SAPIENZA
UNIVERSITÀ DI ROMA

PhD Course in Biochemistry

XXXVIII Cycle (Academic Years 2022-2025)

Dissecting Genotype-Phenotype Associations in Down Syndrome Using WGS and WES Data

Phd Student

Dario Cannella

Supervisor

Prof. Allegra Via

Rome,

15/10/2025

Coordinator

Marialuisa Mangoni

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Summary

5. **Discussion and conclusion**

   5.1 Interpretation of Key Findings and main results
   5.2 Contributions to the Field and Comparison with Existing Literature
   5.3 Main conclusion and future directionS

6. **Bibliography**

7. Glossary of terms and acronyms

1. Introduction

1.1 DS: clinical overview and epidemiology

DS (DS) is the most frequent chromosomal disorder associated with intellectual disability, caused by the presence of a complete or partial extra copy of human chromosome 21 (HSA21). Its name came from John Langdon Down, an English doctor who described the characteristics of the syndrome in 1866 [1]. The link between DS and the 21th chromosome was found for the first time in 1959 [2] and has been an important landmark for the development of genetic medicine. The first mouse models designed to investigate DS appeared in 1990 [3]. A decade later, in 2000, an international research consortium reported the full nucleotide sequence of the long arm of human chromosome 21 [4].

The lifetime prevalence is increasing as the population grows. This is mainly due to the improvements in childhood survival of individuals with DS. As an example, in the USA the DS population has grown from ~ 50 000 in 1950 (3.3 per 10 000 individuals) to approximately 212 000 in 2013 (6.7 per 10 000 individuals) [5]. Moreover, still in the USA, the life expectancy of people with DS increased from a mean of 26 years old and a median of 4 years in 1950 to 53 years and 58 years respectively in 2010 [6]. Accurate global estimates are still difficult to determine, as more birth registries are needed and additional data on both historical and current survival of individuals with DS across different countries are required [7].

DS is primarily recognised for causing intellectual disability and characteristic physical features including short stature, muscle hypotonia, and atlantoaxial instability. However,

DS affects multiple body systems, particularly the musculoskeletal, neurological, and cardiovascular systems. Individuals with DS demonstrate increased susceptibility to various health conditions, including hypothyroidism, autoimmune diseases, obstructive sleep apnoea, epilepsy, hearing and vision impairments, haematological disorders, anxiety disorders, and early-onset Alzheimer's disease (AD) (as shown in Fig 1.1.1) [7]. Conversely, certain conditions such as solid tumours exhibit inverse comorbidity patterns and appear to occur less frequently in individuals with DS compared to the general population [8].

The social and health impacts of DS are multifaceted, encompassing physical, mental, and social well-being. Individuals with DS often report that their health status extends beyond physical conditions to include mental health and social integration [9]. Quality of life in children and adolescents with DS is significantly influenced by the presence of medical comorbidities and social factors such as friendships, with persistent health issues often leading to decreased well-being [10]. Moreover, social inclusion is a critical component of life quality for adults, promoting community participation and positive social relationships; however, challenges such as discrimination and social isolation remain prevalent [10]. Multidisciplinary support programs, which include medical care, physiotherapy, speech therapy, and psychological support, have been shown to improve health outcomes, decrease hospitalizations, and enhance family satisfaction, underscoring the necessity of coordinated care models [11]. The syndrome's impact on families is profound, often requiring parents to reduce work commitments and cope with increased emotional and social burdens. Effective multidisciplinary and social support programs are essential to mitigate these challenges and improve overall family well-being [12] [13].

In conclusion, DS represents a unique paradigm for investigating the relationship between genomic alterations and complex phenotypic expression. The well-characterised trisomy 21 provides a defined genetic framework to examine how

chromosomal dosage imbalance translates into diverse clinical manifestations across multiple organ systems. Furthermore, the increased prevalence of specific comorbidities in DS compared to the general population provides crucial insights into how trisomy 21 influences disease susceptibility patterns, making this condition an invaluable model for understanding how chromosomal abnormalities shape human development and health outcomes.

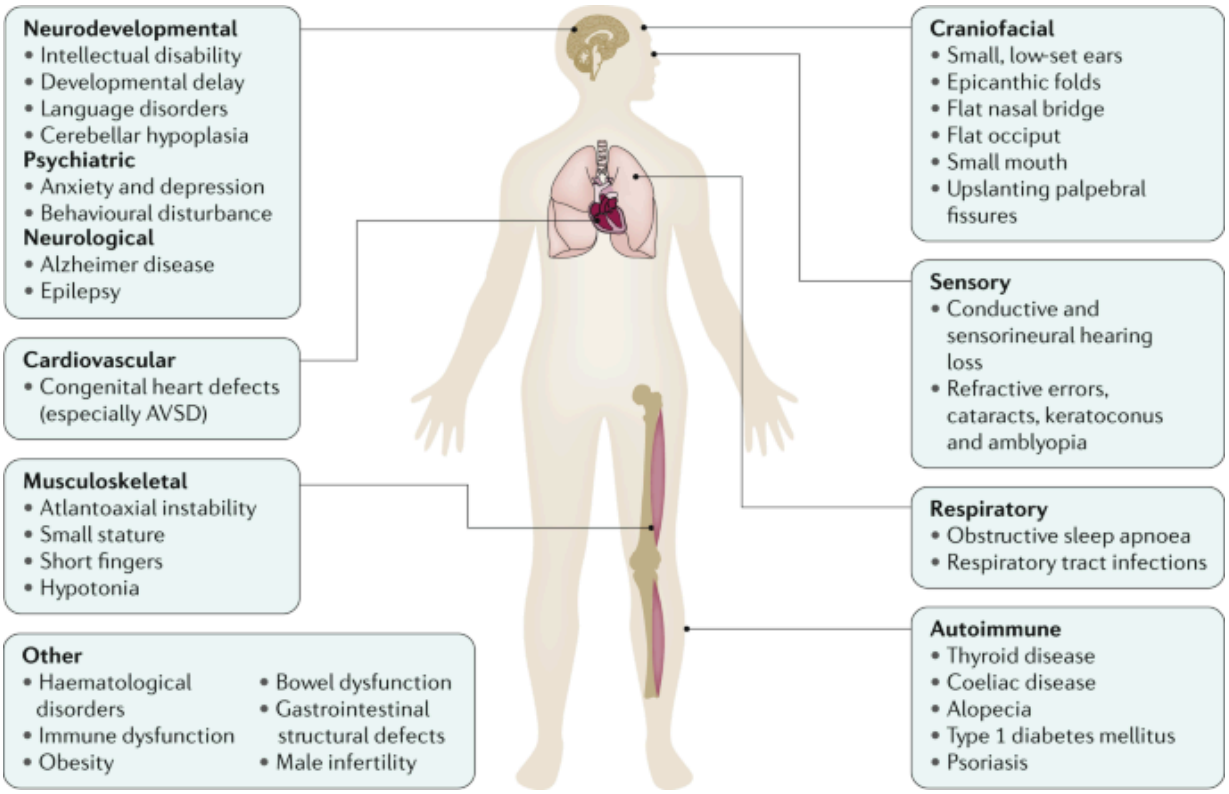

Fig 1.1 The wide spectrum of DS comorbidities [7].

The figure summarizes the wide range of multisystem comorbidities observed in DS patients, affecting neurological, cardiovascular, musculoskeletal, sensory, and immune functions.

1.2 Genomic architecture of DS

Chromosome 21 is recognized as the smallest human autosome, spanning approximately 46.7–48 million base pairs and representing about 1.5–2% of the total cellular DNA and has been the second human chromosome to be fully sequenced after the 22th [4] . This chromosome encodes more than 500 genes - approximately 200 protein coding genes, a small number of known micro-RNAs , and more than 300 genes that may represent functional RNAs, novel proteins, or transcriptional noise [14], which significantly contribute to neurological, immune, and cardiovascular development process.

DS most commonly arises from "free" trisomy 21 (≈ 90%), caused by nondisjunction during maternal meiosis, which is strongly connected with an advanced maternal age due to the cohesin a protein complex that mediates sister chromatid cohesion, homologous recombination, and DNA looping deterioration, spindle instability, and recombination errors [15]. Robertsonian translocations (≈ 3–4%), result from chromosomal rearrangements, sometimes leading to hereditary forms of the disorder, whereas mosaicism (≈1–2%) is due to post-zygotic mitotic errors. These cytogenetic forms introduce significant variability in clinical phenotype and recurrence risk, emphasizing the necessity of cytogenetic analysis for genetic counseling [16].

Historically, a DS Critical Region (DSCR) within 21q22 was proposed as essential for the manifestation of the syndrome, particularly intellectual disability and characteristic facies. More recent evidence argues against the idea of a single critical region; instead, current models emphasize a distributed network of dosage-sensitive genes along the entire chromosome contributing collectively to diverse DS phenotypes [17].

The pathophysiology of DS is nowadays largely attributed to a gene dosage effect, leading to a 1.5-fold increased expression of genes on chromosome 21[18].

Several genes within 21q have been strongly associated with distinct clinical traits, including DYRK1A (neurological and cognitive development), APP (early-onset Alzheimer's disease) SOD1 (oxidative stress), RCAN1, ETS2, COL6A1, DSCR1 [19], and numerous other protein-coding and regulatory non-coding RNAs [7].

Gene dose effects

The most straightforward consequence of trisomy 21 is the increased dosage of individual genes located on HSA21. A notable example is the amyloid precursor protein gene (APP), for which elevated dosage has been shown to contribute to the heightened risk of early-onset Alzheimer's disease (AD) in individuals with DS. APP therefore represents a clear "effector" gene, with current evidence demonstrating that increased APP dosage leads to multiple downstream effects including synaptic protein reductions and alterations in endolysosomal processing. Research has revealed that individuals with DS-AD exhibit decreased levels of several synaptic proteins in the frontal cortex, including SNARE proteins (syntaxin 1A and SNAP25), synaptic vesicle proteins (synaptophysin and synapsin 1), and postsynaptic density protein PSD95. Importantly, these synaptic changes follow a hierarchical temporal pattern, with synaptic dysfunction affecting early childhood and young adulthood, followed by AD-related synaptic and neuronal loss in later stages. Crucially, findings from both partial trisomy 21 (PT-DS) cases and Dp16 mice with normalised APP gene dosage demonstrate that increased APP gene dosage is essential for these synaptic alterations. However, it remains unresolved whether the penetrance and severity of AD are driven exclusively by APP overexpression, as studies in the Ts65Dn mouse model show that restoring App copy number to diploid state mitigates some, but not all, phenotypic consequences, suggesting interactions between proteins or genes located on different chromosomes[20].

While APP exemplifies how gene dosage effects can be traced from chromosomal imbalance to specific phenotypic outcomes, the genomic architecture of DS reveals a more complex picture where both direct gene effects and genome-wide transcriptional dysregulation contribute to phenotypic expression.

The consistent 3:2 expression ratios observed across multiple systems, punctuated by extreme outliers, demonstrate that trisomy 21 creates a network-level perturbation that extends far beyond the additive effects of individual triplicated genes.

This transcriptional dysregulation is further compounded by epigenetic alterations, including global hypermethylation patterns and altered histone modifications mediated by triplicated genes such as DYRK1A and DNMT3L. Additionally, post-transcriptional regulation introduces another layer of complexity, as evidenced by the weak correlation between transcript and protein levels for non-HSA21 genes, suggesting substantial buffering mechanisms that differentially affect protein expression. Mitochondrial dysfunction, involving multiple HSA21 genes including SOD1, APP, and RCAN1, represents a convergent pathway where gene dosage effects translate into altered cellular energetics and oxidative stress [21]. This systems-level understanding provides essential context for interpreting the diverse clinical manifestations observed across different organ systems in DS, where the interplay between direct gene dosage effects, epigenetic modifications, and broader transcriptional imbalances shapes the complex pattern of health conditions that characterise this syndrome [22].

1.3 Comorbidities associated with DS

DS, as discussed above, is known to be associated with a wide spectrum of comorbidities. These comorbidities reflect the genetic dose imbalance produced by the supernumerary copy of chromosome 21. Furthermore, this imbalance causes an alteration in all the genomic interactions thus producing a large variability between the individuals affected.

From a clinical perspective, the comorbidities associated with the DS are the principal determinants of morbidity and mortality as well as the quality of life.

More specifically, there is now a greater focus on avoiding comorbidities through early diagnosis, as well as providing guidelines for everyday life. This is because the life expectancy of people with Down's syndrome has nearly doubled in the last 30 years.

Congenital heart defects

Individuals with DS exhibit a 40-50-fold increased risk of congenital heart defects (CHD) compared to the general population, establishing CHD as one of the most prevalent and clinically significant comorbidities in this population. While cardiac abnormalities in DS were first described in 1894, comprehensive clinical characterisation of CHD types, prevalence, and treatment outcomes was not achieved until the 1950s through landmark studies conducted in Baltimore-Washington and New South Wales. These investigations demonstrated that atrioventricular septal defects (AVSDs) and ventricular septal defects (VSDs) account for approximately 76% of cardiac malformations observed in DS. Current data indicate that roughly half of live-born infants with DS present with CHD, contrasting sharply with the 1% incidence in the general population. However, reported prevalence in DS varies considerably across population-based studies, ranging from 23% to 79% [23].

Molecular investigations have identified a putative critical region encompassing approximately 39 genes implicated in cardiac development. The DSCAM gene, which encodes a cell adhesion molecule, has received particular attention as its overexpression may disrupt endocardial cushion fusion and other cellular processes essential for normal cardiac morphogenesis, potentially leading to AVSDs.

The incomplete penetration of CHD in DS—with 40-60% of individuals unaffected—indicates that trisomy 21 alone is insufficient for cardiac malformation development. This observation has prompted new investigation of additional genetic modifiers, including single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) affecting cardiac morphogenesis pathways. The occurrence of CHDs is influenced by a number of environmental factors, including maternal age and consanguinity. This suggests the presence of complex gene-environment interactions that play a role in the expression of cardiac phenotypes in DS (DS)[24].

Gastrointestinal disorders

Structural congenital anomalies such as duodenal atresia, Hirschsprung's disease, annular pancreas, and anal atresia occur in approximately 4-5% of DS cases, with higher associated morbidity and mortality. Non-structural disorders, including chronic constipation, gastroesophageal reflux disease (GERD), and functional gastrointestinal (GI) disorders , are even more prevalent, impacting quality of life substantially. The epidemiology reflects a consistent pattern of GI tract involvement from infancy to adulthood, with some disorders diagnosed prenatally or shortly after birth [25].

Besides chromosome 21 genes, other non-chromosome 21 genes and genetic pathways modulate risk via SNPs and CNVs.

Studies have localized critical regions on chromosome 21 associated with other GI malformations like duodenal stenosis and imperforate anus, although no other specific chromosome 21 genes have yet been confirmed beyond DSCAM for these defects.

While gene pathways controlling neural crest development — such as those involving the RET receptor tyrosine kinase and SOX10 — are crucial for gastrointestinal tract patterning, their disruption by genetic variation can increase susceptibility to malformations such as Hirschsprung's disease [26].
As of today, there isn't any strong evidence that can explain the relationship between the different GI disorders and DS under a genomic perspective.

Immunological abnormalities

Individuals with DS exhibit alterations in both innate and adaptive immunity, increasing their susceptibility to frequent respiratory infections and otitis media. Autoimmune conditions are notably prevalent, with autoimmune hypothyroidism (including Hashimoto's thyroiditis) affecting approximately 20-30% of individuals with DS, and type 1 diabetes occurring in about 1-10%. Autoimmune skin diseases such as alopecia areata are also commonly observed [27].
These immune system abnormalities are linked to the overexpression of chromosome 21 genes, particularly RCAN1, which contributes to cytokine dysregulation and impaired immune responses. It is also known that excess RCAN1 inhibits calcineurin-dependent nerve growth factor (NGF) signaling, leading to impaired development and survival of sympathetic neurons, contributing to autonomic nervous system abnormalities [28].

Hematological disorder

Children with DS show a 10-20-fold increased risk of developing acute leukemia during infancy, with acute lymphoblastic leukemia (ALL) and acute myeloid leukemia of the megakaryoblastic subtype (AML-M7) representing the predominant forms. Approximately 10% of DS newborns present with transient abnormal myelopoiesis (TAM) [29], a disorder characterised by peripheral blood blast cells that typically

resolves spontaneously within several months, though progression to acute myeloid leukaemia may occur [30].

The development of TAM shows strong association with GATA1 gene mutations, which serve both as molecular markers and critical mediators of leukemogenesis in DS[31]. This gene encodes a transcription factor essential for normal hematopoietic development, and its disruption in the context of trisomy 21 creates a permissive environment for malignant transformation.

The hematological profile in DS presents an intriguing paradox: while childhood leukemia risk is markedly elevated, adults with DS exhibit substantially reduced susceptibility to solid tumours compared to the general population. This inverse relationship suggests that the same genetic alterations underlying increased hematopoietic malignancy risk may confer protection against other cancer types, reflecting the complex and tissue-specific effects of chromosomal imbalance on cellular transformation pathways.

Neurological and psychiatric conditions

The neurological profile in DS encompasses a spectrum of developmental and degenerative conditions that impact cognitive function across the lifespan. Neurocognitive development follows a characteristic pattern of delay, with IQ scores typically falling within the 35-70 range, reflecting the underlying disruption of neural development processes including neurogenesis, synaptogenesis, and myelination [32].
Epilepsy represents another significant neurological comorbidity, with infantile seizures showing particularly high prevalence in this population. The seizure phenotype in DS often differs from that observed in the general population, with specific patterns of onset and progression that may reflect the unique neuropathological landscape of trisomy 21 [33].

The most striking neurological feature of DS, however, is the development of Alzheimer's disease. Neuropathological hallmarks of AD - including amyloid plaques and neurofibrillary tangles - are present in nearly all individuals with DS by age 40, representing one of the most consistent genotype-phenotype correlations in human genetics. Despite this universal pathological burden, clinical dementia symptoms exhibit considerable variability in onset and severity, suggesting that additional factors modulate the translation of pathology to clinical phenotype.

This accelerated AD trajectory stems directly from APP gene triplication on chromosome 21, creating a paradigmatic example of gene dosage effects in human disease. The 1.5-fold increase in APP expression drives amyloid-beta accumulation through enhanced protein production, establishing DS as a natural model for understanding the relationship between APP overexpression and neurodegeneration. However, the incomplete correlation between pathological burden and clinical symptoms underscores the complexity of factors governing cognitive decline in this population [7].

Other medical conditions

The clinical complexity of DS extends beyond cardiac and neurological manifestations to embrace a broad range of multisystem conditions that significantly impact quality of life and long-term health outcomes. These comorbidities demonstrate the far-reaching consequences of chromosomal imbalance on diverse physiological processes.

Metabolic and endocrine dysfunction occurs with notable frequency in DS, reflecting disrupted hormonal regulation across multiple axes. Obesity affects 7-23% of children with DS, driven by reduced metabolic rate, diminished physical activity levels, and the high prevalence of concurrent hypothyroidism. The endocrine profile is further complicated by increased susceptibility to thyroid dysfunction, diabetes mellitus,

gonadal dysfunction, and vitamin D deficiency, necessitating systematic endocrinological surveillance throughout the lifespan [34], [35].

Musculoskeletal abnormalities represent another significant domain of concern, with atlantoaxial instability and generalised joint hyperlaxity posing particular risks. These structural anomalies reflect underlying connective tissue abnormalities and require careful monitoring to prevent potentially serious complications, including spinal cord compression.

Sensory impairments affect the majority of individuals with DS, with visual and auditory deficits showing high prevalence rates. Ocular abnormalities including cataracts and strabismus necessitate routine ophthalmological assessment, while conductive and sensorineural hearing losses require ongoing audiological monitoring. These sensory deficits can significantly impact communication development and educational progress if left unaddressed [7].

Sleep disorders, particularly obstructive sleep apnoea, demonstrate high prevalence in DS and create cascading effects on multiple organ systems. The respiratory dysfunction associated with sleep apnoea contributes to cardiovascular strain while simultaneously affecting cognitive performance and daytime functioning. Early recognition and intervention are essential, given the potential for sleep disorders to exacerbate existing comorbidities and compromise overall health status. Taken together, the wide range of comorbidities observed in individuals with DS underscores the systemic consequences of trisomy 21 beyond neurodevelopmental impairment. These conditions not only represent major determinants of morbidity and mortality but also provide critical insights into the mechanisms by which chromosomal imbalance shapes human health. A comprehensive understanding of such comorbidities is therefore essential for both clinical management and the development of targeted therapeutic strategies.

1.4 The role of genomic technologies in understanding complex traits in DS

Most diseases and conditions, including those typical of DS, are determined by the interplay of multiple genetic and environmental factors. Differently from monogenetic traits, complex traits variation is not fully explained by one or a small number of genes but earthen from a complex set of inherited and environmental factors. The genetic inherited factors, which partially explain the phenotypic variation between individuals, is known as heritability [36], and measures the degree of the similarity among relatives [37].

Genetic variance usually predisposes people to diseases rather than causing them. More complications arise when it interacts with other variants in different genes, under particular environmental conditions and pressures. Penetrance is usually defined as the probability of a person having a disease given their genotype. However, for complex diseases, it makes more sense to define penetrance as a function of the total genetic load, with each genetic variant weighted according to the risk it confers.

On top of the aforementioned concepts to be considered, one must account for the biological complexity at the cellular, tissue and organism level involved in disease pathogenesis [38].

Down's syndrome is a paradigmatic example of such complexity: a single genetic event — the trisomy of chromosome 21 — gives rise to a substantial wide phenotypic spectrum and a heterogeneous set of comorbidities. Despite sharing the same chromosomal imbalance, individuals with Down's syndrome exhibit significant variability in clinical manifestations, ranging from congenital heart defects to haematological disorders, immune dysregulation and neurodegeneration.

This raises a fundamental question: why do some individuals with DS develop specific comorbidities, such as congenital heart disease, leukaemia or early-onset Alzheimer's disease, while others do not?

To answer this question, different genomic technologies have been implemented during the last thirty years, which allowed a better understanding of DS.

Classical cytogenetics remains the foundational diagnostic method by identifying trisomy 21 through karyotype analysis, which differentiates between free trisomy, translocation, and mosaic forms. Completing this approach, molecular cytogenetic techniques such as Array Comparative Genomic Hybridization (Array-CGH) and Single Nucleotide Polymorphism (SNP) arrays enable detailed analysis of gene dosage imbalances, copy number variations (CNVs), and critical chromosomal regions, enhancing molecular characterization beyond the conventional diagnosis[39] [40].

Next-generation sequencing (NGS) technologies, including Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS), have transformed DS research by allowing comprehensive identification of coding and non-coding variants, respectively. WES focuses on discovering coding variants implicated in comorbidities, while WGS extends the analysis to regulatory regions and structural variations, providing an all-encompassing genomic landscape of trisomy 21[41].

Transcriptomic profiling through RNA sequencing (RNA-seq) has exposed widespread gene expression deregulation in multiple DS-affected tissues like the brain, heart, and immune system [42]. This deregulation is characterized by not only the predictable overexpression of chromosome 21 genes due to gene dosage but also global transcriptomic alterations, influencing developmental and pathological processes in DS [43]. Moreover, epigenomic investigations, including methylome mapping and chromatin accessibility assays, emphasize the critical role of epigenetic modifications in modulating gene expression in DS [44] [45].

Proteomic and metabolomic analyses have further contributed to systemic phenotyping by identifying biomarkers related to oxidative stress, metabolic dysfunction, and mitochondrial abnormalities that underlie DS clinical manifestations.

These multi-omic approaches collectively facilitate a deeper understanding into the pathogenesis of DS and hold promise for identifying novel therapeutic targets [46].

The integration of multi-omics technologies has provided unparalleled insights into the pathogenesis of Down's syndrome and related disorders. However, a key challenge remains: linking genomic variation to the diverse range of clinical outcomes. Of the available approaches, whole genome sequencing (WGS) and whole exome sequencing (WES) are the most comprehensive strategies for investigating the genetic basis of comorbidities in DS (DS), and will therefore be the focus of the next section.

1.5 Whole genome and whole exome sequencing in genetic association studies

Whole Exome Sequencing (WES) is a high-throughput approach designed to capture and sequence the exome, which comprises approximately 30 million base pairs representing the 1-2% of the human genome that encodes proteins. This targeted capture enables higher sequencing depth within coding regions at a reduced cost compared to whole genome sequencing (WGS), making WES a cost effective tool for detecting rare e non rare variants associated with Mendelian disorders and complex traits primarily influenced by coding variants [47].

Contrary to WES, WGS technique is designed to sequence entire genomes, including both coding and non coding regions, such as regulatory elements, enhancers and non coding regions. WGS provides a comprehensive coverage that enhances resolution and the detection of different types of genetic variation, including single nucleotide variants, copy number variants (CNVs), and structural variants, with a better sensitivity particularly in the non coding regions. Although WGS entails higher costs and data complexity, it gives a more complete genomic landscape, aiding the detection of variants missed by WES [48], in particular those outside protein coding regions.

Regarding their role in genetic association studies, both WES and WGS have been extensively employed to link genotypic variation to phenotype. Empirical studies

indicate that while WGS assays a larger number of variants overall, the yield of genetic association signals - particularly for coding variants - is largely comparable between WES combined with imputation (WES + IMP) and WGS. Importantly, the choice between WES and WGS has a minimal effect (about 1–2%) on discovery yield in association studies, with larger sample sizes and study design considerations having a greater influence on detecting meaningful genotype–phenotype correlations[49].
Thus, WES remains a cost-effective approach for large-scale association studies focusing on coding regions, whereas WGS provides additional opportunities to explore non-coding variation and complex structural variants [50] [51].

Both the WES and WGS have been implemented in a wide spectrum of comorbidities associated with the DS.
Both technologies offer a better description of the genomic landscape and possible genetic interaction between the supernumerary chromosome 21th and the rest of the genome. Recent literature reports detail WES analyses uncovering novel potentially pathogenic variants that co-occur in genes known to interact in acute megakaryoblastic leukemia, highlighting the need for techniques capable of providing a comprehensive overview of genomic-scale events [52].
In another study, WES enabled the identification of rare variants in different genes among 81 unrelated probands with DS affected by atrioventricular septal defects, a condition known to occur more frequently in individuals with DS [53].

In contrast to WES, WGS expands beyond coding regions, enabling discovery of regulatory and intergenic contributions that underlie complex traits and multifactorial comorbidities in DS. WGS analyses have highlighted the involvement of non-coding regions in modulating gene expression relevant to AD [54], and other neurodegenerative [55] processes prevalent in DS populations.

These genomic approaches collectively support nuanced insights into the genetic architecture of DS comorbidities, bridging single-gene pathogenic variants and complex regulatory mechanisms, thereby aiding precision medicine strategies and managing for the DS population.

2. Aim and structure of the thesis work

The scientific community has reached a consensus regarding the interplay between DS (DS) and comorbidities directly associated with genes located on chromosome 21. This observation suggests that the heterogeneous nature of DS comorbidities may arise from interactions between supernumerary genes on chromosome 21 and genes distributed throughout the rest of the genome. However, no consensus has yet been established on mechanisms underlying comorbidities that are not directly linked to chromosome 21.

This PhD project aims to advance our understanding of comorbidities both associated with chromosome 21 and those not directly related to it. To this end, it will comprehensively exploit whole-exome and whole-genome sequencing data from individuals with DS to identify and interpret novel potential causative variants.

In this thesis, the analyses are focused exclusively on single nucleotide polymorphisms (SNPs). This choice was guided by two main considerations. First, the overarching aim of the project is to generate new insights into the genetic basis of Down syndrome and its comorbidities through an analytical framework that can be applied to both whole-genome and whole-exome sequencing data. Given that exome sequencing targets only the coding regions of the genome, the study prioritised SNPs and gene–gene interactions as the most consistent and interpretable source of information across both datasets. Consequently, structural variants—often located in non-coding regions—were not included in the present analyses.

The thesis is organized into three sections, each building on the previous one:

1. The primary objective of Section 1 is to conduct an exploratory analysis of the whole-genome sequence of 17 individuals affected by DS. The aim was to assess whether it was possible to detect genotype–phenotype associations within a limited sample size. To this end, a framework for functional annotation, variant prioritisation, and statistical analysis was designed and applied. The most prevalent phenotypes were identified, and the distribution of prioritised variants was analysed. The most important outcome of this phase was methodological: it provided a robust framework for subsequent analyses.

2. The second section applies the framework developed in Section 1 to the Biorepository and Integrated Genome (BIG) cohort, for which whole-exome sequencing data from more than 135 samples were available. Here, variant annotation and prioritisation procedures were again performed, with needed adaptations in statistical testing. This larger cohort allowed a more systematic evaluation of genotype–phenotype associations, including an investigation of whether ancestry composition influences susceptibility to specific comorbidities. Chi-square and Fisher's tests were applied to explore associations between prioritized variants and phenotypes, leading to the identification of candidate variants for each analyzed phenotype. Gene sets were then derived by counting the average number of variants per gene, highlighting those enriched in variants associated with specific comorbidities.
3. Section 3 is dedicated to the discussion of the results, with emphasis on the potential role and interactions of the identified genes. This section integrates the methodological advances from the exploratory study with the biological insights gained from the larger cohort, and reflects on how these findings can inform the design of future studies on DS.

Together, these three phases form a coherent trajectory: starting from a pilot analysis to establish methods, scaling up to a larger dataset for association testing, and culminating in a synthesis of results and perspectives. In this way, the thesis not only contributes

new insights into the genetic basis of comorbidities in DS, but also provides a methodological foundation for future research in the field.

3. Methods

3.1 Whole genome sequencing data and variant identification

3.1.1 Study cohort and sample collection

For this study the starting biological material consisted of peripheral blood samples and, for some individuals , parental samples (mother, father, or both) were also available. Participants included both males and females, with age ranging from 4 to 19 years. Recruitment was carried out through telephone interviews, and personal information was collected only after obtaining informed consent from the families. All procedures related to sample collection, DNA extraction, library preparation, and sequencing were performed by the technical staff of the Bambino Gesù Children's Hospital. Downstream analyses, starting from variant calling, were carried out by the author.

3.1.2 Experimental protocol

The DNA samples used in this study were obtained from children with DS and processed at the Ospedale Pediatrico Bambino Gesù laboratories in collaboration with the GenomeUP company. Experimental procedures were carried out using the following workflow and instrumentation:

1. DNA extraction from blood (QIAsymphony);
2. DNA quantification (QUBIT);
3. Library preparation for WGS (Whole-genome sequencing) process:
    - enzymatic fragmentation;
    - end repair;
    - addition of adapters (whole-genome "KAPA EvoPlus" kit by Roche)
4. Cleanup (with Freshly-prepared 80% ethanol);
5. Real time PCR
6. Sequencing step(Illumina Novaseq6000).

In this study the DNA samples were extracted from blood samples belonging to the children affected by DS using the QIAsymphony instrument [56].

The quantification process occurs thanks to the use of the Qubit fluorometric quantification instrument [57]. The kit used for the NGS library construction workflows that requires DNA fragmentation, end repair, A-tailing and adapter ligation is the "KAPA EvoPlus" kit by Roche. Libraries for Illumina sequencing may be prepared from a wide range of DNA samples and inputs (1 ng – 1 μg) in 1.5 – 3 hrs. The kit contains all the enzymes and reaction burrs required for:

1. enzymatic fragmentation to produce dsDNA fragments;
2. end repair and A-tailing to produce end-repaired, 5'-phosphorylated, 3'-dA-tailed dsDNA fragments;
3. adapter ligation, during which dsDNA adapters with 3'-dTMP overhangs are ligated to 3'-dA-tailed molecules.

The enzymatic fragmentation was performed starting with the mixing of 500 ng of gDNA with the FragTail Ready mix, vortexing and centrifuging briefly. After putting the sample on ice, we have mixed the Frag and A-tailing reaction and incubate it in a

thermocycler precooled to 4°C and set the lid temperature to 65°C. For the adapter ligation step we have added:

- 5 microlitres of a unique KAPA UDI adapters
- 10 microlitres of the ligation ReadyMix to each tube containing 60 microlitres
- 75 microlitres was the total amount inside each tube

After mixing and centrifuging samples were incubated at 20°C for 15 minutes on a thermocycler. Once the adapters have been ligated purified our samples were purified out thanks to the KAPA Hyper Pure beads. Have been added 60 microlitres of beads (75+60=135 microlitres), mixed, centrifuged and incubated at room temperature for 5 minutes to allow the sample library to bind to the beads. Then samples were placed on a magnet to capture the beads and Incubated until the liquid was clear. Once the liquid is clear the supernatant has been carefully removed , and, while keeping the sample on a magnet, 200 microlitres of freshly-prepared 80% ethanol have been added. The samples were then incubated at room temperature for 30 seconds, after which the ethanol was removed. This washing step was repeated once, and finally, the beads were allowed to dry at room temperature until all residual ethanol had completely evaporated. The samples were then removed from the magnetic rack, and the beads were resuspended in 25 µL of 10 mM Tris-HCl. The mixture was incubated at room temperature for 2 minutes to allow library elution from the beads. Since this protocol is PCR-free, it was not possible to perform sample quality control using the TapeStation system. To overcome this limitation, real-time PCR was employed for library quantification. The *KAPA Library Quantification Kit* (Roche) was used, and of the 96 available wells, 93 were occupied while 3 remained empty.Specifically, 24 libraries were analyzed in triplicate (72 wells), 6 standards were analyzed in triplicate (18 wells), and 3 wells were used for no-template controls (NTCs), for a total of 93 occupied wells (72 + 18 + 3 = 93). After the necessary calculations were completed, the real-time PCR procedure was performed according to the manufacturer's protocol.

Were needed:

- 20 microlitres of master mix (PCR reagents)
- Primers (non universal, because it is a PCR-free protocol)
- Rox tube (for the signal normalization)

After mixing it all 16 microlitres have been added for each well plus 4 microlitres of sample: for a total of 20 microlitres per well. The mix has been added in each well going from the standard point 1 to the standard point 6, so from the less to the most concentrated point. Once done, we loaded the PCR machine and waited for the results (1 hour and a half). The usual steps of a PCR are:

1. whole-genome, shotgun sequencing;
2. whole exome or targeted sequencing;
3. RNA-seq (starting with cDNA).

The enzymes provided in this kit are temperature sensitive and are shipped on dry ice or ice packs, depending upon the country of destination. Enzymes and reaction buffers must be stored at a temperature range between -15°C to -25°C in a constant-temperature freezer. The starting material for the WGS PCR-free protocol is 500 ng high quality gDNA. Ready to use kits are available for the fragmentation, Atailing (FragTail ReadyMix) and adapter ligation processes (adapter stock and Ligation ReadyMix). For any further details see the KAPA EvoPlus kit protocol:

Once the libraries have been prepared and diluted they were loaded into the "Novaseq6000" sequencer by Illumina [34] for the sequencing process. The "Novaseq6000" is one of the newest sequencing machines released by the Illumina company. It represents a cutting-edge solution for the scalability and flexibility that it owns. With this sequencer it is possible to mix and match flow cell types and also to run

one or two flow cells at a time. The entire process of WGS requires less than 2 days (44 hours).

3.1.3 Bioinformatics protocol

The NGS bioinformatics pipeline is composed of:

- Conversion of the BCL files into FASTQ format files
- Quality control
- Alignment to the reference genome
- Mark duplicates
- Base recalibration
- Variant calling
- VCF annotation

The file produced by the sequencer is in BCL format and must be converted into a FASTQ file. Once sequencing ends, the physicochemical signals obtained during the reaction by the sequencer are decoded to generate sequences composed of nucleotide bases.This conversion is performed by specific algorithms during the base calling step, resulting in a FASTQ file. The FASTQ file consists of four distinct lines:

- The first line begins with "@" and contains the sequence identifier and an optional description;

- The second line contains the nucleotide sequence;

- The third line begins with "+" and may include comments;

- The fourth line encodes the quality values, base by base, for the sequence in line 2, and contains the same number of symbols as there are nucleotides in the sequence [58].

The sequenced data generated by the NovaSeq6000 are stored in a FASTQ file with the corresponding quality scores. The certainty of each base call is recorded as a 'Phred' quality score (see Table 3.1.3), which measures the probability that a base is called incorrectly and measures the base quality in DNA sequencing.

The quality score of a given base, Q, is defined by the equation:

$$Q = -10log_{10}(e)$$

where 'e' is the estimated probability of the base call being wrong.

Phred-scaled quality scores [59] in general can range anywhere from 0 to 60 (Table 3.1). A higher score indicates a higher probability that a particular decision is correct, while conversely, a lower score indicates a higher probability that the decision is incorrect.

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10.000	99.99%
50	1 in 100.000	99.999%
60	1 in 1.000.000	99.9999%

Table 3.1.3: Phred quality score table

Data preprocessing (quality control, alignment, mark duplicates and base recalibration steps), variant calling and annotation steps are performed thanks to the use of the "nf-core/sarek" pipeline [60] version 3.1.1. nf-core/sarek is a workflow designed to detect variants on whole genome or targeted sequencing data. The pipeline is built using the "Nextflow" [61] tool to run tasks across multiple compute infrastructures in a very portable manner.

The tools includes by default in the nf-core/sarek pipeline are:

1. Sequencing quality control: FASTQC
2. Alignment: Bwa-mem 2
3. Mark duplicates: GATK Markduplicates (Picard Tools)
4. Base recalibration: GATK BaseRecalibrator
5. Variant calling: HaplotypeCaller
6. Annotation: Ensemble Variant Effect Predictor (VEP)
7. Reports: MULTIQC.

FASTQC [62] performs quality control checks on raw data and it gives a simple overview of the overall quality of the sequence data. Differently from FASTP [63], FASTQC doesn't modify data. FASTP is one of the tools that can be used for the trimming step, which is still part of the quality control. FASTP can be considered as an all-in-one solution for trimming:

- It gives a comprehensive profile both before and after filtering data
- It filters out low quality reads
- It trims all reads in front and tail
- It cuts the adapters
- It supports long reads

A good practice is to perform a FASTQC analysis both before and after the trimming step, to see how the quality of the reads has changed. The next step is the alignment of the reads to the reference genome in order to generate a BAM file (Binary Alignment Map). This step is performed aligning the reads to the reference genome in order to find the precise location in the genome of each read. The tool used for this step is "bwa-mem2" [64] which is designed for long reads. The reference genome used for the alignment step is the human genome (Homo sapiens) , in particular the GRCh38 assembly. Before the Variant Calling step for the generation of the Variant Call Format file (VCF) [65] two steps must be performed:

- Marking duplicates: Duplicate reads can result from a single amplification cluster, incorrectly detected as multiple clusters by the optical sensor of the sequencing instrument. This is done using GATK Markduplicates [66]
- Base quality score recalibration: It is a process in which machine learning is applied to model the errors produced by the machines empirically and adjust the quality scores accordingly. This is done using GATK BaseRecalibrator (and GATK ApplyBQSR) [67] .

One of the most time-consuming processes is the variant calling for the generation of the VCF file. During this step SNPs and small insertions and deletions (indels) from Next-generation sequencing data are identified. To do so the tool "HaplotypeCaller" [68] by GATK is used.

A VCF file is a text file format that contains metainformation lines, a header line and data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position. As part of the header line, 8 columns are mandatory:

1. CHROM: chromosome
2. POS: position
3. ID: identifier

4. REF: reference bases
5. ALT: alternate bases
6. QUAL: quality
7. FILTER: filter status
8. INFO: additional information

A VCF file must be annotated in order to gain important information about the variants: for example to know if a variant is pathogenic or benign. The "Ensembl Variant Effect Predictor" (VEP) [69] is the tool used for the annotation step. VEP integrates the annotations within the INFO column of the original VCF file. VEP annotation:

- Location of the variants on genome
- Known variants based on database matching
- Gene and transcript names affected by the variants
- Consequences of variation on the protein sequence (missense, frameshift, stop gain/lost etc)

To aggregate and collect the results produced by nf-core/sarek into a single report we have used "MULTIQC" [70] version 1.13.

3.2 BIG study cohort and phenotypic data collection

3.2.1 Cohort composition

The study cohort includes 135 pediatric patients with a confirmed diagnosis of Trisomy 21, stratified by the presence or absence of selected comorbidities and by ancestry. Participants were enrolled through four institutions in Tennessee, USA: Le Bonheur Children’s Hospital (LBCH, Memphis), Regional One Health (ROH, Memphis), East Tennessee State University (ETSU, Johnson City), and the Family Resilience Initiative (FRI, Memphis). Blood samples, obtained either from surplus clinical collections or

through research-specific draws, were linked to de-identified electronic health records to enable genomic and phenotypic analysis.

To study the relationship between DS and its comorbidities, controls were defined for each analysis as patients with T21 who did not have the specific comorbidities under investigation.

Gender distribution in the cohort is roughly balanced, with 76 males and 59 females, which aligns with the expected incidence ratio of T21 (approximately 3 males for every 2 females) [71]. Patients' age ranges from 0 years (1 month) to 32.7 years (age at last visit). Although this biorepository shows considerable genetic diversity, no further stratification was performed due to the limited cohort size.

3.2.2 Phenotypic data sources and comorbidity stratification

All the collected clinical associated data have been extracted from the Electronic Health Records (EHR) and then converted into a limited dataset and finally mapped to a common data model. All the phenotypes have been encoded with PheCodes from the ICD9/10 codes, in order to make them easier to manipulate with statistical analyses. Disease phenotype were defined using PheCodes; congenital malformation of heart was identified by PheCode CM_763; ventricular septal defects using PheCode CM_763.31; acute lower respiratory infections with PheCode RE_460.2; and acute upper respiratory infection using PheCode RE_469.1. For each patient there are multiple records, one for each visit to the hospital.

3.3 BIG whole exome sequencing data and variant identification

The data analyzed in this study, as well as all data stored in the BIG biorepository, were generated and processed in the BIG laboratories and not by the authors. In this study, the authors only perform annotation and subsequent analyses on variants that have already been called. The data analyzed in this study, as well as all data stored in the BIG biorepository, were generated and processed in the BIG laboratories and not by the authors. Variant calling and primary data analysis were performed by the research team of Prof. Vincenza Colonna at the University of Tennessee Health Science Center, Department of Genetics, Genomics and Informatics, in Memphis. In this study, the author worked directly on the unannotated VCF files from the entire biobank, performing variant annotation and subsequent analyses.

3.3.1 DNA sequencing:

The 135 samples analyzed in this study were processed, as well as all the 13152 samples in BIG, with NEB/Kapa reagents, captured with Twist Comprehensive Exome Capture design, enhanced by Regeneron-designed spikes targeting sequencing genotyping sites. In the biorepository 95.2% achieved an average sequencing depth of at least 20X, and 99.3% of the samples had > 90% of their bases covered at 20X or greater, all this data underlines the quality of the data.

The genotype spike targets roughly 1.4M variants more in the human genome. Finally the genotyping call rate was 99% and all samples were sequenced on an Illumina NovaSeq 6000 system on S4 flow cells sequencer using 2 x 75 paired-end sequencing [72].

3.3.2 Variant identification

The DNA sequences were aligned by the Burrows-Wheeler Aligner (BWA) MEM [73] to the GRCh38 assembly of the human reference genome in an alt-aware manner. Duplicates have been marked using Picard and mapped reads were sorted using sambamba [74]. Variants have been called with a custom exome model from DeepVariant v0.10.0 [75] , and the GLnexus v1.2.6 tool was used for joint variant calling. Variants have been annotated using Varian Effect Predictor (VEP113) [69].
Phasing was performed using ShapeIT [76]. The cohort dataset is composed of 6,726,045 variants.

3.3.3 Variant prioritization

To establish a possible association between all the selected comorbidities and Trisomy 21, the two complete variant datasets underwent a prioritization process. More specifically, we designed three different filters to remove those variants considered unlikely to be meaningful for the association study. The first filter selected only variants with a frequency < 0.05 in gnomADg (applied in WGS analyses) and those variant with REVEL [77] score > 0.7 and a CADD [78] score ≥ 24 and the second one instead filtered those variants with an exome frequency < 0,0005 gnomADe [79] and an "IMPACT" effect predicted by VEP as "HIGH" (applied in WES analyses). The second one selected variants predicted as "probably damaging" by the software PolyPhen [80] and applied the same exome frequency filter (applied in WES analyses) .The third also maintained the same allele frequency filter and retained variants predicted as "deleterious" by the bioinformatic tool SIFT [81] (applied in WES analyses).

REVEL[77] is a VEP plugin that provides a pathogenicity prediction score for missense variants, integrating multiple computational algorithms to estimate the likelihood that a variant is deleterious. CADD [78] is a VEP plugin that computes a deleteriousness score for genomic variants, combining evolutionary and functional data to assess the potential functional impact of a variant. Both PolyPhen and SIFT are bioinformatic tools widely used to predict the functional effects of missense variants, but with two different approaches: SIFT relies on how conserved the nucleotide (or corresponding amino acid) is across orthologous proteins in different species, while PolyPhen bases its predictions primarily on protein structure information rather than conservation across species.

The results of all these filters were merged to form a final dataset on which we conducted association studies.

3.4 Ancestry inference and population structure

Ancestry inference for the BIG biorepository [72], from which the present cohort of 135 patients with Trisomy 21 was derived, was previously performed as part of a larger population-scale study. Both global and local ancestry were inferred using RFMix v2.0[82], based on reference panels from the 1000 Genomes Project [83] and the Human Genome Diversity Project (HGDP) [84], and further supported by PCA and ADMIXTURE-based clustering. Individuals were assigned to discrete ancestry categories (EUR, AFR, AMR, EAS, or admixed groups) based on global ancestry proportions, following thresholds defined in the original study [72].

In the present analysis, these ancestry labels were used as provided, without additional deconvolution. For detailed ancestry deconvolution methodology, refer to “Buonaiuto, S., Marsico, F., Mohammed, A. *et al.* Insights from the Biorepository and Integrative Genomics pediatric resource. *Nat Commun* 16, 4750 (2025).”.

3.5 Genotype-phenotype association analyses

To study the association between Trisomy 21 and its comorbidities across both cohorts for each PheCode (phenotype), we applied the Chi-square test when applicable) and the Fisher's exact test when the Chi-Square test was not feasible, at the single-variant level using the R software environment for statistical computing [85].

We performed statistical tests at the single-variant level for each PheCode (phenotype). For each comorbidity, we selected the patients affected by that condition and labeled them as "cases", while those without the comorbidity at hand were labeled as "controls". Then, for each variant, we computed a 2×3 contingency table where the rows represented "cases" and "controls" and the columns corresponded to the three genotypes: reference homozygous, reference/alternate heterozygous, and alternate homozygous. On these tables, the Chi-Square test is applied to assess the association between genotype and phenotype when all counts are greater than five. When any count is less than five, Fisher's exact test is used. Correction for multiple testing and data visualization strategies are described in the following section.

3.6 Statistical methods and multiple testing correction

All statistical analyses were conducted in R (version 4.4.2) [85].

To address the issue of multiple hypothesis testing, we applied the Benjamini-Hochberg [86] procedure to adjust the p-values obtained from the association tests described in the previous section. This correction was performed separately for each comorbidity, considering all variants tested for that specific phenotype.

Associations were then evaluated based on the adjusted p-values, adopting a false discovery rate (FDR) threshold of 0.05. For each phenotype, a Manhattan plot was generated to visualize the distribution of association signals across the exome. Manhattan plots were generated using the adjusted p-values obtained through Benjamini–Hochberg correction. A significance threshold was set at $-\log_{10}(p) \approx 6.5$, corresponding to the adjusted p-value cutoff for the selected FDR level.

To further evaluate the robustness of the observed associations, a statistical power analysis was performed to estimate both the achieved power and the sample size required to reach 80% power for genome-wide significance ($\alpha = 5\times10^{-8}$). For each variant, Z-scores were computed from the two-sided p-values, and power was estimated as

$$Power = 1 - \theta(Z_{crit} - Z)$$

where Z_{crit} corresponds to the critical value at $\alpha = 5\times10^{-8}$. The required sample size (N_{needed}) to achieve 80% power was derived from the relation

$$N_{needed} = N_{obs}\, x\, (Z_{crit}/Z)^2$$

Variants with undefined or zero minor allele frequencies (MAF) were assigned a default MAF of 0.005 to ensure numerical stability.
This analysis was implemented in a custom R script using the *data.table* package, producing per-variant estimates of observed power and sample size requirements.

The analysis was carried out using several R packages, including *readr,* *data.table*, and *dplyr* for data handling, *stringr* for identifier processing, and *ggplot2* and *gridExtra* for graphical output.

Prior to testing, quality control steps were performed to ensure consistency in file formats and correct encoding of both genotypes and phenotypes.

Variants with missing values or inconsistent formatting were excluded from the analysis.

To assess mutational load across the genome, we quantified for each gene the number of detected variants normalized by its genomic length (kb), obtaining a Variants_per_kb metric that corrects for gene size bias. The resulting values were compared to the genome-wide mean variant density to compute a Relative_variation_index (RVI), representing the fold enrichment in mutational density (RVI > 1 indicates above-average variation). Genes with RVI values exceeding the upper interquartile threshold (Q3 + 1.5

× IQR) were classified as high-burden genes, reflecting loci with a statistically significant excess of variants. An additional extreme cutoff (Q3 + 3 × IQR) was used for graphical representation.

This analysis was implemented in a custom R script , which automatically identified the relevant columns, computed interquartile statistics, filtered outlier genes based on the IQR criterion, and generated log-scaled plots illustrating the distribution of RVI values. Output tables reported, for each gene, its symbol (*Gene_Names*), normalized mutation density (*Variants_per_kb*), and relative enrichment score (*Relative_variation_index*).

3.7 Functional annotation and interpretation of variants and overburden genes

To functionally annotate the variants included in the final dataset analyzed, we used the Variant Effect Predictor (VEP, version 113) [69] developed by Ensembl [87]. VEP is run on the complete set of variants, and annotation is performed on the canonical transcripts. Core annotations—including gene name, transcript ID, consequence type, impact, clinical significance, molecular consequence, and exon—are included, along with variant frequencies from gnomADe [79]. Finally, additional functional predictions are integrated through the use of key plugins. We used polyphen-2 [80]to predict the possible impact of amino acids substitutions on the structure and function of proteins, using physical and evolutionary comparative considerations and classifying them from benign to damaging. We used SIFT [81] to evaluate the degree of conservation of amino

acid positions in protein alignment derived from closely related sequences (species) and classify substitutions as tolerated and deleterious.

Only variants predicted as “probably_damaging” by Polyphen and “deleterious” by SIFT, or classified by VEP as “HIGH” impact and with a "missense_variant" consequence, always with also a frequency in Gnomade < 0.05, were retained for further genotype-phenotype analyses, as detailed in Section 3.2.3.

The output from VEP was also used to support the biological plausibility of candidate gene and variants association identified in statistical analyses, especially for prioritizing variants with potential functional relevance.

All annotations were obtained in tabular format and parsed using custom R scripts.

Finally, the VarElect [88] web application was employed to evaluate potential direct or indirect relationships between over-burdened genes and each comorbidity, analyzed case by case within the BIG cohort.

To further assess the robustness of the detected associations, we evaluated gene-level burden signals using a custom R script that identified genes with an unusually high relative variation index (RVI) compared to the cohort distribution. These genes were subsequently prioritized for biological interpretation through VarElect analysis. In parallel, we performed a post-hoc power analysis to estimate, for each variant, the achieved statistical power and the sample size that would have been required to reach a conventional genome-wide significance threshold ($\alpha = 5\times10^{-8}$). This analysis provided an additional measure of confidence in the interpretability of our findings and the adequacy of the cohort size.

4. Results

The results presented in this thesis follow the structure delineated in Section 2. Aim and structure of the thesis work.

Section 4.1 is dedicated to an exploratory analysis of whole-genome sequencing data from 17 individuals with DS . The objective of this analysis is to assess the feasibility of detecting genotype–phenotype associations in a small cohort. Additionally, it aims to establish a methodological framework for variant annotation, prioritisation, and statistical analysis.

Section 4.2 applies this framework to a larger cohort of 137 individuals from the Biorepository and Integrated Genome (BIG) project, leveraging whole-exome sequencing data to systematically investigate genotype–phenotype associations and explore the influence of ancestry on susceptibility to specific comorbidities.

Finally, Section 4.3 integrates the findings from both studies, emphasising the biological interpretation of identified variants and gene sets, and discussing their potential implications for understanding comorbidities in DS and guiding future research.

4.1 Initial exploratory study of WGS data for 17 individuals

4.1.1 Cohort description and phenotypic stratification of WGS data

The exploratory study cohort comprises 17 mostly pediatric patients with DS from the Ospedale Bambino Gesù (OPBG) in Rome, Italy. Whole-genome sequencing for all patients was performed with the valuable support of the GenomeUp company. The entire sample is characterized by 52 different comorbidities, so each patient has an average of 3 different registered comorbidities at the time of recruiting. The entire set of comorbidities ranges from pathologies that are well known to be associated with DS such as obesity, celiac disease, congenital heart defects, recurrent respiratory infections, intellectual disability and also comorbidities like scoliotic deviation, chronic kidney disease, hepatic angioma; the entire set of co-morbidities is reported in Table 4.1which reports, for each disease, its ICD-10 code (International disease code - 10th version) and its description . We spotted three different comorbidities displaying significantly high prevalence; these are identified through H52, H52.2 and E03.9 ICD-10, which correspond to hypermetropia, astigmatism and hypothyroidism, respectively, as reported in Figure 4.1. The gender distribution comprises eight males and nine females.
This small set was used to do exploratory analyses and initial associations between individuals affected by DS and comorbidities.

List OPBG cohort Comorbidity	
ICD10	**Description**
K59.0	Obstinate constipation
H52.0	Hypermetropia
L73.9	Folliculitis
M41.2	Left convex lumbar scoliosis
M21.1	Bilateral bowleg (genu varum)
Q70	Bilateral syndactyly of the 2nd-3rd toe
Q21.3	Repaired tetralogy of Fallot
E03.9	Autoimmune thyroidism (hypothyroidism
H52.2	Astigmatism
H00.02	Chalazion of the left eye
R32	Nocturnal enuresis
K07.3	Class I malocclusion tending towards class III with dental crowding
Q25.1	Surgically corrected patent ductus arteriosus
J31.0	Recurrent respiratory infections
K59.00	Constipation (unspecified)
Q87.0	OMPC syndrome (other specified congenital malformations)
Q24.9	Outcomes of corrected congenital heart disease
H50.0	Convergent strabismus
M21.4	Flatfoot and knee valgus
Q78.8	Fibrodysplasia ossificans progressiva
E66.9	Obesity (unspecified)
G47.3	Suspected obstructive sleep apnea
N83.2	Left adnexal cyst (ovarian cyst
K76.0	Severe hepatic steatosis (fatty liver
Q42.1	History of surgically treated anal atresia
G40.9 R50.9	Suspected seizure crisis with fever and apyrexia
N13.3	Left pyelectasis (hydronephrosis unspecified)
Q55.2	Retractile testes
Q21.1	Spontaneous closure of the patent ductus arteriosus
F80.9	Language deficit (specific developmental disorder of speech and language
K07.4	Class III malocclusion
F72.0	Severe intellectual disability
Q78.2	IAO osteopetrosis
R13	Dysphagia for solid foods
H90.4	Profound left-sided sensorineural hearing loss
H55.0	Nystagmus
I44.2	Surgically corrected complete atrioventricular canal
H71	History of cholesteatoma
K90.0	Celiac disease
E00.9	Congenital hypothyroidism
M41.8	Scoliotic deviation
E50.8	Follicular keratosis
K76.4	Hepatic angioma
D22.9	Melanocytic nevi (unspecified)
Q66	Hallux valgus
E66	Overweight
H90	Conductive hearing loss
R73.0	Suspected initial insulin resistance
K02.9	Dental caries
H53.9	Punctate cortical opacities
N18.9	Chronic kidney disease
N13.7	Vesicoureteral reflux

Table 4.1.1 - Comorbidities present in the OPBG cohort. *ICD-10: International disease code 10th*

version.

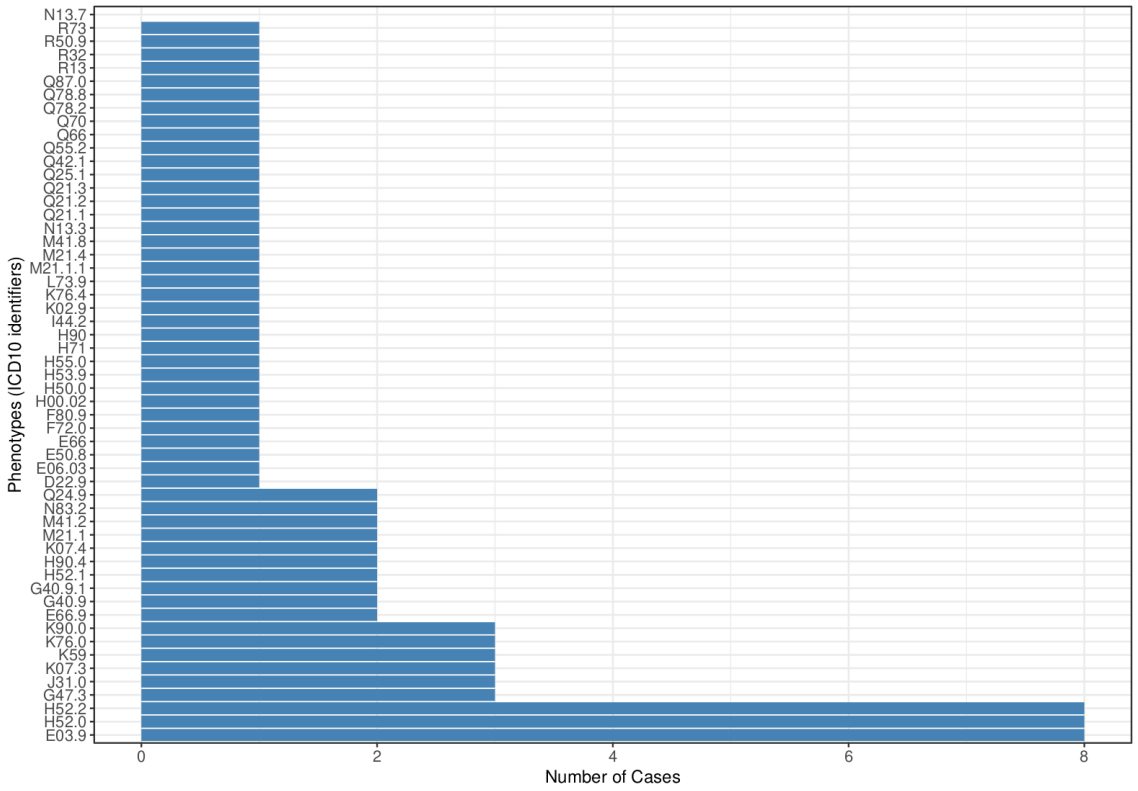

Figure 4.1.2 - Distribution of comorbidities in DS patients.

The most prevalent phenotypes are H52, H52.2, E03.9. corresponding to hypermetropia, astigmatism and hypothyroidism, respectively.

4.1.2 Summary of detected variants and genetic landscape of WGS data

The most compelling aspect of Trisomy 21, and the central aim of this PhD project, is to elucidate why this condition is accompanied by such a wide spectrum of comorbidities occurring at a higher prevalence compared to the non-syndromic population. To address this question, single nucleotide polymorphisms, deletions, and insertions were identified from the WGS data of the 17 patients.

The analyses revealed a total of 11,854,727 variants across the OPBG cohort, the majority being single nucleotide variants (SNPs) (80.15%), followed by insertions (9.59%) and deletions (9.95%). Among the variants predicted with the Variant Effect Predictor (VEP, see Methods) to have the most severe functional consequences, 4,039 were classified as stop-gained mutations, 6,729 as frameshift variants, 790 as stop-lost, 887 as start-lost, 5,874 as splice acceptor, and 5,749 as splice donor variants. Variants with a moderate predicted impact were predominantly missense variants, totaling 286,830. The distribution of variants is presented in Figure 4.1.3 and 4.1.4. As expected, the first two chromosomes exhibit the highest number of variants in absolute values, reflecting their larger size and higher gene content, as shown in Figure 4.1.4. . As expected, chromosomes 1, 2, display the highest number of variants per million base pairs across all chromosomes per megabase (Mb) in Figure 4.1.3. Conversely, chromosomes 22 and 21, being the smallest, display the lowest number of variants in absolute values and per Mb.

Overall, the number of variants per chromosome appears to correlate with chromosome length and average gene density.

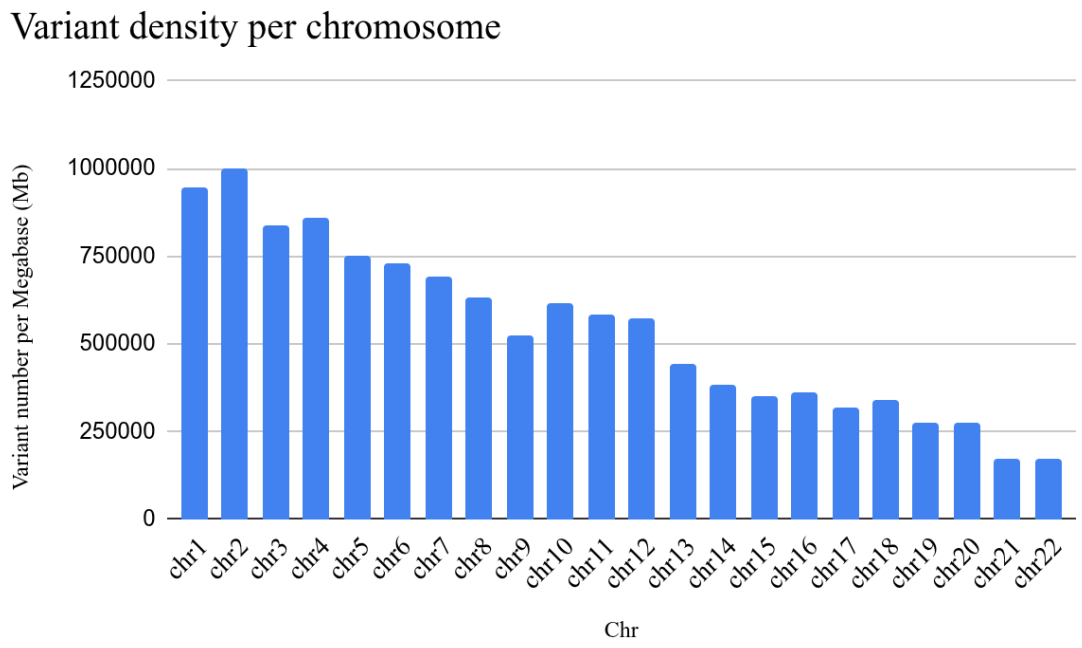

Figure 4.1.3 - Number of variants per million base pairs (Variants/Mb) for each human chromosome in the OPBG cohort.

On the Y axis is shown the variant count per Mb and on the X axis the chromosome number.

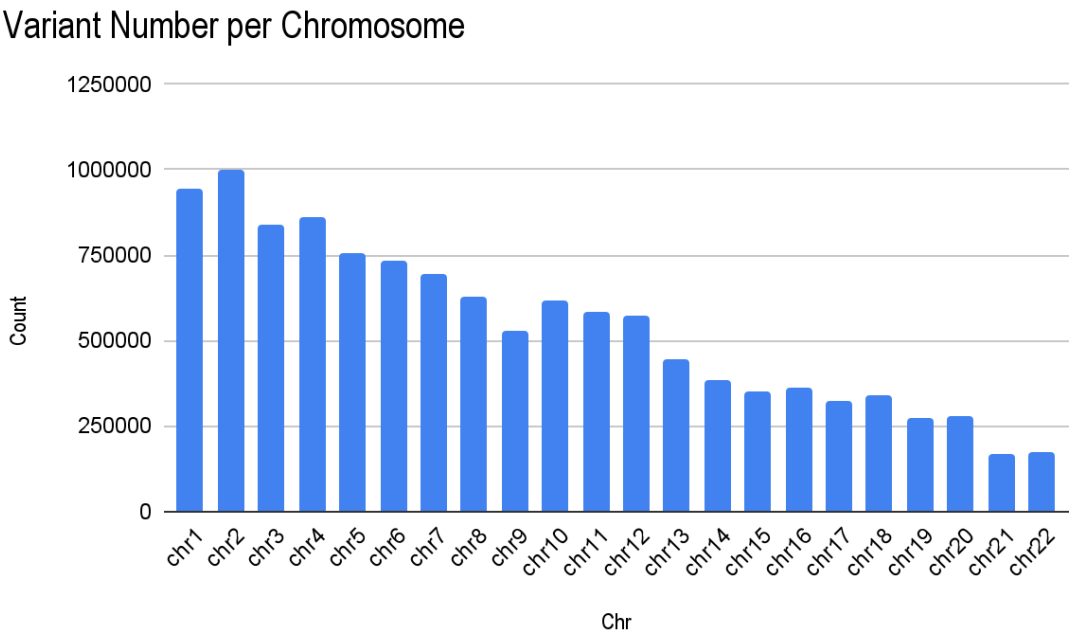

Figure 4.1.4 - Distribution of the variant count across the OPBG cohort. *On the Y axis is shown the variant count and on the Y axis is shown the chromosome numbers.*

The tool Variant Effect Predictor (VEP) (see Methods) was also used for the functional impact assessment thanks to the associated plugins Polyphen-2 and SIFT, which predict the effect of mutations at the protein level. SIFT analyses classified 21,325 (20.92%) mutations as "deleterious", 14,830 as "deleterious with low confidence" (14.55%), 17,371 as "tolerated" (17,04%), and 48,407 (47.49%), suggesting that 35,47% of annotated variants may impact the protein function.

Polyphen-2 prediction classified 11,049 (11.03%) mutations as “probably_damaging”, 11,794 (11.77%) as “possibly_damaging” , 70,508 (70.40%) as “benign” and 6,847 (6.83%) having "unknown_effect”. Even though both algorithms indicate that the majority of variants are likely benign, these predictions must be interpreted within the specific genetic context of DS, as the tools do not account for the unique genomic background of trisomy 21, which may alter the functional consequences of variants identified in this DS cohort.

This limitation highlights the importance of a context-specific interpretation when assessing variant pathogenicity across chromosomally distinct populations and will be further discussed in the Discussion section.

4.1.3 Phenotype-genotype association analyses.

To identify genetic variants associated with the comorbidities identified in the OPBG cohort, classical statistical tests such as Fischer Test and Chi Square have been used between the variant genotypes of the patients with and without a specific ICD-10 code (for a specific comorbidity). The three most prevalent phenotypes were selected - these are H52, H52.2 and E03.9 (see Table 4.3.1); all of them are already known to have a higher prevalence in the DS community.

In order to assess genotype-phenotype associations , both Fisher's exact test and the chi-square test were applied. Fisher's exact test was employed in instances where the number of observations in the contingency table was less than five, while the chi-square test was utilised under standard conditions.

We implemented two complementary approaches to prioritize and identify potential pathogenetic variants the first one on the WGS data and the refind (2° one) one for the WES data :

1. The first approach applied three stringent filtering criteria: variants were required to have a minor allele frequency (MAF, i.e., the frequency of the less common allele in the general population) below 5 in 100 in the gnomADg genome database (gnomADg_AF < 0.05) to focus on very rare variants; a REVEL pathogenicity score greater than 0.7; and a CADD score exceeding 24, to identify variants with a high predicted functional impact;
2. The second approach was based on multiple functional prediction tools, such as Polyphen-2, SIFT and VEP prediction of the molecular impact (“IMPACT vcf field”) according to the Ensembl’s standardized consequence definitions and also the gnomADe frequency..

The major difference between the two aforementioned filter criteria is the use of CADD and REVEL in the first one and the use of SIFT and Polyphen for the second one.
CADD is a VEP plugin that computes a deleteriousness score for genomic variants, combining evolutionary and functional data to assess the potential functional impact of a variant, REVELis a different plugin that provides a pathogenicity prediction score for missense variants, integrating multiple computational algorithms to estimate the likelihood that a variant is deleterious.

All the variants are reported in Table 4.1.5 with the following annotations: genomic position, reference and alternative alleles, rsID, predicted consequence, gnomADe allele frequency, associated gene symbol, pathogenicity scores (CADD and REVEL), and number of carriers.
Fisher's exact and chi-square tests were applied to assess potential associations between these variants and the two comorbidity groups (refractive disorders H52 and H52.2 and hypothyroidism E03.9). No significant associations were observed.
The second approach, instead, revealed variants also based on multiple functional prediction tools. Polyphen-2 identified 383 probably damaging variants (score > 0.85), while SIFT predicted 2,889 deleterious variants (score <0.05).

The frequency of causative alleles at the variable sites was analyzed in the two case groups with comorbidities: patients with disorders of refraction and accommodation (ICD-10 code H52, n=10) and patients with hypothyroidism (ICD-10 code E03.9, n=8). These frequencies were compared against control subjects using three statistical approaches, respectively Fischer Test, Chi Square. None of the variants analyzed showed statistically significant associations ($p > 0.05$) with either refractive disorders or hypothyroidism across all three statistical tests.

This lack of significant associations suggests that the identified potentially damaging variants may not play a major role in the development of these specific comorbidities in our cohort.

In parallel, we explored the phenotypic plausibility of the most recurrently mutated genes by submitting them to the VarElect prioritization tool (GeneCards Suite) (see Methods). VarElect ranks genes based on their direct or indirect associations with user-defined phenotypes through a combination of literature-mining, gene–gene interaction networks, and functional annotations.Although these associations do not provide statistical evidence, they highlight potentially relevant biological connections that may warrant further functional investigation.

Surely the small sample size (n=10 and n=8) substantially limited the statistical power of detecting genetic associations.

Missense variants identified with gnomADg frequencies and predictive scores							
Chr-Pos-Ref-Alt	**Rs-Id**	**Consequence**	**gnomADg**	**Symbol**	**Cadd**	**Revel**	**carriers**
chr6:110442732:T:C	rs41288594	missense	1 x 10^-2	SLC22A16	25.0	0,87	2
chr1:48228922:G:A	rs61746559	missense	6.702 × 10^-6	SLC5A9	29.0	0,85	2
chr21:34096306:G:A	rs35707420	missense	6.697 × 10^-6	SLC5A3	28.0	0,83	2
chr19:44781009:A:C	rs35106910	missense	4 x 10^-2	CBLC	24.0	0,81	2
chr5:103003107:A:G	rs35658696	missense	5 x 10^-2	PAM	27.0	0,79	2
chr3:122415185:C:T	rs61736421	missense	6.696 × 10^-6	WDR5B	25.0	0,80	2
chr16:84237137:G:A	rs3803641	missense	1 x 10^-2	KCNG4	27.0	0,80	2
ch3:132684574:C:T	rs34391943	missense	1 x 10^-2	NPHP3	25.0	0,79	2
chr9:94567343:C:T	rs61755092	missense	6.698 × 10^-6	FBP2	26.0	0,78	2
chr4:121380442:A:C	rs34270076	missense	1 x 10^-2	QRFPR	27.0	0,78	2
chr1:103575299:G:A	rs140978983	missense	3 x 10^-2	AMY2B	27.0	0,78	2
chr11:10487356:G:T	rs117706710	missense	1 x 10^-2	AMPD3	25.0	0,77	2
chr2:224808020:C:T	rs200998922	missense	6.7045 × 10^-6	DOCK10	27.0	0,74	2
chr4:55106779:A:G	rs34231037	missense	3 x 10^-2	KDR	24.0	0,74	2
chr15:42287751:G:A	rs145853612	missense	1 x 10^-2	GANC	27.0	0,73	2
chr9:5361143:G:C	rs72703655	missense	1 x 10^-2	PLGRKT	29.0	0,73	2
chr4:103091208:A:G	rs140261412	missense	1 x 10^-2	BDH2	25.0	0,72	2
chr10:13283784:C:T	rs62619919	missense	8 x 10^-3	PHYH	21.0	0,70	2
chr22:30470066:G:A	rs114438349	missense	2 x 10^-2	SEC14L3	31.0	0,80	3
chr22:30468668:A:G	rs115278158	missense	1 x 10^-2	SEC14L3	26.0	0,73	3
chr16:47663707:A:G	rs16945474	missense	4 x 10^-2	PHKB	24.0	0,88	4
chrX:38408967:A:G	rs1800328	missense	4 x 10^-2	OTC	26.0	0,83	5

Table 4.1.5 - List of missense variants identified in the OPBG cohort.

Annotations include: genomic position, reference and alternative alleles, rs identifiers, predicted consequence, gnomADg allele frequency, associated gene symbol, pathogenicity scores (CADD and REVEL), and number of carriers.

4.2 Application of the framework to WES data from the Biorepository and Integrated Genome (BIG) cohort

4.2.1 Cohort description and phenotypic stratification of WES data

The study cohort consists of 135 pediatric patients with DS from the Biorepository and Integrative Genomics (BIG) initiative [72]. BIG is a comprehensive research platform that combines genomic, phenotypic, and environmental data from individuals representing diverse demographic backgrounds, including underrepresented and mixed populations.

The BIG Initiative employs detailed ancestry stratification and identifies ancestry-specific functional variants of clinical significance while maintaining substantial community engagement and genomics education components.

The cohort includes 22 patients of African ancestry, 81 of European ancestry, 18 with African-European admixture, 4 with European-American admixture, and 10 with multiway ancestry involving more than two ancestral backgrounds.

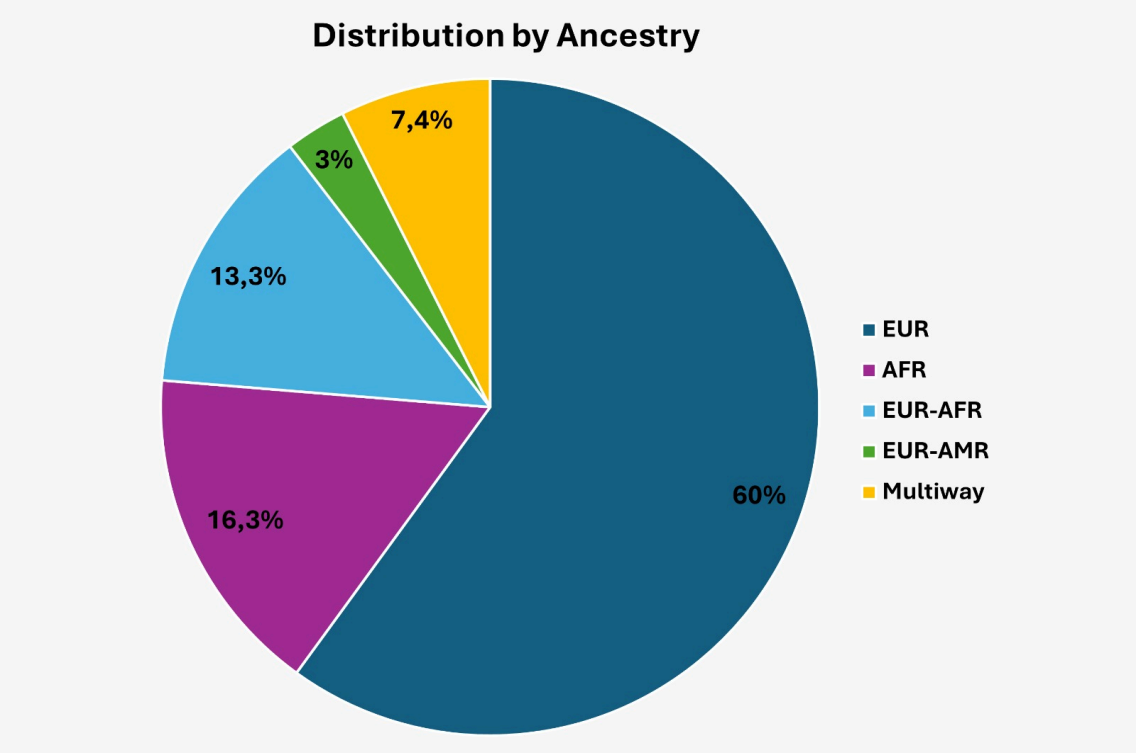

Figure 4.2.1 - Ancestry distribution within the BIG cohort. *AFR: African, EUR: European, EUR-AFR: European-African, EUR-AMR: European-American. Multiway indicates more than two different ancestries.*

Gender distribution comprises 59 females (44%) and 76 males (56%).

The dataset encompasses over 1,000 unique phenotypes, with patients averaging 41 distinct diagnoses each. Phenotype distributions according to PheCode categorization (see Methods) were stratified by ancestry group as shown in *Fig 4.4.1*. No statistically significant associations were found between diagnostic codes and genetic ancestry within the cohort.

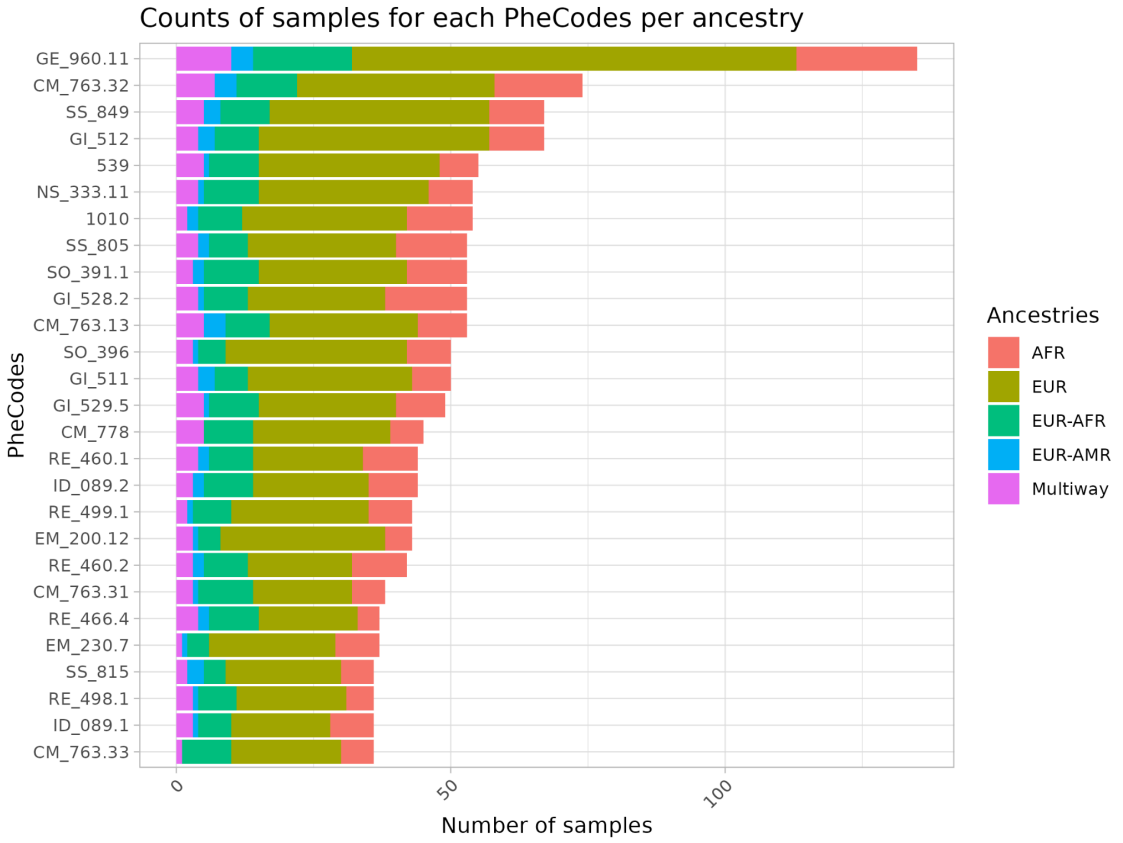

Figure 4.2.2 - Counts of samples for each Phecodes per ancestry .

This figure has been done for only those PheCodes that have a prevalence higher than 35 in the cohort.

The group of comorbidities with the highest frequency in the BIG cohort are: congenital heart diseases (CHD) such as atrial septal defect (ASD), patent ductus arteriosus (PDA), atrioventricular septal defect (AVSD), and ventricular septal defects (VSD). Interestingly, the CHD is known to be the leading cause of mortality and morbidity during the first two years of life in the DS population [89], with 40% [90] to 63.5% patients having CHD [91].

Figure 4.2.3 shows the distribution of the top 20 most common phenotypes in the BIG cohort across different ancestry groups. Notably, the PheCode GI_512, which corresponds to "Aphagia or Dysphagia", appears to be much more prevalent among individuals of European ancestry**.** This observation is particularly interesting when considering additional metadata: this phenotype is also significantly more frequent among individuals labeled as "Non-Health Disparity", as opposed to those from ancestries categorized under "Health Disparity". While it is difficult to formulate a strong hypothesis at this stage, this pattern might reflect cultural or healthcare access differences rather than purely genetic factors.

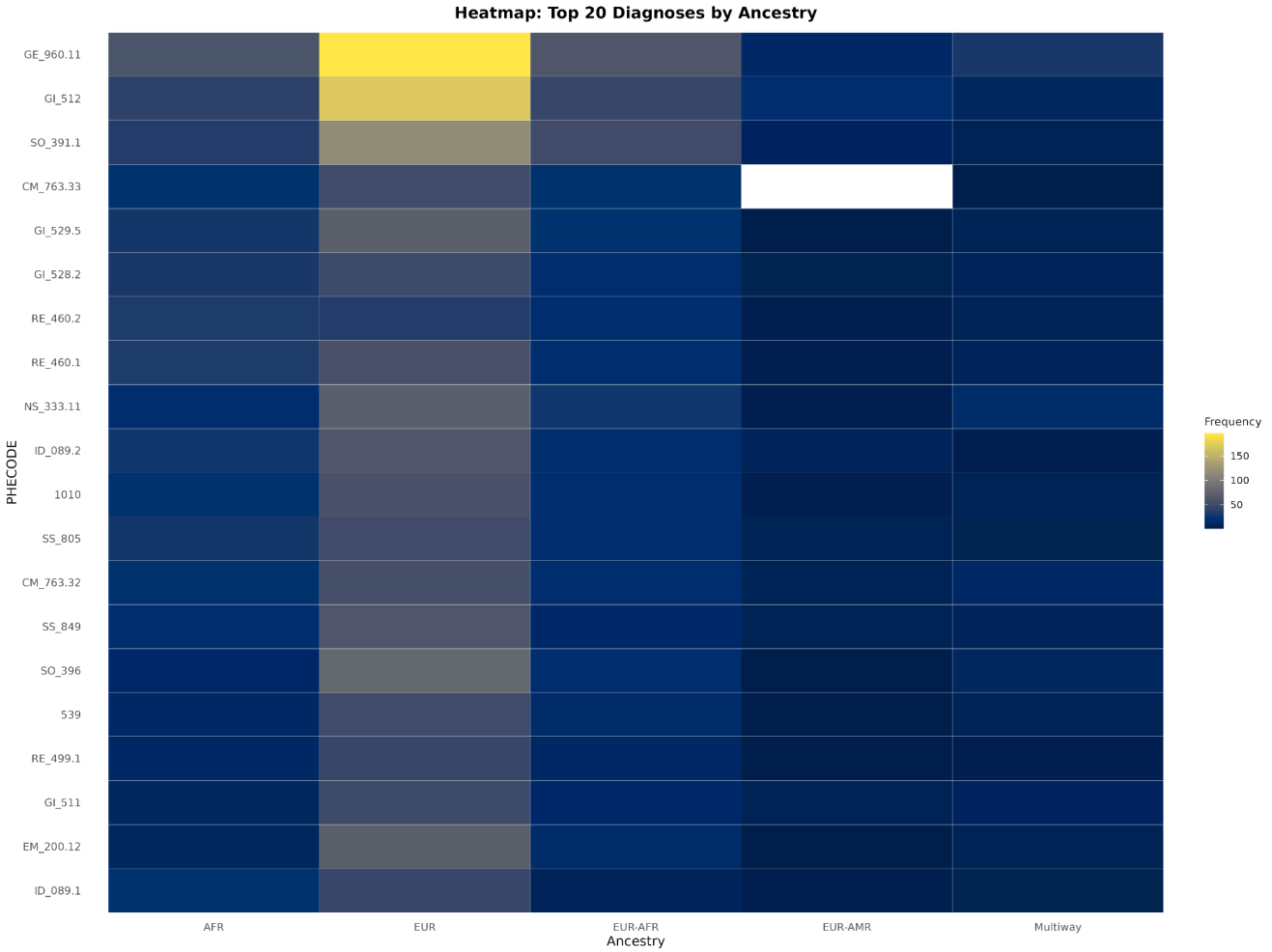

Figure 4.2.3 - Top 20 diagnoses by ancestry.

Heatmap representing the distribution of the 20 most prevalent phenotypes by ancestry. The X-axis shows the different ancestries, while the Y-axis lists the top 20 phenotypes. Colors indicate phenotype frequency, with yellow representing higher frequencies and blue representing lower ones.

4.2.2 Summary of detected variants and exonic landscape

This investigation sought to elucidate the genetic basis underlying the increased susceptibility to comorbidities observed in individuals with DS. Given the well-documented high prevalence of numerous medical conditions in DS populations relative to euploid individuals, we hypothesized that genetic variants beyond the characteristic trisomy 21 may contribute to this phenotypic heterogeneity. To test this hypothesis, we leveraged whole-exome sequencing (WES) data to perform comprehensive single-nucleotide variant analysis across the complete BIG DS cohort, comprising 135 DS patients from the Biorepository and Integrative Genomics initiative. Among the sequenced samples, 95.2% achieved an average sequencing depth of at least 20X, and 99.3% of the samples had > 90% of their bases covered at 20X or greater.

The analysis revealed 6,726,045 variants across the BIG DS cohort, with the majority comprising single nucleotide variants (92.82%), followed by deletions (3.91%), insertions (2.13%), substitutions (1.13%), and indels (0.0007%).

Among variants predicted to have the most severe functional consequences, Variant Effect Predictor (VEP) classified approximately 38,000 as stop-gained mutations, 49,000 as frameshift variants, 2,000 as stop-lost, 5,000 as start-lost, 11,000 as splice acceptor vDS variants, and 13,000 as splice donor variants. Variants with moderate predicted impact were predominantly missense variants, totaling approximately 1.3 million.

Variant distribution is reported in Figure 4.2.5. As expected the first two chromosomes display the highest number of variants, reflecting their larger size and higher gene content.Conversely, chromosome 21, being the smallest chromosome, has the lowest absolute number of variants; however, when calculated per megabase, is the chromosome 13, as shown in Figure 4.2.4. Overall, the number of variants per chromosome appears to correlate with chromosome length and average gene density beside some .

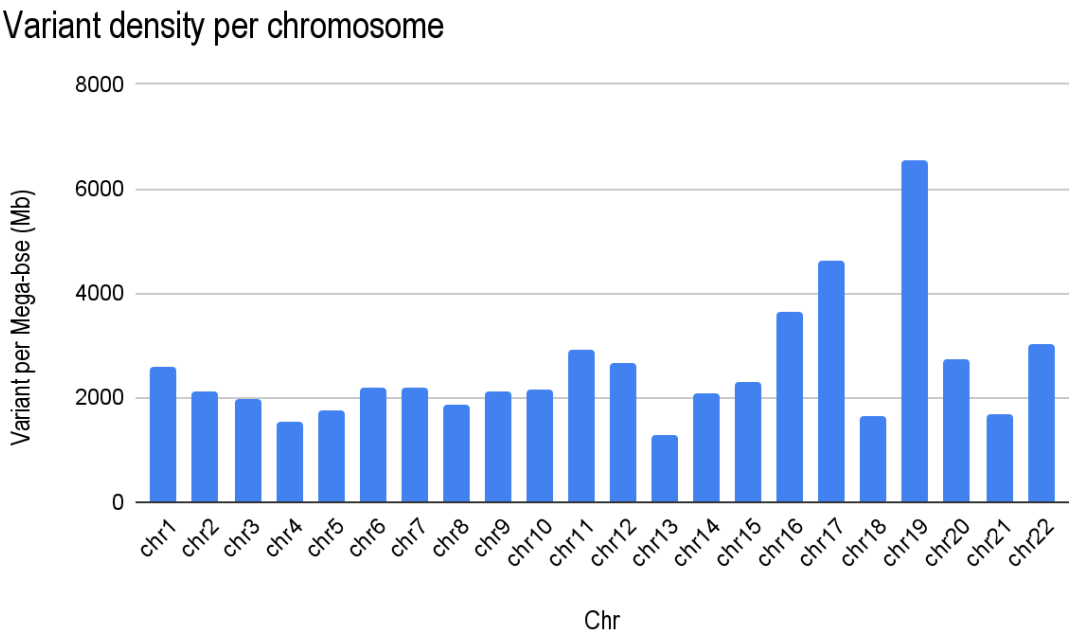

Figure 4.2.4 - Number of variants per million base pairs (Variants/Mb) for each human autosome.
On the Y axis is shown the variant count per Mb and on the X axis the chromosome number.

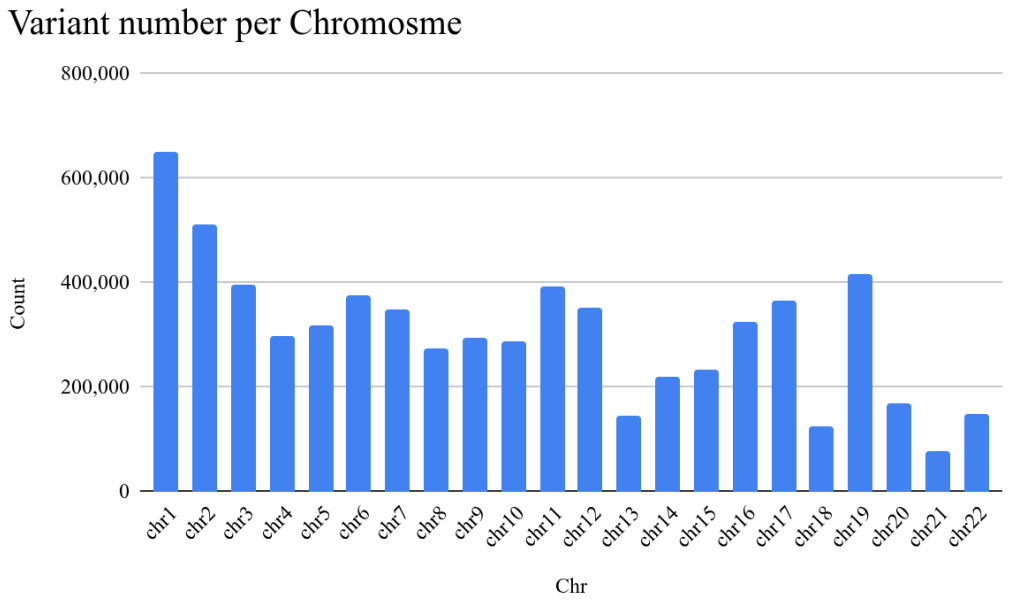

Figure 4.2.5 - Distribution of the variants count across the BIG DS cohort.

On the Y axis is shown the variants count while on the X axis the chromosome number.

Functional impact assessment of variants was conducted using VEP annotations and associated plugins, including PolyPhen-2 and SIFT, which predict the effects of mutations at the protein level. SIFT analysis classified 1,494,384 variants (33%) as "deleterious", 1,239,183 (27%) as "deleterious_low_confidence", 1,470,815 (33%) as "tolerated", and 554,913 (12%) as "tolerated_low_confidence", suggesting that approximately 60% of annotated variants may potentially impact protein function. PolyPhen-2 predictions categorized 1,044,618 variants (24%) as "probably damaging," 813,654 (18%) as "possibly damaging", 2,507,907 (57%) as "benign", and 238,515 (5%) with "unknown effect" .

While both predictive algorithms indicate that the majority of variants are likely benign, these assessments must be interpreted within the specific genetic context of DS. Standard prediction tools do not account for the unique genomic background of trisomy 21, which may modify the functional consequences of variants identified in our DS cohort. This limitation underscores the importance of context-specific interpretation when evaluating variant pathogenicity in chromosomally abnormal populations and will be further taken into consideration in the discussion section.

4.2.3 Association analyses with comorbidities

In order to identify genetic variants associated with the comorbidities identified in the BIG DS cohort, classical statistical tests such as Chi Square and Fischer Test were implemented between variant genotypes of the patients with and without a given PheCode . Among the 57 relevant phenotypes selected, many were already known to have a higher prevalence in DS patients.

Phenotypic data was represented as PheCodes, which were generated by extracting diagnostic information from electronic health records (EHRs). This approach enabled standardized and clinically meaningful phenotype definitions suitable for downstream genotype-phenotype association analyses, particularly in the context of comorbidities associated with DS. We applied both Fisher's exact test and the chi-square test to assess associations between genotypes and phenotypes. Fisher's exact test was used in cases where any observation count in the contingency table was less than five, whereas the chi-square test was applied under standard conditions. No covariates were included in the analysis, given the genetic homogeneity of the BIG cohort, following ancestry inference and phenotype-based stratification based on PheCodes.

Following rigorous quality control procedures, we retained rare variants (minor allele frequency < 0,0005 in GnomADe) predicted to have functional consequences. Selection criteria included missense variants and those annotated as "functionally deleterious" by SIFT, "probably damaging" by PolyPhen-2, or classified as HIGH IMPACT by Variant Effect Predictor (VEP). Nevertheless, variants predicted to have an IMPACT by VEP as "moderate" have also been analysed, given that the variants filtered by SIFT plus frequency and PolyPhen-2 plus frequency did not include the VEP filter "HIGH" impact category.

This filtering strategy yielded 1,401,432 prioritized variants for downstream analysis.
Individual variant–phenotype associations were assessed without gene-based aggregation. Multiple testing correction was performed using the Benjamini–Hochberg procedure to control the false discovery rate. However, due to the limited sample size, this correction inflated most p-values to 1, excluding those that were close to reaching genome-wide significance that were still under the 0,05 threshold. Therefore, the p-values displayed in the Manhattan plots (which is a graphical tool used in genome-wide association studies to display the statistical significance of associations between genetic variants and traits across the genome it plots the negative $\log_{10}$ of *p*-values against genomic positions, where prominent peaks indicate loci with strong evidence of association) correspond to the uncorrected values and were used to evaluate general association trends rather than definitive statistical significance. Phenotypes showing the strongest trends were subsequently investigated through a gene burden analysis to further explore potential genes–phenotype relationships.

While one variant approached statistical significance across all tests, no associations exceeded the corrected significance threshold.
This result is likely due to factors such as sample size, phenotypic heterogeneity, or the rarity of the variants under study—issues that are further addressed in the discussion section.
Several variants approached the significance threshold; those circled in red in the plots shown below represent the closest signals. Among all analyzed phenotypes, the associations related to cardiovascular comorbidities—particularly prevalent in DS populations—were the ones nearest to reaching significance..
For acute lower respiratory infection (RE_460.2), the strongest signal corresponded to rs16836525, a missense variant on chromosome 1 (position 154926423) within the

phosphomevalonate kinase gene (PMVK), which functions in cholesterol metabolism pathways as shown in Figure 4.6.1.

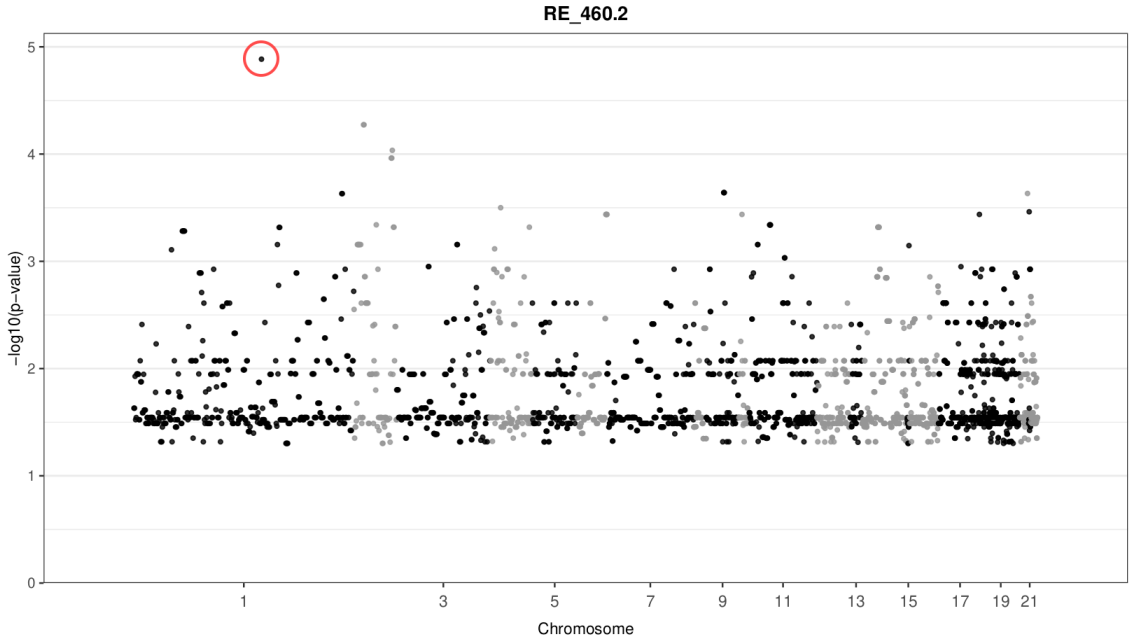

Figure 4.6.1 - Manhattan plot of chi-square variant p-values generated for the Acute lower respiratory infection phenotype (RE_460.2).

Different shades of grey represent the individual chromosomes.

The X-axis represents the chromosomes, while the Y-axis shows the negative base-10 logarithm of the p-values ($-\log_{10}(p)$), the red circle highlights the SNP with the strongest statistical power.

Acute upper respiratory infections (RE_469.1) yielded two proximal signals on chromosome 1 (position 173851884) within the aspartyl-tRNA synthetase 2 gene (DARS2), essential for mitochondrial protein synthesis. Notably, this variant lacks annotation in ClinVar, suggesting potential novelty in the context of DS comorbidities.

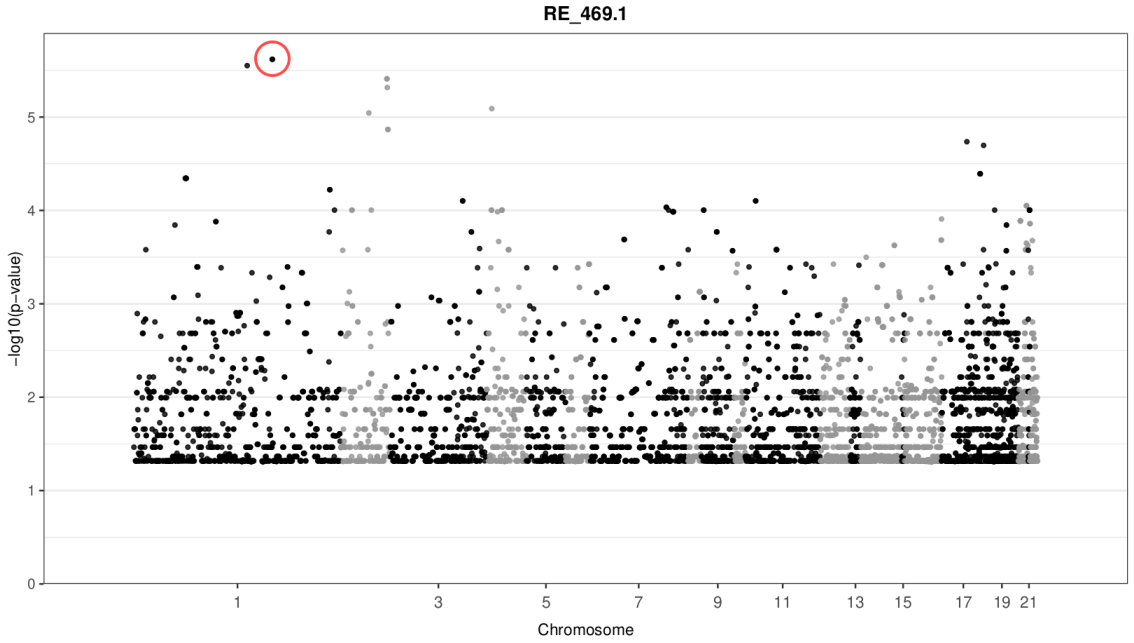

Figure 4.6.2 - Manhattan plot of chi-square variant p-values generated for the Acute upper respiratory infections phenotype (RE_469.1).

Different shades of grey represent the individual chromosomes.

The X-axis represents the chromosomes, while the Y-axis shows the negative base-10 logarithm of the p-values ($-\log_{10}(p)$), the red circle highlights the SNP with the strongest statistical power.

Cardiovascular phenotypes also yielded promising associations. Ventricular septal defect (CM_763.31) showed the strongest signal from chromosome 17 (position 50274511) within TMEM92, encoding transmembrane protein 92 (rs9894445).

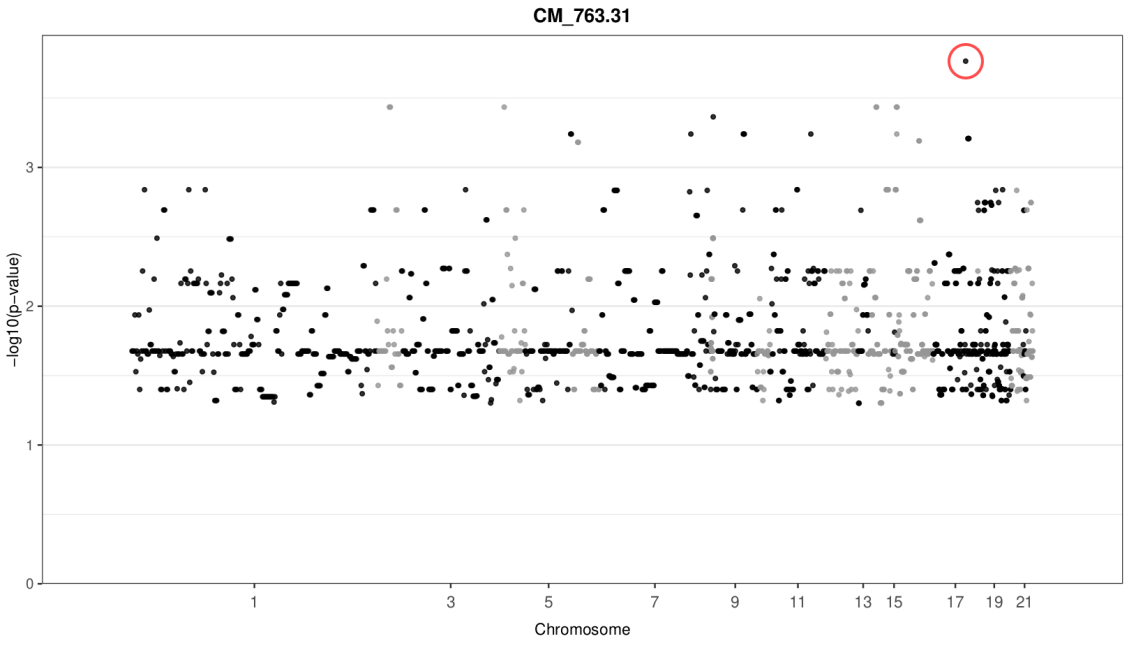

Figure 4.6.3 - Manhattan plot of chi-square variant p-values generated for the Ventricular septal defect phenotype (CM_763.31).

Different shades of grey represent the individual chromosomes.

The X-axis represents the chromosomes, while the Y-axis shows the negative base-10 logarithm of the p-values ($-\log_{10}(p)$), the red circle highlights the SNP with the strongest statistical power.

The most significant signal across the entire cohort emerged from atrioventricular septal defect analysis (CM_763), localizing to chromosome 1 position 156586027 known as rs12090808 within the tetratricopeptide repeat domain 24 gene (TTC24).

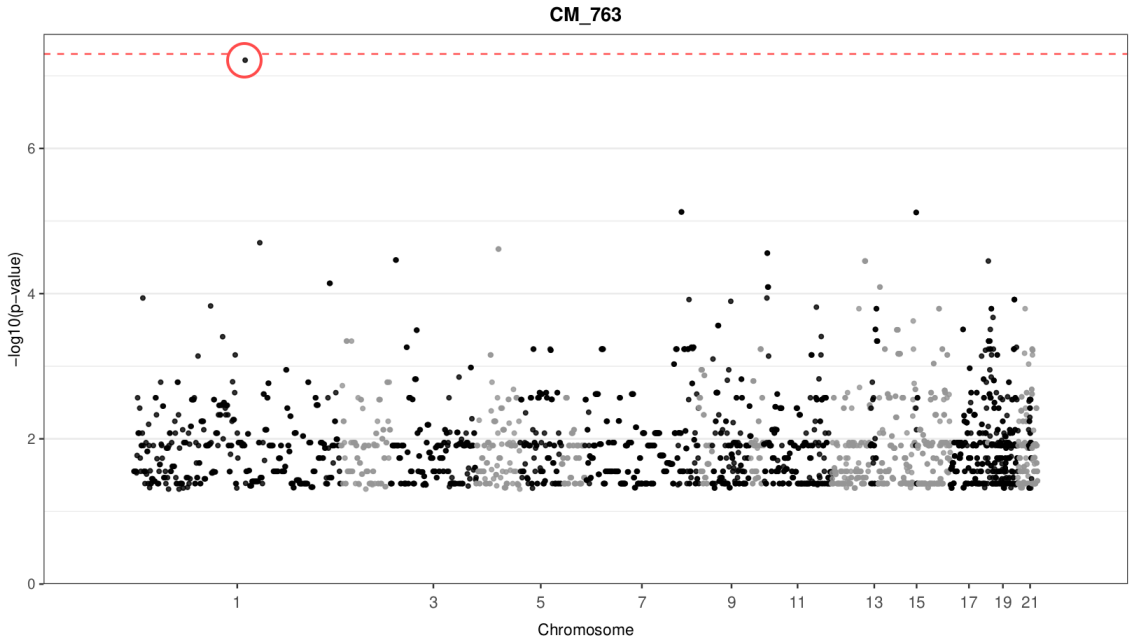

Figure 4.6.4 - Manhattan plot of chi-square variant p-values generated for the Atrioventricular Septal Defect (CM_763).

Different shades of grey represent the individual chromosomes.

The X-axis represents the chromosomes, while the Y-axis shows the negative base-10 logarithm of the p-values ($-\log_{10}(p)$), the red circle highlights the SNP with the strongest statistical power.

TTC24 currently lacks documented associations with DS or atrioventricular septal defects. Its encoded protein shows notable interactions with three key components of the polymerase-associated factor 1 complex (PAF1C)—LEO1, PAF1, and CDC73—which acts as a transcriptional regulator during early cardiac development. Additionally, TTC24 interacts with MT-ND4, an essential subunit of the mitochondrial electron transport chain [92]. All the interactions described here were identified using STRING, a database that integrates known and predicted protein–protein interactions from multiple sources. These protein interactions may have mechanistic relevance for cardiovascular congenital malformations in DS, particularly given the high metabolic demands of cardiac tissue during fetal development, see the Figure 4.6.5 in the next page.

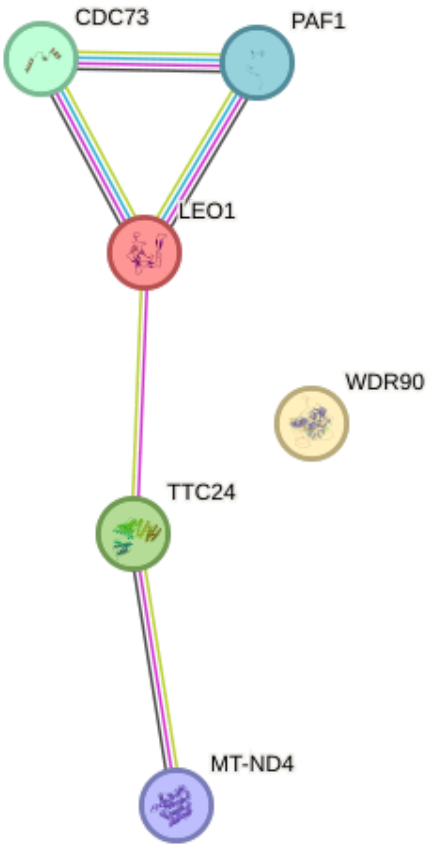

Figure 4.6.5 - STRING interaction network preview for the TTC24 gene.

The link colours describe the nature of the interaction: cyan and magenta represent known interactions retrieved from curated databases; green, red, and blue indicate predicted interactions, specifically gene neighborhood, gene fusions, and gene co-occurrence, respectively. Light green, black, and light blue represent interactions inferred from text mining, co-expression, and protein homology ***[88]****.*

4.3 Identification of candidate genes

4.3.1 Rationale and interpretation beyond statistical significance

Despite the comprehensive association analyses performed across the BIG cohort, no variants reached a genome-wide statistical significance after multiple testing corrections. These results are due to a combination of factors such as the limited sample size and/or the wide genetic heterogeneity beneath the phenotypes. Despite these limitations, several genes emerged as suggestive candidates based on the accumulation of rare and possibly damaging and deleterious variants, especially in the four selected comorbidities under study.

To identify genes that accumulate a disproportionate number of variants while accounting for gene length, a normalization analysis was performed. For each gene, the number of detected variants was divided by its genomic length (in kilobases), yielding a *Variants_per_kb* value that corrects for size-related biases. This metric was then compared to the genome-wide mean density of variants to compute the *Relative_variation_index*, which quantifies the degree of mutational enrichment relative to the global average (values >1 indicating higher-than-expected variation). The resulting tables report for each gene its symbol (*Gene_Names*), normalized mutation density (*Variants_per_kb*), and relative enrichment score (*Relative_variation_index*).

Subsequently, genes exhibiting an exceptionally high *Relative_variation_index* were identified as *high-burden genes* using an interquartile range (IQR)-based outlier detection approach, thereby highlighting those loci with the strongest evidence of mutational overload.

Although these findings should be taken with caution, they suggest a set of previously unknown loci, which represent a meaningful and precious insight in the design of

further studies and hypotheses on how the trisomy 21 context may amplify the occurrence of comorbidities. Here in the following tables we are presenting only the top 5 five ranking genes exhibiting a meaningful increase in variant burden related to the overall gene-level average.

4.3.2 Initial screening to quantify the top burden genes across comorbidity subgroups.

In the following figure are reported and explained only some of the overmutated genes for brevity sake we have selected only the first 5 top ranked ones.

In the CM_763 phenotype (Atrioventricular septal defects), our screening identified 143 genes exhibiting a meaningfully higher number of variants compared to the variant number per kilobase across all the genes (table 4.3.1). These genes include TRAJ47[93] (T cell receptor alpha joining 47) which joins segments of the T-cell receptor alpha chain; contributes to TCR diversity through V(D)J recombination in T lymphocytes, OR6K6 [94] (Olfactory receptor family 6 subfamily k member 6) that encodes for a member of the olfactory receptor family of G protein–coupled receptors involved in odorant detection and olfactory signal transduction, TRBV6-8 and TRBV6-7 (T-cell receptor beta variable 6-8 and 6-7)[95] the first one encodes for a variable segment of the T-cell receptor beta loc us which contributes to antigen recognition by T

lymphocytes, while the second one encodes for a variable segment of the TCR beta locus involved in generating T-cell receptor diversity, the last one the gene OMP[96] (Olfactory marker protein) encodes for a cytoplasmic protein expressed in mature olfactory sensory neurons, involved in olfactory signal transduction.

Burden Table CM_763		
Gene_Names	**Variants_per_kb**	**Relative_variation_index**
TRAJ47	18	102
OR6K6	6	34.7
TRBV6-8	5	27.1
TRBV5-7	4	24.9
OMP	4	23.6

Table 4.3.1 - Burden table for the CM_763 phenotype (Atrioventricular Septal Defects).
*This table reports genes **Variant_per_kb**: number of variants per 1000 bp; **Relative_variation_index**: loci with the strongest evidence of mutational overload, as identified by an interquartile range (IQR)-based outlier detection approach.*

For the CM763.31 phenotype (ventricular septal defects) we found a total of 82 and there are also similar genes include TRAJ3 TRBV5-7 and OR6K6 which are mostly the same genes found before with slightly differences while the last one LCE3A[97] encodes for a structural protein involved in epidermal differentiation and formation of the skin barrier. (Figure 4.3.2).

Burden Table CM_763.31		
Gene_Names	**Variants_per_kb**	**Relative_variation_index**
TRAJ3	32	158
OR6K6	7	33.5
TRBV5-7	4	21.0
OR7G3	4	20.9
LCE3A	4	18.1

Table 4.3.2 - Burden table for the CM_763.31 phenotype (ventricular septal defects).
*This table reports genes **Variant_per_kb**: number of variants per 1000 bp; **Relative_variation_index**: loci with the strongest evidence of mutational overload, as identified by an interquartile range (IQR)-based outlier detection approach.*

For phenotype RE_469.1 - acute upper respiratory infections, we found a total of 286, some of them were already found in the previous analyses such as OR6K6 (see above) and a gene TRAJ36 which is another gene of the same family of the previous TRAJ3 and TRAJ47 and having also a similar functions, instead the gene IGHV4-31[95] encodes for a variable segment of the immunoglobulin heavy chain locus contributing to antibody diversity in B cells, OR10A5[98] encodes for a different G protein–coupled olfactory receptor involved in odorant recognition and signal transduction, the last one is IGLC7[99] (Immunoglobulin lambda constant 7) which encodes for the constant region of the immunoglobulin lambda light chain, essential for antibody structure and stability see figure 4.3.3.

Burden Table RE_469.1		
Gene_Names	**Variants_per_kb**	**Relative_variation_index**
TRAJ36	17	103
IGHV4-31	10	58.5
OR6K6	8	45.9
OR10A5	5	31.3
IGLC7	5	27.1

Table 4.3.3 - Burden table for the RE_469.1 phenotype (acute upper respiratory infection).
This table reports genes with a variant burden at least five times higher than the genome-wide average.

The last phenotype RE_460.2 (acute lower respiratory infections) is characterized by 146 overmutated genes; here we report only the 5 top ranked ones. IGHV4-28 and IGHV2-70D[99], both of which are part of the same protein family of immunoglobulins mentioned before so are important for the antibody diversification in B cells, the OR2S2 gene encodes for G protein–coupled olfactory receptor involved in odorant molecule detection and signal transduction, the TAF11L2[100] gene is predicted to encode instead for putative transcription factor related to TAF11, possibly involved in regulation of RNA polymerase II transcription, the last one is the gene KRTAP4-5 which encodes for a structural protein of the hair shaft cortex that crosslinks with keratin intermediate filaments to strengthen hair fibers. (Figure 4.3.4).

Burden Table RE_460.2		
Gene_Names	**Variants_per_kb**	**Relative_variation_index**
IGHV4-28	4	27.8
IGHV2-70D	4	27.1
OR2S2	4	26.9
TAF11L2	3	23.6
KRTAP4-5	3	23.4

Table 4.3.4 - Burden table for the RE_460.2 phenotype (acute lower respiratory infections).
*This table reports genes **Variant_per_kb**: number of variants per 1000 bp; **Relative_variation_index**: loci with the strongest evidence of mutational overload, as identified by an interquartile range (IQR)-based outlier detection approach.*

Genes identified with higher variant burden in Acute Upper Respiratory Infections (RE_469.1):

By analyzing the phenotype RE_469.1, we identified 287 genes exhibiting a high variant burden.
In the analyzed data, the gene CSF3 (Colony Stimulating Factor 3) appears as the only gene with a direct interaction with the phenotype combination “Down syndrome + upper respiratory infections”. CSF3 encodes a granulocyte colony-stimulating factor that regulates neutrophil proliferation and maturation. In DS individuals, alterations in CSF3-mediated pathways may contribute to impaired innate immune responses, increasing susceptibility to upper respiratory infections [101].

Among the indirect interactions, the genes identified that interact with the main implicated genes include SERPINA2, NFKB1, STAT3, CXCL8, CCL5, TNF, MMP1, MMP9, IL6, IL10, JAK3, STAT1, IFNG, HMOX1, HP, HPX, CBS, ARSA.

CXCL5, a key chemokine for neutrophil recruitment, is indirectly connected to CXCL8, CCL5, CCL11, TNF, MMP9, and MMP1, highlighting a network of cytokines and proteases that may modulate local airway inflammation moreover is known to be involved different lung disorders such as alveolitis and Pulmonary Sarcoidosis (information From GeneCards website).

CCL11 interacts with CCL5, CXCL8, IL10, TNF, and IL6, indicating involvement in leukocyte trafficking and antiviral immune regulation, which are critical for susceptibility to upper respiratory infections in DS[102].

CNTF, known for its neurotrophic function, interacts with IL6, IL6R, IL6ST, LEP, and HP, suggesting it may indirectly influence inflammatory signaling and epithelial responses in the respiratory tract.
Different MUC genes (mucins), interact with JAK3, STAT1, IFNG, SERPINA1, and TNF, suggesting potential effects on mucosal barrier function and antiviral signaling, particularly relevant in DS individuals[103].
These genes are directly or indirectly related to airway protection, underscoring the multifactorial reasons under the higher susceptibility of DS people to these diseases.

Genes identified with higher variant burden in Acute lower respiratory infection phenotype (RE_460.2).

For the phenotype RE_460.2, we identified 147 genes with a variant burden markedly above the expected threshold, some of which had already emerged in the analysis of acute upper respiratory infections (RE_469.1). This recurrence across two related phenotypes strengthens the reliability of our approach and suggests the existence of shared biological pathways underlying respiratory vulnerability in DS.

In the analyzed dataset, FOXE3 (Forkhead Box E3) emerges as the only gene with a direct interaction with the phenotype combination “Down syndrome + lower respiratory infections”. FOXE3 encodes a transcription factor primarily involved in ocular development, but it may also influence epithelial differentiation and immune regulation in respiratory tissues. Alterations in FOXE3-associated pathways could contribute indirectly to susceptibility to lower respiratory infections in individuals with Down syndrome[104].

Among the indirect interactions, several genes identified by the user interact with the main implicated genes, including MIRLET7C, JAK3, NFKB1, STAT1, ICOSLG, IFNG, KMT2A, LINC02605, NFKBIA, CTLA4, HLA-DRB1, TNFRSF1A, CD40, IL6R, LEP, PIK3R1, IL7R, ADA, VAV1, HCK, LRRK2, PTEN, STAT3, XIAP, CARMIL2, MAF, APOA1. These interactions suggest complex networks linking transcriptional regulation, immune signaling, and inflammation.

TNFRSF18, which interacts with MIRLET7C, JAK3, NFKB1, STAT1, and ICOSLG, is implicated in T cell co-stimulation and apoptosis regulation. Its indirect network suggests a role in modulating adaptive immune responses, which may affect the severity or frequency of lower respiratory infections in Down syndrome patients [105].

MUC4, a membrane-associated mucin, interacts with JAK3, STAT1, IFNG, KMT2A, and LINC02605, indicating potential effects on mucosal barrier integrity, antiviral responses, and cytokine signaling [106].

SUMO4, a regulator of protein sumoylation, interacts with NFKB1, NFKBIA, CTLA4, HLA-DRB1, and STAT1, highlighting its potential role in controlling NF-κB-mediated inflammation and T cell activation during infection [107].

Other indirect genes, including TNFRSF4, CSH1/2, CTSG, CEACAM3, MUC6, IVL, KRT36, TSSK2, IFNL2, ID3, KIF2B, GSTP1, WDR38, KLF14, and USH1G, are involved in networks of cytokine signaling, immune regulation, epithelial structure, apoptosis, and oxidative stress. Their interactions with key genes such as JAK3, STAT1, NFKB1, IFNG, IL6R, BRWD1, and KMT2A suggest that multiple pathways converge to influence lower respiratory tract susceptibility in Down syndrome.

Overall, although FOXE3 serves as the primary direct gene, the indirect network illustrates a complex system of immune signaling, mucosal defense, and transcriptional regulation. These findings provide a functional framework for understanding vulnerability to lower respiratory infections in Down syndrome and may guide experimental validation and potential therapeutic strategies.

Discussion

5.1 Contributions to the field and comparison with existing literature

Research on DS and its associated comorbidities has historically been dominated by descriptive, clinical and epidemiological studies, with the objective of characterising prevalence, natural history and health outcomes in affected individuals. These investigations have provided a detailed catalogue of the most frequent comorbidities, including congenital heart disease, gastrointestinal malformations, haematological disorders, immune dysfunction and neurodegenerative conditions such as early-onset Alzheimer's disease. This establishes DS as a multisystem disorder with substantial impact on morbidity, mortality and quality of life.

From a genetic perspective, the majority of research has concentrated on chromosome 21 itself, under the classical gene-dosage hypothesis. It has been demonstrated that specific dosage-sensitive genes have been recurrently implicated in distinct DS-associated traits, such as DYRK1A (neurocognitive development), APP

(Alzheimer's disease), SOD1 (oxidative stress), and DSCAM (congenital heart disease). For decades, the concept of a "DS Critical Region" (DSCR) [108] has further reinforced this gene-centric perspective, proposing that a limited set of loci on 21q22 could account for the major features of the syndrome. Despite the fact that more recent evidence has shifted towards a distributed model, attributing phenotypic variability to the combined effect of multiple genes across the entire chromosome, the focus has remained largely restricted to HSA21.

Conversely, genome-wide approaches have been constrained. While studies in the general population have widely employed genome-wide association studies (GWAS), burden tests and rare variant analyses to dissect the genetic architecture of complex traits, these strategies have only rarely been applied to individuals with DS. Investigations of comorbidities in DS have rarely considered the contribution of variants located outside HSA21, nor have they systematically assessed the cumulative effect of rare coding variants across genes or pathways. A small number of studies have begun to explore the impact of copy number variations (CNVs) or the exome-wide burden of mutations. However, the available evidence remains sparse and fragmented.

Consequently, despite the well-defined primary genetic cause of DS, our understanding of how background genetic variation across the genome modulates phenotypic expression and comorbidity risk in trisomy 21 remains limited. The field has been characterised by an absence of integrative, genome-wide frameworks capable of transcending the descriptive clinical spectrum and gene-centric approaches that are exclusively focused on HSA21.

This thesis details the research conducted during the three years of the doctoral programme. The objective of this research was to explore the genetic underpinnings of DS and its associated comorbidities. To this end, a comprehensive genetic association analysis was performed on whole genome and exome data from two distinct cohorts.

This analysis was followed by an initial investigation into the gene burden variant across the entire genome focusing on the interaction of the top over-burden genes in relation to the DS plus comorbidities phenotype. This approach was undertaken to ascertain the potential indirect interplay between the 21st chromosome and the rest of the genome in individuals with DS.

The aforementioned objective and approach represent the main novelty of this work, offering a new perspective that may unveil previously unexplored genes and proteins of interest. By starting from an agnostic framework and leveraging gene network analyses, this strategy aims to better capture the broader and more complex interactions underlying the pathogenesis of DS comorbidities. This is particularly relevant given that most studies in the scientific community still have thus far concentrated almost exclusively on genes located on chromosome 21. This approach starts to consider the multisystemic dimension and polygenic dimension of DS.

The results of this work contribute to bridging a long-standing gap in the field by extending the genetic investigation of DS comorbidities beyond chromosome 21. It has been established that a number of the identified signals may provide a biological basis for clinical observations that individuals with DS are particularly susceptible to respiratory conditions and immune-related pathologies, as well as congenital heart malformations; this knowledge is thus reinforced.
Concurrently, the emergence of novel candidate genes located outside HSA21 introduces new hypotheses regarding additional genomic contributors and potential modifier effects that require further exploration. These findings emphasise the polygenic and multisystemic nature of comorbidity pathogenesis in DS, and point to the need for future studies combining larger cohorts, functional validation and integrative multi-omic analyses to clarify the interplay between dosage imbalance on chromosome 21 and genome-wide genetic variation.

5.2 Main finding

In Ospedale Bambino Gesù cohort:

The work presented in this thesis provided the first genome-wide gene burden analysis in the context of DS and its comorbidities, identifying mostly and novel possible candidate genes, some of which could contribute directly or indirectly to the phenotype under study as ventricular septal defect, upper and lower respiratory infections and atrio-ventricular septal defects.

Besides the biological finding, in this thesis we propose and describe a possible methodological framework which is usable to test and find new candidate genes and variants associated with DS and its comorbidities and its robustness is supported by similar results for similar phenotype starting from whole exome data.

Although no individual variants prioritized from the whole genome sequence data reached statistical significance, in our cohort of 17 DS patients, the analysis nonetheless identified biologically plausible candidate genes through variant prioritization and pathway-based interpretation. Well-established chromosome 21 genes such as SLC5A3 [109] and HMGN1 [110] provided important internal positive controls, confirming that the analytical pipeline was capable of recovering relevant signals. Beyond these, additional candidates including KDR [111], GANC [112], PHKB [113], and CBLC [114] emerged, pointing to potential roles in angiogenesis, folate metabolism, and molecular interactions that may exacerbate congenital comorbidities in DS. Importantly, several of these genes also appeared indirectly connected to established DS (e.g., DYRK1A, GATA4, DSCAM), suggesting that they may act within shared pathogenic networks. While the small sample size of the present study limited statistical power, the convergence of the VarElect analyses conducted on all the prioritized variants of the DS phenotype supports the biological plausibility of these findings. Future research will require larger cohorts and experimental validation to clarify the contribution of these candidate genes to DS and associated comorbidities.

In the Biorepository of Integrated Genomes cohort:

The work that has been done on the DS BIG cohort is one of the first systematic attempts to explore variant burden across multiple comorbidities in the DS community using whole-exome sequencing. Although single-variant association testing yielded only suggestive evidence for rs12090808, the burden analysis identified several genes with variant enrichment across phenotypes such atrio-ventricular septal defect, ventricular septal defects, and both upper and lower recurrent respiratory infections.

The integrative analyses combining DS with specific comorbid phenotypes — namely atrioventricular septal defects (AVSD), ventricular septal defect (VSD), and recurrent respiratory infections (both upper and lower) — revealed a complex interplay between developmental and immune networks. In particular, several genes emerged from VarElect analyses as either directly or indirectly associated with the DS genotype through known molecular interactors or shared functional pathways.

For the DS + AVSD group, VarElect indirect associations centred around GATA1, RUNX1, PTPN11, and DYRK1A. GATA1, located on chromosome X, plays a critical role in erythroid and megakaryocytic differentiation and is often dysregulated in DS-related transient myeloproliferative disorders[115]. RUNX1 is a transcription factor involved in haematopoiesis whose triplication in DS has been implicated in both leukaemogenesis and abnormal cardiac development [116],[117]. PTPN11 encodes the SHP2 phosphatase, a key regulator of RAS–MAPK signalling frequently mutated in Noonan syndrome and associated with congenital heart defects [118]. DYRK1A, one of the canonical dosage-sensitive genes on chromosome 21, is known to influence cardiogenesis through its interaction with GATA4 and DSCAM, both of which contribute to atrioventricular septum morphogenesis [119].

In contrast, the DS + VSD combination highlighted genes with broader roles in cardiac structural development and transcriptional regulation, such as GATA4, NKX2-5,

FOXF1, and FOXC2. Variants in NKX2-5 and GATA4 have been repeatedly associated with septal defects and are considered central hubs in cardiac morphogenesis [120]. The forkhead family members FOXF1 and FOXC2 contribute to mesodermal differentiation and outflow tract formation, suggesting that DS-related perturbations in transcriptional regulation could synergise with dosage imbalance to amplify cardiac malformation risk. For respiratory comorbidities, both the DS + upper and lower respiratory infection datasets converged on immune and inflammatory networks, particularly involving CSF3, TNF, IL6, STAT1, NFKB1, and JAK3. CSF3, the only gene with a direct VarElect link for upper airway infections, encodes granulocyte colony-stimulating factor, a critical regulator of neutrophil function. Dysregulation of granulopoiesis in DS may contribute to the recurrent bacterial infections commonly observed in this population [121]. Indirect interactions further implicated the SERPINA family, CXCL5, and CCL11, reflecting an altered chemokine environment that may impair effective immune responses. The lower respiratory infection network exhibited a similar but more extensive inflammatory signature, integrating genes such as MUC4, TNFRSF4, SUMO4, with immune regulators such as JAK3, STAT1, IFNG, and NFKB1. The presence of non-coding elements, such as MIRLET7C and LINC02605, is particularly noteworthy, as these molecules map within chromosome 21 and have been shown to modulate interferon signalling, epithelial barrier integrity, and immune cell differentiation [122] [123].

All interaction findings are summarized in the Fig 5.2.1

Results Overview			
Phenotype Combination	**Key Genes**	**Functional Pathways**	**Representative Interactors / Evidence**
DS + AVSD	GATA1, RUNX1, PTPN11, DYRK1A, GATA4, DSCAM	Cardiogenesis, RAS–MAPK, chromatin regulation	Noonan/DS overlap; cardiac malformation
DS + VSD	GATA4, NKX2-5, FOXF1, FOXC2, GATA5	Transcriptional control of cardiac septation	Developmental regulation
DS + Upper respiratory infections	CSF3, TNF, IL6, STAT3, NFKB1, SERPINA1	Cytokine signalling, innate immunity	Immune dysregulation
DS + Lower respiratory infections	FOXE3, JAK3, STAT1, IFNG, MIRLET7C, LINC02605	Interferon signalling, epithelial integrity	Inflammatory response

Fig 5.2.1 *Summary of Down syndrome comorbidities,* ***highlighting key genes****,* ***associated functional pathways****, and* ***representative interactions.***

5.3 Immune and Inflammatory Networks in Respiratory Comorbidities and DS Insights into DS heart defects Gene Interactions.

In the burden gene analyses we identified the highest signals in both hearth and respiratory related phenotype (CM_763, CM_763.31 and RE_469.1) from the different TRAJ genes (47, 3, 36) join segments of the T-cell receptor alpha chain; contributes to TCR diversity through V(D)J recombination in T lymphocytes. However interestingly all the aforementioned genes encodes for elements that contribute to the diversity of T-cell while other genes like IGHV(4-31, 4-28 and 2-70D) encode for a variable segment of the immunoglobulin heavy chain locus contributing to antibody diversity in B cells. These findings suggest that individuals with Down syndrome may exhibit an overmutation or dysregulation of genes involved in both T- and B-cell receptor diversity, potentially contributing to the altered immune function and increased susceptibility to infections and autoimmune conditions characteristic of the syndrome.
Nevertheless, the signals from these genes have to be taken with caution because they are very small genes less than 1Kb so the measure of their mutation rate since is elaborated per Kb could be over-estimated.

The overlap between immune and developmental pathways represents a central theme in DS comorbidities. The enrichment of inflammatory mediators (TNF, IL6, NFKB1, STAT1) across both upper and lower respiratory infections suggests that baseline immune dysregulation in DS amplifies susceptibility to recurrent infections. This is consistent with the chronic interferon hyperactivation signature previously reported in DS, where constitutive activation of the JAK–STAT axis leads to increased pro-inflammatory cytokine production [124] [125].

Taken together, the VarElect-driven integration of developmental and immune genes delineates a multidimensional network in which canonical trisomic genes (e.g. DYRK1A, RUNX1) interact with extrachromosomal modifiers (e.g. PTPN11, STAT1, JAK3) to shape both cardiac morphogenesis and immune homeostasis in DS.

5.4 Limitations

Despite providing a biologically coherent framework for interpreting DS-related comorbidities, this study presents several methodological and analytical limitations that must be acknowledged. The most evident limitation concerns the restricted cohort size, which substantially reduced statistical power and limited the detection of genome- or exome-wide significant associations. This constraint was clearly reflected in the Manhattan plots, where no signals reached conventional significance thresholds. Consequently, the analytical strategy was refined to focus on the identification of over-mutated genes, followed by an integrative exploration using VarElect to predict and infer biologically plausible gene–phenotype relationships.

To better contextualise the impact of sample size, a power analysis was conducted for each association test, illustrating at least one thousand observations (including both cases and controls) would be required to achieve sufficient power to detect robust

associations at the genome-wide level. In contrast, the available sample sizes — 135 individuals in the larger subset and 17 in the smallest — were insufficient to confidently support genome-scale inference.

These constraints emphasise that the present results should be interpreted as exploratory and hypothesis-generating rather than confirmatory; however, within this framework, the findings remain encouraging and highlight several biologically meaningful patterns that warrant deeper investigation. Notably, the reliance on tools such as VarElect—while inherently influenced by database coverage and more extensively characterised genes—also underscores the potential of integrative approaches to reveal novel gene–phenotype links in Down syndrome, even when operating under limited sample sizes. Likewise, although complementary in vitro or in vivo validation was beyond the scope of this work, its absence points to clear and promising avenues for future research. Overall, the limitations identified here do not diminish the relevance of the observed signals; instead, they underline the originality of the study and reinforce the value of expanding this effort through larger cohorts, functional experiments, and the integration of additional omics layers, particularly transcriptomics, to strengthen and validate the emerging hypotheses.

5.5 Future Perspectives

Future research should prioritise the expansion of cohort sizes to enhance statistical power and improve the reliability of association analyses. Increasing the number of both cases and controls will be essential to validate the preliminary signals observed here and to enable genome- or exome-wide significance to be reached with greater confidence.

A second important step will involve a more refined phenotypic stratification of DS-associated conditions. By systematically examining additional binary combinations of Down syndrome and comorbid traits, it will be possible to determine whether the over-mutated genes identified in this study are influenced by overlapping or secondary

clinical manifestations. Such stratification would also allow a more nuanced understanding of how specific phenotypic constellations modulate the molecular architecture of DS. Integrating complementary *omics* layers—particularly transcriptomic and epigenomic data—represents another crucial direction for future investigation. This approach would clarify whether over-mutated genes are not only genetically altered but also transcriptionally and functionally active in individuals with DS comorbidities.

Finally, once the most promising candidate genes and interactions are identified, *in vitro* and *in vivo* validation studies should be conducted, ideally using trisomic cellular models or mouse systems. These experiments would help verify whether the predicted gene–gene and gene–phenotype interactions observed computationally indeed occur at the biological level. Collectively, these steps would provide a more comprehensive and mechanistic understanding of the interplay between chromosomal dosage, gene network dysregulation, and comorbidity expression in DS. Ultimately, this integrative approach positions Down syndrome not merely as a chromosomal imbalance but as a complex molecular phenotype shaped by intersecting developmental, metabolic, and immunological pathways.

Bibliography

[1] ‘Microsoft Word - Down.1866b.doc’. Accessed: Aug. 19, 2025. [Online]. Available: https://www.romolocapuano.com/wp-content/uploads/2013/07/Langdon-Down-1866.pdf

[2] J. Lejeune, M. Gautier, and R. Turpin, ‘[Study of somatic chromosomes from 9 mongoloid children]’, *C R Hebd Seances Acad Sci*, vol. 248, no. 11, pp. 1721–1722, Mar. 1959.

[3] M. T. Davisson, C. Schmidt, and E. C. Akeson, ‘Segmental trisomy of murine chromosome 16: a new model system for studying Down syndrome’, *Prog Clin Biol Res*, vol. 360, pp. 263–280, 1990.

[4] M. Hattori *et al.*, ‘The DNA sequence of human chromosome 21’, *Nature*, vol. 405, no. 6784, pp. 311–319, May 2000, doi: 10.1038/35012518.

[5] G. de Graaf, F. Buckley, and B. Skotko, ‘People living with Down syndrome in the USA’:.

[6] G. de Graaf, F. Buckley, and B. G. Skotko, ‘Estimation of the number of people with Down syndrome in the United States’, *Genet Med*, vol. 19, no. 4, pp. 439–447, Apr. 2017, doi: 10.1038/gim.2016.127.

[7] S. E. Antonarakis *et al.*, ‘Down syndrome’, *Nat Rev Dis Primers*, vol. 6, no. 1, p. 9, Feb. 2020, doi: 10.1038/s41572-019-0143-7.

[8] H. Hasle, I. H. Clemmensen, and M. Mikkelsen, ‘Risks of leukaemia and solid tumours in individuals with Down’s syndrome’, *Lancet*, vol. 355, no. 9199, pp. 165–169, Jan. 2000, doi: 10.1016/S0140-6736(99)05264-2.

[9] S. L. Santoro, M. Cabrera, K. Haugen, K. Krell, and V. L. Merker, ‘Indicators of health in Down syndrome: A virtual focus group study with patients and their parents’, *J Appl Res Intellect Disabil*, vol. 36, no. 2, pp. 354–365, Mar. 2023, doi: 10.1111/jar.13065.

[10] F. Haddad, J. Bourke, K. Wong, and H. Leonard, ‘An investigation of the determinants of quality of life in adolescents and young adults with Down

syndrome', *PLOS ONE*, vol. 13, no. 6, p. e0197394, giu 2018, doi: 10.1371/journal.pone.0197394.
[11] V. J. T. Peters, 'Multidisciplinary Care for Children with Down syndrome in the Netherlands: A Modular Perspective', *Medical Research Archives*, vol. 11, no. 3, Mar. 2023, doi: 10.18103/mra.v11i3.3531.
[12] T. L. Rutter, R. P. Hastings, C. A. Murray, N. Enoch, S. Johnson, and C. Stinton, 'Psychological wellbeing in parents of children with Down syndrome: A systematic review and meta-analysis', *Clinical Psychology Review*, vol. 110, p. 102426, June 2024, doi: 10.1016/j.cpr.2024.102426.
[13] R. A. Phelps, J. D. Pinter, D. J. Lollar, J. G. Medlen, and C. D. Bethell, 'Health Care Needs of Children With Down Syndrome and Impact of Health System Performance on Children and Their Families', *Journal of Developmental & Behavioral Pediatrics*, vol. 33, no. 3, pp. 214–220, Apr. 2012, doi: 10.1097/DBP.0b013e3182452dd8.
[14] M. Gupta, A. R. Dhanasekaran, and K. J. Gardiner, 'Mouse models of Down syndrome: gene content and consequences', *Mamm Genome*, vol. 27, no. 11–12, pp. 538–555, Dec. 2016, doi: 10.1007/s00335-016-9661-8.
[15] S. E. Antonarakis, R. Lyle, E. T. Dermitzakis, A. Reymond, and S. Deutsch, 'Chromosome 21 and down syndrome: from genomics to pathophysiology', *Nat Rev Genet*, vol. 5, no. 10, pp. 725–738, Oct. 2004, doi: 10.1038/nrg1448.
[16] V. PLAIASU, 'Down Syndrome – Genetics and Cardiogenetics', *Maedica (Bucur)*, vol. 12, no. 3, pp. 208–213, Sept. 2017.
[17] L. E. Olson, J. T. Richtsmeier, J. Leszl, and R. H. Reeves, 'A Chromosome 21 Critical Region Does Not Cause Specific Down Syndrome Phenotypes', *Science*, vol. 306, no. 5696, pp. 687–690, Oct. 2004, doi: 10.1126/science.1098992.
[18] K. Gardiner, 'Gene-dosage effects in Down syndrome and trisomic mouse models', *Genome Biol*, vol. 5, no. 10, p. 244, 2004, doi: 10.1186/gb-2004-5-10-244.
[19] K.-H. Baek *et al.*, 'Down syndrome suppression of tumor growth and the role of the calcineurin inhibitor DSCR1', *Nature*, vol. 459, no. 7250, pp. 1126–1130, June 2009, doi: 10.1038/nature08062.
[20] X.-Q. Chen and X. Zuo, 'New insights into the effects of APP gene dose on synapse in Down syndrome', *Neural Regen Res*, vol. 19, no. 5, pp. 961–962, Aug. 2023, doi: 10.4103/1673-5374.382245.
[21] M. Vilardell *et al.*, 'Meta-analysis of heterogeneous Down Syndrome data reveals consistent genome-wide dosage effects related to neurological processes', *BMC Genomics*, vol. 12, p. 229, May 2011, doi: 10.1186/1471-2164-12-229.
[22] M. C. Pelleri *et al.*, 'Integrated Quantitative Transcriptome Maps of Human Trisomy 21 Tissues and Cells', *Front Genet*, vol. 9, p. 125, Apr. 2018, doi: 10.3389/fgene.2018.00125.
[23] K. Dimopoulos *et al.*, 'Cardiovascular Complications of Down Syndrome: Scoping Review and Expert Consensus', *Circulation*, vol. 147, no. 5, pp.

425–441, Jan. 2023, doi: 10.1161/CIRCULATIONAHA.122.059706.
[24] H. Zhang, L. Liu, and J. Tian, ‘Molecular mechanisms of congenital heart disease in down syndrome’, *Genes & Diseases*, vol. 6, no. 4, pp. 372–377, Dec. 2019, doi: 10.1016/j.gendis.2019.06.007.
[25] M. M. Elgendy, J. Cortez, F. Saker, M. A. Mohamed, and H. Aly, ‘Prevalence and Outcomes of Gastrointestinal Anomalies in Down Syndrome’, *American Journal of Perinatology*, vol. 41, pp. 2047–2052, May 2024, doi: 10.1055/s-0044-1786874.
[26] K. Ludwig *et al.*, ‘Congenital anomalies of the tubular gastrointestinal tract’, *Pathologica*, vol. 114, no. 1, pp. 40–54, Feb. 2022, doi: 10.32074/1591-951X-553.
[27] L. Nespoli, G. R. Burgio, A. G. Ugazio, and R. Maccario, ‘Immunological features of Down’s syndrome: a review’, *Journal of Intellectual Disability Research*, vol. 37, no. 6, pp. 543–551, 1993, doi: 10.1111/j.1365-2788.1993.tb00324.x.
[28] C. Martínez-Cué and N. Rueda, ‘Signalling Pathways Implicated in Alzheimer′s Disease Neurodegeneration in Individuals with and without Down Syndrome’, *International Journal of Molecular Sciences*, vol. 21, no. 18, p. 6906, Jan. 2020, doi: 10.3390/ijms21186906.
[29] ‘Pathology Outlines - Transient abnormal myelopoiesis associated with Down syndrome’. Accessed: Aug. 22, 2025. [Online]. Available: https://www.pathologyoutlines.com/topic/leukemiaTAM.html
[30] A. V. Gosavi, P. S. Murarkar, D. N. Lanjewar, and R. V. Ravikar, ‘Transient Leukemia in Down Syndrome: Report of Two Cases with Review of Literature’, *Indian J Hematol Blood Transfus*, vol. 27, no. 3, pp. 172–176, Sept. 2011, doi: 10.1007/s12288-011-0079-x.
[31] S. Triarico *et al.*, ‘Hematological disorders in children with Down syndrome’, *Expert Rev Hematol*, vol. 15, no. 2, pp. 127–135, Feb. 2022, doi: 10.1080/17474086.2022.2044780.
[32] D. BARCA *et al.*, ‘Intellectual Disability and Epilepsy in Down Syndrome’, *Maedica (Bucur)*, vol. 9, no. 4, pp. 344–350, Dec. 2014.
[33] S. Tapp, T. Anderson, and J. Visootsak, ‘Neurodevelopmental outcomes in children with Down syndrome and infantile spasms’, *J Pediatr Neurol*, vol. 13, no. 2, pp. 74–77, June 2015, doi: 10.1055/s-0035-1556768.
[34] S. Molinari *et al.*, ‘Endocrine, auxological and metabolic profile in children and adolescents with Down syndrome: from infancy to the first steps into adult life’, *Front. Endocrinol.*, vol. 15, Apr. 2024, doi: 10.3389/fendo.2024.1348397.
[35] K. A. Metwalley and H. S. Farghaly, ‘Endocrinal dysfunction in children with Down syndrome’, *Ann Pediatr Endocrinol Metab*, vol. 27, no. 1, pp. 15–21, Mar. 2022, doi: 10.6065/apem.2142236.118.
[36] D. S. Falconer, ‘QUANTITATIVE GENETICS’.
[37] P. M. Visscher, W. G. Hill, and N. R. Wray, ‘Heritability in the genomics

era--concepts and misconceptions', *Nat Rev Genet*, vol. 9, no. 4, pp. 255–266, Apr. 2008, doi: 10.1038/nrg2322.
[38] S. J. Rowe and A. Tenesa, 'Human Complex Trait Genetics: Lifting the Lid of the Genomics Toolbox - from Pathways to Prediction', *Curr Genomics*, vol. 13, no. 3, pp. 213–224, May 2012, doi: 10.2174/138920212800543101.
[39] A. Ćatović and S. Kendić, 'CYTOGENETIC FINDINGS AT DOWN SYNDROME AND THEIR CORRELATION WITH CLINICAL FINDINGS', *Bosn J Basic Med Sci*, vol. 5, no. 4, pp. 61–67, Nov. 2005, doi: 10.17305/bjbms.2005.3236.
[40] A. L. Beaudet, 'The Utility of Chromosomal Microarray Analysis in Developmental and Behavioral Pediatrics', *Child Dev*, vol. 84, no. 1, p. 10.1111/cdev.12050, 2013, doi: 10.1111/cdev.12050.
[41] M. Stoyanova, D. Yahya, M. Hachmeriyan, and M. Levkova, 'Diagnostic Yield of Next-Generation Sequencing for Rare Pediatric Genetic Disorders: A Single-Center Experience', *Med Sci (Basel)*, vol. 13, no. 2, p. 75, June 2025, doi: 10.3390/medsci13020075.
[42] C. R. Palmer, C. S. Liu, W. J. Romanow, M.-H. Lee, and J. Chun, 'Altered cell and RNA isoform diversity in aging Down syndrome brains', *Proceedings of the National Academy of Sciences*, vol. 118, no. 47, p. e2114326118, Nov. 2021, doi: 10.1073/pnas.2114326118.
[43] I. De Toma, C. Sierra, and M. Dierssen, 'Meta-analysis of transcriptomic data reveals clusters of consistently deregulated gene and disease ontologies in Down syndrome', *PLoS Comput Biol*, vol. 17, no. 9, p. e1009317, Sept. 2021, doi: 10.1371/journal.pcbi.1009317.
[44] N. El Hajj *et al.*, 'Epigenetic dysregulation in the developing Down syndrome cortex', *Epigenetics*, vol. 11, no. 8, pp. 563–578, May 2016, doi: 10.1080/15592294.2016.1192736.
[45] I. S. Muskens *et al.*, 'The genome-wide impact of trisomy 21 on DNA methylation and its implications for hematopoiesis', *Nat Commun*, vol. 12, no. 1, p. 821, Feb. 2021, doi: 10.1038/s41467-021-21064-z.
[46] C. Lanzillotta *et al.*, 'Proteomics Study of Peripheral Blood Mononuclear Cells in Down Syndrome Children', *Antioxidants*, vol. 9, no. 11, p. 1112, Nov. 2020, doi: 10.3390/antiox9111112.
[47] S. H. Lelieveld, M. Spielmann, S. Mundlos, J. A. Veltman, and C. Gilissen, 'Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein‐Coding Regions', *Hum Mutat*, vol. 36, no. 8, pp. 815–822, Aug. 2015, doi: 10.1002/humu.22813.
[48] A. Belkadi *et al.*, 'Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants', *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 17, p. 5473, Mar. 2015, doi: 10.1073/pnas.1418631112.
[49] S. M. Gaynor *et al.*, 'Yield of genetic association signals from genomes, exomes

and imputation in the UK Biobank', *Nat Genet*, vol. 56, no. 11, pp. 2345–2351, Nov. 2024, doi: 10.1038/s41588-024-01930-4.
[50] S. H. Lelieveld, M. Spielmann, S. Mundlos, J. A. Veltman, and C. Gilissen, 'Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein‐Coding Regions', *Hum Mutat*, vol. 36, no. 8, pp. 815–822, Aug. 2015, doi: 10.1002/humu.22813.
[51] A. Kumar, S. Adhikari, M. Kankainen, and C. A. Heckman, 'Comparison of Structural and Short Variants Detected by Linked-Read and Whole-Exome Sequencing in Multiple Myeloma', *Cancers (Basel)*, vol. 13, no. 6, p. 1212, Mar. 2021, doi: 10.3390/cancers13061212.
[52] S. I. Nikolaev *et al.*, 'Exome sequencing identifies putative drivers of progression of transient myeloproliferative disorder to AMKL in infants with Down syndrome', *Blood*, vol. 122, no. 4, pp. 554–561, July 2013, doi: 10.1182/blood-2013-03-491936.
[53] L. C. A. D'Alessandro *et al.*, 'Exome sequencing identifies rare variants in multiple genes in atrioventricular septal defect', *Genet Med*, vol. 18, no. 2, pp. 189–198, Feb. 2016, doi: 10.1038/gim.2015.60.
[54] L. Xicota *et al.*, 'Whole genome‐wide sequence analysis of long‐lived families (Long‐Life Family Study) identifies MTUS2 gene associated with late‐onset Alzheimer's disease', *Alzheimers Dement*, vol. 20, no. 4, pp. 2670–2679, Feb. 2024, doi: 10.1002/alz.13718.
[55] R. Cocoș and B. O. Popescu, 'Scrutinizing neurodegenerative diseases: decoding the complex genetic architectures through a multi-omics lens', *Hum Genomics*, vol. 18, p. 141, Dec. 2024, doi: 10.1186/s40246-024-00704-7.
[56] 'Sample to Insight - QIAGEN'. Accessed: Sept. 17, 2025. [Online]. Available: https://www.qiagen.com/it
[57] 'Qubit Fluorometric Quantification - IT'. Accessed: Sept. 17, 2025. [Online]. Available: https://www.thermofisher.com/uk/en/home/industrial/spectroscopy-elemental-isotope-analysis/molecular-spectroscopy/fluorometers/qubit.html
[58] 'Base Calling'. Accessed: Sept. 17, 2025. [Online]. Available: https://genohub.com/bioinformatics/10/
[59] 'Phred-scaled quality scores', GATK. Accessed: Sept. 17, 2025. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred-scaled-quality-scores
[60] M. Garcia *et al.*, 'Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants', Sept. 04, 2020, *F1000Research*. doi: 10.12688/f1000research.16665.2.
[61] 'A DSL for parallel and scalable computational pipelines | Nextflow'. Accessed: Sept. 17, 2025. [Online]. Available: https://www.nextflow.io/
[62] 'Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput

Sequence Data'. Accessed: Sept. 17, 2025. [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[63] *OpenGene/fastp*. (Sept. 13, 2025). C++. OpenGene - Open Source Genomics Toolbox. Accessed: Sept. 17, 2025. [Online]. Available: https://github.com/OpenGene/fastp
[64] *bwa-mem2/bwa-mem2*. (Sept. 12, 2025). C++. bwa-mem2. Accessed: Sept. 17, 2025. [Online]. Available: https://github.com/bwa-mem2/bwa-mem2
[65] 'Base Quality Score Recalibration (BQSR)', GATK. Accessed: Sept. 17, 2025. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR
[66] 'MarkDuplicates (Picard)', GATK. Accessed: Sept. 17, 2025. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360037052812-MarkDuplicates-Picard
[67] 'ApplyBQSR', GATK. Accessed: Sept. 17, 2025. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360037055712-ApplyBQSR
[68] 'HaplotypeCaller', GATK. Accessed: Sept. 17, 2025. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller
[69] W. McLaren *et al.*, 'The Ensembl Variant Effect Predictor', *Genome Biol*, vol. 17, no. 1, p. 122, June 2016, doi: 10.1186/s13059-016-0974-4.
[70] 'MultiQC | Seqera'. Accessed: Sept. 17, 2025. [Online]. Available: https://seqera.io/multiqc/
[71] N. V. Kovaleva, I. V. Butomo, and A. Körblein, '[Sex ratio in Down syndrome. Studies in patients with confirmed trisomy 21]', *Tsitol Genet*, vol. 35, no. 6, pp. 43–49, 2001.
[72] S. Buonaiuto *et al.*, 'Insights from the Biorepository and Integrative Genomics pediatric resource', *Nat Commun*, vol. 16, no. 1, May 2025, doi: 10.1038/s41467-025-59375-0.
[73] H. Li and R. Durbin, 'Fast and accurate short read alignment with Burrows–Wheeler transform', *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, July 2009, doi: 10.1093/bioinformatics/btp324.
[74] A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins, 'Sambamba: fast processing of NGS alignment formats', *Bioinformatics*, vol. 31, no. 12, pp. 2032–2034, June 2015, doi: 10.1093/bioinformatics/btv098.
[75] R. Poplin *et al.*, 'A universal SNP and small-indel variant caller using deep neural networks', *Nat Biotechnol*, vol. 36, no. 10, pp. 983–987, Oct. 2018, doi: 10.1038/nbt.4235.
[76] O. Delaneau, B. Howie, A. J. Cox, J.-F. Zagury, and J. Marchini, 'Haplotype Estimation Using Sequencing Reads', *The American Journal of Human Genetics*, vol. 93, no. 4, pp. 687–696, Oct. 2013, doi: 10.1016/j.ajhg.2013.09.002.
[77] N. M. Ioannidis *et al.*, 'REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants', *Am J Hum Genet*, vol. 99, no. 4, pp.

877–885, Oct. 2016, doi: 10.1016/j.ajhg.2016.08.016.
[78] ‘CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions | Nucleic Acids Research | Oxford Academic’. Accessed: Nov. 12, 2025. [Online]. Available: https://academic.oup.com/nar/article/52/D1/D1143/7511313?login=false
[79] ‘A genomic mutational constraint map using variation in 76,156 human genomes | Nature’. Accessed: July 15, 2025. [Online]. Available: https://www.nature.com/articles/s41586-023-06045-0
[80] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, ‘Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2’, *Curr Protoc Hum Genet*, vol. 0 7, p. Unit7.20, Jan. 2013, doi: 10.1002/0471142905.hg0720s76.
[81] P. C. Ng and S. Henikoff, ‘SIFT: predicting amino acid changes that affect protein function’, *Nucleic Acids Res*, vol. 31, no. 13, pp. 3812–3814, July 2003, doi: 10.1093/nar/gkg509.
[82] B. K. Maples, S. Gravel, E. E. Kenny, and C. D. Bustamante, ‘RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference’, *The American Journal of Human Genetics*, vol. 93, no. 2, pp. 278–288, Aug. 2013, doi: 10.1016/j.ajhg.2013.06.020.
[83] A. Auton *et al.*, ‘A global reference for human genetic variation’, *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015, doi: 10.1038/nature15393.
[84] A. Bergström *et al.*, ‘Insights into human genetic variation and population history from 929 diverse genomes’, *Science*, vol. 367, no. 6484, p. eaay5012, Mar. 2020, doi: 10.1126/science.aay5012.
[85] ‘R: The R Project for Statistical Computing’. Accessed: July 15, 2025. [Online]. Available: https://www.r-project.org/
[86] Y. Benjamini and Y. Hochberg, ‘Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing’, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, Jan. 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.
[87] S. C. Dyer *et al.*, ‘Ensembl 2025’, *Nucleic Acids Res*, vol. 53, no. D1, pp. D948–D957, Jan. 2025, doi: 10.1093/nar/gkae1071.
[88] G. Stelzer *et al.*, ‘The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses’, *Current Protocols in Bioinformatics*, vol. 54, no. 1, p. 1.30.1-1.30.33, 2016, doi: 10.1002/cpbi.5.
[89] B. Sanaa, D. Abdenasser, and E. H. Ayoub, ‘Congenital heart disease and Down syndrome: various aspects of a confirmed association’, *Cardiovasc J Afr*, vol. 27, no. 5, pp. 287–290, 2016, doi: 10.5830/CVJA-2016-019.
[90] D. Levenson, ‘The AJMG SEQUENCE: Decoding news and trends for the medical genetics community’, *American Journal of Medical Genetics Part A*, vol. 149A, no. 4, p. fm vii-fm x, 2009, doi: 10.1002/ajmg.a.32867.
[91] H. B. Laursen, ‘Congenital heart disease in Down’s syndrome.’, *Br Heart J*, vol.

38, no. 1, pp. 32–38, Jan. 1976, doi: 10.1136/hrt.38.1.32.
[92] D. Szklarczyk *et al.*, 'The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest', *Nucleic Acids Res*, vol. 51, no. D1, pp. D638–D646, Jan. 2023, doi: 10.1093/nar/gkac1000.
[93] M. M. Davis and P. J. Bjorkman, 'T-cell antigen receptor genes and T-cell recognition', *Nature*, vol. 334, no. 6181, pp. 395–402, Aug. 1988, doi: 10.1038/334395a0.
[94] L. Buck and R. Axel, 'A novel multigene family may encode odorant receptors: a molecular basis for odor recognition', *Cell*, vol. 65, no. 1, pp. 175–187, Apr. 1991, doi: 10.1016/0092-8674(91)90418-x.
[95] M.-P. Lefranc *et al.*, 'IMGT, the international ImMunoGeneTics information system', *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D1006-1012, Jan. 2009, doi: 10.1093/nar/gkn838.
[96] F. L. Margolis, 'Olfactory marker protein (OMP)', *Scand J Immunol Suppl*, vol. 9, pp. 181–199, 1982, doi: 10.1111/j.1365-3083.1982.tb03764.x.
[97] R. de Cid *et al.*, 'Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis', *Nat Genet*, vol. 41, no. 2, pp. 211–215, Feb. 2009, doi: 10.1038/ng.313.
[98] L. Buck and R. Axel, 'A novel multigene family may encode odorant receptors: A molecular basis for odor recognition', *Cell*, vol. 65, no. 1, pp. 175–187, Apr. 1991, doi: 10.1016/0092-8674(91)90418-X.
[99] H. W. Schroeder and L. Cavacini, 'Structure and function of immunoglobulins', *J Allergy Clin Immunol*, vol. 125, no. 2 Suppl 2, pp. S41-52, Feb. 2010, doi: 10.1016/j.jaci.2009.09.046.
[100] 'TAF11L2 TATA-box binding protein associated factor 11 like 2 [Homo sapiens (human)] - Gene - NCBI'. Accessed: Nov. 11, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/gene/391742
[101] J. D. Santoro *et al.*, 'Evidence of blood–brain barrier dysfunction and CSF immunoglobulin synthesis in Down Syndrome Regression Disorder', *Annals of Clinical and Translational Neurology*, vol. 12, no. 4, pp. 805–820, 2025, doi: 10.1002/acn3.52299.
[102] V. Callahan *et al.*, 'The Pro-Inflammatory Chemokines CXCL9, CXCL10 and CXCL11 Are Upregulated Following SARS-CoV-2 Infection in an AKT-Dependent Manner', *Viruses*, vol. 13, no. 6, p. 1062, June 2021, doi: 10.3390/v13061062.
[103] M. G. Roy *et al.*, 'Muc5b Is Required for Airway Defense', *Nature*, vol. 505, no. 7483, pp. 412–416, Jan. 2014, doi: 10.1038/nature12807.
[104] S. Y. Khan *et al.*, 'FOXE3 contributes to Peters anomaly through transcriptional regulation of an autophagy-associated protein termed DNAJB1', *Nat Commun*, vol. 7, p. 10953, Apr. 2016, doi: 10.1038/ncomms10953.
[105] S. Santosh Nirmala *et al.*, 'Beyond FOXP3: a 20-year journey unravelling human

regulatory T-cell heterogeneity', *Front Immunol*, vol. 14, p. 1321228, 2023, doi: 10.3389/fimmu.2023.1321228.
[106] K. C. Kim, 'Role of epithelial mucins during airway infection', *Pulm Pharmacol Ther*, vol. 25, no. 6, pp. 415–419, Dec. 2012, doi: 10.1016/j.pupt.2011.12.003.
[107] K. W. Hwang, T. J. Won, H. Kim, H.-J. Chun, T. Chun, and Y. Park, 'Characterization of the regulatory roles of the SUMO', *Diabetes Metab Res Rev*, vol. 27, no. 8, pp. 854–861, Nov. 2011, doi: 10.1002/dmrr.1261.
[108] X. Jiang *et al.*, 'Genetic dissection of the Down syndrome critical region', *Hum Mol Genet*, vol. 24, no. 22, pp. 6540–6551, Nov. 2015, doi: 10.1093/hmg/ddv364.
[109] K. E. McVeigh, J. J. Mallee, A. Lucente, B. L. Barnoski, S. Wu, and G. T. Berry, 'Murine chromosome 16 telomeric region, homologous with human chromosome 21q22, contains the osmoregulatory Na(+)/myo-inositol cotransporter (SLC5A3) gene', *Cytogenet Cell Genet*, vol. 88, no. 1–2, pp. 153–158, 2000, doi: 10.1159/000015509.
[110] S. J. Farley, A. Grishok, and E. Zeldich, 'Shaking up the silence: consequences of HMGN1 antagonizing PRC2 in the Down syndrome brain', *Epigenetics Chromatin*, vol. 15, no. 1, p. 39, Dec. 2022, doi: 10.1186/s13072-022-00471-6.
[111] D. Liu, H. Jia, D. I. R. Holmes, A. Stannard, and I. Zachary, 'Vascular endothelial growth factor-regulated gene expression in endothelial cells: KDR-mediated induction of Egr3 and the related nuclear receptors Nur77, Nurr1, and Nor1', *Arterioscler Thromb Vasc Biol*, vol. 23, no. 11, pp. 2002–2007, Nov. 2003, doi: 10.1161/01.ATV.0000098644.03153.6F.
[112] U. Hossain, A. K. Das, S. Ghosh, and P. C. Sil, 'An overview on the role of bioactive α-glucosidase inhibitors in ameliorating diabetic complications', *Food Chem Toxicol*, vol. 145, p. 111738, Nov. 2020, doi: 10.1016/j.fct.2020.111738.
[113] R. J. Brushia and D. A. Walsh, 'Phosphorylase kinase: the complexity of its regulation is reflected in the complexity of its structure', *FBL*, vol. 4, no. 4, pp. 618–641, Sept. 1999, doi: 10.2741/brushia.
[114] M. Kim, T. Tezuka, Y. Suziki, S. Sugano, M. Hirai, and T. Yamamoto, 'Molecular cloning and characterization of a novel *cbl*-family gene, *cbl-c*', *Gene*, vol. 239, no. 1, pp. 145–154, Oct. 1999, doi: 10.1016/S0378-1119(99)00356-X.
[115] 'Germline Mutations in NKX2-5, GATA4, and CRELD1 are Rare in a Mexican Sample of Down Syndrome Patients with Endocardial Cushion and Septal Heart Defects | Pediatric Cardiology'. Accessed: Nov. 13, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s00246-014-1091-3
[116] A. Riddell *et al.*, 'RUNX1: an emerging therapeutic target for cardiovascular disease', *Cardiovasc Res*, vol. 116, no. 8, pp. 1410–1423, July 2020, doi: 10.1093/cvr/cvaa034.
[117] O. Heidenreich, P. Derevyanko, A. Krippner-Heidenreich, and L. Swart, '3098 – RUNX1/RUNX1T1 ORCHESTRATES THE LEUKAEMIC MICROENVIRONMENT', *Experimental Hematology*, vol. 111, p. S94, Jan.

2022, doi: 10.1016/j.exphem.2022.07.154.
[118] M. F. Faienza *et al.*, ‘Cardiac Phenotype and Gene Mutations in RASopathies’, *Genes*, vol. 15, no. 8, p. 1015, Aug. 2024, doi: 10.3390/genes15081015.
[119] N. Mollo, R. Scognamiglio, A. Conti, S. Paladino, L. Nitsch, and A. Izzo, ‘Genetics and Molecular Basis of Congenital Heart Defects in Down Syndrome: Role of Extracellular Matrix Regulation’, *International Journal of Molecular Sciences*, vol. 24, no. 3, p. 2918, Jan. 2023, doi: 10.3390/ijms24032918.
[120] M. Yamagishi *et al.*, ‘Polycomb-mediated loss of miR-31 activates NIK-dependent NF-κB pathway in adult T cell leukemia and other cancers’, *Cancer Cell*, vol. 21, no. 1, pp. 121–135, Jan. 2012, doi: 10.1016/j.ccr.2011.12.015.
[121] D. Hollard, R. Berthier, and F. Douady, ‘[Granulopoiesis and its regulation]’, *Sem Hop*, vol. 51, no. 10, pp. 643–651, Feb. 1975.
[122] C.-Y. Wang, P. Yang, M. Li, and F. Gong, ‘Characterization of a negative feedback network between SUMO4 expression and NFkappaB transcriptional activity’, *Biochem Biophys Res Commun*, vol. 381, no. 4, pp. 477–481, Apr. 2009, doi: 10.1016/j.bbrc.2009.02.060.
[123] W.-C. Chen, C.-K. Wei, and J.-C. Lee, ‘MicroRNA-let-7c suppresses hepatitis C virus replication by targeting Bach1 for induction of haem oxygenase-1 expression’, *Journal of Viral Hepatitis*, vol. 26, no. 6, pp. 655–665, 2019, doi: 10.1111/jvh.13072.
[124] K. D. Sullivan *et al.*, ‘Trisomy 21 consistently activates the interferon response’, *eLife*, vol. 5, p. e16220, doi: 10.7554/eLife.16220.
[125] P. Araya *et al.*, ‘Trisomy 21 dysregulates T cell lineages toward an autoimmunity-prone state associated with interferon hyperactivity’, *Proc Natl Acad Sci U S A*, vol. 116, no. 48, pp. 24231–24241, Nov. 2019, doi: 10.1073/pnas.1908129116.

Glossary of terms and acronyms

Acronym / Gene	Full Name / Description
DS	Down Syndrome
AD	Alzheimer's Disease
HSA21	Homo sapiens chromosome 21
APP	Amyloid Precursor Protein
SNAP-25	Synaptosomal-Associated Protein 25
PT-DS	Partial Trisomy 21 Down Syndrome
CHD	Congenital Heart Disease
AVSD	Atrioventricular Septal Defect
GI	Gastrointestinal
SNP	Single Nucleotide Polymorphism
CNV	Copy Number Variation
GERD	Gastroesophageal Reflux Disease

NGF	Nerve Growth Factor
ALL	Acute Lymphoblastic Leukemia
AML-L7	Acute Myeloid Leukemia L7
TAM	Transient Abnormal Myelopoiesis
NGS	Next-Generation Sequencing
WGS	Whole-Genome Sequencing
WES	Whole-Exome Sequencing
WES-IMP	Whole-Exome Sequencing – Imputed Data
BIG	Biology of the Integrated Genome
VCF	Variant Call Format
VEP	Variant Effect Predictor
EHR	Electronic Health Record
ICD	International Classification of Diseases
HGDP	Human Genome Diversity Project
OPBG	Ospedale Pediatrico Bambino Gesù (Rome, Italy)
ASD	Atrial Septal Defect
GWAS	Genome-Wide Association Study
PMVK	Phosphomevalonate Kinase (gene)

DARS2	Aspartyl-tRNA Synthetase 2, Mitochondrial
RE_469.1	ICD/Phecode label for Respiratory conditions (likely “Acute bronchitis or bronchiolitis”)
RE_460.2	ICD/Phecode label for Acute lower respiratory infection
CM_763.31	Phecode/ICD label for Ventricular Septal Defect
TMEM92	Transmembrane Protein 92 (gene)
CM_763	Phecode group for Ventricular Septal Defect
TTC24	Tetratricopeptide Repeat Domain 24 (gene)
PAF1C-LEO 1	RNA Polymerase II–Associated Factor 1 Complex, LEO1 subunit
CDC73	Cell Division Cycle 73 (gene)
MT-ND4	Mitochondrially Encoded NADH Dehydrogenase 4
MUC4	Mucin 4 – membrane-associated mucin involved in epithelial protection and signal transduction
OR6K6	Olfactory Receptor Family 6 Subfamily K Member 6 – G protein–coupled receptor involved in olfactory signal transduction
DYRK1A	Dual Specificity Tyrosine-Phosphorylation-Regulated Kinase 1A
SOD1	Superoxide Dismutase 1
DSCAM	Down Syndrome Cell Adhesion Molecule

DSCR	Down Syndrome Critical Region
SLC5A3	Solute Carrier Family 5 Member 3
HMGN1	High Mobility Group Nucleosome Binding Domain 1
KDR	Kinase Insert Domain Receptor (VEGFR2)
GANC	Glucosidase Alpha, Neutral C
PHKB	Phosphorylase Kinase Regulatory Subunit Beta
CBLC	Cbl Proto-Oncogene C (gene)
GATA4	GATA Binding Protein 4
TRAJ47	T Cell Receptor Alpha Joining 47 – contributes to T-cell receptor diversity via V(D)J recombination
TRBV6-8	T Cell Receptor Beta Variable 6-8 – encodes a variable segment of the T-cell receptor β chain
TRBV6-7	T Cell Receptor Beta Variable 6-7 – encodes another β-chain variable region contributing to T-cell receptor diversity
OMP	Olfactory Marker Protein – cytoplasmic protein in mature olfactory sensory neurons, involved in olfactory signalling
TRAJ3	T Cell Receptor Alpha Joining 3 – joins α-chain segments in T-cell receptors
LCE3A	Late Cornified Envelope Protein 3A – structural component in epidermal differentiation and skin barrier formation

TRAJ36	T Cell Receptor Alpha Joining 36 – contributes to antigen recognition and TCR diversity
IGHV4-31	Immunoglobulin Heavy Variable 4-31 – encodes a variable region segment of immunoglobulin heavy chains in B cells
OR10A5	Olfactory Receptor Family 10 Subfamily A Member 5 – G protein–coupled receptor involved in odour recognition
IGLC7	Immunoglobulin Lambda Constant 7 – encodes the constant region of the λ light chain in antibodies
IGHV4-28	Immunoglobulin Heavy Variable 4-28 – heavy chain variable segment contributing to antibody diversity
IGHV2-70D	Immunoglobulin Heavy Variable 2-70D – heavy chain gene involved in B cell antibody variability
OR2S2	Olfactory Receptor Family 2 Subfamily S Member 2 – receptor mediating olfactory signal transduction
TAF11L2	TAF11-Like 2 – putative transcription factor related to TAF11
KRTAP4-5	Keratin Associated Protein 4-5 – structural component of hair cortex, crosslinks keratin filaments
CSF3	Colony Stimulating Factor 3 – regulates neutrophil proliferation and differentiation
SERPINA2	Serpin Family A Member 2 – serine protease inhibitor involved in inflammation regulation

NFKB1	Nuclear Factor Kappa B Subunit 1 – transcription factor controlling immune and inflammatory responses
STAT3	Signal Transducer and Activator of Transcription 3 – mediates cytokine signalling
CXCL8	C-X-C Motif Chemokine Ligand 8 (IL-8) – attracts neutrophils to infection sites
CCL5	C-C Motif Chemokine Ligand 5 (RANTES) – regulates immune cell trafficking
TNF	Tumour Necrosis Factor – proinflammatory cytokine regulating immune and apoptotic processes
MMP1	Matrix Metallopeptidase 1 – enzyme involved in extracellular matrix remodelling
MMP9	Matrix Metallopeptidase 9 – gelatinase involved in tissue remodelling and inflammation
IL6	Interleukin 6 – cytokine involved in inflammation, immune regulation, and haematopoiesis
IL10	Interleukin 10 – anti-inflammatory cytokine modulating immune responses
JAK3	Janus Kinase 3 – tyrosine kinase mediating cytokine receptor signalling
STAT1	Signal Transducer and Activator of Transcription 1 – mediates interferon signalling

IFNG	Interferon Gamma – activates macrophages and promotes antiviral immunity
HMOX1	Heme Oxygenase 1 – catalyses haem degradation, antioxidant and cytoprotective roles
HP	Haptoglobin – binds free haemoglobin to prevent oxidative damage
HPX	Hemopexin – binds free haem to protect against oxidative stress
CBS	Cystathionine Beta-Synthase – enzyme in homocysteine metabolism
ARSA	Arylsulfatase A – lysosomal enzyme degrading sulfatides
CXCL5	C-X-C Motif Chemokine Ligand 5 – recruits neutrophils to inflammatory sites
CCL11	C-C Motif Chemokine Ligand 11 (Eotaxin-1) – involved in eosinophil recruitment and allergic inflammation
CNTF	Ciliary Neurotrophic Factor – supports neuronal survival and modulates inflammatory signalling
IL6R	Interleukin 6 Receptor – mediates IL6 signalling via JAK/STAT
IL6ST	Interleukin 6 Signal Transducer (gp130) – mediates IL6-family cytokine signal transduction
LEP	Leptin – adipokine regulating energy balance and immune function
FOXE3	Forkhead Box E3 – transcription factor essential for lens and epithelial development

MIRLET7C	MicroRNA Let-7c – post-transcriptional gene regulation, located on HSA21
ICOSLG	Inducible T-cell Costimulator Ligand – co-stimulatory molecule in T cell activation
KMT2A	Lysine Methyltransferase 2A (MLL) – histone methyltransferase regulating gene expression
LINC02605	Long Intergenic Non-Protein Coding RNA 2605 – may regulate immune-related genes
NFKBIA	Nuclear Factor Kappa B Inhibitor Alpha – cytoplasmic inhibitor of NF-κB signalling
CTLA4	Cytotoxic T-Lymphocyte-Associated Protein 4 – inhibitory receptor controlling T cell activation
HLA-DRB1	Major Histocompatibility Complex Class II, DR Beta 1 – encodes a class II antigen-presenting molecule
TNFRSF1A	TNF Receptor Superfamily Member 1A – mediates TNF-induced apoptosis and inflammation
CD40	CD40 Molecule – costimulatory receptor on antigen-presenting cells
IL7R	Interleukin 7 Receptor – essential for lymphocyte development and homeostasis
PIK3R1	Phosphoinositide-3-Kinase Regulatory Subunit 1 – regulates growth and metabolism via PI3K signalling

ADA	Adenosine Deaminase – enzyme in purine metabolism and lymphocyte development
VAV1	Vav Guanine Nucleotide Exchange Factor 1 – regulates actin cytoskeleton in immune cells
HCK	HCK Proto-Oncogene, Src Family Tyrosine Kinase – myeloid cell activation
LRRK2	Leucine Rich Repeat Kinase 2 – involved in inflammation and neuronal processes
PTEN	Phosphatase and Tensin Homolog – tumour suppressor regulating cell growth and survival
XIAP	X-linked Inhibitor of Apoptosis Protein – inhibits caspases and prevents apoptosis
CARMIL2	Capping Protein Regulator and Myosin 1 Linker 2 – cytoskeletal regulator in immune synapse formation
MAF	MAF BZIP Transcription Factor – regulates immune and developmental pathways
APOA1	Apolipoprotein A1 – major HDL protein involved in lipid transport and anti-inflammatory effects
SUMO4	Small Ubiquitin-Like Modifier 4 – regulates protein stability and NF-κB activity
TNFRSF18	TNF Receptor Superfamily Member 18 (GITR) – involved in T-cell activation and apoptosis

TNFRSF4	TNF Receptor Superfamily Member 4 (OX40) – mediates T-cell co-stimulation and immune regulation
CSH1 / CSH2	Chorionic Somatomammotropin Hormone 1/2 – placental hormones in growth and metabolic regulation
CTSG	Cathepsin G – serine protease in inflammation and host defence
CEACAM3	Carcinoembryonic Antigen-Related Cell Adhesion Molecule 3 – neutrophil activation receptor
MUC6	Mucin 6 – secreted mucin contributing to mucosal barrier and immune protection
IVL	Involucrin – structural protein in keratinocytes, essential for epithelial barrier formation
KRT36	Keratin 36 – intermediate filament protein in epithelial structure
TSSK2	Testis-Specific Serine/Threonine Kinase 2 – involved in spermatogenesis and signalling
IFNL2	Interferon Lambda 2 – type III interferon in antiviral defence
ID3	Inhibitor of DNA Binding 3 – transcriptional regulator modulating cell differentiation
KIF2B	Kinesin Family Member 2B – motor protein involved in microtubule dynamics
GSTP1	Glutathione S-Transferase Pi 1 – detoxification enzyme in oxidative stress response

WDR38	WD Repeat Domain 38 – protein potentially involved in ciliary function
KLF14	Kruppel-Like Factor 14 – transcription factor regulating lipid metabolism and inflammation
USH1G	Usher Syndrome Type 1G Protein – scaffold protein in cell adhesion and ciliary function

Funding

This project has been funded by grants of the European Union - NextGenerationEU, (Scholarship DM 352), Mission 4 Component 2, CUP: B53C22001780004.

Acknowledgment

I would like to express my sincere gratitude to **GenomeUp** for the opportunity to collaborate and to deepen my understanding of the company's world, as well as for granting access to their data, which greatly enriched my research experience.

My heartfelt thanks also go to the **Colonna and Garrison Labs in Memphis (a special thanks to Silvia)** for providing valuable datasets and, even more importantly, for allowing me to learn firsthand how a genomic bioinformatics unit is structured and operates.

Last but not least, I wish to thank **Allegra Via**, both for her invaluable mentorship and for her constant emotional support throughout this journey.