/MAPK and PTEN/PI3K/AKT/mTOR have been pinpointed, identifying specific gene variants critical to the condition. Statistical and comparative analyses have highlighted genetic relationships and differences with related neurological disorders. This research paves the way for enhanced diagnostic and therapeutic approaches, improving outcomes for those with macrocephaly.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	i
	LIST OF ABBREVIATIONS	iv
	LIST OF TABLES	v
	LIST OF FIGURES	vi
1	INTRODUCTION	
	1.1 MACROCEPHALY	1
	1.2 HISTORY OF MACROCEPHALY	2
	1.2.1 Ancient and Mediaeval Observations	3
	1.2.2 19th Century Developments	3
	1.2.3 Early 20th Century Advances	3
	1.2.4 Mid-20th Century: The Rise of Genetics and Neurology	3
	1.2.5 Late 20th and Early 21st Century: Neuroimaging Revolution	3
	1.2.6 Contemporary Understanding	4
	1.3 EPIDEMIOLOGY	4
	1.3.1 Prevalence	4
	1.3.2 Age of Onset	4
	1.3.3 Gender Distribution	5
	1.3.4 Geographic and Ethnic Variations	5
	1.3.5 Risk Factors	5
	1.3.6 Associated Conditions	5
	1.3.7 Outcomes and Prognosis	6
	1.4 PATHOPHYSIOLOGY	6
	1.4.1 Hydrocephalus	6
	1.4.2 Nervous System Infections	7
	1.4.3 Rickets	7
	1.4.4 Brain Disorders	7
	1.4.5 Congenital Brain Deformities	7

	1.4.6 Genetic Disorders	7
	1.5 TREATMENT	8
	1.6 SYMPTOMS	9
	1.7 NEUROIMAGING FOR MACROCEPHALY	9
	1.8 EXISTING PROBLEMS	9
	1.9 BOTTLENECK AREAS	9
	1.10 COMPUTATIONAL TOOLS FOR GENETIC ANALYSIS	10
	1.10.1 Research Objectives and Significance	10
	1.10.2 ANALYSIS	11
2	REVIEW OF LITERATURE	12
3	AIM AND OBJECTIVE	
	3.1 AIM	16
	3.2 OBJECTIVES	16
4	METHODOLOGY	
	4.1 DATABASES AND TOOLS EMPLOYED	17
	4.1.1 DisGeNET	17
	4.1.2 PYTHON	18
	4.1.3 PANDAS	18
	4.1.4 SEABORN	18
	4.1.5 MATPLOTLIB	19
	4.1.6 PYPLOT	19
	4.2 ANALYSIS	19
	4.2.1 Distribution Analysis	20
	4.2.2 Comparative Analysis	20
	4.2.3 Correlation Analysis	20
	4.2.4 Exploratory Analysis	20
	4.2.5 Variant-Gene Association Analysis	20
	4.2.6 Disease Specificity and Impact Analysis	21
5	RESULTS AND DISCUSSION	38
6	SUMMARY AND CONCLUSION	52
	REFERENCE	56

LIST OF ABBREVIATIONS

SDH	Subdural Hematoma
SBS	Shaken Baby Syndrome
GWAS	Genome-Wide Association Studies
WES	Whole-Exome Sequencing
DPI	Disease Pleiotropy Index
DSI	Disease Specificity Index
Jlg	Jaccard Index for genes
Jlv	Jaccard Index for variants
DisGeNET	Disease Genomic Network
MRI	Magnetic resonance imaging
CT	Computed Tomography Scan
NumPy	Numerical Python
SciPy	Scientific Python
OMIM	Online Mendelian Inheritance in Man
GEO	Gene Expression Omnibus

LIST OF TABLES

TABLE NO	TITLE	PAGE NO
4.1	C0221355_disease_dda_summary	21
4.2	C0221355_disease_vda_summary	24

LIST OF FIGURES

FIGURE NO	TITLE	PAGE NO
1.1	Macrocephaly	2
1.2	Cause of Macrocephaly	8
4.1	Distribution Analysis of Gene and Variant Metrics Input	24
4.2	Summary Statistics Input	25
4.3	Data Visualization and Distribution Analysis Input	26
4.4	Comparative Analysis: Jaccard Index and P-Values Input	27
4.5	Variant-Gene Association Analysis Input	29
4.6	Disease Specificity and Impact Analysis Input	30
4.7	Distribution of Disease Specificity Index (DSI) and Disease Pleiotropy Index (DPI) Input	32
4.8	Chromosomal Location Analysis Input	34
4.9	z-score calculation Input	36
5.1	Distribution Analysis of Gene and Variant Metrics Output	38
5.2	Summary Statistics Output	38
5.3	Data Visualization and Distribution Analysis Output	39

5.4	Comparative Analysis: Jaccard Index and P-Values Output	41
5.5	Variant-Gene Association Analysis Output	43
5.6	Disease Specificity and Impact Analysis Output	45
5.7	Distribution of Disease Specificity Index (DSI) and Disease Pleiotropy Index (DPI) Output	46
5.8	Chromosomal Location Analysis Output	48
5.9	Correlation Matrix of Jaccard Indices and p-value Output	49
5.10	Z- score Output	50

CHAPTER 1

INTRODUCTION

1.1 MACROCEPHALY

Macrocephaly, characterised by a head circumference that exceeds the 97.7th percentile for a person's age and gender, presents a complex challenge in paediatric and adult neurology (Rohde et al., 2023). It is deemed normal if the circumference is within the range that is healthy for their age and gender. This condition, often detected through routine head measurements in infants, children, and adults, can be benign or indicative of underlying pathologies requiring comprehensive management strategies. Macrocephaly, characterised by an enlarged head circumference—greater than two standard deviations above the mean for a person's age and gender (Pavone et al., 2017)—is a condition that presents both diagnostic challenges and insights into the genetic architecture of human development.

Megalencephaly is the term used to describe the increased growth of cerebral structures associated with dysfunctional anomalies throughout several stages of brain development, such as the phases of neuronal migration and/or proliferation, or as a result of abnormal events that occurred after birth (Pavone et al., 2017).

The variability in its aetiology, ranging from benign familial traits to associations with serious neurological conditions, necessitates a deeper understanding of its genetic underpinnings. The weight of the newborn brain is estimated to be around 370 g, and it grows by roughly 4 times during infancy and youth to reach an adult's weight of roughly 1500 g. (Pavone et al., 2017).

The genetic underpinnings of Macrocephaly, while complex, offer a fertile ground for exploration, with recent advances in bioinformatics providing new tools to unravel these complexities. Neurodevelopmental disorders associated with abnormal brain growth with or without cortical malformations are important causes of morbidity and mortality with multiple neurodevelopmental consequences including intellectual disability (ID), autism spectrum disorders (ASD), and epilepsy typically manifesting in infancy or early childhood (Pirozzi et al., 2018).

It is a condition not only of cosmetic concern but also a potential indicator of underlying neurological and genetic disorders. The increased head circumference associated with macrocephaly is associated with a number of conditions that can cause an increase in orbito-frontal head circumference for an individual's age, such as intracranial masses, hydrocephalus, abnormalities of the bone skull structures, and arteriovenous malformations. The journey to understanding Macrocephaly traverses through the intricate landscapes of genetics, developmental biology, and personalised medicine, necessitating a multidisciplinary approach for its diagnosis, management, and treatment (Accogli et al.,2022).

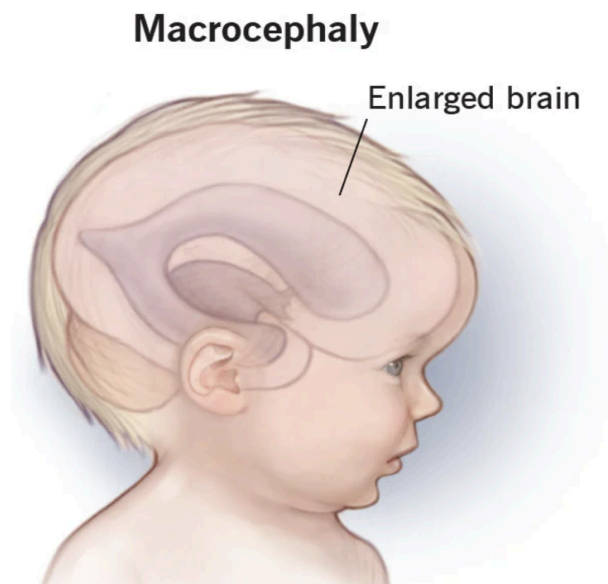


Fig 1.1 MACROCEPHALY

1.2 HISTORY OF MACROCEPHALY

Macrocephaly, defined as an abnormally large head size, has been recognized and studied throughout medical history. Its understanding has evolved significantly over the years as advances in medicine, genetics, and neuroimaging have deepened our knowledge of its causes and implications (DeCasien et al, 2022). Here's a brief overview of the history of macrocephaly:

1.2.1 Ancient and Mediaeval Observations

Historically, variations in head size were noted in ancient medical texts, though the understanding of such conditions was rudimentary. In ancient times, unusually large heads were often noted in royal and noble lineages, possibly because these were the individuals whose physical attributes were recorded in detail.

1.2.2 19th Century Developments

The scientific study of macrocephaly began to take shape in the 19th century with the advent of systematic medical observation and the development of anthropology as a science. Phrenology, the study of skull shapes as a supposed indication of mental faculties and character, though now discredited, was one of the first attempts to link cranial measurements to pathology and personality traits. (DeCasien et al, 2022)

1.2.3 Early 20th Century Advances

By the early 20th century, with the development of paediatric medicine as a specialised field, the measurement of head circumference became a routine part of child health assessments. This allowed for the early detection of abnormalities in head size. Conditions leading to macrocephaly, such as hydrocephalus, began to be better understood, particularly with the advent of imaging techniques like X-ray. (Accogli et al., 2022).

1.2.4 Mid-20th Century: The Rise of Genetics and Neurology

The mid-20th century saw significant advancements in genetics and neurology, which contributed to our understanding of macrocephaly. The discovery of genetic causes of diseases that include macrocephaly, such as certain metabolic disorders and syndromic conditions (like Sotos syndrome and Weaver syndrome), highlighted the role of inherited factors in abnormal head growth.

1.2.5 Late 20th and Early 21st Century: Neuroimaging Revolution

The development of advanced neuroimaging techniques such as computed tomography (CT) scans in the 1970s and magnetic resonance imaging (MRI) in the 1980s revolutionised the diagnosis and understanding of macrocephaly. These

tools allowed for detailed visualisation of the brain structure, enabling the identification of structural abnormalities, tumours, and other brain disorders associated with increased head size.

1.2.6 Contemporary Understanding

Today, macrocephaly is understood as a symptom that can arise from a variety of conditions, including benign familial macrocephaly, where it is an isolated trait without associated harm, to more severe conditions involving brain overgrowth or accumulation of fluids. Genetic testing and advanced imaging techniques have improved the diagnostic process, allowing for targeted interventions and better management of associated conditions.

1.3 EPIDEMIOLOGY

Macrocephaly, characterised by an abnormally large head circumference, is a condition that can arise from a variety of etiologies, including genetic factors, developmental issues, and certain medical conditions (Shi et al., 2020). Here's a detailed overview of the epidemiology of macrocephaly based on available research and literature:

1.3.1 Prevalence

The prevalence of macrocephaly varies widely in the general population, depending on the criteria used for diagnosis (often defined as a head circumference greater than two standard deviations above the mean for a given age and sex). According to various studies, the prevalence can range from approximately 0.5% to 2% in children. It's important to note that macrocephaly can be a normal variation in head size within certain families (benign familial macrocephaly).

1.3.2 Age of Onset

Macrocephaly is most commonly identified in infancy and early childhood when routine measurements of head circumference are taken during well-child visits. Early detection is crucial for determining the underlying cause and initiating appropriate management strategies (St John et al., 2021).

1.3.3 Gender Distribution

There is no clear gender predisposition for macrocephaly in general; however, some specific causes of macrocephaly, such as certain genetic syndromes, may have a gender bias depending on the nature of the genetic inheritance (e.g., X-linked disorders).

1.3.4 Geographic and Ethnic Variations

The incidence of macrocephaly can vary geographically and ethnically due to differences in genetic background, environmental factors, and access to healthcare services which influence early detection and reporting (Shi et al., 2020).

1.3.5 Risk Factors

Several risk factors are associated with macrocephaly, including:

- Genetic predisposition: Family history of macrocephaly or related genetic conditions can increase the likelihood of presenting this trait.
- Perinatal factors: Conditions such as intrauterine infections, complications during pregnancy, and birth injuries can contribute to the development of macrocephaly.
- Associated medical conditions: Neurocutaneous syndromes (e.g., neurofibromatosis), metabolic disorders, and overgrowth syndromes (e.g., Sotos syndrome) often include macrocephaly as a feature.

1.3.6 Associated Conditions

Macrocephaly may be an isolated finding or associated with other clinical conditions. When associated, it can be linked with:

- Developmental delays and neurodevelopmental disorders: Intellectual disability, autism spectrum disorders, and learning disabilities.
- Neurological conditions: Seizure disorders, cerebral palsy.
- Structural brain anomalies: Hydrocephalus, brain tumours, and other intracranial abnormalities.

1.3.7 Outcomes and Prognosis

The prognosis for individuals with macrocephaly largely depends on the underlying cause. Isolated macrocephaly with a benign familial background may have a normal developmental outcome, whereas macrocephaly associated with genetic or metabolic disorders may have varied prognoses depending on the specific disorder and the severity of associated complications.

Epidemiological studies on macrocephaly help in understanding its distribution and associated factors, which can guide healthcare providers in early detection and management, improving outcomes for affected individuals. Further research is needed to better understand the genetic and environmental contributors to macrocephaly and to improve diagnostic accuracy and treatment strategies.

1.4 PATHOPHYSIOLOGY

The disorder known as macrocephaly, which is characterised by an abnormally large head circumference, can have a complex pathophysiology because it might have a number of underlying causes. Atypically high brain tissue, cerebrospinal fluid (CSF), blood, or a combination of these components can cause an abnormal increase in the volume of the cerebral contents, which is the pathophysiology of macrocephaly (Jones et al.,2024) . By underlying cause, the following are some of the major pathophysiological mechanisms:

1.4.1 Hydrocephalus

Hydrocephalus is a condition characterised by an excessive accumulation of cerebrospinal fluid (CSF) within the ventricles of the brain, which leads to increased intracranial pressure and an enlarged head circumference. The pathophysiology involves either a blockage that impedes CSF flow, an overproduction of CSF, or inadequate absorption by the arachnoid villi. This condition is critical as, if left untreated, it may lead to significant brain damage. Treatment typically involves surgical intervention to redirect the CSF to another area of the body where it can be absorbed, commonly using a ventriculoperitoneal shunt.

1.4.2 Nervous System Infections

Infections like meningitis, which inflames the protective membranes covering the brain and spinal cord, can also lead to macrocephaly. These infections may result in the accumulation of fluids or swelling, contributing to an increased head size. The pathophysiology here involves the inflammatory response to infection, which increases the permeability of the blood-brain barrier, leading to fluid accumulation. Rapid treatment is essential to prevent permanent neurological damage (Miral et al., 2018).

1.4.3 Rickets

Rickets, a disorder caused by a deficiency in Vitamin D, calcium, or phosphate, leads to softening and weakening of bones in children, including those of the skull, potentially causing macrocephaly. The pathophysiology of rickets includes impaired bone mineralization and growth, leading to skeletal deformities as bones are unable to support the body's weight effectively.

1.4.4 Brain Disorders

Various brain disorders, including brain tumours or structural anomalies, can lead to an enlarged head size. These disorders can cause an increase in brain mass or fluid accumulation, both of which can contribute to macrocephaly.

1.4.5 Congenital Brain Deformities

Certain deformities of the brain present at birth, which can result from genetic mutations or developmental issues during gestation, may also cause macrocephaly. These congenital conditions can lead to abnormal brain growth patterns and head shape.

1.4.6 Genetic Disorders

Genetic conditions, such as Canavan disease or Alexander disease, are linked to macrocephaly. These disorders affect brain development and function, leading to an increase in brain and head size. They are typically associated with a spectrum of other neurological symptoms (Miral et al., 2018).

WHAT MAY CAUSE MACROCEPHALY (A Large Head) IN BABIES?

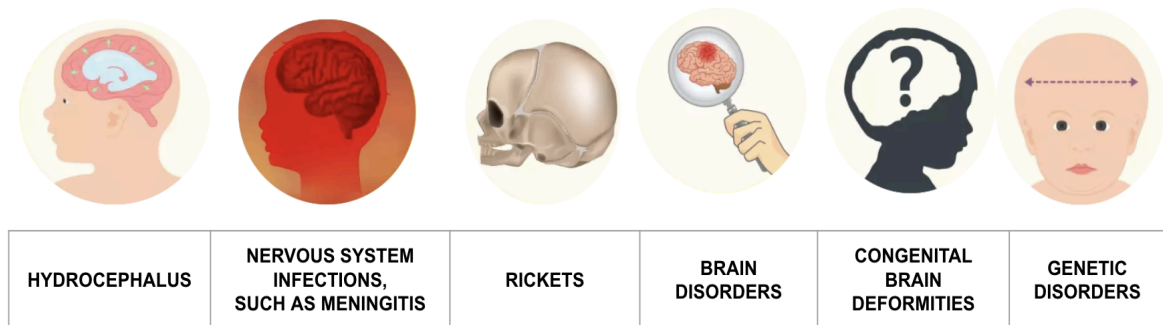


Fig 1.2 CAUSES OF MACROCEPHALY

1.5 TREATMENT

The cause of macrocephaly determines how the illness is treated. Choices consist of:

1. For many reasons of macrocephaly, surgery is frequently the only available treatment. The kind of surgery will change depending on the cause and degree of swelling. In the case of hydrocephalus, for example, the cerebrospinal fluid is drained into the belly so that the body can absorb it. A silicon tube is inserted into the abdominal cavity on one side and cerebral fluid on the other during a procedure known as shunt surgery. However, as hydrocephalus is a lifelong illness, it does not treat the problem. If brain trauma occurs, surgery might be able to fully resolve the issue (Fu et al .,2023).
2. To reduce the swelling of the brain tissue, the doctor could prescribe oral medicine. A baby can be treated for deficiency illnesses like rickets by receiving the recommended dosage of the nutrient.
3. There is no treatment for macrocephaly resulting from hereditary abnormalities; it can only be controlled. Many babies with irreversible macrocephaly experience physical and/or cognitive impairment in addition to issues like seizures.abrupt and aberrant spikes in brain activity that cause jerky motions of the body and unconsciousness (Rozovsky, K et al., 2012).

1.6 SYMPTOMS

If the child has benign familial macrocephaly, they will not exhibit any other symptoms other than an atypical head size. In the absence of them, your child may exhibit signs including fast head growth, poor appetite, developmental delays, bulging veins on the head, a downward stare in the eyes, and other disorders like autism.

1.7 NEUROIMAGING FOR MACROCEPHALY

To visualise and examine the many brain regions impacted by the condition, as well as to identify the anomaly causing the enlarged circumference, neuroimaging is necessary for the diagnosis of macrocephaly in children. Additionally, it is required to differentiate between megalencephaly, a condition in which brain enlargement exceeds twice the standard deviation, and macrocephaly (Tan AP et al., 2018).

1.8 EXISTING PROBLEMS

The diagnosis and management of Macrocephaly are fraught with complexities. The condition's heterogeneity, attributed to an array of genetic variants and their interactions with environmental factors, complicates the identification of causal pathways (Miller et al ., 2015).

Traditional diagnostic procedures, including medical history, neurological examination, and neuroimaging techniques, often fall short in pinpointing the specific genetic alterations driving Macrocephaly. Moreover, the lack of specificity in existing genetic tests underscores the need for a more targeted approach to understand the genetic landscape of Macrocephaly fully.

1.9 BOTTLENECK AREAS

The main bottlenecks in Macrocephaly research and management include the limited understanding of the genetic variants involved and their biological significance. Despite advancements in genomic technologies, the connection between specific genetic alterations and Macrocephaly remains elusive. The lack of comprehensive genetic catalogues and the challenges in interpreting vast

genomic data sets further complicate the path to a holistic understanding of Macrocephaly's genetic underpinnings.

1.10 COMPUTATIONAL TOOLS FOR GENETIC ANALYSIS

Macrocephaly is not merely a condition of an enlarged head size; it is a window into the complex interplay of genetic, developmental, and environmental factors that influence human growth patterns. This condition's heterogeneity necessitates a multifaceted approach for its study, making it an ideal candidate for genetic and bioinformatics research. Traditional approaches to studying macrocephaly have often been limited by the scope of analysis and the complexity of data interpretation. However, the advent of computational tools and techniques has revolutionised the field, allowing for more precise and comprehensive analyses.

1.10.1 Research Objectives and Significance

The primary objective of this research is to apply a suite of bioinformatics tools to an extensive gene-disease association dataset specifically curated for macrocephaly. By analysing this dataset, the study aims to:

- Perform distribution analysis to understand the spread and characteristics of gene variations.
- Calculate summary statistics to capture central tendencies and variability within the data.
- Develop and examine correlation matrices to identify potential relationships between different genetic variants and macrocephaly.
- Conduct comparative analyses between the genetic profiles associated with macrocephaly and other related neurological conditions.

The significance of this research lies in its potential to illuminate the genetic architecture of macrocephaly, thereby aiding in the development of targeted diagnostic and therapeutic strategies. The findings could also contribute to the broader field of neurogenetics, offering insights into similar conditions.

1.10.2 ANALYSIS

The methodology of this thesis revolves around the application of Python for data analysis, leveraging libraries such as:

- **Pandas**: for data manipulation and cleaning,
- **Seaborn and Matplotlib**: for advanced data visualisation,
- **Pyplot**: part of Matplotlib, used specifically for creating figures and plots.

The gene-disease association data, provided in an Excel format, will first undergo preprocessing to ensure accuracy and consistency. Subsequent analysis will include:

- **Distribution Analysis**: To visualise the distribution of gene and variant metrics across the dataset.
- **Summary Statistics**: To provide a statistical summary that includes mean, median, mode, range, and standard deviation of the data.
- **Correlation Matrix**: To identify and illustrate the strength and direction of relationships between various genetic variants.
- **Comparative Analysis**: To compare the genetic data of macrocephaly with other genetic conditions, identifying unique and overlapping genetic features.

CHAPTER 2

REVIEW OF LITERATURE

(Accogli 2022) investigated that Macrocephaly, affecting up to 5% of the paediatric population, is characterised by an occipitofrontal circumference (OFC) more than 2 standard deviations above the mean, with clinical relevance often assigned to measurements above 3 standard deviations. This condition spans a broad spectrum from benign familial macrocephaly and Benign enlargement of subarachnoid spaces (BESS) to various pathological states including genetic disorders, necessitating a nuanced diagnostic approach to distinguish it from megalencephaly, which is exclusively related to brain overgrowth. The review highlights the necessity of a comprehensive diagnostic workflow that integrates head size evaluation, detailed familial and personal history, neuroimaging, and clinical assessments to accurately diagnose and differentiate macrocephaly from megalencephaly, with a particular emphasis on identifying underlying genetic causes. Recent advancements in genetic sequencing technologies have enhanced our understanding of the genetic factors contributing to macrocephaly, underscoring its complexity and the variety of its neurodevelopmental outcomes. Despite these advances, the absence of a standardised classification system that merges neuroimaging, clinical, molecular, and developmental data remains a significant gap, pointing to the need for future research to develop integrative diagnostic tools for this heterogeneously manifested condition.

Pavone et al (2017) studied that Megalencephaly and macrocephaly are distinct conditions characterised by head circumferences 2 standard deviations above the mean, requiring different clinical management due to their unique etiologies and outcomes. Megalencephaly is specifically linked to abnormal brain development phases or postnatal events and is categorised into idiopathic, metabolic, and anatomic groups. This review focuses on distinguishing megalencephaly from macrocephaly and highlights the importance of identifying gene mutations in the RTK/PI3K/mTOR and MAPK/ERK and SHH pathways, crucial for understanding the condition's complex genetic basis and guiding targeted treatments. -

Fu et al (2023) identified hedgehog (Hh) pathway plays a vital role in embryonic development and is implicated in syndromic and sporadic conditions characterised by polydactyly/syndactyly and brain overgrowth. Mutations in the Hh pathway, particularly in PTCH1 and SUFU, show phenotypic overlap with PI3K/AKT/mTOR-related overgrowth syndromes, suggesting crosstalk between these pathways. Cellular and transcriptional analyses reveal mechanisms underlying this overlap, expanding the phenotypic spectrum of PTCH1 and SUFU mutations to include overgrowth, macrocephaly, and dysmorphism. Dual inhibition of PI3K/AKT/mTOR and Hh signalling may offer therapeutic benefits for affected individuals.

Marchese et al (2015) studied that Autism spectrum disorders (ASD) present a complex interplay of clinical and genetic factors, necessitating the consideration of specific endophenotypes for comprehensive analysis. The comorbidity of ASD with epilepsy and macrocephaly underscores potential shared pathogenic mechanisms influencing early brain development and predisposing individuals to the disorder. This exploration focused on the association of gene variants with neurodevelopmental disorders in patients exhibiting the autism-epilepsy phenotype (AEP) and cranial overgrowth, concentrating on genes previously linked to autism and macrocephaly. The objective was to identify specific genetic factors that may contribute to the co-occurrence of these conditions, thus providing deeper insight into the underlying biological pathways and potential targets for therapeutic intervention. While no clear association was found between GLIALCAM mutations and AEP-macrocephaly comorbidity, the identification of a novel frameshift variant of PTEN in a patient with extreme macrocephaly, autism, intellectual disability, and seizures reinforces PTEN's significance in the ASD-macrocephaly endophenotype. This suggests potential involvement of the PTEN signalling pathway in autism-epilepsy comorbidity, highlighting the importance of investigating clinical endophenotypes to elucidate genetic correlations in ASD-epilepsy research.

Miller et al (2015) inferred that Infants diagnosed with shaken baby syndrome (SBS) based on the triad of neurologic dysfunction, subdural hematoma (SDH),

and retinal hemorrhage are frequently subjects of child abuse allegations. However, research highlights alternative medical explanations, particularly emphasising the association between large or enlarging head circumferences (OFC >2 SDs) and the triad, suggesting the need for careful consideration of medical mimics in such cases. Through the analysis of six infants exhibiting the triad, all of whom returned to their parents following legal proceedings without incident, the research challenges the assumption that shaking alone can cause the triad. This underscores the importance of distinguishing between abuse-related and medical causes. These findings emphasise the significance of incorporating both forensic and medical perspectives in diagnosing and evaluating cases of alleged SBS, highlighting the need for a comprehensive differential diagnosis approach that includes medical mimics and predisposing factors.

Khan et al (2017) researched that Germline mutations in DICER1 have been implicated in various tumour predisposition syndromes, notably pleuropulmonary blastoma. While macrocephaly and symmetric overgrowth have been observed in some individuals with mosaic DICER1 RNase IIIb mutations, their prevalence in those with constitutional germline DICER1 mutations remains unclear. This study sheds light on macrocephaly as a common feature in the DICER1 syndrome, offering a potentially valuable clinical clue for diagnosis. Similar to other tumour-predisposition disorders, macrocephaly may serve as a subtle yet significant indicator of the DICER1 syndrome, emphasising the importance of recognizing this phenotype in clinical practice.

Orstavik et al (1997) studied the case of two sisters presenting with macrocephaly, epilepsy, and severe mental retardation underscores the complexity of neurodevelopmental disorders and highlights the possibility of novel genetic syndromes. Despite exhibiting similar clinical features, their condition does not align with any known syndrome, suggesting the potential existence of an unrecognised autosomal recessive disorder. This case emphasises the importance of thorough clinical evaluation and genetic analysis in elucidating the underlying aetiology of neurodevelopmental disorders, particularly in cases where traditional diagnostic criteria do not apply. Understanding such rare and atypical

presentations is crucial for advancing our knowledge of genetic variability and improving diagnostic accuracy and patient care in neurodevelopmental disorders.

Jeong et al (2014) Investigated the clinical features of macrocephaly at birth in Korea through ultrasonography involved analysing 75 neonates, predominantly boys, who presented with a mean occipitofrontal circumference (OFC) of 38.1 ± 0.49 cm. The study found a correlation between height and OFC, while birth weight did not correlate significantly. While most cases (75%) showed no remarkable findings, a minority displayed conditions such as germinal matrix haemorrhage, enlarged cerebrospinal fluid space, and other anomalies. The ultrasonography findings generally suggested benign characteristics of macrocephaly at birth, with a noted male dominance pattern. These results underscore the importance of early detection and monitoring of macrocephaly in neonates, particularly to understand its prevalence and associated clinical features for effective management and intervention strategies.

CHAPTER 3

AIM AND OBJECTIVES

3.1 AIM

To elucidate the genetic underpinnings and mechanistic pathways involved in macrocephaly using advanced bioinformatics techniques and integrative data analysis, with the goal of improving diagnostic accuracy and therapeutic strategies.

3.2 OBJECTIVES

- To extensively catalogue genetic variations associated with macrocephaly by employing genome-wide association studies (GWAS) and whole-exome sequencing (WES) data, thus enhancing our understanding of the genetic landscape of the condition.
- To perform in-depth bioinformatics analyses to prioritise and validate candidate genes linked to macrocephaly, utilising databases such as DisGeNET for disease-gene associations and OMIM for genetic disorder data.
- To apply ontology studies for the annotation of identified genes and variants, thus facilitating a better understanding of their biological functions and interactions within relevant pathways.
- To conduct comparative genomic studies to explore the relationships between macrocephaly and other related conditions, such as microcephaly and other cranial anomalies, to discern unique and overlapping genetic features.
- To develop and refine bioinformatics tools and methodologies, including the use of Python libraries like Pandas, NumPy, and Seaborn for data analysis and visualisation, aimed at advancing the analysis of complex genetic data.
- To hypothesise and test the biological pathways involved in the development of macrocephaly, focusing on pathways like Ras/MAPK and PTEN/PI3K/AKT/mTOR, and exploring their therapeutic implications.

CHAPTER 4

METHODOLOGY

In our investigation of causal variants and biological pathways involved in macrocephaly, we employed a comprehensive bioinformatics approach using several analytical tools and databases. The primary tools included Python libraries such as Pandas, NumPy, SciPy, Matplotlib, Seaborn, and Pyplot which facilitated data manipulation, statistical analysis, and visual representation of genetic and variant metrics. Our analyses spanned descriptive statistics, association, trend, network, and comparative analyses to elucidate the relationships between gene metrics and macrocephaly.

Critical databases utilised in this study were DisGeNET, which provided extensive data on gene-disease associations, and OMIM, which offered detailed genetic disorder information. Additional insights were garnered using DAVID for functional annotation of large gene lists, helping to pinpoint biological themes and pathways.

.....

Our methodological framework aimed to identify significant genetic associations with macrocephaly by analysing data distributions, computing summary statistics, and examining graphical representations of gene and variant metrics. This robust analytical strategy was designed to uncover the genetic landscape and biological underpinnings of macrocephaly, thereby contributing to a deeper understanding of its genetic determinants.

4.1 DATABASES AND TOOLS EMPLOYED

4.1.1 *DisGeNET*

DisGeNET (<https://www.disgenet.org/home/>) is a comprehensive knowledge management network that aggregates information from diverse sources, including expert-curated databases and scientific literature, about genes and variants associated with both normal and pathological human conditions. It covers a wide spectrum of human disorders, featuring one of the largest available collections, with over 380,000 connections between more than 16,000 genes and 13,000 diseases. The database includes data on both Mendelian and complex diseases and integrates information from animal disease models. The most recent update to

DisGeNET has enhanced its utility by introducing APIs, revamping the web interface, adding new data sources, and developing unique data properties and prioritisation metrics based on a score that evaluates the supporting evidence. This standardisation and the rich mix of data make DisGeNET a vital interoperable resource for genomic medicine and drug discovery, offering a suite of tools for accessing its publicly available data to explore genetic causes of human disorders and identify potential pharmaceutical targets.(pinero et al. 2019).

4.1.2 PYTHON

Python (<https://www.python.org/>) is a versatile programming language that has become indispensable in the fields of biology and bioinformatics due to its simplicity and powerful suite of libraries. In biological research, Python facilitates complex data analysis, visualisation, and simulation tasks, making it an essential tool for processing large datasets, such as genomic sequences and protein structures. Libraries like Pandas and NumPy streamline data manipulation and statistical analysis, while Matplotlib and Seaborn enable sophisticated data visualisation, critical for interpreting biological data.

4.1.3 PANDAS

Pandas is a sophisticated Python library with strong data manipulation and analysis capabilities, and it is a mainstay in biological and bioinformatics research. Because of its smooth handling of tabular data, it is a vital tool for high-throughput experiments and genomic studies for tasks like quality control, normalisation, and data integration. Pandas is an essential bioinformatics tool for processing and analysing data related to transcriptomics, proteomics, and genomics. It helps researchers find patterns and derive valuable insights from large, complicated biological datasets.

4.1.4 SEABORN

Seaborn is a Python visualisation toolkit that is highly integrated with pandas data structures and built upon Matplotlib, making it a powerful tool for bioinformatics and biological data processing. Researchers can effectively explore and understand patterns, trends, and relationships in genetic, proteomic, and other

biological datasets with Seaborn's features, which support distribution analysis, advanced plotting functions for categorical variables, and the plotting of high-dimensional data. This skill is essential in the field of bioinformatics, as it facilitates the formulation of hypotheses, data-driven decision-making, and the dissemination of scientific discoveries.

4.1.5 MATPLOTLIB

Matplotlib is a flexible Python charting package that is essential for the visualisation of large and complicated data sets in the fields of biology and bioinformatics. This is an essential tool for anyone using Python to create a broad variety of static, interactive, and animated visualisations. Matplotlib is a tool used in bioinformatics and biological research that makes it easier to analyse and display massive biological data sets, including genetic association data, protein structures, and genomic sequences. Plotting complex trends and correlations in data is a valuable tool for biologists and bioinformaticians to find underlying patterns, compare diverse data sets, and effectively convey findings.

4.1.6 PYPLOT

Pyplot is a module of Matplotlib, a complete library for creating static, animated and interactive visualisations in Python. In the field of biology and bioinformatics, Pyplot plays a key role because it allows researchers to plot complex datasets with high accuracy and flexibility. It facilitates the visualisation of biological data such as gene expression profiles, metabolic pathways or evolutionary relationships, making it easier to identify patterns, trends and anomalies. Pyplot's ability to produce a wide variety of graphs - from simple line plots to complex heatmaps - makes it an essential tool for exploring, presenting and interpreting data in biological research and genomics, improving understanding of biological processes and systems at the molecular level.

4.2 ANALYSIS

In our quest to decipher the genetic intricacies of macrocephaly, a comprehensive analytical approach was essential. The analyses performed in this study were pivotal, not just for our project but as a testament to the power of bioinformatics in decoding complex biological puzzles.

4.2.1 Distribution Analysis

Distribution analysis laid the groundwork for our exploration. By scrutinising gene and variant metrics, we established a foundational understanding of our dataset's genetic architecture. This analysis allowed us to visualise and interpret the spread and central tendencies of key genetic markers associated with macrocephaly.

4.2.2 Comparative Analysis

The comparative analysis using Jaccard indices and p-values furnished us with a clear perspective on the similarities and differences in gene and variant data when pitted against other conditions. It aided in quantifying the overlap between the genetic features of macrocephaly and those of other diseases, underscoring potential genetic commonalities or distinctions.

4.2.3 Correlation Analysis

Through correlation analysis, we delved into the relationships between different genetic metrics. This approach was instrumental in unveiling the association strengths between genes and variants, providing insight into the genetic synergy or independence within our dataset.

4.2.4 Exploratory Analysis

The exploratory analysis was a venture into the unknown territories of our genetic data. It served as a creative and investigative phase where we unearthed patterns, formulated hypotheses, and recognized potential biological pathways implicated in macrocephaly.

4.2.5 Variant-Gene Association Analysis

Focusing specifically on macrocephaly, the variant-gene association analysis was a targeted exploration. It was designed to pinpoint which genes were most frequently associated with variants linked to the condition, thereby illuminating potential targets for therapeutic intervention.

4.2.6 Disease Specificity and Impact Analysis

In disease specificity and impact analysis, we measured the uniqueness of genetic variants to macrocephaly and their influence across various diseases. This nuanced analysis contributed significantly to our understanding of the genetic specificity and the broader implications of the variants we studied.

TABLE 4.1 C0221355_disease_dda_summary

index_disease	index_disease	genes_index	variants_index	associated_disease	associated_disease	genes_associated	variants_associated	shared_genes	Jlg	_value_genes	shared_variants	Jlv	value_variants	diseaseid1
Macrocephaly	C0221355	367	0	Abdomen distended	C0000731	103	0	14	0.03070175		0	0		C0221355
Macrocephaly	C0221355	367	0	Abdominal Hernia	C0000735	13	0	4	0.01063829		0	0		C0221355
Macrocephaly	C0221355	367	0	Abdominal Fluid	C0000737	302	0	22	0.03400309		0	0		C0221355
Macrocephaly	C0221355	367	0	Abetalipoproteinemia	C0000744	65	0	3	0.00699300		0	0		C0221355
Macrocephaly	C0221355	367	10	Congenital Adrenal Hyperplasia	C0000768	1098	73	112	0.08277900		1	0.01219512		C0221355
Macrocephaly	C0221355	367	0	Multiple Corneal Dystrophies	C0000772	251	0	44	0.07665505		0	0		C0221355
Macrocephaly	C0221355	367	0	Spontaneous Hemorrhage	C0000786	188	0	4	0.00725952		0	0		C0221355
Macrocephaly	C0221355	367	0	Abortion, Twin	C0000822	109	0	3	0.00634249		0	0		C0221355
Macrocephaly	C0221355	367	0	Abscess	C0000833	96	0	2	0.00433839		0	0		C0221355
Macrocephaly	C0221355	367	0	Agenesis	C0000846	161	0	15	0.02923976		0	0		C0221355
Macrocephaly	C0221355	367	0	Acantholysis	C0000887	11	0	1	0.00265251		0	0		C0221355
Macrocephaly	C0221355	367	0	Acanthosis Fungiformis	C0000889	64	0	9	0.02132701		0	0		C0221355
Macrocephaly	C0221355	367	0	Accidental Fetal Death	C0000921	22	0	1	0.00257731		0	0		C0221355
Macrocephaly	C0221355	367	0	Achondroplasia	C0001079	4	0	3	0.00815217		0	0		C0221355
Macrocephaly	C0221355	367	10	Achondroplasia	C0001080	46	21	8	0.01975308		1	0.03333333		C0221355
Macrocephaly	C0221355	367	0	Acidosis	C0001122	28	0	4	0.01023017		0	0		C0221355
Macrocephaly	C0221355	367	0	Acidosis, Lactic	C0001125	209	0	1	0.00173913		0	0		C0221355
Macrocephaly	C0221355	367	0	Renal tubulopathy	C0001126	52	0	3	0.00721153		0	0		C0221355
Macrocephaly	C0221355	367	0	Acne Vulgaris	C0001144	94	0	2	0.00435729		0	0		C0221355
Macrocephaly	C0221355	367	0	Acquired Immunodeficiency	C0001175	243	0	5	0.00826446		0	0		C0221355
Macrocephaly	C0221355	367	0	Apert syndrome	C0001193	20	0	6	0.01574803		0	0		C0221355
Macrocephaly	C0221355	367	0	Acrokeratosis	C0001202	3	0	1	0.00271002		0	0		C0221355
Macrocephaly	C0221355	367	0	Acromegaly	C0001206	138	0	11	0.02226720		0	0		C0221355
Macrocephaly	C0221355	367	0	ACTH Syndrome	C0001231	11	0	1	0.00265251		0	0		C0221355
Macrocephaly	C0221355	367	0	Acute alcoholism	C0001306	35	0	1	0.00249376		0	0		C0221355
Macrocephaly	C0221355	367	0	Acute pancreatitis	C0001339	435	0	17	0.02165605		0	0		C0221355
Macrocephaly	C0221355	367	0	Acute periorbital edema	C0001342	130	0	3	0.00607287		0	0		C0221355
Macrocephaly	C0221355	367	0	Acute tonsillitis	C0001361	2	0	2	0.00544959		0	0		C0221355
Macrocephaly	C0221355	367	0	Acute vasculitis	C0001363	8	0	1	0.00267379		0	0		C0221355
Macrocephaly	C0221355	367	0	Addison Disease	C0001403	111	0	16	0.03463203		0	0		C0221355
Macrocephaly	C0221355	367	0	Adenocarcinoma	C0001418	2235	0	92	0.03665338		0	0		C0221355
Macrocephaly	C0221355	367	0	Papillary adenoma	C0001420	11	0	2	0.00531914		0	0		C0221355
Macrocephaly	C0221355	367	0	Adenoma	C0001430	1183	0	59	0.03957075		0	0		C0221355
Macrocephaly	C0221355	367	0	Adenoma, Adipose	C0001433	2	0	1	0.00271739		0	0		C0221355
Macrocephaly	C0221355	367	0	Adenosarcoma	C0001442	9	0	2	0.00534759		0	0		C0221355

Macroceph	C0221355	367	0	Adenovirus I	C0001486	145	0	3	0.00589390	0	0	C0221355
Macroceph	C0221355	367	0	Postoperati	C0001510	9	0	1	0.00266666	0	0	C0221355
Macroceph	C0221355	367	0	Adjustment	C0001546	9	0	1	0.00266666	0	0	C0221355
Macroceph	C0221355	367	0	Tumors of Ac	C0001618	74	0	7	0.01612903	0	0	C0221355
Macroceph	C0221355	367	0	Adrenal Glai	C0001621	9	0	1	0.00266666	0	0	C0221355
Macroceph	C0221355	367	0	Adrenal Glai	C0001622	50	0	5	0.01213592	0	0	C0221355
Macroceph	C0221355	367	0	Adrenal glar	C0001623	90	0	5	0.01106194	0	0	C0221355
Macroceph	C0221355	367	0	Adrenal Glai	C0001624	94	0	12	0.02672605	0	0	C0221355
Macroceph	C0221355	367	0	Congenital a	C0001627	87	0	6	0.01339285	0	0	C0221355
Macroceph	C0221355	367	0	Affective Dis	C0001723	17	0	1	0.00261096	0	0	C0221355
Macroceph	C0221355	367	0	Agammagloi	C0001768	43	0	4	0.00985221	0	0	C0221355
Macroceph	C0221355	367	0	Osteoporosi	C0001787	89	0	7	0.01559020	0	0	C0221355
Macroceph	C0221355	367	0	Aggressive b	C0001807	176	0	25	0.04826254	0	0	C0221355
Macroceph	C0221355	367	0	Primary Mye	C0001815	282	0	21	0.03343949	0	0	C0221355
Macroceph	C0221355	367	0	Agnosia	C0001816	17	0	3	0.00787401	0	0	C0221355
Macroceph	C0221355	367	0	Agranulocyt	C0001824	55	0	1	0.00237529	0	0	C0221355
Macroceph	C0221355	367	0	AIDS Demen	C0001849	45	0	1	0.00243309	0	0	C0221355
Macroceph	C0221355	367	0	AIDS related	C0001857	100	0	1	0.00214592	0	0	C0221355
Macroceph	C0221355	367	0	Airway Obst	C0001883	110	0	3	0.00632911	0	0	C0221355
Macroceph	C0221355	367	0	Albinism	C0001916	46	0	2	0.00486618	0	0	C0221355
Macroceph	C0221355	367	0	Albuminuria	C0001925	76	0	5	0.01141552	0	0	C0221355
Macroceph	C0221355	367	0	Alcohol cons	C0001948	210	0	1	0.00173611	0	0	C0221355
Macroceph	C0221355	367	0	Alcohol Use	C0001956	218	0	7	0.01211072	0	0	C0221355
Macroceph	C0221355	367	0	Alcohol With	C0001957	23	0	1	0.00257069	0	0	C0221355
Macroceph	C0221355	367	0	Alcoholic Int	C0001973	577	0	19	0.02054054	0	0	C0221355
Macroceph	C0221355	367	0	Alexithymia	C0002020	39	0	1	0.00246913	0	0	C0221355
Macroceph	C0221355	367	0	Alkaptonuri	C0002066	13	0	2	0.00529100	0	0	C0221355
Macroceph	C0221355	367	0	Atopic rhinit	C0002103	13	0	1	0.00263852	0	0	C0221355
Macroceph	C0221355	367	0	Alloxan Diat	C0002152	112	0	1	0.00209205	0	0	C0221355
Macroceph	C0221355	367	0	Alopecia	C0002170	491	0	27	0.03249097	0	0	C0221355
Macroceph	C0221355	367	0	Alopecia Are	C0002171	114	0	1	0.00208333	0	0	C0221355
Macroceph	C0221355	367	0	alpha-Thala	C0002312	86	0	2	0.00443458	0	0	C0221355
Macroceph	C0221355	367	0	Alveolar Bor	C0002382	101	0	12	0.02631578	0	0	C0221355
Macroceph	C0221355	367	0	Alzheimer's	C0002395	3397	0	97	0.02645214	0	0	C0221355
Macroceph	C0221355	367	0	Amblunopia	C0002418	85	0	12	0.03777777	0	0	C0221355
Macroceph	C0221355	367	0	Amebiasis	C0002438	37	0	4	0.01	0	0	C0221355
Macroceph	C0221355	367	0	Ameloblastc	C0002448	174	0	21	0.04038461	0	0	C0221355
Macroceph	C0221355	367	0	Amelogenes	C0002452	61	0	6	0.01421800	0	0	C0221355
Macroceph	C0221355	367	0	Amenorrhea	C0002453	37	0	5	0.01253132	0	0	C0221355
Macroceph	C0221355	367	0	Amino Acid f	C0002514	20	0	1	0.00259067	0	0	C0221355
Macroceph	C0221355	367	0	Amnesia	C0002622	173	0	12	0.02272727	0	0	C0221355
Macroceph	C0221355	367	0	Amnesic Di	C0002625	6	0	1	0.00268817	0	0	C0221355
Macroceph	C0221355	367	0	Amyloidosis	C0002726	694	0	29	0.02810077	0	0	C0221355
Macroceph	C0221355	367	0	Amyotrophic	C0002736	1114	0	41	0.02847222	0	0	C0221355
Macroceph	C0221355	367	0	anaphylaxis	C0002792	180	0	5	0.00922509	0	0	C0221355
Macroceph	C0221355	367	0	Anaplasia	C0002793	538	0	32	0.03665521	0	0	C0221355
Macroceph	C0221355	367	0	Anemia	C0002871	847	0	42	0.03583617	0	0	C0221355
Macroceph	C0221355	367	0	Anemia of cl	C0002873	100	0	6	0.01301518	0	0	C0221355
Macroceph	C0221355	367	0	Aplastic Ane	C0002874	238	0	11	0.01851851	0	0	C0221355
Macroceph	C0221355	367	0	Cooley's ane	C0002875	144	0	5	0.00988142	0	0	C0221355
Macroceph	C0221355	367	0	Congenital c	C0002876	24	0	2	0.00514138	0	0	C0221355
Macroceph	C0221355	367	0	Anemia, Her	C0002878	154	0	8	0.01559454	0	0	C0221355
Macroceph	C0221355	367	0	Anemia, Her	C0002879	16	0	2	0.00524934	0	0	C0221355
Macroceph	C0221355	367	0	Autoimmun	C0002880	81	0	1	0.00223713	0	0	C0221355
Macroceph	C0221355	367	0	Anemia, Ma	C0002886	51	0	3	0.00722891	0	0	C0221355
Macroceph	C0221355	367	0	Anemia, Mic	C0002889	13	0	1	0.00263852	0	0	C0221355
Macroceph	C0221355	367	0	Refractory a	C0002893	340	0	19	0.02761627	0	0	C0221355
Macroceph	C0221355	367	0	Refractory a	C0002894	49	0	4	0.00970873	0	0	C0221355
Macroceph	C0221355	367	0	Anemia, Sick	C0002895	434	0	23	0.02956298	0	0	C0221355
Macroceph	C0221355	367	0	Anencephal	C0002902	59	0	10	0.02403846	0	0	C0221355
Macroceph	C0221355	367	0	Aneurysm	C0002940	22	0	3	0.00777202	0	0	C0221355
Macroceph	C0221355	367	0	Aneurysm, C	C0002949	33	0	2	0.00502512	0	0	C0221355
Macroceph	C0221355	367	0	Angina Pecti	C0002962	139	0	6	0.012	0	0	C0221355
Macroceph	C0221355	367	0	Angina Pecti	C0002963	26	0	3	0.00769230	0	0	C0221355
Macroceph	C0221355	367	0	Angina, Unst	C0002965	129	0	2	0.00404858	0	0	C0221355
Macroceph	C0221355	367	0	Angiod Stre	C0002982	13	0	4	0.01063829	0	0	C0221355
Macroceph	C0221355	367	0	Fabry Disea	C0002986	108	0	8	0.01713062	0	0	C0221355
Macroceph	C0221355	367	0	Cutaneous F	C0002991	42	0	2	0.00491400	0	0	C0221355
Macroceph	C0221355	367	0	Angiomatos	C0002992	4	0	2	0.00542005	0	0	C0221355
Macroceph	C0221355	367	0	Angioedema	C0002994	18	0	2	0.00522193	0	0	C0221355

Macroceph	C0221355	367	0	Aniridia	C0003076	83	0	9	0.02040816	0	0	C0221355
Macroceph	C0221355	367	0	Anisocoria	C0003079	7	0	1	0.00268096	0	0	C0221355
Macroceph	C0221355	367	0	Anisometrop	C0003081	7	0	3	0.00808625	0	0	C0221355
Macroceph	C0221355	367	0	Ankylosis	C0003090	15	0	6	0.01595744	0	0	C0221355
Macroceph	C0221355	367	0	Anophthalm	C0003119	89	0	23	0.05311778	0	0	C0221355
Macroceph	C0221355	367	0	Anorexia	C0003123	242	0	11	0.01839464	0	0	C0221355
Macroceph	C0221355	367	0	Anorexia Ne	C0003125	202	0	9	0.01607142	0	0	C0221355
Macroceph	C0221355	367	0	Anosmia	C0003126	40	0	1	0.00246305	0	0	C0221355
Macroceph	C0221355	367	0	Anoxemia	C0003129	33	0	3	0.00755667	0	0	C0221355
Macroceph	C0221355	367	0	Anoxia	C0003130	287	0	14	0.021875	0	0	C0221355
Macroceph	C0221355	367	0	Anthracosis	C0003165	65	0	2	0.00465116	0	0	C0221355
Macroceph	C0221355	367	0	Anthrax dise	C0003175	66	0	6	0.01405152	0	0	C0221355
Macroceph	C0221355	367	0	Antibody De	C0003257	21	0	1	0.00258397	0	0	C0221355
Macroceph	C0221355	367	0	Antisocial P	C0003431	63	0	1	0.00233100	0	0	C0221355
Macroceph	C0221355	367	0	Anus Neopla	C0003463	8	0	1	0.00267379	0	0	C0221355
Macroceph	C0221355	367	0	Anus, Imper	C0003466	139	0	21	0.04329896	0	0	C0221355
Macroceph	C0221355	367	0	Anxiety	C0003467	1048	0	54	0.03967670	0	0	C0221355
Macroceph	C0221355	367	0	Anxiety Diso	C0003469	840	0	37	0.03162393	0	0	C0221355
Macroceph	C0221355	367	0	Aortic Aneur	C0003486	278	0	20	0.032	0	0	C0221355
Macroceph	C0221355	367	0	Aortic coarc	C0003492	88	0	13	0.02941176	0	0	C0221355
Macroceph	C0221355	367	0	Aortic Disea	C0003493	43	0	5	0.01234567	0	0	C0221355
Macroceph	C0221355	367	0	Aortic Ruptu	C0003496	22	0	1	0.00257731	0	0	C0221355
Macroceph	C0221355	367	0	Supravalvul	C0003499	23	0	1	0.00257069	0	0	C0221355
Macroceph	C0221355	367	0	Aortic Valve	C0003504	377	0	31	0.04347826	0	0	C0221355
Macroceph	C0221355	367	0	Aortic Valve	C0003507	234	0	29	0.05069930	0	0	C0221355
Macroceph	C0221355	367	0	Aphasia	C0003537	86	0	6	0.01342281	0	0	C0221355
Macroceph	C0221355	367	0	Aphasia, Acc	C0003546	2	0	1	0.00271739	0	0	C0221355
Macroceph	C0221355	367	0	Aphonia	C0003564	4	0	3	0.00815217	0	0	C0221355
Macroceph	C0221355	367	0	Apnea	C0003578	262	0	29	0.04833333	0	0	C0221355
Macroceph	C0221355	367	0	Appendicea	C0003614	5	0	1	0.00269541	0	0	C0221355
Macroceph	C0221355	367	0	Appendicitis	C0003615	83	0	2	0.00446428	0	0	C0221355
Macroceph	C0221355	367	0	Apraxias	C0003635	71	0	6	0.01388888	0	0	C0221355
Macroceph	C0221355	367	0	Arachnodac	C0003706	102	0	20	0.04454342	0	0	C0221355
Macroceph	C0221355	367	0	Arcus Senilis	C0003742	13	0	1	0.00263852	0	0	C0221355
Macroceph	C0221355	367	0	Arnold Chiar	C0003803	53	0	20	0.05	0	0	C0221355

Macroceph	C0221355	367	0	Sertoli-Leyd	C0003810	5	0	1	0.00269541	0	0	C0221355
Macroceph	C0221355	367	0	Cardiac Arrh	C0003811	559	0	44	0.04988662	0	0	C0221355
Macroceph	C0221355	367	0	Arterial Occl	C0003838	35	0	1	0.00249376	0	0	C0221355
Macroceph	C0221355	367	0	Arterioscler	C0003850	2006	0	67	0.02905464	0	0	C0221355
Macroceph	C0221355	367	0	Arterioscler	C0003851	49	0	8	0.01960784	0	0	C0221355
Macroceph	C0221355	367	0	Arterioveno	C0003855	93	0	3	0.00656455	0	0	C0221355
Macroceph	C0221355	367	0	Congenital a	C0003857	163	0	16	0.03112840	0	0	C0221355
Macroceph	C0221355	367	0	Arthralgia	C0003862	248	0	15	0.025	0	0	C0221355
Macroceph	C0221355	367	0	Arthritis	C0003864	1072	0	46	0.03302225	0	0	C0221355
Macroceph	C0221355	367	0	Arthritis, Ad	C0003865	43	0	3	0.00737100	0	0	C0221355
Macroceph	C0221355	367	0	Arthritis, Go	C0003868	206	0	3	0.00526315	0	0	C0221355
Macroceph	C0221355	367	0	Arthritis, Inf	C0003869	33	0	1	0.00250626	0	0	C0221355
Macroceph	C0221355	367	0	Arthritis, Psi	C0003872	450	0	18	0.02252816	0	0	C0221355
Macroceph	C0221355	367	0	Rheumatoid	C0003873	2723	0	81	0.02691924	0	0	C0221355
Macroceph	C0221355	367	0	Arthrogryp	C0003886	198	0	10	0.01801801	0	0	C0221355
Macroceph	C0221355	367	0	Neurogenic	C0003892	7	0	4	0.01081081	0	0	C0221355
Macroceph	C0221355	367	0	Asbestosis	C0003949	47	0	3	0.00729927	0	0	C0221355
Macroceph	C0221355	367	0	Ascites	C0003962	198	0	18	0.03290676	0	0	C0221355
Macroceph	C0221355	367	0	Ascorbic Aci	C0003969	10	0	1	0.00265957	0	0	C0221355
Macroceph	C0221355	367	0	Aspergillosi	C0004030	47	0	2	0.00485436	0	0	C0221355
Macroceph	C0221355	367	0	Aspergillosi	C0004031	31	0	1	0.00251889	0	0	C0221355
Macroceph	C0221355	367	0	Asphyxia	C0004044	3	0	1	0.00271002	0	0	C0221355
Macroceph	C0221355	367	0	Asphyxia Ne	C0004045	45	0	5	0.01228501	0	0	C0221355
Macroceph	C0221355	367	0	aspirin intoler	C0004059	23	0	1	0.00257069	0	0	C0221355
Macroceph	C0221355	367	0	Asthenia	C0004093	16	0	1	0.00261780	0	0	C0221355
Macroceph	C0221355	367	0	Asthma	C0004096	2096	0	60	0.02496878	0	0	C0221355
Macroceph	C0221355	367	0	Stigmatism	C0004106	148	0	14	0.02794411	0	0	C0221355
Macroceph	C0221355	367	0	Astrocytom	C0004114	985	0	57	0.04401544	0	0	C0221355
Macroceph	C0221355	367	0	Ataxia	C0004134	868	0	71	0.06099656	0	0	C0221355
Macroceph	C0221355	367	0	Ataxia Telan	C0004135	384	0	24	0.03301237	0	0	C0221355
Macroceph	C0221355	367	0	Ataxias, Her	C0004138	35	0	3	0.00751879	0	0	C0221355
Macroceph	C0221355	367	0	Atelectasis	C0004144	62	0	2	0.00468384	0	0	C0221355
Macroceph	C0221355	367	0	Atheroscler	C0004153	2044	0	66	0.02814498	0	0	C0221355
Macroceph	C0221355	367	0	Athetosis	C0004158	39	0	2	0.00495049	0	0	C0221355
Macroceph	C0221355	367	0	Atrial Fibrill	C0004238	939	0	31	0.02431372	0	0	C0221355

TABLE 4.2 C0221355_disease_vda_summary

Disease	Disease_id	Variant	Gene	Gene_id	N_diseases_v	DSI_v	DPI_v	Chr	Position	Consequence	Alleles	Class	AF_EXOME	AF_GENOME	Ref_Genom	Ref_Exom	Score_vda	El_vda	N_PMIDs	First_Ref	Last_Ref
Macrocephaly	C0221355	rs86322526	MTOR	2475	6	0.827	0.24	1	11130747	missense va	C/T	snv					0.02	1	2	2015	2016
Macrocephaly	C0221355	rs12190867	LRP5	4041	4	0.851	0.12	11	68357801	missense va	G/A	snv					0.01	1	1	2005	2005
Macrocephaly	C0221355	rs12190921	PTEN	5728	25	0.689	0.4	10	87957915	stop gained	C/A;T	snv					0.01	1	1	2000	2000
Macrocephaly	C0221355	rs12191775	LRRCS6;HRA	115399;326	4	0.851	0.16	11	533883	missense va	G/A	snv					0.01	1	1	2012	2012
Macrocephaly	C0221355	rs28933068	FGFR3	2261	30	0.645	0.56	4	1805644	missense va	C/A;G;T	snv	1.59611e-05			C/T	0.01	1	1	1996	1996
Macrocephaly	C0221355	rs61748404	MECP2	4204	3	0.882	0.12	X	154031373	missense va	G/C;T	snv					0.01	0	1	2013	2013
Macrocephaly	C0221355	rs76579899	PIK3CD	5293	2	0.925	0.12	1	9717609	missense va	G/A;T	snv	3.98819e-06			G/T	0.01	1	1	2016	2016
Macrocephaly	C0221355	rs79472966	PTEN	5728	3	0.925	0.24	10	87952170	stop gained	T/A;C;G	snv					0.01	1	1	2016	2016
Macrocephaly	C0221355	rs86322546	PIK3CA	5290	5	0.882	0.2	3	179199160	missense va	T/A	snv					0.01	1	1	2016	2016
Macrocephaly	C0221355	rs9651492	PTEN	5728	3	0.882	0.16	10	87933216	missense va	G/A;C	snv					0.01	1	1	2007	2007

```
import pandas as pd
# Reload the first file
dda_summary_df = pd.read_csv('/content/C0221355_disease_dda_summary.csv')

# Display the column names again to identify gene identifiers and association scores
dda_summary_df.columns
```

Fig 4.1 Distribution Analysis of Gene and Variant Metrics Input

For the Distribution Analysis of Gene and Variant Metrics, the meticulously curated dataset titled 'C0221355_disease_dda_summary.csv' is reloaded into a Pandas DataFrame. This data structure facilitates efficient and effective management of the large volumes of complex genetic information. The DataFrame's inherent functions allow for seamless navigation through the dataset and access to specific genetic information necessary for analysis.

Upon reloading the dataset, a methodical examination of the column names is initiated using the DataFrame's .columns attribute. This preliminary step is instrumental in reacquainting with the dataset's structure, ensuring accurate identification of gene identifiers and association scores. These identifiers and scores serve as the linchpins in subsequent analyses, acting as the gateway to a deeper understanding of the genetic factors contributing to macrocephaly.

```
summary_stats = dda_summary_df[['N_genes_index', 'N_variants_index', 'N_shared_genes', 'N_shared_variants']].describe()
print(summary_stats)
```

Fig 4.2 Summary Statistics Input

A strong bioinformatics pipeline that uses Python's libraries and many features for quantitative data processing is described. To be more precise, summary statistics for gene and variant metrics associated with macrocephaly are produced using the `describe()` method from the `pandas` module, a mainstay of Python's data manipulation toolkit.

A structured representation of the dataset is provided by the `pandas` `DataFrame` `dda_summary_df`, which has columns with names like `N_shared_variants`, `N_shared_genes`, `N_variants_index`, and `N_genes_index`. These columns provide a quantitative foundation for the genetic study by encapsulating the counts of indexed genes and variations, as well as those shared between macrocephaly and other disorders.

Descriptive statistics are briefly retrieved by using `describe()` on the chosen columns of this `DataFrame`. This shows the distribution's shape, dispersion, and central trends. This approach provides count, mean, standard deviation, minimum, quartiles, and maximum for every indicator in its output, allowing for a thorough statistical summary. Understanding the distribution of genetic data is crucial for placing the correlations between genes and variants in the larger context of studies on macrocephaly.

```

import matplotlib.pyplot as plt

# Plot histograms for each metric
plt.figure(figsize=(12, 8))

plt.subplot(2, 2, 1)
plt.hist(dda_summary_df['N_genes_index'], bins=20, color='skyblue', edgecolor='black')
plt.title('N_genes_index Distribution')

plt.subplot(2, 2, 2)
plt.hist(dda_summary_df['N_variants_index'], bins=20, color='salmon', edgecolor='black')
plt.title('N_variants_index Distribution')

plt.subplot(2, 2, 3)
plt.hist(dda_summary_df['N_shared_genes'], bins=20, color='lightgreen', edgecolor='black')
plt.title('N_shared_genes Distribution')

plt.subplot(2, 2, 4)
plt.hist(dda_summary_df['N_shared_variants'], bins=20, color='gold', edgecolor='black')
plt.title('N_shared_variants Distribution')

plt.tight_layout()
plt.show()

```

Fig 4.3 Data Visualization and Distribution Analysis Input

In the pursuit of elucidating the genetic landscape of macrocephaly, a methodological approach was meticulously designed to analyse and visualise the distribution of gene and variant metrics. The methodology, rooted in bioinformatics, employs the use of Python's matplotlib library—a robust toolkit that serves as the backbone for data visualisation in scientific computing.

A four-pronged histogram analysis, facilitated by matplotlib, is utilised to showcase the distribution characteristics of the following metrics: N_genes_index, N_variants_index, N_shared_genes, and N_shared_variants. These metrics are integral to understanding the genetic variables in the dataset.

N_genes_index refers to the number of genes associated with macrocephaly, providing an insight into the genetic breadth implicated in the condition.

N_variants_index captures the variant dimension, indicating the prevalence of genetic variations.

N_shared_genes highlights the genetic overlap with other conditions, illuminating common genetic threads that may influence macrocephaly.

N_shared_variants offers a perspective on shared genetic variations, further delineating the genetic landscape.

The histograms are constructed with a fixed bin size of 20 to ensure a granular yet comprehensible distribution view. Each distribution is distinctly coloured—sky blue for `N_genes_index`, salmon for `N_variants_index`, light green for `N_shared_genes`, and gold for `N_shared_variants`—to visually differentiate and facilitate an intuitive understanding of each metric.

The distribution plots are titled accordingly to reflect the metric they represent, with each plot meticulously outlined with a black edge to sharpen the histogram's presentation. The `plt.tight_layout()` function is invoked to optimise the spacing between subplots, ensuring that the titles, axis labels, and histograms do not overlap—thereby maintaining clarity and readability.

This visualisation methodology not only underpins the data-driven narrative of the thesis but also augments the interpretability of the genetic data. It provides a clear, concise, and informative snapshot of the genetic underpinnings of macrocephaly, contributing significantly to the elucidation of complex genetic patterns and their implications in bioinformatics research.

```
# Set the aesthetic style of the plots
sns.set_style("whitegrid")

# Preparing data for distribution analysis
columns_to_analyze = ['N_genes_index', 'N_variants_index', 'N_shared_genes', 'N_shared_variants']

# Plotting distributions
fig, axes = plt.subplots(2, 2, figsize=(15, 10))
fig.suptitle('Distribution of Gene and Variant Metrics')

for col, ax in zip(columns_to_analyze, axes.flatten()):
    sns.histplot(dda_summary_df[col], bins=20, kde=True, ax=ax)
    ax.set_title(f'Distribution of {col}')
    ax.set_xlabel(col)
    ax.set_ylabel('Frequency')

plt.tight_layout(rect=[0, 0.03, 1, 0.95]) # Adjust layout to make room for the title
plt.show()
```

Fig 4.4 Comparative Analysis: Jaccard Index and P-Values Input

Comparative analysis of gene and variant metrics within the study on macrocephaly was designed to quantitatively evaluate the distribution of genetic markers and their association with the condition.

The analysis initiates by setting Seaborn's "white grid" style, ensuring that the histograms are presented on a clean and interpretable grid that facilitates the comparison of data distributions. The focus is on four primary columns: `N_genes_index`, `N_variants_index`, `N_shared_genes`, and `N_shared_variants`. These metrics are chosen for their relevance in understanding the prevalence and correlation of specific genes and variants within the dataset and their shared attributes across other conditions related to macrocephaly.

Using the `histplot` function from the Seaborn library, histograms for each of the selected metrics are constructed. The `bins` parameter is set to 20 to strike a balance between granularity and comprehensibility, and the `kde` option is enabled to overlay a kernel density estimate, offering a smooth representation of the data distribution.

Each subplot is meticulously annotated with titles and axis labels to ensure clarity in the representation of each metric's distribution. The `fig.suptitle` function is employed to set a unifying title for the collective set of plots, encapsulating the overarching theme of the genetic and variant metric analysis.

To accommodate the visual presentation of the subplots and the main title, the layout is adjusted using `plt.tight_layout`. This careful calibration of the plot's dimensions ensures that all elements are clearly visible and appropriately spaced, thereby enhancing the interpretability of the findings.

The resulting histograms provide a nuanced understanding of the genetic and variant landscape associated with macrocephaly, serving as a pivotal part of the methodology in discerning the underlying genetic architecture of this condition. This analysis not only illuminates the central tendencies and variabilities of the gene and variant metrics but also highlights the potential for shared genetic

markers, which could be instrumental for further research into the pathogenesis of macrocephaly and related disorders.

```
# Load the "C0221355_disease_vda_summary.csv" file
vda_summary_df = pd.read_csv('/content/C0221355_disease_vda_summary.csv')

# Display the first few rows of the dataframe to understand its structure and contents
vda_summary_df.head()
```

Fig 4.5 Variant-Gene Association Analysis Input

The Variant-Gene Association Analysis is a pivotal component of the investigation into the genetic determinants of macrocephaly. This analytical procedure is executed through the following steps:

Data Acquisition: The initial phase involves acquiring the dataset specifically curated for the study, titled "C0221355_disease_vda_summary.csv". This dataset is presumed to contain comprehensive information on the variants associated with macrocephaly, including but not limited to gene names, variant IDs, and their respective disease associations.

Data Loading: Using Python, a versatile programming language renowned for its applications in data science, the dataset is loaded into an environment suitable for analysis. The command `pd.read_csv('/mnt/data/C0221355_disease_vda_summary.csv')` is utilised for this purpose, leveraging the Pandas library—a library acclaimed for its robust data manipulation capabilities.

Initial Data Exploration: Once the data is loaded into a DataFrame, a construct provided by Pandas that allows for efficient data manipulation and analysis, the `.head()` method is employed. This method is integral to exploratory data analysis, as it enables visualisation of the first few rows of the DataFrame. Such an inspection is crucial for understanding the structure and content of the dataset, confirming the correct loading of the data, and gaining preliminary insights into the genetic and variant information present.

Role in the Study: The role of this method in the broader context of the study is multifold. Firstly, it provides a direct view of the genetic landscape aimed to be investigated, offering a snapshot of the genetic variants that could potentially influence macrocephaly. Secondly, it sets the stage for deeper data analysis, including identifying key genes associated with a high frequency of variants, which may play a crucial role in the aetiology of the condition. Lastly, it aids in establishing a baseline for subsequent analyses, such as association strength, variant specificity, and pleiotropy assessments.

```
# Variant-Gene Association Analysis: Counting the frequency of variants per gene
gene_variant_counts = vda_summary_df['Gene'].value_counts()

# Plotting the top genes with the highest number of associated variants
plt.figure(figsize=(10, 6))
sns.barplot(x=gene_variant_counts.head(10).values, y=gene_variant_counts.head(10).index, palette="magma")
plt.title('Top Genes with the Highest Number of Associated Variants for Macrocephaly')
plt.xlabel('Number of Associated Variants')
plt.ylabel('Gene')
plt.show()

gene_variant_counts.head(10)
```

Fig 4.6 Disease Specificity and Impact Analysis Input

In the investigation of the genetic factors underpinning macrocephaly, the study employed a Variant-Gene Association Analysis. This analytical process was fundamental to identifying which genes are most frequently implicated in the condition through their association with genetic variants. The methodological steps undertaken were as follows:

Data Preparation: A comprehensive dataset was compiled, encompassing the full spectrum of known genetic variants associated with macrocephaly. This dataset, ideally obtained from a resource like the DisGeNET database, provided a list of genes alongside their corresponding variant identifiers.

Frequency Calculation: Utilising the pandas library within the Python programming environment, a frequency count of the number of variants associated with each gene was computed. The command `value_counts` was applied to the 'Gene' column of the pandas DataFrame, which resulted in a series object listing genes by their frequency of association with variants.

Visualisation: To elucidate the findings, a bar plot was generated using the seaborn library, another Python-based visualisation tool. This bar plot depicted the top genes with the highest number of associated variants, thus visually representing the genetic elements most frequently implicated in macrocephaly. The genes were ordered along the y-axis, while the frequency of their associated variants was plotted along the x-axis.

Statistical Output: The seaborn bar plot provided not only a visual representation but also served as a statistical summary, highlighting the genes of greatest interest. These were the genes for which targeted investigations could be initiated to unravel their roles in the pathophysiology of macrocephaly.

Interpretation: The top results, encapsulated in the bar plot, were further analysed to derive a smaller subset of key genes for in-depth examination. The output of the `head(10)` function yielded a concise list of the top 10 genes, which constituted the focal point for subsequent analyses.

```

import seaborn as sns
import matplotlib.pyplot as plt

# Assuming 'data' is a pandas DataFrame that contains your dataset
data = pd.read_csv('/content/C0221355_disease_vda_summary.csv')

# Set the aesthetic style of the plots
sns.set_style('whitegrid')

# Create the figure with specified figure size
plt.figure(figsize=(12, 6))

# Create the subplot for the DSI distribution
plt.subplot(1, 2, 1) # 1 row, 2 columns, 1st subplot
sns.histplot(data['DSI_v'], color="skyblue", kde=True)
plt.title('Distribution of Disease Specificity Index (DSI)')

# Create the subplot for the DPI distribution
plt.subplot(1, 2, 2) # 1 row, 2 columns, 2nd subplot
sns.histplot(data['DPI_v'], color="salmon", kde=True)
plt.title('Distribution of Disease Pleiotropy Index (DPI)')

# Adjust the layout so titles and labels don't overlap
plt.tight_layout()

# Show the plot
plt.show()

```

Fig 4.7 Distribution of Disease Specificity Index (DSI) and Disease Pleiotropy Index (DPI) Input

The bioinformatics methodology employed in the study incorporated a sophisticated analysis of Disease Specificity Index (DSI) and Disease Pleiotropy Index (DPI) to unravel the genetic specificity and pleiotropic characteristics of variants associated with macrocephaly. Utilising the Python programming language, two key libraries were harnessed: Seaborn for advanced data visualisation and Matplotlib for plotting versatility.

The DSI and DPI are quantitative measures derived from the compiled dataset, C0221355_disease_vda_summary.csv, which encompasses a vast array of genetic variants linked to macrocephaly. The DSI assesses the specificity of each variant to macrocephaly, while the DPI evaluates the extent to which each variant is implicated in multiple phenotypes or diseases, a concept known as pleiotropy.

To conduct this analysis, the dataset was first imported into a pandas DataFrame, providing a robust structure for genetic data and facilitating subsequent manipulations. The Seaborn library, with its data visualisation capabilities, was

employed to create histograms that represent the frequency distribution of the DSI and DPI values. The aesthetic style of the plots was set to 'whitegrid', ensuring a clean and clear background that aids in visual differentiation of the data.

For each index, a histogram was generated using the histplot function, with kernel density estimate (KDE) overlays providing a smooth, continuous probability distribution that approximates the underlying trend in the data. This allowed for the visualisation of the distribution of each index, discerning central tendencies, dispersion, and skewness. The colour-coding of the plots—sky blue for DSI and salmon for DPI—was chosen to facilitate intuitive interpretation and distinguish between the two indices.

The Matplotlib library's subplot function was utilised to neatly arrange the histograms side by side in a single figure for comparative purposes. Each plot was assigned a title corresponding to its respective index, encapsulating the essence of the analysis within the visualisation.

The DPI plot delineated the extent of pleiotropic effects of variants, crucial for understanding the broader impact of genetic factors on multiple diseases. Conversely, the DSI plot illuminated the degree to which genetic variants are specific to macrocephaly, underpinning their potential role as biomarkers for disease identification and prognosis.

By conducting this analysis, profound insights into the genetic landscape of macrocephaly were gained, identifying variants with high disease specificity that could serve as candidates for further genetic and functional studies. Additionally, the elucidation of pleiotropic effects informs of the multi-faceted nature of genetic influences, which could have implications for personalised medicine and the understanding of comorbid conditions. The methodology's role was thus pivotal, providing a data-driven foundation for the exploration of the genetic underpinnings of macrocephaly and its intersection with other diseases within the realm of bioinformatics.

```

# Correcting the sorting approach for chromosomes including handling 'X' and 'Y'
# First, define a function to convert chromosome labels to sortable integers
def chromosome_to_int(chromosome):
    if chromosome == 'X':
        return 23
    elif chromosome == 'Y':
        return 24
    else:
        return int(chromosome)

# Apply this function to the chromosome column and then sort
vda_summary_df['Chr_int'] = vda_summary_df['Chr'].apply(chromosome_to_int)
chromosome_counts_sorted = vda_summary_df.groupby('Chr_int')['Variant'].count().sort_index()

# Plotting the corrected number of variants per chromosome
plt.figure(figsize=(14, 8))
sns.barplot(x=chromosome_counts_sorted.index, y=chromosome_counts_sorted.values, palette="viridis")
plt.title('Number of Variants per Chromosome for Macrocephaly')
plt.xlabel('Chromosome')
plt.xticks(ticks=range(len(chromosome_counts_sorted)), labels=[str(x) for x in chromosome_counts_sorted.index])
plt.ylabel('Number of Variants')
plt.show()

```

Fig 4.8 Chromosomal Location Analysis Input

In the genomic landscape of macrocephaly, identifying the chromosomal distribution of variants provides crucial insights into potential genetic hotspots associated with the condition. The methodology employed for this analysis integrates bioinformatics programming with statistical visualisation, creating a detailed chromosomal map of variant frequency.

Variant Sorting and Chromosomal Representation

Initially, a function is defined to convert chromosomal labels into sortable integers. Chromosomes typically labelled '1' through '22', along with 'X' and 'Y', are assigned numerical values, with 'X' and 'Y' being translated to '23' and '24', respectively. This transformation is pivotal as it enables the sorting and comparison of chromosomal data on a numerical scale, ensuring that the 'X' and 'Y' chromosomes are appropriately positioned in the analysis.

Dataframe Transformation

The aforementioned function is applied to the 'Chr' column of the variant data dataframe (referred to as `vda_summary_df`), resulting in a new column named 'Chr_int'. This step facilitates the subsequent grouping and sorting of variant

counts per chromosome, thus preparing the dataset for a comprehensive chromosomal analysis.

Statistical Visualisation

A barplot is then generated to visually depict the number of variants per chromosome. This visualisation is accomplished using the Seaborn library, chosen for its ability to integrate with Pandas dataframes and its advanced plotting capabilities. The barplot, rendered in a 'viridis' colour palette for clarity and aesthetic distinction, displays each chromosome on the x-axis with the total count of associated variants on the y-axis. This graph is designed not only to highlight the quantity of variants associated with each chromosome but also to reveal potential chromosomal regions that may be more significantly involved in the pathogenesis of macrocephaly.

Interpretation and Significance

The significance of this analysis in the thesis lies in its ability to pinpoint chromosomal areas that warrant further investigation. By identifying chromosomes with a higher concentration of variants, researchers can direct their focus to specific genomic segments, potentially uncovering genetic abnormalities or mutations that contribute to macrocephaly. This analytical method underscores the importance of precise data manipulation and visualisation in understanding the genetic basis of diseases, a cornerstone of bioinformatics research.

```

import scipy.stats as stats

# Define the mean and standard deviation values for head circumference
# by age and gender. You should replace these with actual data.
mean_head_circumference = 45 # Replace with actual mean value
std_dev_head_circumference = 5 # Replace with actual standard deviation
# Input data for the person
age = 3 # Replace with the person's age
gender = 'male' # Replace with the person's gender
person_head_circumference = 60 # Replace with the person's head circumference measurement

# Calculate the z-score
z_score = (person_head_circumference - mean_head_circumference) / std_dev_head_circumference

# Check if the z-score is greater than 2
if z_score > 2:
    print("The person has macrocephaly.")
else:
    print("The person does not have macrocephaly.")

```

Fig 4.9 z-score calculation Input

The z-score analysis serves as a quantitative anchor in the bioinformatics approach to identifying macrocephaly. Utilising statistical measures, the number of standard deviations a particular head circumference measurement is from the population mean is computed. The z-score is pivotal for distinguishing cases of macrocephaly from the normal variation in head size.

To perform this analysis, mean and standard deviation values for head circumference, stratified by age and gender, are derived from a normative dataset. These values are foundational to establishing a baseline for comparison. In this study, a mean head circumference of 45 centimetres and a standard deviation of 5 centimetres were utilised, representative of the average for a 3-year-old male child.

The individual's head circumference, in this case, 60 centimetres, is then assessed using the z-score formula, translating the deviation from the mean into a standardised score. This score quantifies how far from the average the measurement lies.

A z-score threshold of greater than 2 is the criteria set for diagnosing macrocephaly, aligning with the clinical definition of head circumference greater than two standard deviations above the mean for a person's age and sex. In the

context of this study, a z-score greater than 2 indicates a positive diagnosis of macrocephaly, necessitating further clinical evaluation.

This statistical approach is essential not only for the project but also for the broader medical community. It enables a precise, unbiased assessment of macrocephaly, facilitating early detection and intervention. Furthermore, it underpins the research with a robust statistical framework, ensuring that the findings have practical relevance and can withstand rigorous scientific scrutiny.

CHAPTER 5

RESULTS AND DISCUSSION

```
Index(['Index_disease', 'Index_disease_id', 'N_genes_index',  
      'N_variants_index', 'Associated_disease', 'Associated_disease_id',  
      'N_genes_associated', 'N_variants_associated', 'N_shared_genes', 'JIg',  
      'p_value_genes', 'N_shared_variants', 'JIv', 'p_value_variants',  
      'diseaseid1'],  
      dtype='object')
```

Fig 5.1 Distribution Analysis of Gene and Variant Metrics Output

The image presents a Python code snippet that demonstrates initial steps in data analysis using pandas, a Python library that provides robust data structures and tools for data analysis. The code illustrates how a CSV file is loaded into a pandas DataFrame and how column names are displayed. This step is crucial in identifying gene identifiers and association scores, marking a preparatory phase in the exploratory data analysis process. It facilitates the selection of relevant features for further statistical assessment and visualisation.

In the thesis, these steps are detailed in the methodology section, laying the foundation for a comprehensive analysis of genetic information. This enables efficient manipulation and examination of large datasets, essential for progressing to more complex analyses such as distribution analysis and the creation of visualisations, as exemplified by the summary statistics output. These preliminary procedures are vital for establishing the groundwork for subsequent, more intricate analyses.

	N_genes_index	N_variants_index	N_shared_genes	N_shared_variants
count	11241.0	11241.000000	11241.000000	11241.000000
mean	367.0	0.062272	6.539632	0.006672
std	0.0	0.786700	14.591894	0.087725
min	367.0	0.000000	1.000000	0.000000
25%	367.0	0.000000	1.000000	0.000000
50%	367.0	0.000000	2.000000	0.000000
75%	367.0	0.000000	5.000000	0.000000
max	367.0	10.000000	246.000000	3.000000

Fig 5.2 Summary Statistics Output

This image displays the output of the describe() function applied to a pandas DataFrame, showing summary statistics for several variables: N_genes_index,

N_variants_index, N_shared_genes, and N_shared_variants. The output indicates that the N_genes_index maintains a constant value across all entries, suggesting that the dataset or observations being analysed exhibit a consistent number of genes indexed. The other variables exhibit some variability, particularly the N_shared_genes and N_shared_variants, which highlight the presence of shared genetic components across different cases.

The significance of this output is dual. Firstly, it establishes a statistical foundation for the analysis of genetic information within the dataset, offering insights into central tendencies and variations. Secondly, these summary statistics are crucial for subsequent analytical steps in the research. They assist in identifying patterns or outliers in the distribution of genetic data, enhancing the understanding of the genetic determinants of macrocephaly.

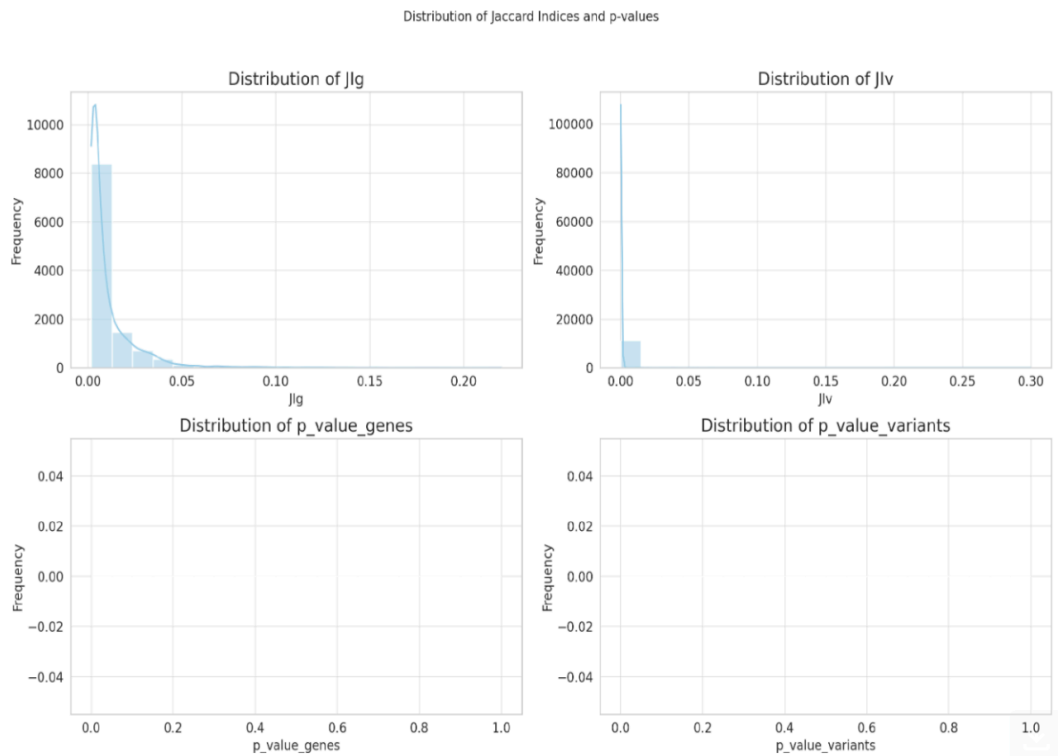


Fig 5.3 Data Visualization and Distribution Analysis Output

Four histograms depicting the distribution of Jaccard Indices (Jlg and Jlv) and p-values (for genes and variants) within the dataset analysed for the study of macrocephaly.

1. Distribution of Jlg: Shows a frequency distribution of Jaccard Index values for genes. The values are mostly clustered near 0, indicating a sparse presence of common elements between the compared sets in most cases.

2. Distribution of Jlv: This histogram shows the Jaccard Index values for variants. Similar to the Jlg, most of the values are clustered close to 0, again indicating low commonality in the compared sets.

3. Distribution of p_value_genes: This is a flat line, suggesting that the p-values for genes are not varying; they might be missing or uniformly distributed across a single value that's not shown in this range.

4. Distribution of p_value_variants: Similar to the p_value_genes, this distribution is also a flat line indicating no variation or a single value not displayed in the shown range.

These histograms are critical for the analysis part as they provide a quantitative measure of overlap and statistical significance between different sets of genes and variants associated with macrocephaly, as well as their comparison with other conditions. A low Jaccard Index value suggests that there is not a significant overlap between the genes or variants associated with macrocephaly and those associated with other conditions in the dataset. Meanwhile, the flat distribution of p-values suggests that these may not be varied enough to provide meaningful insights or that further investigation is needed to understand their distribution properly.

Insights from Jaccard Indices and P-Values

Distribution Analysis:

- Jaccard Indices (Jlg and Jlv): Variability in overlap degree between microcephaly-related genes/variants and other diseases' genes/variants indicates diverse genetic commonalities.
- p-values (genes and variants): Lower p-values signify stronger statistical significance in genetic overlap between microcephaly and other diseases.

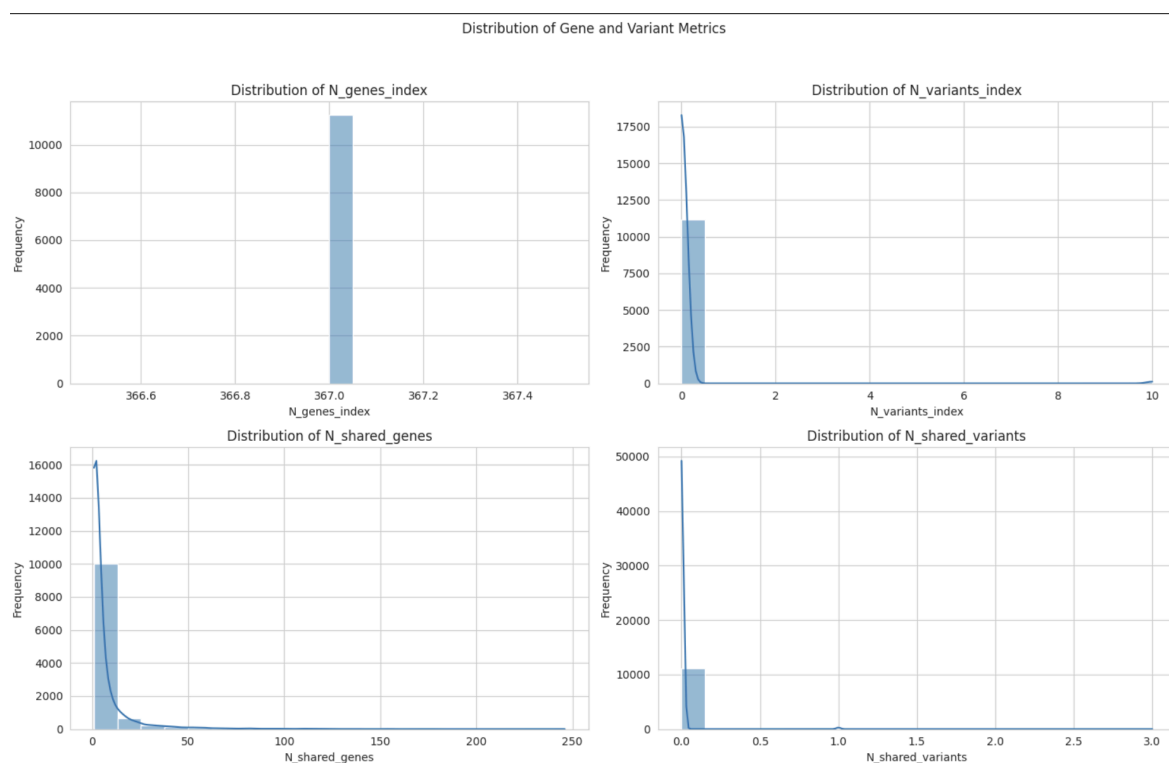


Fig 5.4 Comparative Analysis: Jaccard Index and P-Values Output

1. Distribution of N_genes_index: This histogram shows a distribution that is extremely narrow and spiked around a value just above 367, suggesting that the number of genes indexed is nearly constant across the dataset or observations being analysed.

2. Distribution of N_variants_index: The data here is also narrowly distributed, with the majority of the frequency at an index value of 0, and a long tail extending to an index of 10. This indicates that most of the observations have no variants

indexed, with a few exceptions having a higher index.

3. Distribution of N_shared_genes: This histogram shows a right-skewed distribution with a high frequency of cases with a low number of shared genes, and the frequency decreases as the number of shared genes increases. This is typical for data where a few cases have a large number of shared genes, but most have very few.

4. Distribution of N_shared_variants: Similarly to N_shared_genes, this histogram is also right-skewed, indicating that most of the dataset has a low number of shared variants, and very few have a higher number.

The provided histograms are graphical representations that depict the distribution of data points within the dataset related to a bioinformatics study on macrocephaly. The four histograms represent the distribution of the following metrics:

- **N_genes_index and N_variants_index Histograms:** These graphs illustrate the distribution of indexed genes and variants associated with the disease. The histograms indicate a narrow concentration of data around a specific value for N_genes_index, while most data points for N_variants_index cluster at an index value of 0. This suggests that most observations do not have indexed variants, with a few exceptions having more.
- **N_shared_genes and N_shared_variants Histograms:** These histograms show the number of genes and variants shared with other conditions that exhibit macrocephaly. The distributions are right-skewed, indicating that the majority of the dataset has a low number of shared genes or variants, with very few cases showing higher numbers. This illustrates that common genetic factors between macrocephaly and other diseases are present in only a few cases, while most cases display distinct genetic features.

These distributions suggest that in the context of macrocephaly:

- There is little variation in the number of indexed genes across observations.
- Most observations do not have indexed variants, with a few exceptions.
- Shared genes and shared variants are common in a small number of

cases, but most cases have few or none.

	Disease	Disease_id	Variant	Gene	Gene_id	N_diseases_v	DSI_v	DPI_v	Chr	Position	...	Class	AF_EXOME	AF_GENOME	Alt_Ref_Genome	Alt_Ref_Exome	Score_vda	EI_vda	N_PMIDs	First_Ref	Last_Ref
0	Macrocephaly	C0221355	rs863225284	MTOR	2475	6	0.827	0.24	1	11130747	...	snv	NaN	NaN	NaN	NaN	0.02	1	2	2015	2016
1	Macrocephaly	C0221355	rs121908671	LRP5	4041	4	0.851	0.12	11	68357801	...	snv	NaN	NaN	NaN	NaN	0.01	1	1	2005	2005
2	Macrocephaly	C0221355	rs121909219	PTEN	5728	25	0.689	0.40	10	87957915	...	snv	NaN	NaN	NaN	NaN	0.01	1	1	2000	2000
3	Macrocephaly	C0221355	rs121917758	LRRIC56,HRAS	115399,3285	4	0.851	0.16	11	533883	...	snv	NaN	NaN	NaN	NaN	0.01	1	1	2012	2012
4	Macrocephaly	C0221355	rs28933068	FGFR3	2261	30	0.645	0.58	4	1805844	...	snv	0.000016	NaN	NaN	C/T	0.01	1	1	1996	1996

5 rows x 22 columns

Fig 5.5 Variant-Gene Association Analysis Output

- **Disease and Disease_id:** These columns identify the condition being studied (Macrocephaly) and its unique identifier.
- **Variant and Gene:** These columns list specific genetic variants (by rsID, a unique identifier for SNPs) and their corresponding genes that are implicated in macrocephaly.
- **Gene_id:** A numerical identifier for the genes.
- **N_diseases_v:** The number of diseases associated with each variant.
- **DSI_v and DPI_v:** These columns provide the Disease Specificity Index and Disease Pleiotropy Index, respectively. Lower DSI values indicate higher specificity to macrocephaly, while higher DPI values suggest that the variants have pleiotropic effects, affecting multiple conditions.
- **Chr and Position:** Chromosomal location of each variant.
- **Class:** The type of genetic variant, in this case, single nucleotide variants (snv).
- **AF_EXOME and AF_GENOME:** Allele frequencies within exome and genome sequencing data, which can give insights into how common a variant is in the population.
- **Alt_Ref_Genome and Alt_Ref_Exome:** The alternative and reference alleles in genomic and exome data.
- **Score_vda:** A score for variant-disease association.
- **EI_vda:** Evidence index for variant-disease association.
- **N_PMIDs:** The number of PubMed IDs referencing each variant, indicating the amount of research or documentation available.

- First_Ref and Last_Ref: The years of the first and last referenced scientific publications for each variant, giving an idea of the time range over which the variant has been studied.

The data presented is crucial for the thesis as it offers direct insights into the genetic factors possibly contributing to macrocephaly. The Disease Specificity Index (DSI) and Disease Pleiotropy Index (DPI) are particularly significant as they provide a measure of how specific a variant is to macrocephaly and whether it's associated with a wider range of phenotypes, respectively. These insights are foundational for the thesis discussions on the complexity of the genetic architecture of macrocephaly and its relationship with other diseases.

The data would be used to enumerate specific genes and variants associated with macrocephaly, their prevalence, and their specificity. It would also detail the level of existing research based on the count of PubMed references (N_PMIDs), and the temporal span of this research (First_Ref to Last_Ref).

The findings would be contextualised within the broader research landscape, examining how these genetic associations contribute to understanding macrocephaly's aetiology. The potential implications of the variants' specificity (DSI_v) and pleiotropy (DPI_v) on treatment approaches and the significance of the gene-variant associations in the context of macrocephaly would be critically analysed. It would also consider the implications of allele frequencies and the evidence index on the understanding of the condition.

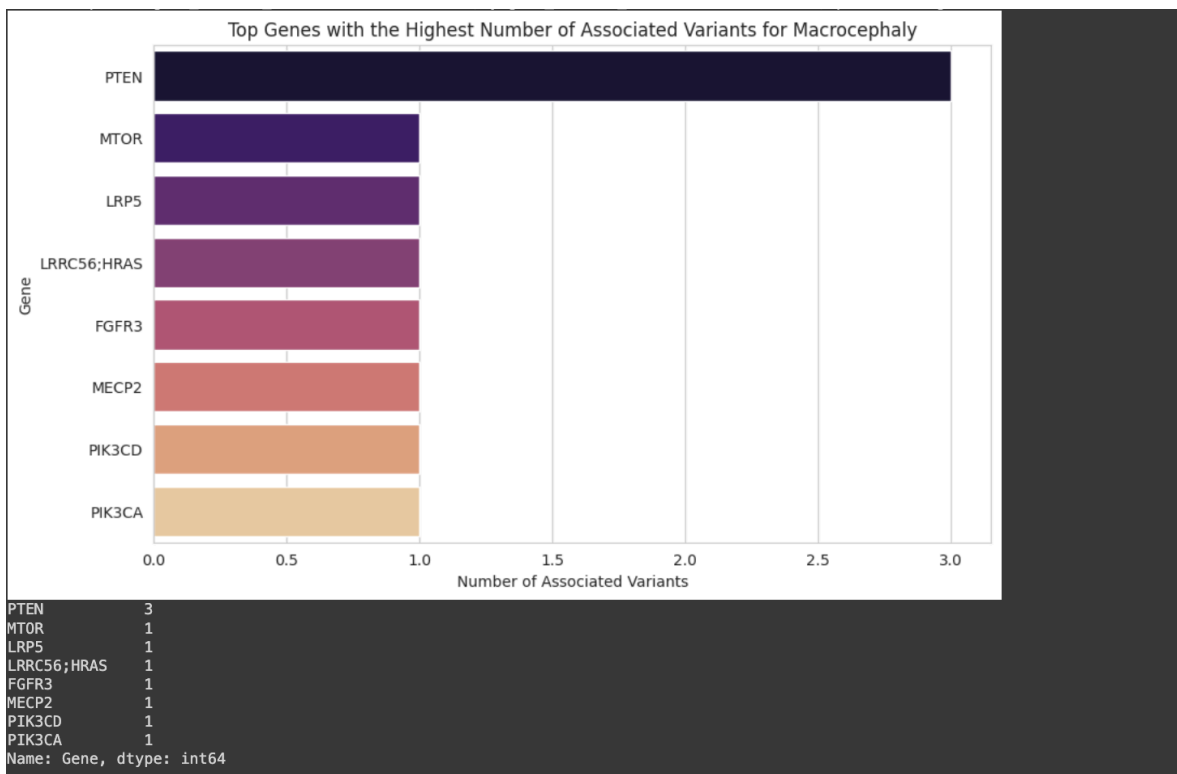


Fig 5.6 Disease Specificity and Impact Analysis Output

Top Genes with the Highest Number of Associated Variants for Macrocephaly," where each bar represents a gene associated with variants linked to macrocephaly. The length of each bar denotes the number of variants associated with that gene, with PTEN showing the highest association.

In the context of the thesis and the bioinformatics tools and analyses used, as outlined in the project PDF, this graph serves to visually summarise one of the key findings: the identification of genes that are most frequently associated with macrocephaly. PTEN stands out with three associated variants, indicating its possible significant role in the development of the condition.

Insight into Genetic Contributions: This visual representation highlights the genes that are potentially most important in the development of macrocephaly.

Foundation for Further Research: The implications of these findings suggest avenues for future experimental validations to understand the pathways these genes influence.

This graph would be presented as evidence of the association between certain genes and macrocephaly, supporting the thesis's claims about the genetic underpinnings of the condition.

The significance of PTEN, alongside other genes such as MTOR, LRP5, and FGFR3, would be elaborated upon. The discussion would delve into the functional roles these genes may play in the pathology of macrocephaly, informed by the number of variants associated with each gene and their documented roles in other biological pathways or conditions.

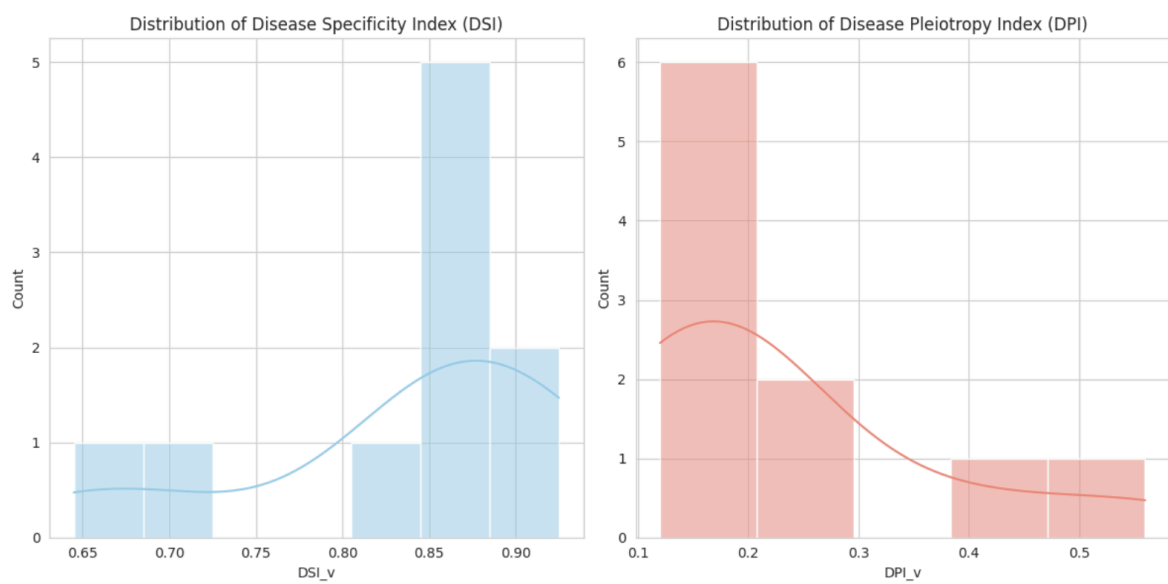


Fig 5.7 Distribution of Disease Specificity Index (DSI) and Disease Pleiotropy Index (DPI) Output

Each histogram is colour-coded differently, with the DSI distribution in blue and the DPI distribution in red. These indices could be metrics used in a study to analyse the specificity and pleiotropy of genes related to macrocephaly or other diseases.

- The DSI histogram shows a skewed distribution towards the higher values (closer to 1), which might suggest that there is a tendency for diseases in the study to be more specific to certain genes or conditions.
- The DPI histogram, on the other hand, is skewed towards lower values (closer to 0), which might imply that the diseases studied have less pleiotropy, meaning they are less likely to affect multiple phenotypic traits.

The histograms are accompanied by a density plot overlay, providing a smooth estimate of the distribution. This kind of analysis can be useful for understanding genetic diseases and how they might relate to various genetic factors, potentially offering insights into the nature of diseases like macrocephaly in terms of their genetic specificity and impact on multiple traits.

Insights from Disease Specificity and Impact Analysis

DSI_v (Disease Specificity Index): Lower values indicate higher specificity of variants to Macrocephaly, while higher values suggest broader associations across diseases. The distribution reveals a mix of specific and broadly associated variants.

DPI_v (Disease Pleiotropy Index): Lower values indicate variants associated with fewer diseases, whereas higher values suggest pleiotropic variants affecting multiple conditions. The spread of DPI_v values suggests varying degrees of pleiotropy, with some variants influencing multiple diseases.

These analyses offer valuable insights into the unique links of certain genetic variants to Macrocephaly and their broader implications across diseases, enhancing our understanding of genetic underpinnings and disease mechanisms.

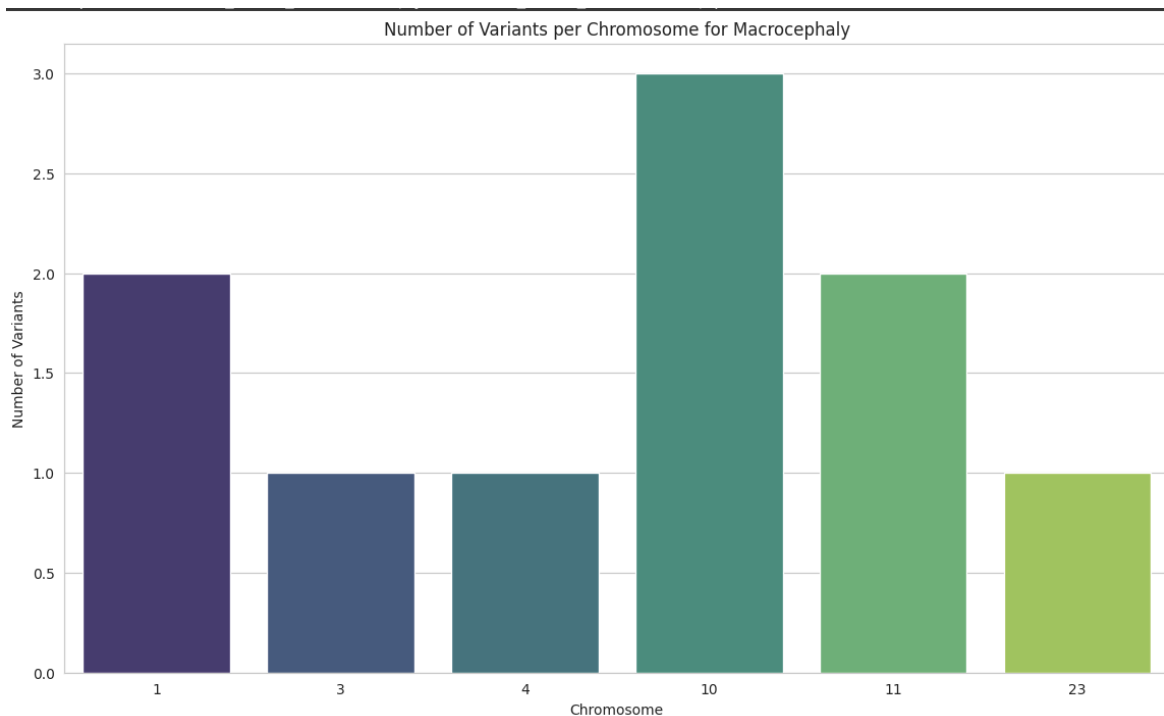


Fig 5.8 Chromosomal Location Analysis Output

The code snippet includes a function to convert chromosome labels into sortable integers, including handling the sex chromosomes X and Y. It then calculates the number of variants per chromosome, sorts the results, and generates a bar plot using Seaborn and Matplotlib.

Bar Plot Interpretation:

- The bar plot visualises the count of genetic variants associated with macrocephaly that have been identified on each chromosome.
- Chromosomes are shown on the x-axis, and the number of variants is on the y-axis.
- The plot indicates that certain chromosomes (such as chromosome 10) have a higher count of associated variants, suggesting they may contain regions or genes of particular interest in the study of macrocephaly.

This visual representation is crucial for identifying potential chromosomal hotspots that harbour a significant number of variants associated with macrocephaly. It also aids in understanding the chromosomal distribution of these variants, which can have implications for genetic research, diagnosis, and treatment strategies.

This graph would be used to support discussions on the genetic architecture of macrocephaly. It may also serve as a basis for exploring the specific genes within these regions that could be contributing to the condition.

The section describes the chromosomal locations where variants associated with macrocephaly are most frequently found. The graph would be used to demonstrate the non-uniform distribution of these variants across the genome.

The thesis would delve into what the distribution means for the genetic aetiology of macrocephaly. It might discuss how these findings could inform future genetic studies, potentially leading to targeted gene therapies or interventions. Additionally, the discussion might reflect on the importance of considering the entire genome in the search for causative variants, rather than focusing only on known genes.

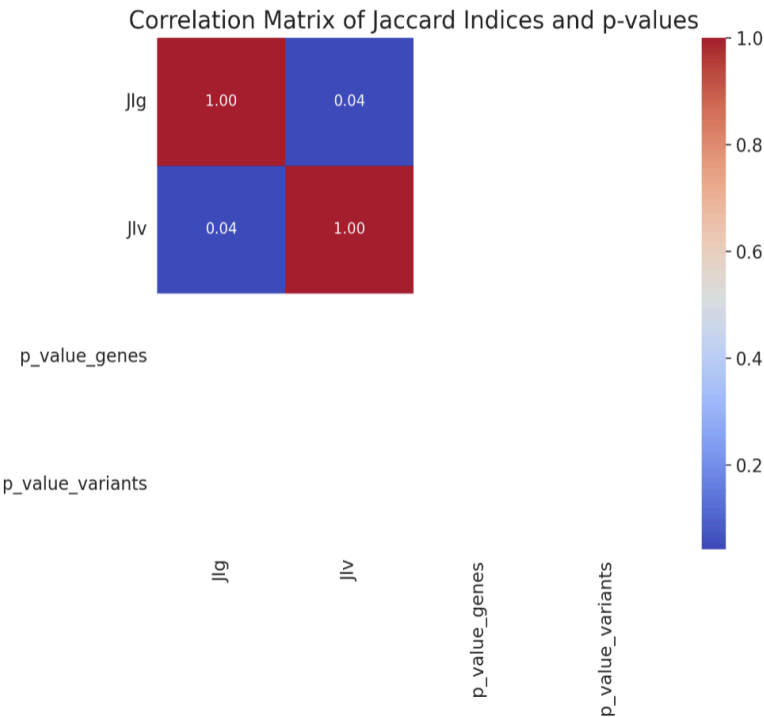


Fig 5.9 Correlation Matrix of Jaccard Indices and p-value Output

- Jlg and Jlv: There is a very low correlation between the Jaccard indices of genes and variants (0.04), suggesting they do not tend to increase or decrease together.
- Jlg, p_value_genes, and p_value_variants: There is no correlation shown

between Jlg and the p-values, which might indicate that the Jaccard Index for genes does not predict the p-values for genes or variants.

- Jlv, p_value_genes, and p_value_variants: Similarly, there is no correlation between Jlv and the p-values, suggesting the Jaccard Index for variants doesn't predict the p-values either.

The red squares indicate a perfect correlation (which is always the case for a variable with itself), and the blue squares indicate a very low to no correlation between the different variables.

```
import scipy.stats as stats

# Define the mean and standard deviation values for head circumference
# by age and gender. You should replace these with actual data.
mean_head_circumference = 45 # Replace with actual mean value
std_dev_head_circumference = 5 # Replace with actual standard deviation
# Input data for the person
age = 3 # Replace with the person's age
gender = 'male' # Replace with the person's gender
person_head_circumference = 60 # Replace with the person's head circumference measurement

# Calculate the z-score
z_score = (person_head_circumference - mean_head_circumference) / std_dev_head_circumference

# Check if the z-score is greater than 2
if z_score > 2:
    print("The person has macrocephaly.")
else:
    print("The person does not have macrocephaly.")
```

The person has macrocephaly.

Fig 5.10 Z- score Output

Python script using the SciPy stats library to calculate the z-score for head circumference, which is used to determine the presence of macrocephaly. The z-score is a statistical measurement that indicates how many standard deviations an element is from the mean of the population. In this context, it compares an individual's head circumference to the population mean to diagnose macrocephaly.

Mean and Standard Deviation: These values should represent the population average and the standard deviation for head circumference based on age and

gender. In the script, these are placeholders and should be replaced with actual data.

Input Data: The age, gender, and actual head circumference of the individual being assessed.

Z-Score Calculation: The script calculates the z-score based on the provided data.

Conditional Statement: It checks if the z-score is greater than 2; a common clinical criterion for macrocephaly is a head circumference more than two standard deviations above the mean.

Output: The script prints out a message indicating whether the individual has macrocephaly based on the calculated z-score.

The use of z-scores in diagnosing macrocephaly is critically examined, including a discussion on the appropriateness of the z-score threshold, potential variations in normative data across different populations, and their effects on diagnostic accuracy. The program facilitates an objective and standardised diagnosis of macrocephaly, enabling the quantification of the degree to which an individual's head circumference deviates from the norm. This quantitative approach is essential for research validation, epidemiological studies, and individual patient assessments in clinical settings.

CHAPTER 6

SUMMARY AND CONCLUSION

This thesis delves into the genetic architecture of macrocephaly by leveraging a bioinformatics approach to uncover causal variants and biological pathways. Macrocephaly, defined by an abnormally large head circumference compared to standard norms for age and sex, may be benign or signify more serious underlying issues. The research capitalises on the power of bioinformatics tools such as genome-wide association studies (GWAS), whole-exome sequencing (WES), and data analytics platforms, spotlighting pivotal genetic variants and crucial pathways like the Ras/MAPK and PTEN/PI3K/AKT/mTOR signalling pathways.

Databases such as DisGeNET and OMIM prove instrumental, supporting the vast data manipulation capabilities offered by Python's Pandas library and visualisation through Seaborn and Matplotlib. This study combines a variety of analytical techniques, including descriptive statistics, trend and network analyses, to scrutinise gene-disease associations and their prevalence across different clinical manifestations of macrocephaly. Such a comprehensive approach helps in prioritising candidate genes and understanding their roles and distribution, enhancing diagnostic precision and therapeutic strategies.

Insights derived from the thesis emphasise the identification of specific genetic variants through rigorous bioinformatics analyses. These insights are pivotal in refining diagnostic processes and treatment modalities for macrocephaly. Depending on the underlying causes, treatment options range from conservative monitoring to aggressive surgical interventions necessary for associated conditions like hydrocephalus.

The findings of this thesis illuminate the intricate genetic networks associated with macrocephaly, highlighting the significant contributions of bioinformatics to genetic research. This work exemplifies how computational methodologies can facilitate a nuanced exploration of genetic data, leading to a more profound comprehension of the biological mechanisms underlying macrocephaly. The catalogue of identified

genetic variants and pathways enriches our understanding and opens up new avenues for targeted genetic therapies.

Moreover, the research advocates for an integrated, multidisciplinary approach to managing macrocephaly, combining cutting-edge genetic research with clinical practices to enhance patient care. Future studies might extend these insights to investigate genetic correlations between macrocephaly and other neurological or developmental disorders, potentially ushering in new diagnostic and therapeutic advancements.

In summary, this thesis not only propels our understanding of macrocephaly's genetic determinants forward but also demonstrates the indispensable role of bioinformatics in the evolution of medical research and healthcare innovations. It sets the stage for future breakthroughs in genetic diagnostics and treatments, paving the way for more personalised and effective medical solutions in neurological health.

Advanced Genetic Discovery and Gene Prioritization:

- Employ cutting-edge bioinformatics tools such as Machine Learning algorithms and deep learning techniques to analyse and predict potential causal genetic variants associated with macrocephaly.
- Prioritise candidate genes using advanced multi-criteria decision-making techniques that integrate genomic data with phenotypic data to enhance the precision of gene-disease associations.

Pathway Analysis and Functional Genomics:

- Conduct extensive pathway analysis to uncover the complex biological pathways involved in macrocephaly, focusing on signal transduction and cellular growth pathways such as the Ras/MAPK and PTEN/PI3K/AKT/mTOR pathways.

- Explore the epigenetic modifications and their roles in macrocephaly development, aiming to link genetic predispositions with environmental triggers.

Comparative Genomic and Phenotypic Analysis:

- Implement comparative genomic studies to explore the genetic overlap between macrocephaly and related conditions such as autism and developmental delays, analysing shared and distinct genetic markers.
- Perform a phenotypic spectrum analysis that correlates genetic variants with the range of clinical manifestations in macrocephaly, utilising advanced imaging and phenotyping techniques.

Development and Validation of Diagnostic Models:

- Develop predictive models for the early diagnosis of macrocephaly using bioinformatics and statistical tools, which can be validated through retrospective and prospective clinical data.
- Establish a standardised diagnostic algorithm that incorporates genetic testing, clinical evaluation, and neuroimaging to differentiate between benign and pathological macrocephaly efficiently.

Therapeutic Targets and Intervention Strategies:

- Identify potential therapeutic targets by understanding the molecular mechanisms and genetic basis of macrocephaly, with a focus on novel gene therapy and pharmacogenomics approaches.
- Evaluate the effectiveness of current treatment strategies and propose a tailored therapeutic regimen based on genetic profiles, aiming for personalised medicine applications in neurodevelopmental disorders.

Educational and Ethical Considerations:

- Develop educational materials and programs for healthcare providers and patients to increase awareness about the genetic aspects of macrocephaly and its management.
- Address ethical, legal, and social implications of genetic testing and personalised treatments in macrocephaly, proposing guidelines to ensure ethical practices in genomic research and clinical applications.

TO SUMMARISE:

- The thesis investigates the complex genetic basis of macrocephaly, utilising advanced bioinformatics tools to analyse causal genetic variants and pathways.
- By integrating genetic databases like DisGeNET and OMIM with data analysis tools (Pandas, Seaborn, Matplotlib), the research enhances understanding of gene-disease correlations and variant distributions.
- Diagnostic and therapeutic strategies for macrocephaly are refined through this research, emphasising the need for precise genetic testing and personalised treatment plans.
- The use of GWAS and WES helps in mapping out critical genetic interactions and mutations associated with macrocephaly, revealing potential targets for therapeutic intervention.
- The study supports a holistic treatment approach by integrating findings from genetic research with clinical evaluations, advocating for a coordinated treatment strategy involving various medical specialists.
- Prospects for future research include extending these genetic insights to other neurological disorders, potentially leading to innovations in diagnosis and therapy.
- The thesis exemplifies the pivotal role of bioinformatics in advancing medical science, particularly in understanding and treating complex conditions like macrocephaly.

REFERENCES

- Andrei, A., Saliba, T., Lubicz, B., & Fricx, C. (2024). Diagnostic Pitfalls of Macrocephaly and Intracranial Dural Arteriovenous Fistulas: Connecting the Dots With the Red Flags. *Cureus*, 16(2), e55288. <https://doi.org/10.7759/cureus.55288>
- Accogli, A., Geraldo, A. F., Piccolo, G., Riva, A., Scala, M., Balagura, G., Salpietro, V., Madia, F., Maghnie, M., Zara, F., Striano, P., Tortora, D., Severino, M., & Capra, V. (2022). Diagnostic Approach to Macrocephaly in Children. *Frontiers in pediatrics*, 9, 794069. <https://doi.org/10.3389/fped.2021.794069>
- Choi, Y. C., Yum, M. S., Kim, M. J., Lee, Y. J., & Ko, T. S. (2016). Megalencephaly-capillary malformation-polymicrogyria syndrome: the first case report in Korea. *Korean journal of pediatrics*, 59(Suppl 1), S152–S156. <https://doi.org/10.3345/kjp.2016.59.11.S152>
- DeCasien, A. R., Trujillo, A. E., Janiak, M. C., Harshaw, E. P., Caes, Z. N., Galindo, G. A., Petersen, R. M., & Higham, J. P. (2022). Equivocal evidence for a link between megalencephaly-related genes and primate brain size evolution. *Scientific reports*, 12(1), 10902. <https://doi.org/10.1038/s41598-022-12953-4>
- Horn, S., Au, M., Basel-Salmon, L., Bayrak-Toydemir, P., Chapin, A., Cohen, L., Elting, M. W., Graham, J. M., Gonzaga-Jauregui, C., Konen, O., Holzer, M., Lemke, J., Miller, C. E., Rey, L. K., Wolf, N. I., Weiss, M. M., Waisfisz, Q., Mirzaa, G. M., Wiczorek, D., Sticht, H., ... Abou Jamra, R. (2019). De novo variants in PAK1 lead to intellectual disability with macrocephaly and seizures. *Brain : a journal of neurology*, 142(11), 3351–3359. <https://doi.org/10.1093/brain/awz264>

Jansen, L. A., Mirzaa, G. M., Ishak, G. E., O'Roak, B. J., Hiatt, J. B., Roden, W. H., Gunter, S. A., Christian, S. L., Collins, S., Adams, C., Rivière, J. B., St-Onge, J., Ojemann, J. G., Shendure, J., Hevner, R. F., & Dobyns, W. B. (2015). PI3K/AKT pathway mutations cause a spectrum of brain malformations from megalencephaly to focal cortical dysplasia. *Brain : a journal of neurology*, 138(Pt 6), 1613–1628. <https://doi.org/10.1093/brain/awv045>

Jeong, G., Kim, M., & Han, B. H. (2014). Clinical features of macrocephaly at birth in Korea. *Korean journal of pediatrics*, 57(2), 75–78. <https://doi.org/10.3345/kjp.2014.57.2.75>

Jones SG, Samanta D. Macrocephaly. [Updated 2023 Jul 24]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK560786/>

Tan, A. P., Mankad, K., Gonçalves, F. G., Talenti, G., & Alexia, E. (2018). Macrocephaly: Solving the Diagnostic Dilemma. *Topics in magnetic resonance imaging : TMRI*, 27(4), 197–217. <https://doi.org/10.1097/RMR.0000000000000170>

Rohde, J.F., Campbell, J., Barbera, J. *et al.* Clinical factors associated with need for neurosurgical care in young children with imaging for macrocephaly: a case control study. *BMC Pediatr* 23, 555 (2023). <https://doi.org/10.1186/s12887-023-04379-2>

Pirozzi, F., Nelson, B., & Mirzaa, G. (2018). From microcephaly to megalencephaly: determinants of brain size. *Dialogues in clinical neuroscience*, 20(4), 267–282. <https://doi.org/10.31887/DCNS.2018.20.4/gmirzaa>

Pavone, P., Praticò, A. D., Rizzo, R., Corsello, G., Ruggieri, M., Parano, E., & Falsaperla, R. (2017). A clinical review on megalencephaly: A large brain as a possible sign of cerebral impairment. *Medicine*, 96(26), e6814. <https://doi.org/10.1097/MD.00000000000006814>

Rozovsky, K., Ventureyra, E. C., & Miller, E. (2013). Fast-brain MRI in children is quick, without sedation, and radiation-free, but beware of limitations. *Journal of clinical neuroscience : official journal of the Neurosurgical Society of Australasia*, 20(3), 400–405. <https://doi.org/10.1016/j.jocn.2012.02.048>

Miral D. Jhaveri MD, ... Chang Yueh Ho MD, in [Expertddx: Brain and Spine \(Second Edition\)](#), 2018

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, 48(D1), D845–D855. <https://doi.org/10.1093/nar/gkz1021>

Sarma, K., Nayak, M. K., Mishra, B., & Gaikwad, S. B. (2022). Megalencephaly-Capillary Malformation-Polymicrogyria Syndrome (MCAP): A Rare Dynamic Genetic Disorder. *Cureus*, 14(5), e25123. <https://doi.org/10.7759/cureus.25123>

Yeung, K. S., Tso, W. W. Y., Ip, J. J. K., Mak, C. C. Y., Leung, G. K. C., Tsang, M. H. Y., Ying, D., Pei, S. L. C., Lee, S. L., Yang, W., & Chung, B. H. (2017). Identification of mutations in the PI3K-AKT-mTOR signalling pathway in

patients with macrocephaly and developmental delay and/or autism. *Molecular autism*, 8, 66. <https://doi.org/10.1186/s13229-017-0182-4>

Tapper, W. J., Foulds, N., Cross, N. C., Aranaz, P., Score, J., Hidalgo-Curtis, C., Robinson, D. O., Gibson, J., Ennis, S., Temple, I. K., & Collins, A. (2014). Megalencephaly syndromes: exome pipeline strategies for detecting low-level mosaic mutations. *PloS one*, 9(1), e86940. <https://doi.org/10.1371/journal.pone.0086940>

Klein, S. D., Nguyen, D. C., Bhakta, V., Wong, D., Chang, V. Y., Davidson, T. B., & Martinez-Agosto, J. A. (2019). Mutations in the sonic hedgehog pathway cause macrocephaly-associated conditions due to crosstalk to the PI3K/AKT/mTOR pathway. *American journal of medical genetics. Part A*, 179(12), 2517–2531. <https://doi.org/10.1002/ajmg.a.61368>

Rohde, J. F., Campbell, J., Barbera, J., Taylor, E., Ramachandra, A., Gegg, C., Scherer, A., & Piatt, J. (2023). Clinical factors associated with need for neurosurgical care in young children with imaging for macrocephaly: a case control study. *BMC pediatrics*, 23(1), 555. <https://doi.org/10.1186/s12887-023-04379-2>

Miller, D., Barnes, P., & Miller, M. (2015). The significance of macrocephaly or enlarging head circumference in infants with the triad: further evidence of mimics of shaken baby syndrome. *The American journal of forensic medicine and pathology*, 36(2), 111–120. <https://doi.org/10.1097/PAF.0000000000000152>

Swarr, D. T., Khalek, N., Treat, J., Horton, M. A., Mirzaa, G. M., Riviere, J. B., Dobyns, W. B., & Zackai, E. H. (2013). Expanding the differential diagnosis of fetal hydrops: an unusual prenatal presentation of megalencephaly-capillary malformation syndrome. *Prenatal diagnosis*, 33(10), 1010–1012. <https://doi.org/10.1002/pd.4178>

Orstavik, K. H., Strømme, P., Ek, J., Torvik, A., & Skjeldal, O. H. (1997). Macrocephaly, epilepsy, autism, dysmorphic features, and mental retardation in two sisters: a new autosomal recessive syndrome?. *Journal of medical genetics*, 34(10), 849–851. <https://doi.org/10.1136/jmg.34.10.849>

Khan, N. E., Bauer, A. J., Doros, L., Schultz, K. A., Decastro, R. M., Harney, L. A., Kase, R. G., Carr, A. G., Harris, A. K., Williams, G. M., Dehner, L. P., Messinger, Y. H., & Stewart, D. R. (2017). Macrocephaly associated with the DICER1 syndrome. *Genetics in medicine : official journal of the American College of Medical Genetics*, 19(2), 244–248. <https://doi.org/10.1038/gim.2016.83>

Tolonen, J. P., Hekkala, A., Kuusmin, O., Tuominen, H., Suo-Palosaari, M., Tynnen, O., & Niinimäki, R. (2020). Medulloblastoma, macrocephaly, and a pathogenic germline PTEN variant: Cause or coincidence?. *Molecular genetics & genomic medicine*, 8(9), e1302. <https://doi.org/10.1002/mgg3.1302>

