Genetics and Epidemiology of Cholesteatoma in the UK BioBank

A thesis submitted in accordance with the requirements of the
Degree of Doctor of Philosophy

by
Emma Wilson

University of East Anglia
2025

Abstract

Cholesteatoma is a skin cyst that grows in the middle ear. It is rare, non-cancerous and locally invasive, frequently resulting in hearing loss due to destruction of the ossicles. Serious and life-threatening sequalae are possible, including facial nerve palsy, meningitis, and abscess of the brain. The only treatment is surgical excision, which can exacerbate hearing loss. There are several theories regarding the origin of cholesteatoma, but its biology is still uncertain. Although not typically considered a heritable disease, observations of family clustering suggest a genetic component. The primary aim of this thesis was to investigate genetic risk of cholesteatoma though genome-wide association study (GWAS) of UK BioBank whole exome data for 1,000 cholesteatoma cases. Single-variant, gene level and gene-set enrichment analyses were performed. This was supported by an epidemiological analysis of demographic factors associated with cholesteatoma and other middle ear disease and a review of global gene expression studies. No single genes or variants met genome-wide significance, but pathways related to cell adhesion, cytoskeletal organisation, ciliary function and calcium binding were enriched for low *p*-value variants. These results were supported by pathway analysis of summary statistics from a Finnish biobank (FinnGen) and a previous cholesteatoma whole exome study of affected families. Dynein binding was also enriched in UK BioBank whole exome data due to rare *DNAH* and *DNAI* family variants, which is promising as *DNAH* variants were also detected in our previous whole exome study and are known to contribute to similar pathologies such as primary ciliary dyskinesia. These results support the existence of a highly polygenic effect on cholesteatoma risk and indicate several pathways for further study.

Acknowledgements

Thank you to my supervisory team, Dan Brewer, Barbara Jennings, Mizanur Khondoker, Carl Philpott and Peter Prinsley. Special thanks to Barabara and Dan for their many readings of various drafts of this thesis and for helping with the publishing of my paper.

I would like to acknowledge the researchers and participants of the UK BioBank without whom this research would not be possible. I also want to acknowledge the participants and investigators of the FinnGen study, which provided a valuable comparison cohort for this thesis.

And of course all the thanks in the world to my family, who are always there to support me.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Cholesteatoma Biology

Cholesteatoma is locally invasive skin cyst occurring in the middle ear. The middle ear is the cavity in which the delicate bones of the ossicular chain are situated. These bones are responsible for conducting vibrations from the tympanic membrane (the ear drum), which separates the middle ear from the external auditory canal to the bony inner ear. The middle ear drains into the nasopharynx via the Eustachian tube, which is responsible for maintaining pressure equilibrium. Expansion of the cyst leads to destruction of the surrounding bone, most commonly the ossicles, resulting in conductive hearing loss. The middle ear is surrounded by the mastoid bone which contains many air cells. The cholesteatoma can invade and erode the mastoid, provoking mastoiditis, and can lead to infection of the brain if the bone is eroded through. The facial nerve passes through the bone close to the middle ear and can also be damaged by the cyst. Surgical removal of the cyst is the only known treatment and is usually performed via mastoidectomy which involves entry into the middle ear via an incision behind the ear, the removal of the cyst and well as the mastoid air cells to reduce risk of recurrence.

1.1.1 Cholesteatoma histology, forms, and features

Histologically, cholesteatoma resembles an epidermoid cyst[1] and is composed of three main layers. The main sac of the cholesteatoma is called the matrix and is essentially ordinary epidermis, consisting of a basal, granular, and lucid layer. The sac is filled with cystic content comprising layered, anucleate keratin squames, sebaceous tissue and necrotic matter shed from the inner layer of the matrix (analogous to the outer layer of the skin) whose accumulation drives expansion of the cyst[2]. The cyst is typically surrounded by a layer of inflamed connective tissue called the perimatrix and has an outwardly pearly, smooth appearance and a layered, undulating structure[1] (**Figure 1c**).

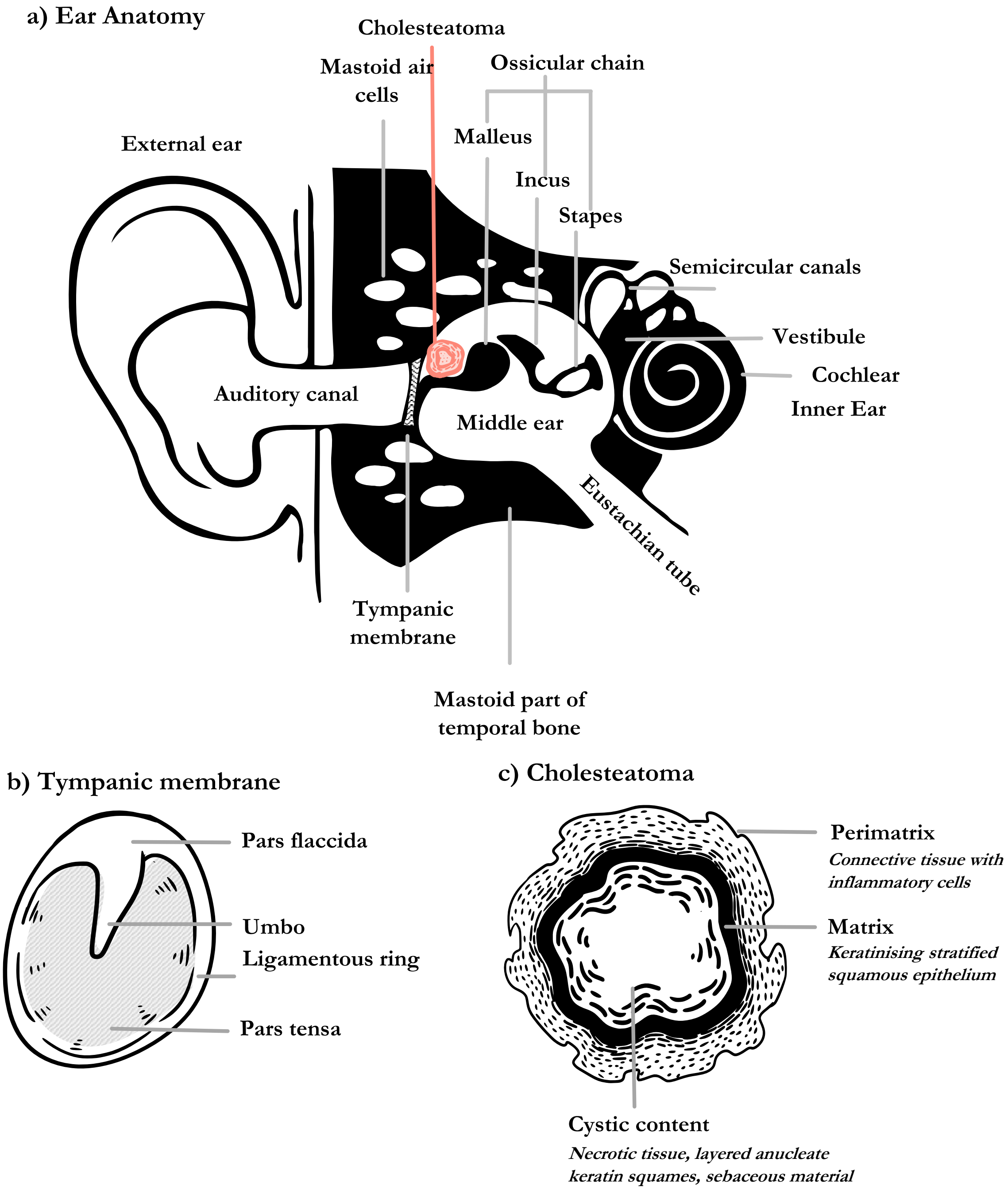
Cholesteatoma is typically located behind the pars flaccida of the tympanic membrane (**Figure 1b**) where it grows into the attic (the space above the ossicular chain, Figure **1a**) but also behind the pars tensa[2]. Very rarely, cholesteatoma occurs between the layers of an intact tympanic membrane[3] or in the external auditory canal; the latter may be confused with the similar but distinct condition keratosis obturans[4]. Typically, only one ear is affected but disease may be bilateral[1].

Cholesteatomas are broadly classified into congenital and acquired types[2]. Congenital cholesteatoma is diagnosed when there is no history of tympanic retraction, perforation or surgery, comprises 4-24% of cholesteatoma cases[5] and is typically seen in young children[2]. Acquired cholesteatoma is more common and is defined as cholesteatoma in a retraction pocket of the tympanic membrane (primary acquired), with tympanic perforation (secondary acquired) or following surgery (iatrogenic).

Structurally, there is little difference between congenital and acquired cholesteatomas, although the dense fibrous layer of the perimatrix tends to be denser in adults whereas congenital cholesteatomas have more granulation tissue[1]. Congenital cholesteatomas may progress with fewer overt symptoms than adult acquired cholesteatoma[1] but pathology is otherwise similar.

Figure 1. Anatomy of the middle ear, tympanic membrane, and cholesteatoma cyst.

A) Cross section of the right ear showing a typical location for cholesteatoma in the 'attic' of the middle ear, above the malleus and resting against the pars flaccida of the tympanic membrane. B) Lateral surface of tympanic membrane (ear drum). The ear drum consists of a thin membrane bordered by a tough ligamentous ring. The umbo is the concave point on the membrane where the malleus attaches to the anterior surface. The pars tensa is pulled taught while the pars flaccida is small and flaccid. C) the cholesteatoma cyst consists of three layers: the perimatrix, an inflamed granulation tissue with a dense connective layer towards the centre; the matrix, a sac of keratinizing stratified squamous epithelium with the basal layer contacting the perimatrix; and the cystic content, a collection of keratin, sebaceous and necrotic debris which makes up the bulk of the cyst.



a) Ear Anatomy

b) Tympanic membrane

c) Cholesteatoma

Perimatrix
Connective tissue with inflammatory cells
Matrix
Keratinising stratified squamous epithelium
Cystic content
Necrotic tissue, layered anucleate keratin squames, sebaceous material

1.1.2 Epidemiology

Cholesteatoma affects between 6.8 and 18.8 people per 100,000 per year depending on the study population (**Table 1**). Cholesteatoma is more common in males by a factor of approximately 1.4[6] and annual incidence is highest in children around 9 years of age[7]. Mean age of onset is >30 years of age and is later in females[6,7]. Congenital cholesteatoma is rarer than acquired and has a mean age of onset of 5.6 years[2]. Some studies[7–9] have found decreasing rates over time, which Djurhuus *et al.* (2015)[9] attribute to increasing treatment of childhood chronic otitis media with ventilation tubes.

Table 1. Incidence of cholesteatoma in a selection of epidemiological studies

Study	Rate (per 100,000 people per year)	Country	Year	N cases	Notes
Im *et al.*, (2020)[8]	6.17-7.15	South Korea	2006-2016	42,705	Surgically treated Rates decreasing
Shibata *et al.* (2015)[10]	6.8-10.0	Japan	2008	40	Found no significant differences in age or sex, but only 40 cases and 175 controls. Highest incidence in 60+ age group.
Britze *et al.*, (2017)[11]	6.8	Denmark	2002-2005	147	Surgically treated 10-year recidivism rate was 0.44 Age < 15 more likely to have recidivism
Kemppainen *et al.* (1999)[6]	9.2	Finland	1982-1981	500	Rates decreased during study period Higher incidence in skilled/specialised workers Male:female ratio 1.4 Median age 38 (males), 45 (females) 22 (4.4%) bilateral
Padgham *et al.*, (1989)[12]	9.4-18.8	Scotland	1966-1986		Study of surgically treated children < 15 Rates stable during study period
Djurhuus *et al.* (2010)[7]	14.3 (males) 9.1 (females)	Denmark	1977-2007	13,606	Surgically treated, registry-based Age-specific peak at 9 years old (21.4 male, 13.6 female). Male:female ratio 1.51. Median age was 32 for males and 35 for females.
Djurhuus *et al.* (2015) [9]	10-15	Denmark	1977-2010	3874	Surgically treated children Rates increased from 1977 to 2002 and decreased from 2002-2010

Cholesteatoma is thought to be more common in white populations and rare in Black and non-Indian Asian populations[2,13,14], although original epidemiological data was not presented in the cited studies. Conversely, the annual incidences reported in **Table 1** are similar for

European and East Asian populations. Ratnesar (1976)[15] reports vanishingly rare cholesteatoma in the Innuit and Innu people of Newfoundland compared to white populations in the same region, despite otherwise high rates of chronic ear disease. Thornton *et al.* (2011)[16] report no difference in the rates of cholesteatoma amongst individuals with chronic otitis media of Tibeto-Mongolian and Indo-Caucasian ancestry in Nepal, nor any difference between geographical regions.

Chronic otitis media (COM) is frequently seen with cholesteatoma, although to what extent COM precedes versus arises from cholesteatoma is unclear. Possibly, a background of chronic inflammation encourages cholesteatoma formation and is often a precursor to tympanic retraction or perforation. Alternatively, ears which are susceptible to chronic disease may also be susceptible to cholesteatoma for overlapping reasons. Symptoms such as inflammation and effusion can also arise from cholesteatoma, so it is possible that a diagnosis of COM indicates underlying cholesteatoma.

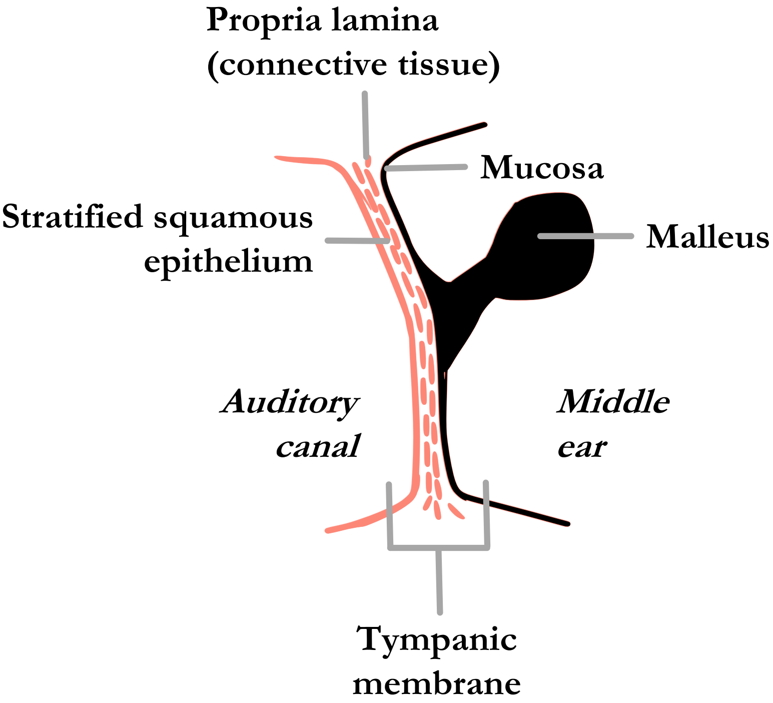
There is also an increased risk of cholesteatoma in persons with primary ciliary dyskinesia, which can result in recurrent ear, sinus and lung infections[17,18]. Both COM and cholesteatoma are more common amongst individuals with some disorders affecting craniofacial morphology, including Down syndrome, Turner syndrome, Branchio-oto-renal syndrome and cleft lip and palate (orofacial cleft)[18]. Chronic ear disease is also more prevalent in males[19], further suggesting a link between susceptibility to general ear disease and cholesteatoma. Whether COM raises the risk of cholesteatoma directly or through shared genetic or environmental factors is not known.

1.1.3 Theories of formation

It has not been conclusively demonstrated how or why cholesteatomas form. Notably, the cholesteatoma matrix consists of keratinising stratified squamous epithelium, or skin tissue, whereas the middle ear is lined with simple cuboidal mucosa. The tympanic membrane itself consists of three layers: an outer surface of stratified squamous epithelial tissue continuous with the skin of the auditory canal; an inner surface of mucosa continuous with the middle ear; and the lamina propria, a layer of connective tissue, between the two[13] (**Figure 2**).

Figure 2. Cross section of tympanic membrane

A cross-section of the tympanic membrane showing the external auditory canal to the viewer's left and the middle ear space to the right. The tympanic membrane consists of stratified squamous epithelium on the external side, continuous with the external auditory canal, and mucosa on the inner side, continuous with the mucosa of the middle ear. The middle layer consists of tough connective tissue.



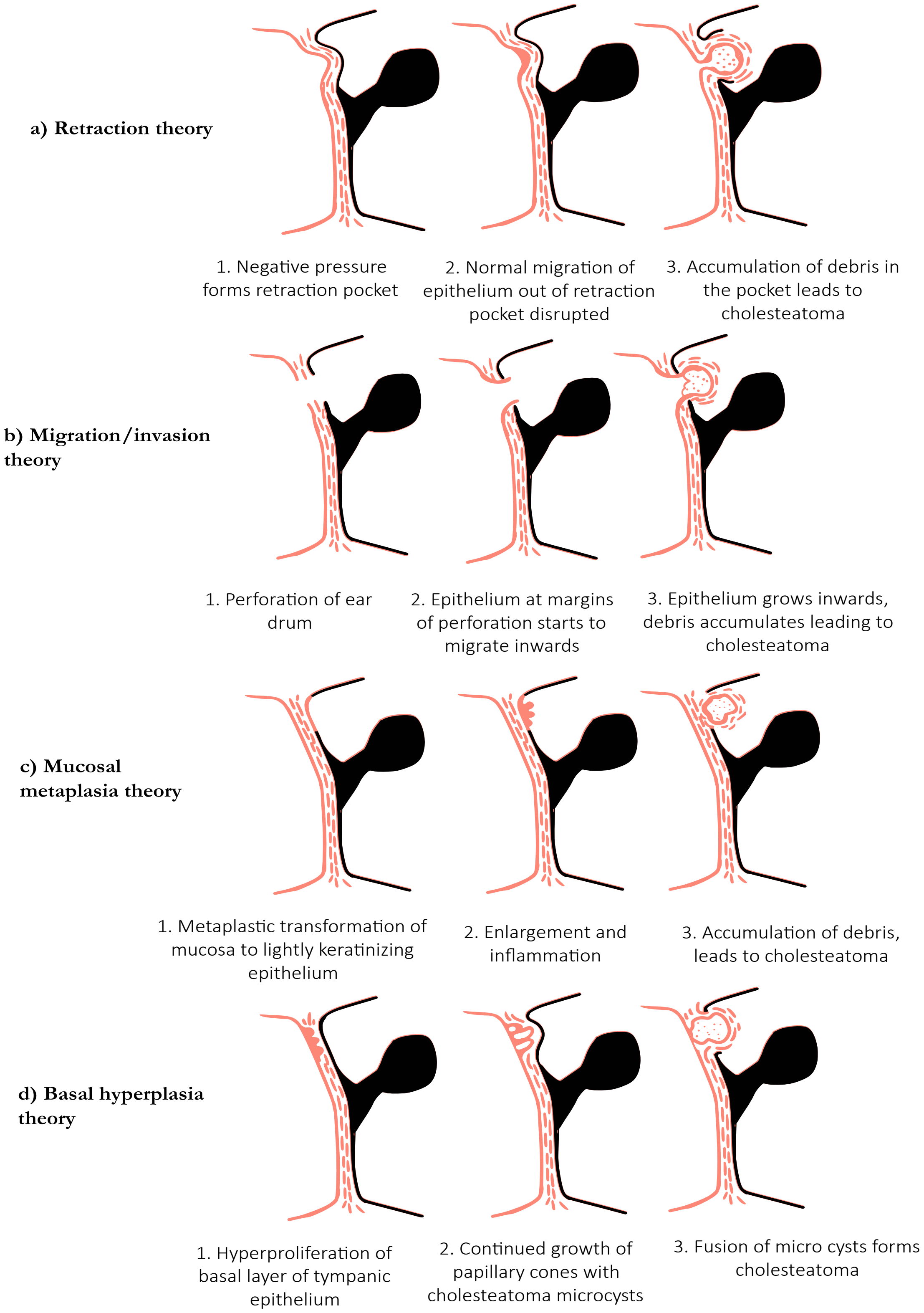
There are four main theories for the origins of cholesteatoma epithelium in the middle ear. First, tympanic retraction (**Figure 3a**) theory or invagination theory proposes that accumulation of keratin debris in a retraction pocket of the tympanic membrane results in the formation of a cholesteatoma. Therefore, the epithelial tissue on the lateral surface of the tympanic membrane forms the matrix of the cholesteatoma. Retraction of the pars flaccida and less commonly the pars tensa can occur due to negative pressure in the ear resulting from chronic infection[20]. Usually, the tympanic membrane maintains a self-cleaning mechanism by the outwards migration of epithelial cells from the lateral surface of the tympanic membrane[21]. Louw (2010)[21] proposes that this process is impaired in cholesteatoma, leading to collection of debris in the retraction pocket with a high turnover of epithelial cells which ultimately results in cholesteatoma. Retraction pockets are common in chronic ear infection, but most will not proceed to cholesteatoma[22,23].

The invasion or migration theory (**Figure 3b**) suggests that epithelium from the margins of a tympanic perforation migrates through the ear drum into the middle ear[2]. Perforation may occur due to chronic infection, trauma, or surgery. Louw suggests that epithelium grows into the middle ear forming mucocutaneous junctions where it meets the middle ear mucosa. Then, rather than growing across the tympanic membrane to heal the perforation, the epithelium migrates inwards.

The mucosal metaplasia **(Figure 3c**) theory proposes that the epithelial tissue does not originate in the epithelium of the tympanic membrane, rather that the mucosa of the tympanic membrane or middle ear undergoes metaplastic transformation into keratinising stratified squamous epithelium[2]. Metaplasia is a general term describing the conversion of any differentiated cell type to another and does not indicate the direct cause of transformation – in this case, the possible cause is unknown but subsequent expansion of the growth and repeated infection could lead to perforation, leading the appearance of typical acquired cholesteatoma[2].

Epidermal basal cell hyperplasia (**Figure 3d**) theory suggests that cholesteatoma microcysts formed in the pars flaccida invade the subepithelial tissue of the middle ear. Prolonged inflammation may provide the conditions stimulating epidermal hyperplasia and papillary cone formation; subject to intense inflammation in the perimatrix, cones become elongated and keratin desquamation towards the centre of the cones form micro-cholesteatomas which eventually fuse under pressure[21].

Figure 3. Four theories of acquired cholesteatoma formation. Adapted from Kuo et al. (2014)[20]



a) Retraction theory

1. Negative pressure forms retraction pocket

2. Normal migration of epithelium out of retraction pocket disrupted

3. Accumulation of debris in the pocket leads to cholesteatoma

b) Migration/invasion theory

1. Perforation of ear drum

2. Epithelium at margins of perforation starts to migrate inwards

3. Epithelium grows inwards, debris accumulates leading to cholesteatoma

c) Mucosal metaplasia theory

1. Metaplastic transformation of mucosa to lightly keratinizing epithelium

2. Enlargement and inflammation

3. Accumulation of debris, leads to cholesteatoma

d) Basal hyperplasia theory

1. Hyperproliferation of basal layer of tympanic epithelium

2. Continued growth of papillary cones with cholesteatoma microcysts

3. Fusion of micro cysts forms cholesteatoma

Finally, congenital cholesteatoma is thought to arise from a remnant of embryonic epithelium in the middle ear. Epithelial tissue may derive from viable epithelial cells in the amniotic fluid;

ectoderm from the external auditory canal may migrate into the middle ear during development; or an epithelial remnant present during early development persists to give rise to cholesteatoma[21]. Congenital epidermoid cysts may occur in the same manner and are most common in the brain and temporal bone, where they may also be described as congenital cholesteatomas[24].

Evaluation of theories of formation

Aspects of each theory for acquired cholesteatoma have been demonstrated in animal models[25]. Common animal models include Mongolian gerbils, rats, chinchillas and guinea pigs; cholesteatoma can be very successfully induced by ligating the external auditory canal in the Mongolian gerbil, leading to the accumulation of keratin debris and a retraction-type cholesteatoma[25]. Meanwhile, injection of talcum powder and fibrin, dimethyl-benzanthrancene, propylene glycol or cortisporin into the middle ear can induce chronic otitis media and cholesteatoma in a range of rodent models[25]. These models support obstructions to middle ear clearance and aberrations in epithelial migration as potential causes for cholesteatoma, though they are induced by rather extreme measures and may not reflect what occurs during a spontaneous cholesteatoma.

Additionally, none of these theories have been conclusively demonstrated in humans. The presence of perforation or retraction as a requisite for acquired cholesteatoma diagnosis, and with most cholesteatomas being acquired, strongly supports the retraction and migration theories. Although it is possible that retraction and perforation may occur secondary to cholesteatoma, retraction pocket formation prior to cholesteatoma is supported by observation[23,26]. While tympanic perforation and retraction appear to be important in many cholesteatomas, they are not present in all cases (including congenital cholesteatomas), nor are they sufficient to cause cholesteatoma as most perforations and retractions resolve without causing it.

However, the theories are not mutually exclusive and may contribute in different parts to different cholesteatomas: histological features amongst cholesteatomas taken from different individuals support different theories of formation with the presence of mucocutaneous junctions in some supporting invasion theory, while papillary cone-like keratin deposits in others support basal cell hyperplasia. Rarely, cholesteatoma occurs between the layers of an intact tympanic membrane (intratympanic cholesteatoma); hyperplasia may explain a lack of perforation or retraction, although most intratympanic cholesteatomas are associated with ear trauma or surgery[27].

Parallels between the theories of cholesteatoma formation and other epidermoid cysts can be seen. Epidermoid cysts are mainly thought to occur when a skin follicle becomes blocked, resulting in build-up of keratin, or through traumatic introduction of epithelium into the dermis[28]. Rupture of an epidermoid cyst leads to inflammation in the surrounding tissue as a response to the presence of keratin[28]. External auditory canal cholesteatomas do not seem to fit well with any of these explanations, although they may be considered another form of epidermoid cyst which happens to arise in the ear canal.

1.1.4 Pathology

Key features of cholesteatoma pathology are progressive expansion of the cyst and destruction of the surrounding bone[2]. Cholesteatoma tissue is typically inflamed, though active infection may or may not be present; inflammation may be in response to spilling of the keratin contents of the cyst, with keratins acting as alarmins which signal an immune response[29]. Repeated infection and damage to the middle ear bony structures are responsible for symptoms including chronic discharge, hearing loss due to destruction or immobilisation of the ossicles, and facial nerve damage to due invasion of the mastoid. The drivers of expansion and precise mechanisms of bone resorption have not been conclusively demonstrated but may involve mechanical pressure, loss of extracellular matrix integrity, osteoclast activation and growth factors expressed in the context of chronic inflammation[2,20]. Paediatric acquired cholesteatoma may be more aggressive than adult cholesteatoma[30], with greater extent and recidivism[31], and childhood cases have been shown to express more inflammatory proteins[32] and have relatively more granulation tissue[33] although many studies do not distinguish between adult and childhood cases.

Extracellular matrix breakdown

The extracellular matrix (ECM) is the scaffold in which cells are organised within tissues such as epithelium. It provides physical structure and facilitates cell-cell communication, migration, and adhesion, and can coordinate proliferation and differentiation[34]. The perimatrix of cholesteatoma is a type of granulation tissue – a connective tissue with inflammatory cell infiltration associated with wound repair which the ECM has an important role in coordinating. During wound-healing, the ECM is first broken down; cells proliferate and differentiate, new blood vessels form (angiogenesis) and tissue remodelling occurs, with apoptosis of old cells and replacement of old ECM collagens[35].

Cholesteatoma shows a degraded ECM in comparison to healthy skin and chronic otitis media granulation tissue[36]. Indeed, cholesteatoma has been described as an impaired wound-healing process[2]. ECM degradation could be involved in excess proliferation, inflammation and altered migration in cholesteatoma. Migration is an important feature because the origin of the epithelial tissue is suspected to be the external auditory canal or external surface of the tympanic membrane; therefore, it must migrate into the middle ear, either via a perforation or through failure to migrate out of a retraction pocket. Additionally, expression of matrix-active proteases such as the matrix metalloproteinases (several of which have been detected upregulated in cholesteatoma) may contribute to bone loss and invasiveness[20].

Osteoclasts

Bone remodelling is a constant process consisting of breakdown by osteoclasts and generation of new bone by osteoblasts[37]. Osteoclast precursor cells are activated by binding of the RANK receptor by RANKL and proceed to release bone matrix-lytic enzymes, the most important being cathepsin-K (CATK) and tartrate-resistant acid phosphatase (TRAP)[38]. Osteoclast involvement in cholesteatoma is controversial: some studies suggest activation of osteoclasts[39] and depleted osteoblast populations[40], though Koizumi *et al.*[41] did not detect osteoclast activity in cholesteatoma-affected bone.

Osteoporosis pathology arises from an imbalance in osteoclast and osteoblast activity leading to poor bone density[37]; as bone resorption in the middle ear associated with cholesteatoma may also be due to osteoclast activity, it is possible that the conditions are related. Thorsteinsson *et al.*[42] found that treatment with bisphosphonates, a class of drugs given to improve bone density in osteoporosis, can also increase risk of external auditory canal cholesteatoma. These drugs interfere with osteoclast function to reduce the turnover rate of bone and also increase the risk of jaw osteonecrosis and atypical femur fracture.

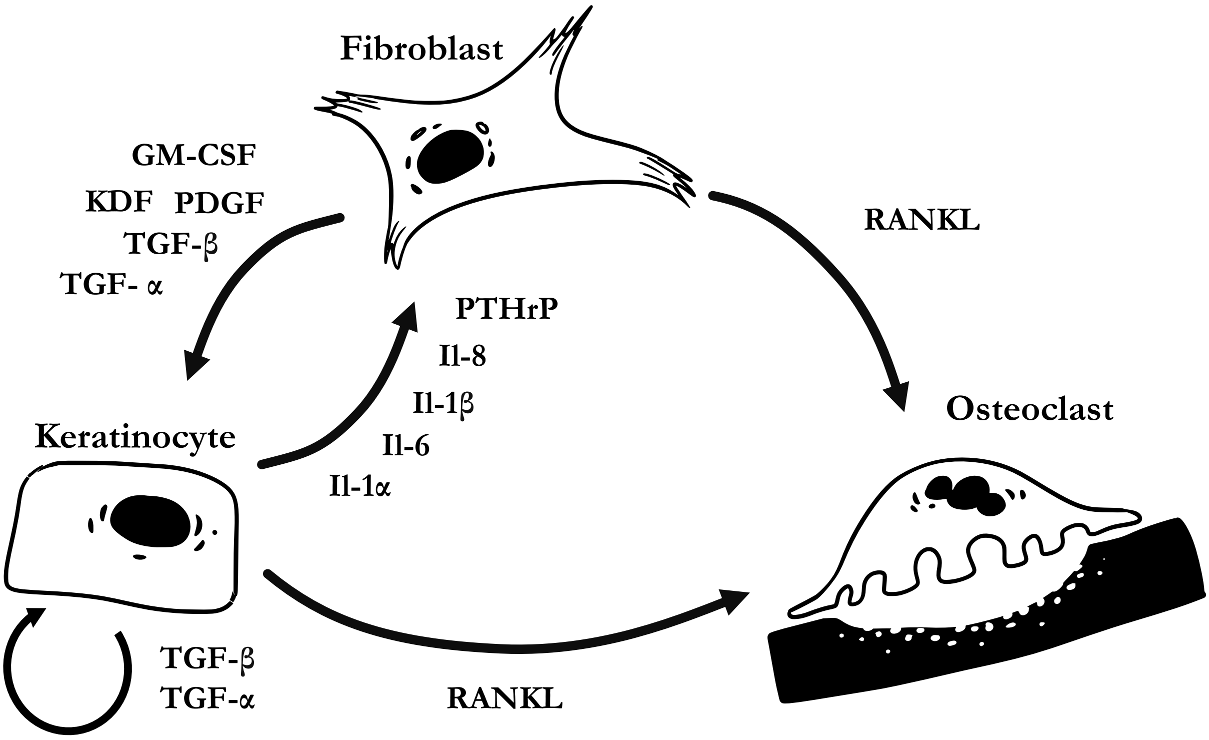
Inflammation

Cholesteatoma tissue is often inflamed, and affected individuals often have a history of chronic otitis media[1] but the relationship between the conditions is not clear. Both may arise from shared risk factors, such as a poorly draining ear, or a history of repeated infection may directly raise cholesteatoma risk. Louw (2010)[21] suggests that chronic inflammation may have a role in the initial establishment of cholesteatoma by triggering aberrant migration, hyperproliferation or mucosal metaplasia.

Inflammation is also implicated in promoting tissue growth and bone loss, and the immune response in cholesteatoma has been described as overly-aggressive[14]. Possibly, inflammatory cytokines provoke excessive proliferation through a positive feedback loop; the keratinocytes of the matrix express several proinflammatory cytokines including interleukins IL1α, IL1β, IL6 and IL8, and parathyroid-hormone-related protein. The matrix fibroblasts in turn produce growth factors including epidermal growth factor, platelet-derived growth factor, keratinocyte growth factor, and transforming growth factor alpha[20]. Imai *et al.*[39] suggest that inflammatory products activate osteoclasts via receptor activator of NF-κB ligand (RANKL) expressed in the perimatrix, possibly due to expression of tumour necrosis factor alpha (TNF- α), PGE2, IL6 and IL1β by the matrix fibroblasts and/or keratinocytes (**Figure 4**).

Figure 4. Interaction between fibroblast, keratinocyte and osteoclast promoting hyperproliferation and bone loss.

Fibroblasts and keratinocytes in the cholesteatoma matrix produce growth factors and inflammatory cytokines which act in a positive feedback loop promoting hyperproliferation. Production of RANKL and TGF-a promote osteoclast activation and bone resorption. Adapted from Kuo (2015)[14] and Imai et al. (2019)[39].



However, several inflammatory proteins widely expressed in cholesteatoma have been shown to be *downregulated* compared to granulation tissue from cholesteatoma-free chronic otitis media (COM)[36], suggesting that inflammation is not excessive but a normal response to infection or tissue damage. Alternatively, under-expression of certain inflammatory proteins could suggest an inadequate or imbalanced immune response which could contribute to pathology. A notable example is SERPINB3, an inflammatory serine protease inhibitor which is expressed in cholesteatoma but downregulated compared to chronic otitis media tissue alongside other inflammatory proteinase inhibitors, possibly contributing to excessive proteinase action and tissue damage[36].

Specific microbes

The middle ear is typically closed to external pathogens, bound by the tympanic membrane at one end and the Eustachian tube at the other, so the microbiome of the middle ear is distinct from both the external auditory canal and the adenoid region. The normal microbiota of the middle ear is not well established but typically dominated by Proteobacteria, Actinobacteria, Firmicutes and Bacteroidetes with child and adult ears varying significantly in the relative proportions[43]. Changes to the middle ear microbiome are noted in otitis media: *Streptococcus* *pneumoniae, Haemophilus influenza*e and *Moraxella catarrhalis* have been detected in otitis media (OM) with effusion and tympanic perforation, though *M. catarrhalis* is not common when perforation is absent[44]. Minami *et al.*[43] found no significant difference in the microbial composition of ears with and without cholesteatoma at the phylum level, while suppurative OM differed significantly from child and adult normal. They noted that *Staphylococcus* and *Peptoniphilus* dominated in some suppurative OM ears with *Peptoniphilus* being particularly frequent with cholesteatoma. Decreased relative abundance of some species in cholesteatoma including *Acidovorax*, *Bacillus* and *Masillia* species has also been demonstrated[45].

Human papillomavirus (HPV) is a group of very common human viruses causing abnormal skin growths called papillomas, including warts and pre-cancerous lesions[46]. Some epidermoid cysts have been found to show signs of HPV infection (termed verrucous cysts)[47] and the typical hyperkeratotic lesion caused by the virus somewhat resembles cholesteatoma. However, the reported prevalence of HPV subtypes in cholesteatoma cysts varies greatly: Chao *et al.* (2000)[48] and Franz *et al.* (2007)[49] detected HPV DNA in 1 of 32 and 29 patients respectively corresponding to a prevalence of ~3%. Viana *et al.* (2021)[50] failed to detect HPV or polyomavirus (another group of tumorigenic viruses) in any cholesteatoma samples (n=26). Skoulakis *et al.* (2018)[51] report 48.3% prevalence (n = 62) compared to 0% prevalence amongst controls, while Rydzewski *et al.*[52] detected HPV-6/11 in 70% of cholesteatoma cases (n=9) compared to 23% of non-cholesteatoma granulation tissue (n=29). This may be partly explained by differences in the specificity of HPV subtypes targeted by studies, although Viana *et al.* and Skoulakis *et al.* both targeted a broad spectrum of subtypes.

Variable rates of HPV in cholesteatoma could also reflect the variable rate of background infection although increased prevalence in comparison to a control group[51,52] suggests a true increase in prevalence with cholesteatoma. It is possible that HPV provokes epithelial hyperplasia in the basal layer of the tympanic membrane consistent with hyperplasia theory and the presence of papillary cones in some cholesteatomas[26].

Roles in pathology and aetiology

As cholesteatoma tissue includes an inflamed granulation layer, a degree of ECM breakdown and inflammation is always present. Due to the roles of the ECM and inflammation in pathologically relevant processes such as cellular proliferation, migration and osteoclast activity, it is likely that these features are clinically important. Indeed, several inflammatory pathways have been suggested as potential therapeutic targets in cholesteatoma[53]. However, whether there is underlying dysfunction in these processes leading to cholesteatoma is not known – inflammation, ECM breakdown and osteoclast activity may be downstream of cholesteatoma establishment. The conflicting evidence for a role for HPV in cholesteatoma suggests that viruses may be involved in some cases, but it does not seem a likely causative agent in all cases; there is also little evidence to suggest any role for specific pathogens in pathology and cholesteatoma may or may not be actively infected.

1.1.5 Potential for a genetic role

Familial clustering

Although cholesteatoma is not traditionally considered a heritable condition, there have been several observations of family clustering[54–57], including three reports of congenital cholesteatomas in identical twins[58–60]. Furthermore, an online survey of 857 individuals found family history in 10.4% of cases[61], while family history of cholesteatoma or chronic otitis media was reported in 64% of 12 cases in a Kibbutz of 3056 individuals[62]. A recent study of surgical records in Sweden indicated an increased risk of cholesteatoma amongst first-degree relatives of those already treated with an odds ratio of 3.9[63]. Interestingly, many case reports of cholesteatoma within families are congenital cholesteatomas[55,57–59] despite congenital cholesteatoma being the rarer form. Prinsley (2019)[54] presents 15 families with multiple affected members, though he does not distinguish between congenital and acquired, noting most had tympanic abnormalities more consistent with the acquired form. Additionally, Collins *et al.* (2020)[61] found a positive association between family history and bilateral disease with an odds ratio of 2.15. Therefore, it seems that both congenital and acquired cholesteatoma may have a hereditary aspect.

A potential genetic association via increased risk of cholesteatoma amongst persons with orofacial cleft is also possible: in a Danish study of 8,593 persons with orofacial cleft[64], hazards of cholesteatoma were 20-fold whilst hazards of cholesteatoma in 6,989 siblings of persons with cleft lip/palate were increased 2.1-fold. The authors suggest that this is due to accumulation of sub-clinical muscular defects in these siblings, posing a genetic mechanism for increased susceptibility.

Genetic studies

Very few genetic studies of cholesteatoma have been performed to date. Outside of the Genetics of Cholesteatoma project, only three such studies have been performed. The earliest of these was a deletion in tumour suppressor *APC* identified in a single 6-year-old boy with familial adenomatous polyposis and cholesteatoma, and his mother[65]. Familial polyposis coli is a disease which causes multiple gut polyps and other cystic lesions, usually due to mutations in *APC*. As such, it is possible that this *APC* deletion was also associated with cholesteatoma in this case, but with a sample size of 1, this is a very low level of evidence. James *et al.*[66] identified the connexin gap junction genes *GJB2* and *GJB6* in a study of 98 affected children. Variants of these genes are associated with some forms of congenital deafness and hyperkeratosis[67]. However, the sample size was still small for a genetic study, the variant was not present in all cases, and there was no control population. A more recent study[68] from identified 12 rare, deleterious variants in 1 of 6 whole-exome tested saliva samples from persons with COM with cholesteatoma. Affected genes were *RTN4, RAB5A, CRYBG1, RGS22, APBB1IP, HEPHL1, BHLHE41, ARID3A, C5AR1, SPTLC3, CPT1B* and *FAM227A*. The authors identified disrupted processes primarily related to endoplasmic reticulum function, including endocytosis, protein transport, apoptotic processes, and rhythmic processes. However, all variants were found in one individual: the remaining 5 samples had no qualifying variants according to their criteria.

There have also been several studies of up- and down-regulated non-coding RNAs in cholesteatoma[69–74]. Non-coding RNAs target messenger RNA to regulate the translation of certain genes into proteins. Jovanovic *et al.*, 2022[75] reviewed eight studies of non-coding RNA in cholesteatoma and suggest that dysregulated miRNAs miR-21 and LET-7 are the 'most highlighted'. miR-21 downregulates *PTEN* and *PDCD4*, suppressors of tumour formation and progression, while LET-7 downregulates *HMGA2* which has been hypothesised to balance reduction of *PTEN/PDCD4* through increased apoptosis. This may contribute to the self-sustained growth in cholesteatoma with a lack of malignancy. Other pathways thought to be disrupted by dysregulated non-coding RNAs include EGFR/Akt/NF-κB/cyclinD1 and PI3K/Akt/PKB[75].

Cholesteatoma is typically non-cancerous and although some features are shared with neoplastic lesions (e.g. excessive growth and angiogenesis), malignance is very rare. In a case report of squamous cell carcinoma of the temporal bone arising from cholesteatoma which had been surgically removed 54 years prior, a total of 5 other similar cases were identified since 1951[76]. Amongst these reports, most cases of carcinoma occurred many years after removal of cholesteatoma. The relationship is thought to be due to the impact of chronic inflammation rather than malignancy of the cholesteatoma itself[76]. Dysregulation of some tumour-associated genes has been detected[77,78], however there is very little literature concerning cholesteatoma somatic mutations which would help differentiate it from cancerous lesions. Albino *et al.* (1998)[79] found evidence for aneuploidy in only 1 of 10 cholesteatoma samples, concluding that there is little evidence for genomic instability consistent with malignant neoplasms. However, in a recent analysis of 17 middle ear cholesteatomas, somatic variants in *MYC* and *NOTCH1* were detected in 14 samples and correlated with bone destruction[80]. *MYC* and *NOTCH1* are proto-onco genes, genes involved in normal cellular growth and differentiation; mutation of these genes can contribute to cancer. Overall, cholesteatoma may display some cancer-like properties but shows less genomic instability than normal cancers and should not be considered a malignant nor pre-malignant neoplasm.

1.2 Genetics

1.2.1 The role of Genetics in disease

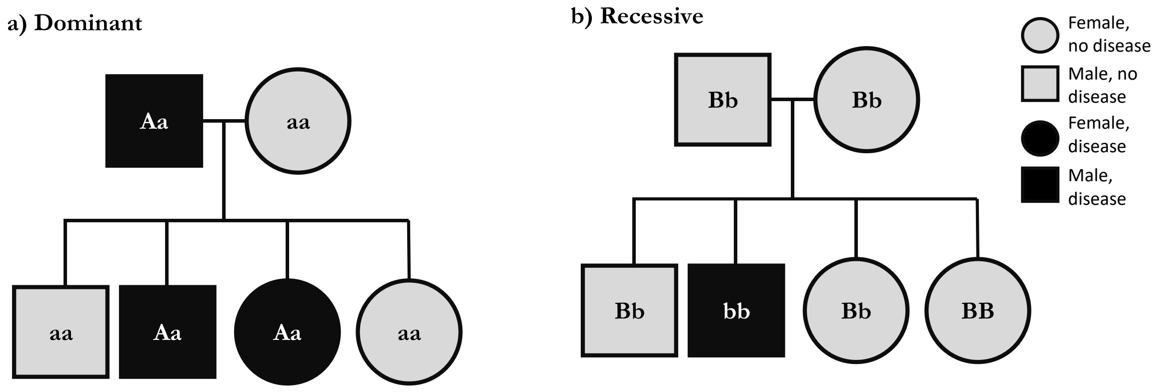
Many diseases have both genetic and environmental factors. The heritability of a trait is defined as the proportion of variance in the phenotype that is explained by genetic variance[81]. A highly heritable trait will have larger genetic contributions than environmental. Not only does heritability vary between diseases, but the number of genetic variants, their contribution to disease risk, the types and locations of the variants can also differ, from simple monogenic disorders where changes to a single gene will always result in disease, to complex, polygenic diseases encompassing many risk variants with small effect. The landscape of genetic variants underlying a disease can be described as its genetic architecture.

Inheritance patterns for monogenic diseases

Some diseases may be caused by deleterious variants in a single gene. A well-known example is cystic fibrosis (CF), which is caused by variants affecting the *CFTR* gene[82]. *CFTR* encodes a transmembrane conductance regulator, a type of ion transporter. Damaging variants in *CFTR* prevent the proper function of the ion transporter, leading to decreased chloride secretion and sodium resorption by epithelial cells. This leads to thickened mucus secretion in all organs, with the lungs severely affected, increasing susceptibility to pulmonary disease. Because human genetic code is diploid, a single working copy of CFTR is adequate to prevent disease: two defective copies must be inherited, making this an autosomal recessive disease (**Figure 5b**).

Figure 5. Dominant and recessive monogenic inheritance.

Examples of family trees with dominant and recessive disease traits. Dominant genotypes are denoted with capital letters, recessive with lowercase. A) Dominant inheritance occurs when the dominant allele (A) is causal. One copy of A is sufficient to cause disease. B) Recessive inheritance occurs when the recessive allele (b) is causal. Two copies are required to cause disease.

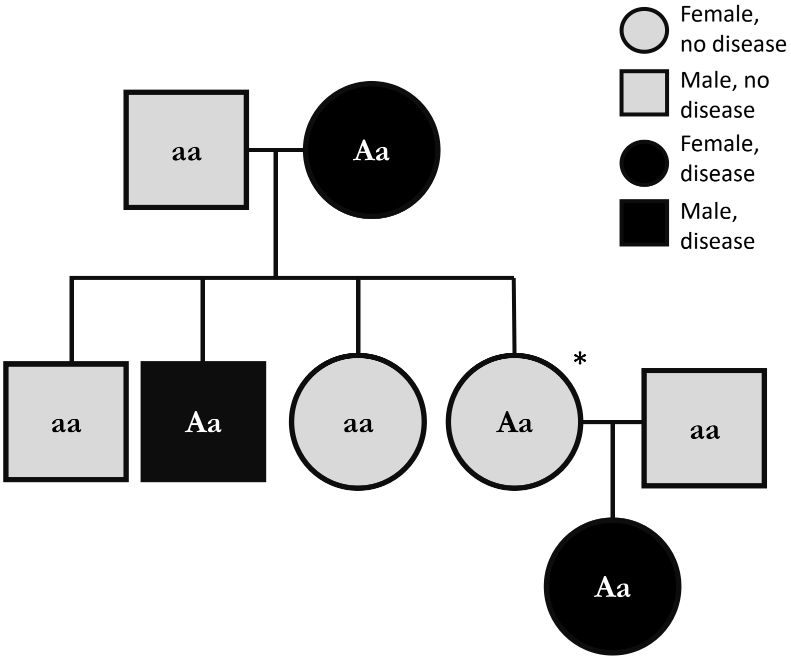


An autosomal dominant disease requires only one defective copy of a causal gene (**Figure 5a**). One example is Huntingdon disease, which occurs when one defective copy of *HTT* is inherited[83]. The pathogenic allele encodes a protein with an expanded repeat region and the presence of this protein causes pathology: therefore, only one copy is required to cause disease. Huntingdon disease and cystic fibrosis are autosomal diseases because the causal genes are not located on the sex chromosomes. A sex-linked disease occurs when the causal variant is located on X or Y. Most examples are X-linked recessive, requiring that males carry a single copy while females must carry two copies of a causal variant.

Although causal variants are often inherited, they may also occur spontaneously in the gametes, resulting in a de-novo mutation in the child. Both 'variant' and 'mutation' refer to a genetic sequence which differs amongst the population, but a variant is more likely be described as a mutation if it is rare or new. I refer to all mutations as variants.

Figure 6. Reduced penetrance of a dominant monogenic trait

*The disease allele (A) is dominant. However, some Aa individuals do not have the disease phenotype. The individual marked with * carries a copy of A but does not have disease; her daughter inherits a copy of A and does have disease. This is a disease with dominant inheritance and incomplete penetrance.*



Even simple monogenic disorders such as cystic fibrosis and Huntingdon disease may be caused by different variants affecting the same causal gene. For example, around 2000 different *CFTR* variants have been identified in cystic fibrosis, although most of these are of uncertain relevance to disease and only about 200-300 may actually be pathogenic[84,85]. Different variants can lead to disease through multiple pathways, from production of functional CFTR which degrades too quickly to failure to synthesise any CFTR at all[82]. Furthermore, diseases may not be fully penetrant. Penetrance refers to the proportion of individuals with the disease genotype who show the phenotype (**Figure 6**). Cystic fibrosis generally has high penetrance with carriers of two defective copies of CFTR typically showing disease. However, this is not true for all monogenic disorders: some individuals with the same variants may show different forms of disease or none at all[86]. There are many reasons why penetrance may be reduced, and the following are some examples: there may be a large environmental component; different causal alleles may have different effects of phenotype and interact differently; there may be age and sex-dependent effects; or there may be undiscovered causal loci, making the disease di-or oligogenic[86,87].

Complex, polygenic diseases

Many common diseases such as type 1 and 2 diabetes, cardiovascular disease, and asthma are thought to have polygenic risk. Rather than a single or small number of genes causing disease, conditions may be associated with many variants across the genome, each of which contribute a small amount to disease risk[88]. High-risk individuals are those carrying many risk alleles, although they may or may not actually develop disease – such individuals are simply at elevated risk compared to the general population. Otosclerosis, a condition affecting the bone of the stapes, is inherited in an autosomal dominant-like pattern but family-based studies have failed to map disease to a single locus[89]. Recent meta-analysis of three biobanks identified 1,452 common variants affecting 27 distinct loci associated with otosclerosis[90]. Meanwhile, over 200 susceptibility loci have been identified for coronary artery disease[91].

1.2.2 Types of genetic variant

There are many different types of genetic variation. The simplest and most common is the single nucleotide polymorphism, or SNP (sometimes called single nucleotide variant, SNV) where only one base pair differs between individuals[92]. An indel is a short insertion or deletion of base pairs into the sequence (**Figure 7**). SNPs and indels are well-studied in disease and most reported trait associations are for these variant types – they are common, relatively easy to detect and well-represented on most genotyping arrays[93].

Figure 7. Common variant types affect a single or few bases.

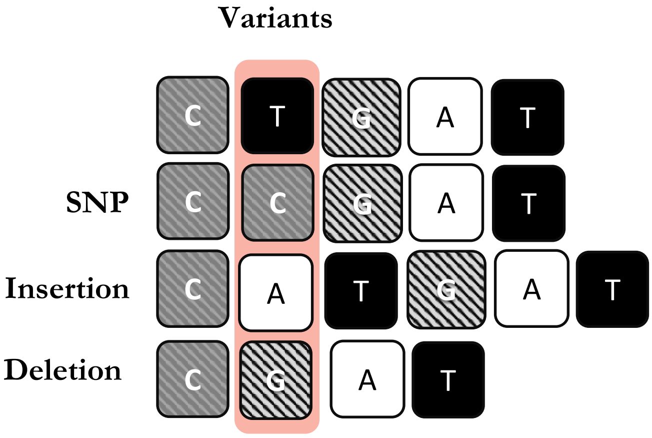
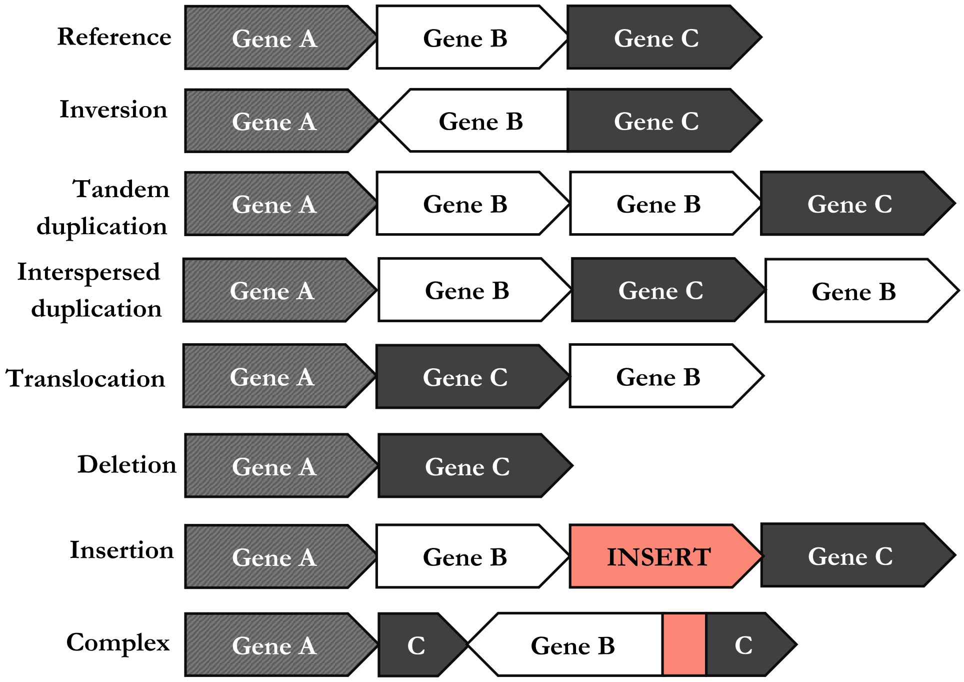


Figure 8. Structural variants affect larger regions of the genomes. Some types of structural variants and illustrative examples are given.



Structural variants are changes to larger stretches of DNA, approximately 1 kilobases or longer[94]. This can include deletion, duplication or inversion of entire genes or gene regions (**Figure 8**). Although they are poorly studied in comparison to SNPs and indels, up to 9.5% of the human genome is associated with copy number variation[95] and there is evidence that they may be associated with diseases such as Crohn's disease, type 1 diabetes and rheumatoid arthritis[96].

Coding variants

For sequences inside protein-coding regions, each three bases constitute a codon, and corresponds to a different amino acid in the final protein. Because some amino acids can be encoded by multiple codons, some changes to the sequence do not alter the corresponding amino acid. These are synonymous variants. When the amino acid sequence is changed, the variant is missense. If a sequence change results in the stop codon, which signals the end of the protein, the final transcript will be truncated, causing a premature stop variant. Indels can cause any of these variant types as well as a frameshift variant, where the number of bases inserted or deleted results in all following codons being shifted by one or two places, changing the entire following sequence. Any variant predicted to cause the protein to stop functioning is a loss-of-function variant. Conversely, gain-of-function variants are changes thought to add a new functionality to the final protein.

Non-coding variants

98% of the genome is non-coding, meaning it does not translate into a protein product. About a quarter of non-coding DNA lies in the introns (non-coding regions within genes) and the rest is intergenic[97]. Non-coding DNA has many functions, including regulatory elements, non-coding RNAs and ribosomal RNAs. However, there are large portions of DNA with no apparent function or whose function is unknown, including pseudogenes (non-functional remnants of genes), transposable elements and repeats, the latter being mostly found in the telomeres and centromeres[95].

As only ~2% of the genome codes for proteins, it may not be surprising that >90% of trait-associated variants discovered by GWAS to date are non-coding[98,99]. However, their roles in disease are poorly described in comparison to coding variants, given the obscure function of many forms of non-coding DNA. Changes to regulatory elements in non-coding regions may alter gene expression and contribute to disease, although which genes are associated with regulatory elements is not always known. In an analysis of 920 publications covering 6,011 trait-associated SNPs, Maurano *et al.* (2012)[99] found that non-coding SNPs were enriched in regulatory elements, with most SNPs either affecting regulatory regions or in linkage disequilibrium with them.

Linkage disequilibrium, the correlation of variants located physically close to each other, means that non-coding variants may also tag causal variants in coding regions if the two are linked. In a study of 21 common disease traits, Yong *et al.* (2020)[100] created predictors which combined the effect of many SNPs across the genome. In their study, 50-60% of predictor SNPs were genic, and these SNPs explained 40-90% of predictor variance. This suggests that genic variants may have a greater relative contribution to phenotype than non-coding, but this is very variable depending on phenotype.

1.2.3 Genotyping data

Genotyping

There are two main approaches to genomic analysis: genotyping or microarray testing and next generation sequencing. Genotyping arrays consist of a number of probes (small stretches of DNA covering a region where there is known variation) and can only measure the presence or absence of a particular variant at the probed sites. Hundreds of thousands to a few million probes are typically included, mostly comprised of SNPs and short indels, although some arrays also detect some copy number variants[93]. Although arrays can only directly measure variants represented by their probe sets, many millions of additional variants can be imputed from high-coverage reference databases, including very rare variants with minor allele frequency < 0.5%[93]. However, imputation is an estimate and not likely to be completely accurate, and the rarest variants may be absent or poorly represented in reference databases and therefore difficult or impossible to impute.

Next generation sequencing

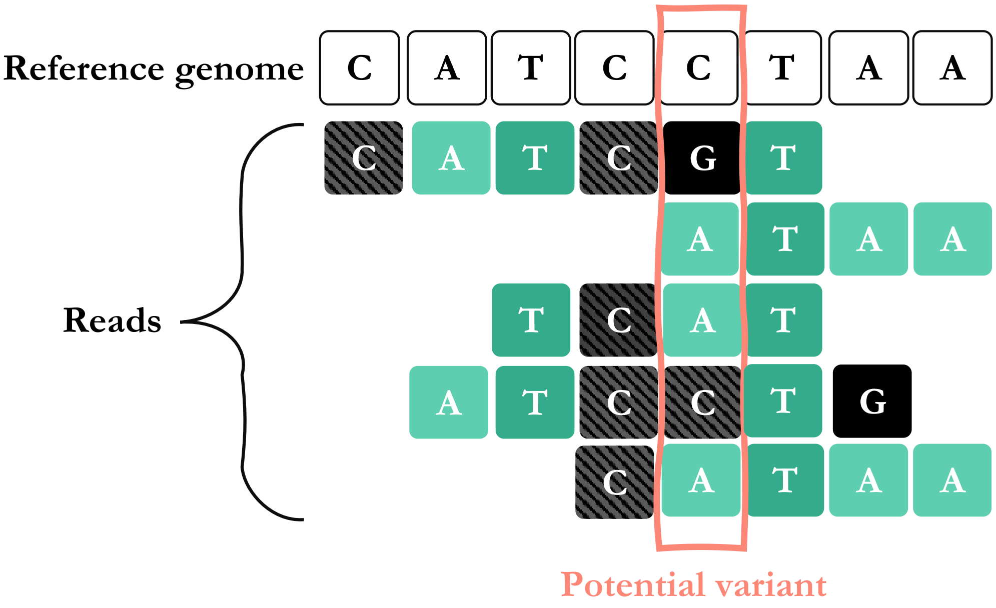
Next generation sequencing (NGS) involves directly measuring all base pairs in the genome or a subset of it. In the approach used by Illumina and similar technologies, this is achieved by fragmenting DNA into short pieces and annealing them to flow cells. Complementary DNA is synthesised using the polymerase chain reaction (PCR) with fluorescently tagged bases. These emit a flash of light as they are added to the growing complementary strand. A computer records the flashes which are coloured according to their nucleotide base and in this way reads the DNA sequence in each fragment. This process is used by Illumina

The result of this process is a large number of short, overlapping segments of sequence (reads) which must be reassembled into a full genome. This is achieved by aligning them to a reference genome. The most recent reference genome assembly from the Genome Reference Consortium is GRCH38 and along with its predecessor GRCH37 form the standard for human genetics reference genomes. These reference genomes are not an average, nor a representation of a 'disease-free' genome, but were drawn from a small number of donors: 93% of GRCh38 is drawn from approximately 11 individuals, with one male contributing to 70% of the primary assembly[101]. However, these reference genomes include large gaps in hard-to-sequence regions such as the telomeres, and representation of large regions of variability is poor[102]. For these reasons, new graph-based genomes representing all regions and multiple alternative haplotypes are being developed by the Telomere-to-Telomere consortium[102] and Human Pangenome Consortium[103]. However, these are not yet in wide use and GRCH38 remains the standard.

Variant calling for NGS data

NGS always generates error and determining true variation from such errors is a non-trivial task. As reads are overlapping, most sites will be represented by multiple reads (the number being the read depth of that site, see **Figure 9**). A simple way to determine if a variant is real is to check the proportion of reads agreeing on the variant at the site. Quality metrics such as the site's read depth may also be used to decide whether to call a variant, where poor quality sites are assumed not to be true variants. Methods such as FreeBayes[104] and GATK[105] use additional information from reference databases to determine the likelihood of variants being true based on the genetic background in which they usually appear. Because variant calling in NGS involves deciding whether there is sufficient evidence that the read sequence differs from the reference, sites with low read depth or large amounts of error may be considered to agree with the reference genome.

Figure 9. Alignment of reads to a reference genome can identify possible variants. *Reads, short fragments of sequence determined from sample DNA, are aligned to a reference genome to reconstruct the original sequence. Where the reads vary from the reference genome, there may be a variant. However, not all reads may agree on the sequence at this site. The number of reads representing a given site is the read depth and generally a higher read depth means a greater certainty about the sequence at that site.*



1.2.4 Trio analysis and family studies

Causal variants for simple, monogenic disorders can often be determined through family studies or trio analysis. Examining the affected individual alone is usually insufficient to determine which genetic variants are likely to be causal as a single individual carries many variants, of which a proportion are predicted to be deleterious – with no obvious ill effects for their carrier[86]. In a trio study, the genetics of an affected child and their parents are examined to determine the causal gene. For example, if disease acts in a dominant fashion but is due to a *de novo* mutation, the causal variant will be present in a child but not in their parents. If disease is recessive, the affected child will be homozygous for variants affecting a gene for which their parents are heterozygous[106].

Additional family members may also be used: if disease is autosomal dominant, variants must be present only in the affected family members and absent in unaffected family members (variants co-segregate with disease). As large swathes of the genome are shared between family members, many variants can quickly be excluded.

Family or trio studies may primarily be used to identify candidate genes when an individual has a rare disorder whose cause is not known[106]. Other families showing the same disease may also carry different risk variants. This is both a benefit and a drawback of family studies: it is beneficial as it removes genetic heterogeneity and allows easier identification of causal alleles, but different families may carry different causal variants.

These types of studies work best when disease is based on few genes. Some digenic and pseudo-digenic disorders have been identified through family studies, where variants in two different genes are required for disease or where a second gene modifies disease appearance and behaviour[107]. When disease risk is raised by small or modest contributions from many genes (i.e. It is highly polygenic), family studies may fail to identify risk variants as large sample sizes are needed to detect these small effects.

1.2.5 Genome-wide association studies

The goal of the genome-wide association study (GWAS) is to test variants across the genome for statistical associations with a phenotype. GWAS is typically applied to common, complex diseases whose genetics cannot be easily established by family studies and whose case numbers are sufficient to provide statistical power for discovery. GWAS may be used to identify causal variants to explain the mechanisms of disease and indicate possible therapeutic targets. There has also been increasing interest in using genome-wide data for diagnosis and screening through the use of polygenic risk scores which sum the effects of a large number of variants across the genome[108,109].

The first GWAS was performed in 2005, and identified complement factor H as a risk locus for age-related macular degeneration, highlighting an immunological role in the disease[110]. Since then, over 5,000 GWAS have been performed and catalogued (EBI GWAS Catalog), identifying over 300,000 variant-trait associations. In most GWAS, each variant is tested individually for an association with the outcome with a linear (for continuous traits) or logistic (for binary traits) regression in the following general form:

$$\boldsymbol{Y} \sim \boldsymbol{W\alpha} + \boldsymbol{X\beta} + \boldsymbol{g} + \boldsymbol{e}$$

Uffelmann et al. (2021)[88]

Where $\boldsymbol{Y}$ is the phenotype, $\boldsymbol{W}$ a matrix of covariates, $\boldsymbol{X}$ a vector of genotypes of all samples at a given locus, and $\boldsymbol{g}$ and $\boldsymbol{e}$ are error terms. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are calculated when fitting the regression and tells us the effect sizes of the covariates and genotypes[88]. This test is repeated across the genome, allowing calculation of effect sizes and *p*-values for all variants and identification of those most likely to be causal or contribute to risk. This makes GWAS suitable for discovery of causal variants where genetics is not yet known or expected to be complex.

It is common to reduce the number of variants considered in a GWAS by filtering for impact and frequency. This reduces the number of variables being tested, therefore should reduce noise. Often, only variants with a minor allele frequency of at least 1-10%, depending on the size of the study population, are included to ensure sufficient sample size for statistical power[88]. Rare variants may require very large sample sizes or techniques to aggregate them at the gene level to provide sufficient power. This may be- in the form of a burden test, which calculates the sum of variants affecting a given gene, which works well when all variants in a gene have similar effects on phenotype[111]. Methods such as SKAT perform regressions for variants within each gene, allowing for different variants to have differing effect sizes[112]. The result is again a gene-level test but can account for differing effect sizes and directions.

Because many statistical tests are being performed across the entire genome, false discovery is a risk and the *p*-value threshold for significance must be adjusted appropriately. A standard method is Bonferroni correction, which reduces the threshold *p*-value by the number of tests performed; as the human genome contains approximately 1 million independent variants, a *p*-value of $5 \text{x} 10^{-2}$ is adjusted to $5 \text{x} 10^{-8}$ (depending on factors such as the population size and the number of independent variants analysed)[88]. Variants are generally not independent from one another due to linkage disequilibrium, meaning variants located physically close to one another are often correlated. Rare variants are unlikely to be in linkage disequilibrium with common variants and so may represent a greater number of independent tests[88].

Population stratification

An important confounder in genetic studies is population stratification, which refers to differences in gene frequencies between populations. Stratification occurs on both a global scale, where gene frequencies differ between ancestries, and within populations reflecting historical human migration patterns, genetic drift and non-random mating[113]. This can result in false associations if the trait also varies in prevalence across the study population. Causal variants may also differ in frequency between populations, leading to the same trait being associated with different variants in different populations[113]. Population stratification is often controlled by taking the first genetic principal components and using these as covariates in the association test[88,114]. Modern GWAS software such as REGENIE[115] and SAIGE[116] typically perform multiple steps to correct for population stratification as well as sample imbalance and other sources of error. An initial step fits a null model which captures the genetic background of the population and is used in the second step to correct the *p*-values accordingly (see *Modern GWAS methods*).

1.2.6 Polygenic risk scores

Genetic risk for complex diseases may comprise a small number of variants with high impact on outcome as well as a large number of variants with small contributions to disease risk. For common diseases, the polygenic component may outweigh the contributions of rare, highly penetrant mutations[117]. Researchers have sought to quantify this risk through the creation of polygenic risk scores (PRS). PRS combine the effects of variants across the genome by summing them with weights according to their effect size. PRS have been calculated for several common, complex conditions such as coronary artery disease, type 2 diabetes and inflammatory bowel disease[118]. Even where single or small sets of genes are known to explain a large proportion of genetic risk, for example *BRCA1/2* and breast cancer, PRS have been found to classify a similar number of high-risk individuals on the population level, even if the individual effect is modest[118]. While some commercial companies offer polygenic testing for diseases like breast cancer, PRS are not widely used in any clinical setting. One exception is the ongoing HEART study[119,120], which combines PRS with clinical data to identify persons with elevated risk of a cardiac event. Observational data from the UK BioBank suggests that PRS alone may be as good a predictor as known clinical risk factors and in combination led to reclassification of 13.7% of individuals between high and low risk categories[121]. Although such tests are not yet widespread in clinical settings, they demonstrate the ability of polygenic risk scores to quantify disease risk due to polygenic effects.

Besides simple summation of weighted variants, machine learning approaches to disease classification can also describe disease risk with more complex genetic architecture. There are many approaches to classification, including support vector machines, Bayesian classifiers and neural networks[122]. A common approach in bioinformatics is Random Forest, where many decision trees are generated using random resampling of the data[123] (see *Random forests*). Importantly, these methods use genetic data in combination rather than testing genes or variants separately. This makes them suitable when disease risk is based on a non-linear combination of risk variants.

1.3 Study Aims and Wider Project

1.3.1 The Genetics of Cholesteatoma Project

Although cholesteatoma is not traditionally considered a heritable condition, recent observations of family clustering and family history in ~10 of cases[61] support a significant genetic component in some individuals. Observations of such family clustering in East Anglia[54] led to the establishment of the Genetics of Cholesteatoma project (GoC) with the aim of studying cholesteatoma heritability.

Two genetic studies have been performed by the GoC to date: Prinsley *et al.* (2019)[124] identified rare loss of function variants of *EGFL8* and *BTNL9* in a multiply-affected family. EGLF8 is an epidermal growth factor-like protein and the variant identified by Prinsley *et al.* is associated with the hyperproliferative inflammatory skin condition psoriasis. This is particularly interesting as several gene expression studies have identified upregulated psoriatic proteins: this includes the *S100A* family of inflammatory proteins, *PI3*, *DEFB4* and *SERPINB4*[125]. According to Macias *et al.* (2013)[125], only two of fourteen psoriatic proteins have not been shown to be dysregulated in cholesteatoma. *BTNL9* (butyrophilin-like protein 9) has signalling receptor binding activity and is involved in T-cell signalling and cytokine production. As inflammation may be intimately involved with cholesteatoma progression and pathology, changes to immune genes could contribute to disease risk. Several additional genes were detected with missense variants of predicted moderate impact on protein function; however, as only one family was considered, it is difficult to draw conclusions that can be generalised. As neither of the variants detected were rare enough to be the sole causes of cholesteatoma, their involvement (if any) may be polygenic[124].

The recent GoC follow-up study[126] of 21 individuals from 10 affected families identified additional candidate variants in *NEB*, *DNAH7*, *DENND2C*, *NBEAL1*, *PRRC2C* and *SHC2*. Gene ontology analysis in this study identified enriched deleterious variants in several pathways: calcium binding, microtubule function (primarily due to *DNAH* and *KIF* family variants), and extracellular matrix organisation. Deleterious variants affecting ECM proteins could result in improper ECM formation and aberrant downstream processes, consistent with cholesteatoma gene expression (see *Semi-systematic review of global gene expression studies*). Multiple families carried variants in *DNAH7*, one of a family of dynein axonemal heavy chain proteins involved in force generation on the cytoskeleton and ciliary/flagellar motility. Ciliary function may be important in cholesteatoma as it is involved in clearance of mucous and debris from the ear; impairment may lead to increased infection and possibly a failure of the tympanic membrane's self-cleaning mechanism. Variants in *DNAH1* and *DNAH5* are associated with primary ciliary dyskinesia, which increases susceptibility to chronic otitis media and cholesteatoma[17]. Additional families carried variants in other DNAH family members and in some KIF family members – *KIF* genes encoding kinesins, force generating proteins on the cytoskeleton acting in the opposite direction. Interestingly, this study did not implicate immune function, suggesting that upregulated immune genes in cholesteatoma tissue are downstream of any possible genetic causes.

As studies so far do not agree on a single set of genes or variants, penetrance is low, and risk factors are diverse, it is likely that cholesteatoma genetics are heterogeneous and complex. There is some evidence that cholesteatoma may be associated with defects in the inflammatory response, as it may be a direct sequalae of chronic ear inflammation, although the precise relationship between cholesteatoma and otitis media is not known. *BTNL9*, identified in the pilot study, had immune function[124], although the second GoC study of a larger number of individuals did not identify this as an enriched process[126]. Meanwhile, gene expression studies point to breakdown of the ECM, while our whole exome study suggest a role for ciliary function. Genetically mediated risk factors are also known via chromosomal disorders which affect craniofacial morphology[18]. With many potential pathways to disease, it is possible that different factors contribute different amounts to disease risk between affected individuals and families, making genetic analysis challenging and calling for larger genetic studies than have previously been performed to increase statistical power.

1.3.2 Aims and objectives

My PhD project will perform the first genome wide association study of cholesteatoma (excluding generic PheWAS which include cholesteatoma but do not report specifically on the disease). This is to identify variants, genes or pathways associated with cholesteatoma and to further investigate its genetic architecture. While previous studies in the GoC project have identified several candidate genes with diverse functions, these along with complex environmental risk factors, point to a complex, heterogeneous disease. Therefore, large, controlled genetic study is required. This project, which uses ~1,000 cases from UK BioBank (UKBB) data, will therefore provide much-needed insight into the mechanisms of disease which may lead to non-surgical preventative treatments or aid in targeted monitoring of at-risk individuals.

The study's aims are as follows:

- To review current knowledge on molecular biology of cholesteatoma through systematic review of global gene expression studies.
- To explore cholesteatoma epidemiology in the UK BioBank to identify important demographic factors and comorbid disease.
- To identify genetic variants associated with cholesteatoma through genome-wide association study of whole exome data.
- To identify affected processes and pathways through gene set enrichment analysis.
- To use machine learning to classify cases and non-cases to explore genetic architecture and determine whether disease risk can be predicted from these genes.
- To further understanding of the genetic mechanisms underpinning cholesteatoma risk and inform future study directions in cholesteatoma treatment and monitoring.

2 Semi-systematic review of global gene expression studies

2.1 Background

2.1.1 Rationale

Differential gene expression analysis is a method used to identify differentially expressed genes (DEGs) between two or more sample sets. A range of methods are used to analyse the RNA or protein products in pathological specimens, comparing them to an appropriate healthy tissue to identify over- and under-expressed genes[127]. Differential gene expression analysis has been applied to cholesteatoma with the aim of understanding its pathological features, namely its aggressive, invasive growth and bone erosion. Changes in gene expression associated with disease may drive pathology or be a consequence of it; either way, these changes may provide insight into disease biology.

Candidate gene-based approaches have investigated the expression of various genes including interleukins and matrix metalloproteinases (MMPs). Selection of targets in such studies may be based on knowledge of biology of similar diseases: an early study[128] investigating MMPs in cholesteatoma followed the discovery of a role for the protein family in other osteolytic diseases such as osteoarthritis. Likewise, study of RANKL in cholesteatoma follows knowledge of its role in osteoclast activity[129]. Fewer studies have taken an approach whereby a large number of genes are tested unselectively to identify those with the greatest differential expression. These studies may reveal previously unexplored genes with important roles in cholesteatoma biology. However, individual studies often have small sample sizes, meaning results may not be generalisable. Interpretation of the large number of differentially expressed genes detected within studies is often based on prior knowledge of cholesteatoma pathology, including gene expression studies of candidate genes, which may be both subject to publication bias and biased towards well-studied gene families. By performing a systematic review of global gene expression studies, the most consistently dysregulated genes can be determined and candidates which have yet to be studied may be identified.

2.1.2 Aims and objectives

In this chapter, I reviewed nine papers where significant results for all genes tested were reported in order to determine which genes are consistently up- and down-regulated in cholesteatoma. This review was conducted in accordance with the PRISMA 2020 systematic review guidelines[130]. As I did not have a second researcher to aid in literature search, this review is semi-systematic rather than systematic. I then performed gene ontology analysis on the set of genes detected in two or more papers as well as up- and down-regulated genes compared to skin and mucosa to build a detailed profile of transcriptional changes characteristic of cholesteatoma. Expression studies often have small sample sizes and differ in their approaches, including choice of control tissue, which is of particular importance as it is unclear which tissue represents a healthy analogue to cholesteatoma due to its uncertain aetiology[1]. Comparing the changes in expression with different tissue controls can further enhance our understanding of cholesteatoma biology. Identification of any consistently up- or down-regulated genes across tissue comparisons may identify new genes to investigate as therapeutic targets or biomarkers for invasiveness and will complement the results of genetic investigations.

The following questions were addressed in this review:

- Which genes are consistently dysregulated across global gene expression studies?
- Which pathways and processes are enriched amongst consistently dysregulated genes?
- Which pathways and processes are enriched amongst up- and downregulated genes compared to skin and mucosa, and do these differ?

2.2 Methods

2.2.1 Data Collection

Search strategy and eligibility criteria

I performed a literature search on web of science for the term **cholesteatoma AND (expression OR regulat*) AND (protein OR gene)** using the web of science core collection and MEDLINE databases.

To be included in this review, a study must meet the following criteria:

- Articles must be in the English language.
- Studies must test gene expression of human middle ear cholesteatoma tissue compared to an appropriate control tissue. Genes tested must be associated with a specific protein product (i.e., not non-coding RNAs).
- Studies must contain a global gene expression element: I define this as testing a large number of genes across the genome in order to screen for expression differences without applying any presupposed knowledge of cholesteatoma biology to select targets. Articles which were clearly testing only a specific set of gene or proteins based on title or abstract were excluded.
- Studies must at least present a table of all significantly differentially expressed genes or proteins in the text body or as supplemental material.

Studies may use any of several approaches to measuring differences in gene expression, e.g., by targeting RNA or performing proteomic analysis. I placed no early limit on date of studies, and the literature search was performed in June 2024. Abstract screening was performed to eliminate papers that did not test human cholesteatoma tissue or were not gene expression studies, and abstracts that specified that gene expression testing was performed for a specific set of genes.

Data extraction

I downloaded global gene expression results for papers where full data were provided as supplementary tables. For papers where results tables were given in text, I copied results tables into an excel file. I noted the comparison tissue, number of participants, type of study and any genotype arrays or other detection methods used. I assess risk of bias using a modified version of the Newcastle-Ottawa scale[131], considering three main areas: selection of participants, comparability of case and control tissues, and data analysis (replacing the ascertainment of outcome in the Newcastle-Ottawa scale of case-control studies).

2.2.2 Data synthesis

For the eight papers where a table of significantly differentially expressed genes was given, I extracted genes and fold-changes meeting each paper's significance criteria. Britze *et al.* present two levels of stringency for reporting, termed group A (identified in 3 of 3 tests, mean fold change >= 2 times the standard error of the mean, p-value < 0.05) and group B (identified in 2 of 3 tests, individual fold change >= 2 times the standard error of the mean). I used both group A and B genes. Where possible, I converted all fold-changes to log2 fold changes and counted the number of papers and tissue comparisons for which each gene was differentially expressed. Log2 fold changes could not be calculated for Tokuriki *et al.* as fold change was not reported. I used the National Center for Biotechnology Information (NCBI) human genome release 38 version p14 (GRCh38) information file to convert gene symbols to their authoritative symbol for consistent naming across papers. 19 genes had ambiguous names; the symbol given was a synonym for one or more genes and it was not possible to determine which was the correct gene. Of these, 5 may have been present in 2-3 papers, depending on the actual identity of genes: *LOR/LORICRIN/LOXL2* and *BCL5/BCL6* may have been present in 3 papers each. The 19 ambiguous genes were excluded from further analyses.

Processing of Shimizu raw data

Raw data included gene barcode counts per cell for 3 individuals (cholesteatoma and skin samples were taken from each). I compared cholesteatoma and non-cholesteatoma gene counts across all cells for each individual by Wilcoxon rank sum test, (Mann-Whitney U test). The rank sum test is a non-parametric test for comparing distributions which is accurate even when data are not normally distributed. I calculated fold change by comparing the Poisson mean for case versus control tissue for each sample, then acquired the mean ratio across all three. I also fit a generalised linear model with logit link and Poisson distribution, using case/control status as the outcome and the participant of origin as a covariate. Analyses were performed in MATLAB 2023b[132] using the *ranksum*, *poissfit* and *fitglm* functions.

I retained genes meeting the following criteria:

- All samples showed a fold-change in the same direction between case and control and at least 2 of the distribution comparisons were significant according to Wilcoxon rank sum test (Bonferroni adjusted for the total number of genes tested (n=29213).
- The gene was also significant according to generalized linear regression (adjusted for the total number of genes tested). This test was not used alone as many of the genes were poorly conditioned and returned warning messages from the *fitglm* function.

g:Profiler gene set analysis

To identify the common disrupted processes in cholesteatoma, I performed gene set enrichment analysis (GSEA) with g:Profiler. Such analyses identify which pathways are overrepresented amongst the most significant results, aggregating the effects of genes involved in the same biological processes. G:Profiler performs a hypergeometric test to determine enriched pathways and corrects p-values using g:SCS, a method taking into account the hierarchical nature of GO terms and defined in Reimand *et al.* (2007)[133].

I tested for functional enrichment using g:Profiler for the following:

- All genes detected in 3 or more papers regardless of comparison tissue or direction.
- Genes detected in 2 or more papers with a consensus direction of expression *for the same comparison tissue*. The number of times a gene was up-regulated or down regulated a within a tissue were summed, with up-regulation counted as +1 and down-regulation as -1, so genes must have acquired >2 or a <-2 score to be counted. Because only skin and mucosa were tested multiple times, tests were performed for genes up- and down-regulated compared to mucosa and skin.

Genes were analysed by g:Profiler as an unordered query using the Gene Ontology (GO) cellular compartment (CC), molecular function (MF), and biological process (BP) databases as well as the human phenotype ontology database. I used the g:Profiler web service available at https://biit.cs.ut.ee/gprofiler/. The g:Profiler version released on 13-02-2024 (reference genomes: Ensembl 111, Ensembl Genomes 57. GO release: 2024-01-17[*]) was used.

GO terms are hierarchical, so enrichment in a given process is likely to result in enrichment of parent or child processes which can result in a large number of pathways being returned. To address this problem, g:Profiler highlights terms which drive significance in order to identify the most biologically relevant and I report highlighted terms only. More detail is given at g:Profiler – a web server for functional enrichment analysis and conversions of gene lists (ut.ee).

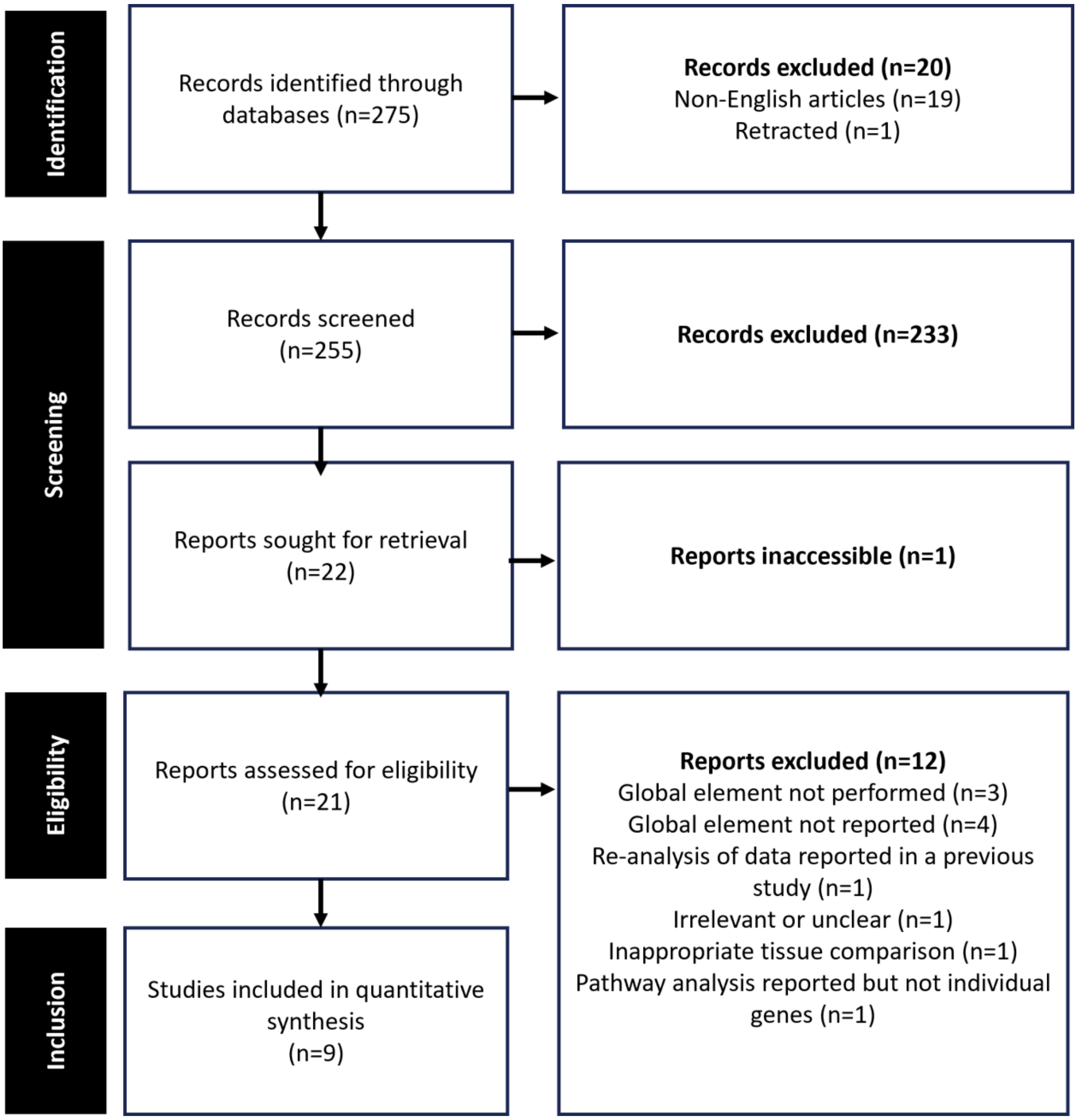
[*] See https://github.com/geneontology/go-announcements/issues/665 for details

2.3 Results

2.3.1 Selection of reports for inclusion in quantitative and qualitative analysis

275 records were retrieved with no duplicates (**Figure 10**). Non-English language articles were excluded (n=19; abstracts also indicated no global gene expression component) and one retracted paper was also excluded. 233 records were excluded during screening for testing specific gene targets, testing non-human or non-cholesteatoma tissues, or being otherwise irrelevant. Of 22 reports selected for retrieval, 21 were available. Eight were excluded for either not containing any global gene expression component (n=3), performing a global element but only reporting a subset of validated proteins (n=4), or the global element was unclear or not reported in full (n=2). One was excluded for using the same data as a prior study. 2 additional papers were excluded from quantitative analysis but contain useful information for qualitative comparison: Yoshikawa *et al.* (2006)[134] compared cholesteatoma and skin fibroblasts before and after exposure to interleukin but not to each other, while Zeng *et al.* (2024)[135] do not report the results of the global element of their study but do report pathway enrichment analysis of differentially expressed genes.

Figure 10. Identification of studies via database searches


Identification
Records identified through databases (n=275)
Records excluded (n=20)
Non-English articles (n=19)
Retracted (n=1)
Screening
Records screened (n=255)
Records excluded (n=233)
Reports sought for retrieval (n=22)
Reports inaccessible (n=1)
Eligibility
Reports assessed for eligibility (n=21)
Reports excluded (n=12)
Global element not performed (n=3)
Global element not reported (n=4)
Re-analysis of data reported in a previous study (n=1)
Irrelevant or unclear (n=1)
Inappropriate tissue comparison (n=1)
Pathway analysis reported but not individual genes (n=1)
Inclusion
Studies included in quantitative synthesis (n=9)

Study characteristics

The reports included for quantitative analysis used various methods including gel electrophoresis, western blot, mass spectrometry and RNAseq to test a wide array of genes for differential expression without first narrowing to a particular set of genes of interest (**Table 2**). Studies varied in number of participants between 3 and 21 and in the type of control tissue. Control tissues include retroauricular skin, external auditory canal skin, middle ear mucosa, tympanic membrane, non-cholesteatoma chronic otitis media (COM) granulation tissue, and congenital cholesteatoma. The total number of participants across all papers was 75, with cholesteatoma samples taken from each and a total of 131 control samples taken across all participants (67 skin, 20 mucosa; other tissues examined in one paper only). Eight papers reported lists of the significantly differentially expressed genes, while the remaining paper reported the raw data from which differentially expressed genes could be calculated. Raw data was only available for four papers.

Table 2. Summary of global differential gene expression/transcriptomic papers. *Analyses aside from Jovanovic et al. were paired, drawing cholesteatoma and control tissue from the same patient. Where control tissues differed in sample size to number of participants, tissue N is given. The two studies that included a global gene expression component not suitable for quantitative analysis are marked with a dagger (†). Differentilly expressed gene (DEG) number calculated during synthesis marked with an asterisk (*)*

Paper	N participants	Control tissue (n where different)	Method		N genes tested	DEGs
Tokuriki *et al*. (2003)[136]	8	Retroauricular skin	Microarray	Atlas 1.2 array	1176	18
Klenke *et al*. (2012)[77]	17	External auditory canal skin	Microarray	Whole Human Genome (4x44) Oligo Microarray,	~41,000 (31,000 genes)	1,145
Macias *et al*. (2013)[125]	13	External auditory canal skin	Microarray	3D-Gene Human Oligo chip 25k	24,267	282
Jovanovic *et al*. (2020)[36]	2	COM granulation tissue (n=4)	Microarray	Illumina iScan HumanHT-12 v4 Expression BeadChip	47,231	169
Britze *et al*. (2014)[78]	9	Tympanic membrane, External auditory canal skin, Neck of cholesteatoma, Middle ear mucosa	Proteomic	NanoLC-MS/MS MaxQuant (version 1.2.2.5) Andromeda search engine	20,255 protein sequences	295
Randall *et al*. (2015) [137]	12	Middle ear mucosa (n=8) Retroauricular skin (n=9) Bone (n=8)	Proteomic	NanoLC-MS/MS Proteome Discoverer 1.4	540,261 sequences	58
Gao *et al*. (2023)[138]	8	Auditory canal skin	Proteomic	MS/MS with Proteome Discoverer	20,395 (including contaminants)	923
Shimizu *et al*. (2023)[139]	3	Retroauricular skin	scRNAseq	Illumina NovaSeq 6000 platform BD Rhapsody Analysis Pipeline		893*
Baschal *et al*. (2019)[140]	3	Middle ear mucosa (N=4)	RNAseq			1,806
Yoshikawa et al. (2006)[134 †]	6	Non-IL stimulated cholesteatoma fibroblasts	Microarray	human genome U133A probe array (GeneChip),		
Zeng et al. (2024)[135†]	3	Skin	RNAseq	FeatureCounts Deseq2		

Certainty of evidence and risk of bias

Per-paper bias assessment with modified Newcastle-Ottawa scale is given in supplementary information (**SI Table 1**).

Selection: In all studies, case tissue was well-defined due to the obvious and recognisable nature of cholesteatoma. In all cases, participants were drawn from those undergoing surgery for middle ear disease, so may represent the more extreme end of disease severity. Most papers specified that acquired cholesteatoma tissue was taken but only Tokuriki *et al.* and Britze *et al.* reported age and sex, and Klenke *et al.* reported age. This may make results difficult to generalise and any differences based on sex, age or cholesteatoma type cannot be determined. Control tissue was generally taken from the same individual at the same time. For tissues in close proximity to cholesteatoma (e.g. mucosa, tympanic membrane), presence of disease may alter gene expression so these may not represent healthy control tissues.

Appropriateness of controls: Aside from Jovanovic *et al.*, all studies used paired control tissues drawn from the same individual, which should control for confounders such as age and sex of participants, as well as any lesser confounders. Jovanovic *et al.* did not provide age/sex of case and control participants, so it is unclear if any matching was performed to consider important confounders.

Bias in analytical methods: Papers using genotyping arrays are only measuring a specific set of human genes which will tend to be biased towards well-studied genes. This was most notable in Tokuriki *et al.* where only 1,176 genes were tested by their microarray. Proteomic studies are similarly biased towards previously measured protein transcripts as all use UniProt databases to identify expressed proteins from mass spectrometry images and are less sensitive to small abundances than RNAseq. All studies were at low risk of reporting bias due to the hypothesis-free nature of global gene expression approaches. It is unlikely that a study will find no DEGs, though this is more likely to affect studies using small arrays. There may be some publication bias if a study does not detect any previously studied DEGs or has a specific hypothesis that a given pathway will be enriched amongst their global gene expression results but it is not.

2.3.2 Consistently dysregulated genes

Across all papers, a total of 3,747 differentially expressed genes were reported (1,090 upregulated compared to skin; 1,113 downregulated compared to skin), of which 624 were reported in more than 1 paper (**Table 3**). No differentially expressed genes (DEGs) were detected in all 9 papers, which may be due to differences in approach, including differences in the number of gene tested, tissue comparisons made, reporting criteria, small sample sizes, and variability of biological samples. However, 20 DEGs were detected in 4 or more papers: *SERPINB3* was detected in 7; *S100A7*, *S100A9* and *S100A8* in 6; *COCH*, *SERPINB4*, *CEACAM6* and *PI3* in 5; *KRT8*, *KRT7*, *TNXB*, *BLMH*, *CTSC*, *SLPI*, *SPRR1B*, *LCN2*, *BPIFA1*, *SERPINB7*, *TACSTD2* and *MMP9* in 4 (**Figure 11**).

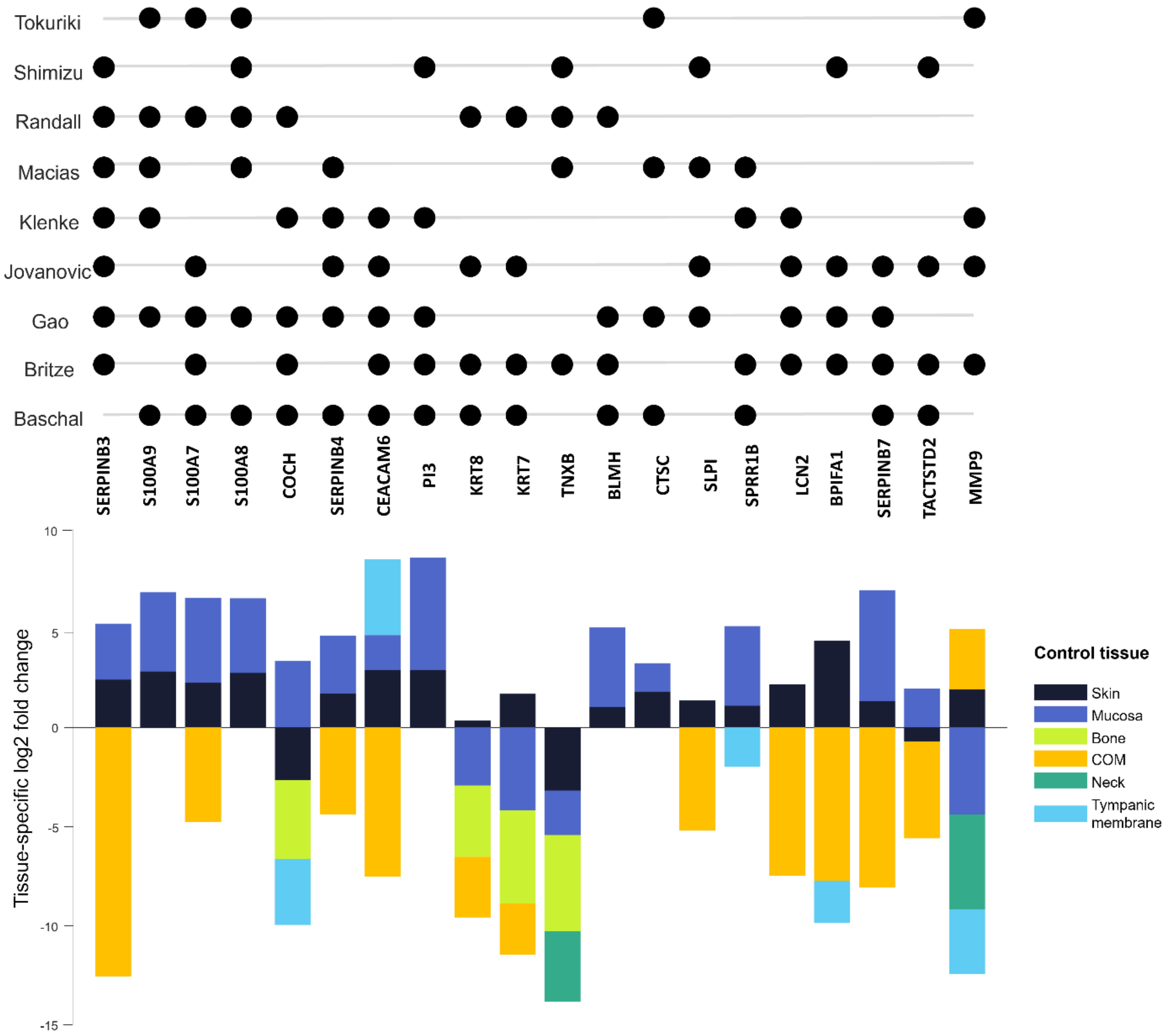
Table 3. Number of papers differentially expressed genes/proteins were detected in, regardless of tissue comparison or direction.

N papers	N genes	Genes
7	1	*SERPINB3*
6	3	*S100A9, S100A7, S100A8*
5	4	*COCH, SERPINB4, CEACAM6, PI3*
4	12	*KRT8, KRT7, TNXB, BLMH, CTSC, SLPI, SPRR1B, LCN2, BPIFA1, SERPINB7, TACSTD2, MMP9*
3	109	*FKBP10, SOD2, SLC25A5, AOC3, FLG2, DSC3, HMGCS1, COL3A1, COL6A6, COL1A2, BPIFB1, GDA, COL8A1, LUM, S100A7A, PRSS23, IGFBP2, S100A2, CARHSP1, BAX, RXRA, LAMA5, CAV1, HSPA1A, ASS1, GATA3, LONRF1, EPCAM, TFAP2C, PRXL2A, PLP1, RHPN2, PGM5, FOS, PHGDH, GPC3, FABP4, FGFBP2, SP5, KRT19, CDH1, ALDH1A3, MAL2, CLDN1, PIP, CXCL17, EYA2, PERP, SYBU, C1orf116, CRABP2, KLF4, FAM83A, CDS1, C9orf152, CXADR, CLDN7, MMP13, FMOD, PFN2, SERPINB12, RNASE7, CYB5R2, ATP5PD, CYB5A, CTSV, HAL, SERPINB13, SDR9C7, PSAPL1, ARG1, NCCRP1, GGH, KRT78, CALML5, KRT10, TAGLN, ANPEP, NPC2, AGRN, DCTN5, GAN, LGALS3BP, ACP3, CNFN, MAN2B1, LNPEP, BGN, SOD3, TTC39B, IL36G, GJB2, CASP14, CYCS, SERPINB2, TMPRSS11D, PNP, DAAM1, MAB21L4, SERBP1, EIF1AX, QPCT, TNS3, FAM83B, DEGS1, NIBAN1, YBX3, TCN1*
2	495	*IDH2, TXNDC5, RPL14, MDH2, ELANE, CTSG, APCS, ACOT1, FLG, FABP5, MPZ, FTL, FADS2, FASN, ITGAM, KRT79, TPSAB1, OGN, FBN1, UPK1B, NIPBL, LAMP5, MS4A7, IGFL1, KRT16, CCN2, SFRP2, S100A12, SIX1, VCAN, CDH11, INHBA, UPP1, GLIPR1, CYTOR, PLBD1, SULF2, NBPF15, CLEC7A, NUCB2, PTN, RGS3, INMT, NUCKS1, SYNCRIP, FHL2, SLC25A45, HOOK2, PAK6, CDC42EP4, LYPD6B, ZFP36, ALDH2, H3-3B, EFS, TLCD3A, GARNL3, THEM5, MATN2, KRT15, HSPA2, PHLPP1, CLEC3B, GPC1, CCL15, ISOC1, APCDD1, ADIRF, SELENBP1, IRX5, COBL, SLC12A2, SLURP1, BCAT2, CGNL1, TFF3, INHBB, LRP4, EDNRB, RBP4, CFD, DNER, F10, DCT, TPPP3, OSR1, HMGCS2, TF, CRAT, PAMR1, ATP6V1B1, TYRP1, KRT2, PI16, GAL, STMN2, AGR2, CAPN13, MUC4, DEFB1, AKR1C2, VTCN1, MUC1, CDH3, ALDH3B2, FOXA1, PROM1, CYP4X1, C19orf33, ANXA8, GABRP, CFB, WFDC2, CYP24A1, MUC20, TFAP2A, SLC34A2, KLF5, VSIG2, PTPRF, ISL1, MMP10, CYP2F1, RAB25, PDZK1IP1, SLC4A11, CP, DSP, DUOX1, FGFR3, SOX2, MSMB, ADGRF1, SLC44A4, GSTA1, CDH6, ECRG4, CD24, RAB17, SAA1, PPL, AQP3, PLEKHS1, DST, SIX4, HAS3, LMO3, TP63, PAX9, SBSPON, SCARA5, CHI3L2, SPP1, CTSH, CTSD, ITGB1, PRDX2, ZFP36L1, KRT4,*

N papers	N genes	Genes
		SBDS, NIT2, ASAH1, ACAT1, ITGB4, AK4, GSTM3, RPS12, CD59, DNAJB1, C1QBP, PGRMC2, ECM1, KRT77, TGM3, ACOX1, KRT23, TGM1, KRT80, PRELP, DMD, POF1B, CAVIN1, SPTBN1, LAMC1, FLNA, APOH, LMNB2, SULT1A1, COL18A1, ABI3BP, APOA2, RAP2C, SNU13, SAE1, NID2, AEBP1, COL4A2, STS, FBXO2, NIPSNAP2, SAP18, DTD1, CNN1, COL4A1, MYLK, HTRA1, GLRX, C6, ANK3, CDH13, CLIP1, NCKAP1, ITGA2, CUL3, IFI30, IMPA2, SDR16C5, KLK5, ARSF, KRTDAP, MBP, PPM1L, PRCP, TINAGL1, SCEL, TTLL12, TMOD1, CHI3L1, VWA1, GNAI1, NDUFA5, ITGA6, KRT18, DCN, LAMB2, OLFML1, F2, EHD2, MYO1C, FBLN5, BPIFC, HOPX, WFDC12, GGCT, KLK13, ELOVL4, GPLD1, KLK6, LYPD5, PTGER3, SERPINA12, LIPN, C5orf46, CSTA, CES1, MSMO1, ACER1, GSDMA, KLK11, KLK8, DBI, AADACL2, PEG10, IDE, SPRR1A, PDCD4, TOB2, MMP7, KLK7, EPHX3, PLXND1, SERPINB8, SPRR4, HPSE, NDRG2, TM7SF2, AFTPH, TOB1, YOD1, SPRR2D, ALOXE3, CDKN1A, FARP1, KLK10, WFDC5, CILP, TGM2, PAIP2B, CLPX, NRARP, RNF227, GATM, FOSL1, EMP2, JUNB, IGF2, KLK14, GNA15, TENT5C, GJB4, EPN2, ALDH1A1, RNF139, ABHD5, KEL, COL11A1, PHLDA1, ALOX12B, KIF21A, AVPR1A, IL1RN, ANXA9, GJB5, CSRNP1, IER2, ZNF740, ATP6V1G1, CHST1, PLEK, TMEM45B, NIPAL1, LCE1A, KLHL18, LGALSL, NLRX1, C15orf48, TMX4, NDFIP2, MEIS2, WDR3, STAB1, MMP15, IL36B, SIX2, SP6, SOWAHC, ENSA, CFI, ATP11B, EPHX1, RPS27A, SERPINA9, PSMA1, RDH12, IL18, LRRC2, EPS8L1, SLC16A7, IER3, QPRT, RARB, PRRG4, CPA4, RCOR1, DYNC1LI1, MALL, BCL2L1, TTC39A, NDUFS5, ANXA6, RBBP6, CLIC3, RPL37A, SERTAD1, PIP5K1B, NOP16, DUSP7, PSMB1, TUBB2A, PPP1R14C, GALNT1, SORBS3, EFNA1, TIFA, LONRF3, SELENOP, ID4, CRYBG1, LXN, VIM, C7, SCRIB, HNRNPAB, S100P, URB2, HPGD, RLIG1, GJB3, VPS4B, CCL27, KHDRBS3, IFFO2, SRSF3, KRT1, HMCES, EPOP, MDK, GOLM1, ITGB2, GMFG, NT5C3A, FBXO3, RAB5A, RAB38, LEO1, KCTD4, TNN, CDYL2, BAG1, LYZ, RET, CANX, RAPGEF1, CD38, EIF1, PSMB2, USP47, EVI2A, SMPDL3A, FGL2, TEX264, ITPKC, PDK4, IQCA1, TENT5B, DNAJA1, POLD1, PKDCC, ALDH3A2, LPAR6, TMEM230, DOCK5, CYP4F22, BCAS2, UPK3BL1, DMTN, AP1AR, CCDC6, BICD2, CERS3, PDE7B, MEMO1, KLF10, NOLC1, HEXIM1, GSN, TGFBR3, EIF5, CTSB, ANK2, DPT, AOX1, MFAP2, ITIH5, THY1, COBLL1, LAMB1, ANGPTL5, PNO1, BAG5, ARID2, IL33, RAI14, TPM1, ITPRID2, SLC25A4, RARRES1, F13A1, MMP11, C6orf132, CIRBP, SMARCD2, GPM6B, TRAFD1, NUP35, PYCR1, LMO7, FAT1, FHL1, CDC42BPG, SORBS1, CRABP1, NAT10, MYH11, LAMC2, PAX3, PTPRD, AHCYL2, HSPB6, MYL9, MGST2, MMP1, SYPL1, PRPH, BAK1

Figure 11. Differentially expressed genes detected in 4+ papers.

The top panel shows which papers a given gene was detected in. The bottom panel shows log2-fold expression change per tissue comparison on the bottom. The height of each stacked segment shows the average across that specific comparison tissue. Tokuriki data are included in the top panel but not in averages as fold changes were not available.



While direction of differential expression was generally the same across papers with regards to the same control tissue (skin or mucosa), there was variation in direction of differential expression between control tissues. Many genes which were upregulated in cholesteatoma relative to normal skin or mucosa were downregulated in comparison to chronic otitis media tissues, including *SERPINB3, SERPINB4, SERPINB7, LCN2, S100A7, CEACAM6,* and *SLPI*. The most consistently upregulated genes compared to skin and mucosa were *SERPINB3*, *S100A9, S100A7, S100A8,* and *PI3*. The most consistently downregulated gene compared to skin and mucosa was *TNXB*, while *KRT8* was generally downregulated compared to skin, mucosa and bone with the exception of one paper where it was upregulated compared to skin.

2.3.3 Dysregulated pathways and processes amongst genes common to multiple papers

Dysregulated genes detected in 3 or more papers (regardless of tissue comparison or direction) were enriched for GO biological processes, molecular functions, and compartments with four main themes (**Table 4**):

- Structural roles and the ECM (*extracellular matrix structural constituent, extracellular region, collagen-containing extracellular matrix, collagen binding, collagen fibril organization, peptidase regulator activity, peptide cross-linking, Golgi lumen p*=8.81×10^{-26}-0.0348)
- Skin development (*cornified envelope, structural constituent of skin epidermis, tissue development p*=1.25×10^{-8}-0.0356).
- Cell lifecycle (*tissue development, cell adhesion, cell migration, locomotion, autocrine signaling, cell population proliferation, regulation of apoptotic signaling pathway, regulation of cell motility p*=0.00424-0.00448).
- Immune response (*RAGE receptor binding, defense response, tertiary granule lumen, p*=0.00728-0.0135*)*.

Table 4. Enriched processes amongst dysregulated genes detected in 3 or more papers regardless of direction or tissue comparison.

Showing g:Profiler highlighted processes only. American spellings for GO terms are retained for consistency with original data sources.

source	Term Name	Adjusted p-value	Term Size	Genes
GO:MF	peptidase regulator activity	9.32x10^{-6}	227	*SERPINB3, SERPINB4, PI3, CTSC, SLPI, SERPINB7, CAV1, GPC3, SERPINB12, CTSV, SERPINB13, SERPINB2*
	extracellular matrix structural constituent	4.69x10^{-5}	167	*TNXB, COL3A1, COL6A6, COL1A2, COL8A1, LUM, LAMA5, FMOD, AGRN, BGN*
	transition metal ion binding	1.01x10^{-2}	1124	*S100A9, S100A7, S100A8, LCN2, MMP9, SOD2, AOC3, FLG2, GDA, S100A7A, S100A2, RXRA, GATA3, KLF4, MMP13, ARG1, ANPEP, LNPEP, SOD3, QPCT*
	RAGE receptor binding	1.35x10^{-2}	10	*S100A9, S100A7, S100A8*
	collagen binding	3.48x10^{-2}	69	*COCH, TNXB, MMP9, LUM, MMP13*
	structural constituent of skin epidermis	3.56x10^{-2}	36	*PI3, KRT7, KRT78, KRT10*
GO:BP	tissue development	3.26x10^{-9}	2010	*SERPINB3, S100A7, KRT7, TNXB, SPRR1B, SERPINB7, TACSTD2, MMP9, FLG2, COL3A1, COL1A2, COL8A1, BAX, RXRA, LAMA5, CAV1, GATA3, EPCAM, PGM5, FOS, PHGDH, GPC3, KRT19, ALDH1A3, CLDN1, EYA2, CRABP2, KLF4, CXADR, MMP13, PSAPL1, KRT78, CALML5, KRT10, TAGLN, CNFN, BGN, GJB2, CASP14, YBX3*
	response to biotic stimulus	5.10x10^{-5}	1586	*S100A9, S100A7, S100A8, SERPINB4, PI3, KRT8, SLPI, LCN2, BPIFA1, MMP9, SOD2, BPIFB1, BAX, CAV1, HSPA1A, ASS1, GATA3, FOS, GPC3, FABP4, CDH1, CLDN1, KLF4, CXADR, RNASE7, ARG1, NPC2, IL36G, GJB2*
	cell adhesion	2.73x10^{-4}	1512	*S100A9, S100A8, CEACAM6, TNXB, TACSTD2, AOC3, FLG2, DSC3, COL3A1, COL6A6, COL8A1, IGFBP2, LAMA5, CAV1, ASS1, GATA3, EPCAM, PGM5, CDH1, CLDN1, PERP, KLF4, CXADR, CLDN7, ARG1, LGALS3BP, PNP*
	regulation of peptidase activity	3.52x10^{-4}	244	*SERPINB3, S100A9, S100A8, SERPINB4, MMP9, BAX, CAV1, PERP, KLF4, SERPINB13, CYCS*
	response to toxic substance	2.43x10^{-3}	238	*S100A9, BLMH, SOD2, BAX, ASS1, PRXL2A, FOS, CDH1, CLDN1, SOD3*
	collagen fibril organization	3.73x10^{-3}	65	*TNXB, FKBP10, COL3A1, COL1A2, LUM, FMOD*
	regulation of apoptotic signaling pathway	4.48x10^{-3}	382	*S100A9, S100A8, CTSC, MMP9, SOD2, SLC25A5, BAX, CAV1, HSPA1A, EYA2, CTSV, YBX3*
	defense response	7.28x10^{-3}	1791	*S100A9, S100A7, S100A8, SERPINB4, PI3, CTSC, SLPI, LCN2, BPIFA1, MMP9, AOC3, BPIFB1, CAV1, HSPA1A,*

source	Term Name	Adjusted p-value	Term Size	Genes
				ASS1, GATA3, PLP1, FOS, FABP4, CLDN1, CXCL17, KLF4, CXADR, RNASE7, ARG1, LGALS3BP, IL36G
	autocrine signaling	1.11x10^{-2}	7	*SERPINB3, S100A9, S100A8*
	Locomotion	1.68x10^{-2}	1234	*SERPINB3, S100A9, S100A7, S100A8, CEACAM6, TNXB, TACSTD2, MMP9, SOD2, COL3A1, LAMA5, CAV1, GATA3, PLP1, CDH1, CLDN1, CXCL17, KLF4, CXADR, CLDN7, PFN2*
	cell population proliferation	2.00x10^{-2}	2006	*SERPINB3, CEACAM6, TNXB, SERPINB7, TACSTD2, MMP9, SOD2, SLC25A5, COL3A1, COL8A1, IGFBP2, BAX, LAMA5, CAV1, HSPA1A, GATA3, EPCAM, TFAP2C, FOS, GPC3, CLDN1, KLF4, FAM83A, CLDN7, ARG1, NCCRP1, PNP, FAM83B*
	cell migration	3.12x10^{-2}	1496	*SERPINB3, S100A9, S100A7, S100A8, CEACAM6, TNXB, TACSTD2, MMP9, SOD2, COL3A1, S100A2, BAX, LAMA5, CAV1, GATA3, PLP1, GPC3, CDH1, CLDN1, CXCL17, KLF4, CXADR, PFN2*
	peptide cross-linking	3.41x10^{-2}	28	*PI3, SPRR1B, COL3A1, KRT10*
	regulation of cell motility	3.54x10^{-2}	996	*SERPINB3, S100A7, CEACAM6, TNXB, TACSTD2, MMP9, SOD2, COL3A1, LAMA5, CAV1, GATA3, PLP1, CDH1, CLDN1, CXCL17, KLF4, CLDN7, PFN2*
GO:CC	extracellular region	8.81x10^{-26}	4213	*SERPINB3, S100A9, S100A7, S100A8, COCH, SERPINB4, CEACAM6, PI3, KRT8, KRT7, TNXB, BLMH, CTSC, SLPI, LCN2, BPIFA1, SERPINB7, TACSTD2, MMP9, SOD2, FLG2, DSC3, COL3A1, COL6A6, COL1A2, BPIFB1, COL8A1, LUM, S100A7A, PRSS23, IGFBP2, BAX, LAMA5, HSPA1A, ASS1, EPCAM, PRXL2A, PHGDH, FABP4, FGFBP2, KRT19, CDH1, ALDH1A3, MAL2, PIP, CXCL17, C1ORF116, CRABP2, CXADR, MMP13, FMOD, PFN2, SERPINB12, RNASE7, CTSV, SERPINB13, PSAPL1, ARG1, NCCRP1, GGH, KRT78, CALML5, KRT10, ANPEP, NPC2, AGRN, LGALS3BP, ACP3, MAN2B1, LNPEP, BGN, SOD3, IL36G, SERPINB2, TMPRSS11D, PNP, SERBP1, QPCT, NIBAN1, TCN1*
	collagen-containing extracellular matrix	9.22x10^{-12}	425	*S100A9, S100A7, S100A8, COCH, TNXB, CTSC, SLPI, MMP9, COL3A1, COL6A6, COL1A2, COL8A1, LUM, LAMA5, GPC3, FMOD, SERPINB12, AGRN, LGALS3BP, BGN, SOD3*
	cornified envelope	1.25x10^{-8}	59	*PI3, SPRR1B, FLG2, DSC3, SERPINB12, KRT10, CNFN, CASP14, SERPINB2*
	lateral plasma membrane	1.31x10^{-3}	76	*TACSTD2, EPCAM, CDH1, CLDN1, CLDN7, GJB2*
	apicolateral plasma membrane	2.27x10^{-3}	23	*KRT8, KRT19, CXADR, CLDN7*
	tertiary granule lumen	4.24x10^{-3}	55	*MMP9, FLG2, GGH, QPCT, TCN1*
	Golgi lumen	7.58x10^{-3}	103	*LUM, GPC3, FMOD, AGRN, BGN, SOD3*

source	Term Name	Adjusted p-value	Term Size	Genes
	costamere	3.87×10^{-2}	18	*KRT8, PGM5, KRT19*

2.3.4 Enriched processes amongst up- and down-regulated genes vs skin and mucosa

ECM disruption and enriched upregulated immune genes across tissues

ECM-related terms were enriched in up- and down-regulated gene sets for both tissue comparisons (**Table 5-Table 8**). *Collagen-containing extracellular matrix, extracellular matrix constituent* and *costamere* were downregulated in both tissue comparisons, while *extracellular region* was upregulated compared to mucosa and downregulated compared to skin.

Certain immune terms were enriched amongst upregulated genes for both tissue comparisons. *Toll-like receptor 4 binding, neutrophil aggregation* and *RAGE receptor binding* were enriched in upregulated genes compared to both skin and mucosa.

Enriched pathways in DEGs compared to mucosa

Terms enriched in the upregulated-vs-mucosa set included epidermal development terms such as *structural constituent of skin epidermis, keratohyalin granule, cornified envelope* and *epidermis development* (**Table 6)**. Such terms were not enriched in the upregulated-vs-skin set.

Few terms were enriched in the downregulated set of genes compared to mucosa (**Table 5**), probably because there were fewer mucosa comparisons to draw from. *Negative regulation of wound healing*, *collagen metabolic process* and *negative regulation of fibrinolysis* may support ECM dysregulation and disrupted cellular processes.

Enriched pathways in DEGs compared to skin

Metal ion binding terms were enriched in the set of genes upregulated compared to skin: this includes *calcium-dependent protein binding, calcium ion binding, metal ion sequestering activity*, and *sequestering of zinc ion* (also upregulated compared to mucosa) (**Table 8**).

Table 5. Enriched processes amongst downregulated genes compared to mucosa

	Term	Adjusted p-value	Term size	Genes	N
GO:BP	multicellular organism development	$3.16\text{x}10^{-3}$	4643	*KRT8, KRT19, COL1A2, APOH, AGRN, F2, APCS, OGN, ANPEP, TNXB*	10
	negative regulation of wound healing	$7.85\text{x}10^{-3}$	71	*APOH, F2, APCS*	3
	collagen metabolic process	$2.27\text{x}10^{-2}$	101	*COL1A2, F2, TNXB*	3
	negative regulation of fibrinolysis	$2.54\text{x}10^{-2}$	13	*APOH, F2*	2
GO:MF	sulfur compound binding	$3.39\text{x}10^{-5}$	268	*APOH, AGRN, F2, SULT1A1, TNXB*	5
	extracellular matrix structural constituent	$2.81\text{x}10^{-4}$	167	*COL1A2, AGRN, OGN, TNXB*	4
GO:CC	collagen-containing extracellular matrix	$2.99\text{x}10^{-8}$	425	*COL1A2, APOH, AGRN, F2, APCS, OGN, TNXB*	7
	extracellular exosome	$6.45\text{x}10^{-8}$	2109	*KRT8, KRT19, COL1A2, APOH, AGRN, F2, APCS, OGN, ANPEP, TNXB*	10
	Golgi lumen	$1.65\text{x}10^{-3}$	103	*AGRN, F2, OGN*	3
	costamere	$3.58\text{x}10^{-3}$	18	*KRT8, KRT19*	2
	apicolateral plasma membrane	$5.92\text{x}10^{-3}$	23	*KRT8, KRT19*	2

Table 6. Enriched processes amongst upregulated genes compared to mucosa

	Term	Adjusted p-value	Term size	Genes	N
GO:BP	structural constituent of skin epidermis	1.24x10^{-8}	36	*PI3, FLG, KRT78, KRT77, KRT80, KRT10*	6
	RAGE receptor binding	2.18x10^{-4}	10	*S100A7, S100A9, S100A8*	3
	serine-type endopeptidase inhibitor activity	3.45x10^{-4}	103	*SERPINB3, PI3, SERPINB7, SERPINB12, SERPINB13*	5
	fatty acid binding	5.34x10^{-4}	48	*S100A9, S100A8, FABP5, ACOX1*	4
	Toll-like receptor 4 binding	6.71x10^{-3}	4	*S100A9, S100A8*	2
	arachidonic acid binding	2.34x10^{-2}	7	*S100A9, S100A8*	2
	epidermis development	7.51x10^{-1}1	393	*S100A7, FLG2, FLG, FABP5, KRT78, CALML5, CNFN, KRT77, KRT80, KLK5, KRTDAP, SCEL, KRT10*	13
	intermediate filament organization	2.62x10^{-4}	73	*KRT78, KRT77, KRT23, KRT80, KRT10*	5
GO:MF	autocrine signaling	2.77x10^{-4}	7	*SERPINB3, S100A9, S100A8*	3
	sequestering of zinc ion	4.91x10^{-3}	2	*S100A9, S100A8*	2
	neutrophil aggregation	4.91x10^{-3}	2	*S100A9, S100A8*	2
	peptide cross-linking	2.53x10^{-2}	28	*PI3, FLG, KRT10*	3
	cornified envelope	9.87x10^{-1}2	59	*FLG2, PI3, FLG, SERPINB12, CNFN, KRT77, SCEL, KRT10*	8
GO:CC	extracellular region	1.38x10^{-8}	4213	*BLMH, S100A7, FLG2, SERPINB3, S100A9, S100A8, PI3, FABP5, SERPINB7, SERPINB12, CTSV, SERPINB13, PSAPL1, ARG1, NCCRP1, GGH, KRT78, CALML5, ACP3, LNPEP, KRT77, KLK5, KRTDAP, SCEL, KRT10*	25
	intermediate filament cytoskeleton	6.96x10^{-4}	254	*S100A8, KRT78, KRT77, KRT23, KRT80, KRT10*	6
	keratohyalin granule	3.39x10^{-3}	4	*FLG2, FLG*	2

Table 7. Enriched processes amongst downregulated genes compared to skin

source	Term Name	Adjusted p-value	Term Size	Genes	N
GO:BP	developmental process	$6.06x10^{-6}$	6453	*TNXB, PRXL2A, PLP1, PGM5, PHGDH, FABP4, SP5, COCH, DSC3, ANPEP, KRT79, FBN1, CAV1, GPC3, ACAT1, COL18A1, COL4A1, MYLK, CDC42EP4, MATN2, KRT15, HSPA2, PHLPP1, GATA3, CLEC3B, APCDD1, ADIRF, IRX5, COBL, SLC12A2, LRP4, EDNRB, DCT, TPPP3, OSR1, HMGCS2, TYRP1, KRT2, PI16, STMN2, ITGB4, GSTM3, ABI3BP, ANK3, CDH13, ITGA6, GSN, ANK2, IL33, FHL1, MYH11, PAX3, HSPB6, FASN, RXRA, HSPA1A, TFAP2C, FOS, TACSTD2*	59
	supramolecular fiber organization	$1.52x10^{-2}$	842	*TNXB, PGM5, KRT79, COL18A1, AEBP1, CDC42EP4, KRT15, COBL, CGNL1, RHPN2, TPPP3, KRT2, STMN2, CLIP1, GSN, DPT, SORBS1, MYH11, HSPA1A, TACSTD2*	20
	extracellular matrix organization	$1.66x10^{-2}$	324	*TNXB, TPSAB1, CAV1, COL18A1, NID2, AEBP1, COL4A1, MATN2, ABI3BP, DPT, MYH11*	11
	cellular response to chemical stimulus	$5.81x10^{-7}$	2683	*PRXL2A, FABP4, SP5, SOD3, FBN1, CAV1, COL4A1, MYLK, GNAI1, CDC42EP4, GATA3, CLEC3B, SLC12A2, BCAT2, EDNRB, OSR1, HMGCS2, AK4, GSTM3, ANK3, ITGA6, GSN, AOX1, IL33, SORBS1, FASN, RXRA, HSPA1A, FOS*	29
	ketone body metabolic process	$2.14x10^{-6}$	10	*ACAT1, HMGCS2, TYRP1*	3
	cytoskeleton organization	$5.76x10^{-4}$	1512	*TNXB, PGM5, KRT79, GNAI1, CDC42EP4, KRT15, COBL, CGNL1, RHPN2, TPPP3, KRT2, STMN2, ANK3, CLIP1, GSN, ANK2, SORBS1, MYH11, HSPA1A, TACSTD2*	20
	nephric duct morphogenesis	$1.00x10^{-2}$	12	*GPC3, GATA3, OSR1*	3
	nephron epithelium development	$2.49x10^{-2}$	123	*GPC3, ACAT1, GATA3, EDNRB, OSR1, TACSTD2*	6
	farnesyl diphosphate biosynthetic process, mevalonate pathway	$2.66x10^{-2}$	2	*HMGCS2, HMGCS1*	2
GO:MF	extracellular matrix structural constituent	$4.54x10^{-2}$	167	*TNXB, COL6A6, FBN1, COL18A1, NID2, AEBP1, COL4A1, MATN2, ABI3BP, DPT*	10
	collagen binding	$4.79x10^{-2}$	69	*TNXB, COCH, NID2, AEBP1, ABI3BP*	5
	hydroxymethylglutaryl-CoA synthase activity	$4.81x10^{-2}$	2	*HMGCS2, HMGCS1*	2
GO:CC	collagen-containing extracellular matrix	$1.30x10^{-10}$	425	*TNXB, SOD3, COCH, COL6A6, TPSAB1, FBN1, GPC3, COL18A1, NID2, AEBP1, COL4A1, MATN2, CLEC3B, ITGB4, ABI3BP, CDH13, DPT, ANGPTL5*	18
	extracellular region	$3.16x10^{-10}$	4213	*TNXB, PRXL2A, PHGDH, FABP4, FGFBP2, SOD3, COCH, DSC3, COL6A6, ANPEP, KRT79, TPSAB1, FBN1, ACAT1, COL18A1, NID2, AEBP1, COL4A1, GNAI1, NIBAN1, ALDH2, H3-3B, MATN2, KRT15, HSPA2, CLEC3B, ADIRF, SLC12A2, CFD, F10, PAMR1, KRT2, PI16, ITGB4, GSTM3, CD59, ABI3BP, CDH13, GSN, DPT, AOX1, ITIH5, ANGPTL5, IL33, MYH11, HSPB6, PRPH, FASN, HSPA1A, TACSTD2*	50

source	Term Name	Adjusted p-value	Term Size	Genes	N
	anchoring junction	2.96×10^{-3}	905	*PGM5, DSC3, CAV1, CDC42EP4, CGNL1, ITGB4, CD59, ANK3, CDH13, ITGA6, GSN, ANK2, FHL1, SORBS1, HSPA1A*	15
	costamere	1.87×10^{-2}	18	*PGM5, ANK3, ANK2*	3
	cytoplasm	2.93×10^{-2}	12345	*PRXL2A, PGM5, PHGDH, FABP4, SOD3, DSC3, ANPEP, KRT79, FBN1, CAV1, GPC3, ACAT1, COL18A1, AEBP1, COL4A1, MYLK, GNAI1, NIBAN1, AOC3, CDC42EP4, ALDH2, TLCD3A, GARNL3, KRT15, HSPA2, PHLPP1, CLEC3B, ISOC1, ADIRF, COBL, SLC12A2, BCAT2, RHPN2, CFD, F10, DCT, TPPP3, OSR1, HMGCS2, CRAT, TYRP1, KRT2, STMN2, AK4, GSTM3, CD59, ANK3, CDH13, CLIP1, NDUFA5, GSN, ANK2, AOX1, IL33, SLC25A4, CIRBP, FHL1, SORBS1, MYH11, HSPB6, PRPH, ACOT1, FADS2, FASN, RXRA, HSPA1A, LONRF1, TFAP2C, FOS, TACSTD2, ATP5PD, CYB5A, HMGCS1*	73
	protein complex involved in cell adhesion	3.90×10^{-2}	59	*TNXB, PLP1, ITGB4, ITGA6*	4

Table 8. Enriched processes amongst upregulated genes compared to skin

source	Term Name	Adjusted p-value	Term Size	Genes	N
GO:BP	biological process involved in interspecies interaction between organisms	1.06×10^{-6}	1724	*SERPINB3, S100A9, S100A8, PI3, SERPINB4, SLPI, BAX, LCN2, S100A7, RNASE7, IL36G, GJB2, UPK1B, S100A12, CLEC7A, NUCKS1, BPIFA1, NPC2, CTSB, MGST2, BAK1, BPIFB1*	22
	antimicrobial humoral response	1.68×10^{-4}	131	*S100A9, PI3, SLPI, S100A7, RNASE7, S100A12, BPIFA1*	7
	autocrine signaling	1.52×10^{-3}	7	*SERPINB3, S100A9, S100A8*	3
	defense response	2.31×10^{-3}	1791	*S100A9, S100A8, PI3, SERPINB4, SLPI, LCN2, S100A7, CTSC, RNASE7, IL36G, S100A12, INHBA, CLEC7A, PTN, BPIFA1, MGST2, BPIFB1, LGALS3BP*	18
	positive regulation of endopeptidase activity	5.05×10^{-3}	138	*SERPINB3, S100A9, S100A8, BAX, CLEC7A, BAK1*	6
	cellular response to X-ray	7.09×10^{-3}	11	*NIPBL, SFRP2, NUCKS1*	3
	proteolysis	9.28×10^{-3}	1573	*SERPINB3, S100A9, S100A8, SERPINB4, BAX, BLMH, CTSC, TMPRSS11D, CLEC7A, MMP13, HTRA1, CTSB, MMP11, MMP1, BAK1, PRSS23*	16
	collagen catabolic process	1.59×10^{-2}	45	*MMP13, CTSB, MMP11, MMP1*	4
	B cell negative selection	1.61×10^{-2}	2	*BAX, BAK1*	2

source	Term Name	Adjusted p-value	Term Size	Genes	N
	sequestering of zinc ion	1.61×10^{-2}	2	*S100A9, S100A8*	2
	neutrophil aggregation	1.61×10^{-2}	2	*S100A9, S100A8*	2
	immune system process	2.20×10^{-2}	2776	*S100A9, S100A8, PI3, SERPINB4, SLPI, BAX, LCN2, S100A7, CTSC, RNASE7, IL36G, S100A12, INHBA, IGFBP2, CLEC7A, PTN, BPIFA1, IFI30, PTPRD, BAK1, BPIFB1*	21
	regulation of apoptotic signaling pathway	2.27×10^{-2}	382	*S100A9, S100A8, BAX, CTSC, SFRP2, INHBA, PYCR1, BAK1*	8
	cell migration	2.37×10^{-2}	1496	*SERPINB3, S100A9, S100A8, BAX, CEACAM6, S100A7, NIPBL, SFRP2, S100A12, CDH11, S100A2, CLEC7A, PTN, FAT1, LAMC2*	15
	cellular response to radiation	2.48×10^{-2}	182	*BAX, NIPBL, SFRP2, NUCKS1, MMP1, BAK1*	6
	positive regulation of programmed cell death	3.37×10^{-2}	532	*S100A9, S100A8, BAX, CTSC, SFRP2, INHBA, CLEC7A, HTRA1, BAK1*	9
	post-embryonic camera-type eye morphogenesis	4.83×10^{-2}	3	*BAX, BAK1*	2
GO:MF	RAGE receptor binding	8.94×10^{-6}	10	*S100A9, S100A8, S100A7, S100A12*	4
	calcium-dependent protein binding	8.16×10^{-5}	79	*S100A9, S100A8, S100A7A, S100A7, S100A12, S100A2*	6
	serine-type endopeptidase activity	6.92×10^{-4}	180	*CTSC, TMPRSS11D, MMP13, HTRA1, MMP11, MMP1, PRSS23*	7
	peptidase regulator activity	3.21×10^{-3}	227	*SERPINB3, PI3, SERPINB4, SLPI, CTSC, SFRP2, RARRES1*	7
	calcium ion binding	6.12×10^{-3}	726	*S100A9, S100A8, S100A7A, S100A7, GJB2, S100A12, CDH11, S100A2, SULF2, MMP13, FAT1*	11
	Toll-like receptor 4 binding	2.97×10^{-2}	4	*S100A9, S100A8*	2
	metal ion sequestering activity	4.95×10^{-2}	5	*LCN2, S100A7*	2
GO:CC	extracellular space	2.34×10^{-21}	3303	*SERPINB3, S100A9, S100A8, PI3, S100A7A, SERPINB4, SLPI, BAX, LCN2, CEACAM6, TCN1, BLMH, S100A7, CTSC, RNASE7, ASAH1, IL36G, TMPRSS11D, UPK1B, NIPBL, COL8A1, IGFL1, SFRP2, CDH11, INHBA, IGFBP2, GLIPR1, PLBD1, SULF2, PTN, BPIFA1, MMP13, NPC2, HTRA1, GLRX, CTSB, RARRES1, MMP11, FAT1, LAMC2, PTPRD, MMP1, BPIFB1, PRSS23, LGALS3BP*	45

source	Term Name	Adjusted p-value	Term Size	Genes	N
	extracellular matrix	9.00×10^{-9}	555	*S100A9, S100A8, PI3, SLPI, S100A7, CTSC, COL8A1, SFRP2, MMP13, HTRA1, CTSB, MMP11, LAMC2, MMP1, LGALS3BP*	15
	BAK complex	2.16×10^{-3}	2	*BAX, BAK1*	2

2.4 Discussion

The studies identified in this review showed heterogeneity in tissue comparisons and analytical methods. The overlap between dysregulated genes detected across papers was small, perhaps as a result of these factors combined with the individually small sample sizes. Gene expression may also vary due to differences in cholesteatoma type (e.g., congenital or acquired), presence or absence of active infection, the size and location of cholesteatoma and the relative amounts of cell types sampled. For most studies, no such information was available. Shimizu *et al.* (2023)[139] identified 11 different cell types within their cholesteatoma samples based on clustering analysis and suggest that cell types change over time. Therefore, different papers may have sampled different relative amounts of cholesteatoma cell types such as keratinocytes and fibroblasts.

Terms related to ECM structure and function were enriched for both up- and downregulated gens compared to skin and mucosa, as was peptidase activity and regulation, suggesting widespread ECM dysfunction. Whilst inflammatory pathways were enriched across skin and mucosa tissue comparisons, several inflammatory proteins were also downregulated compared to chronic otitis media tissue.

Upregulated genes compared to mucosa were enriched for pathways associated with epidermal development, but these were neither up- nor downregulated compared to normal skin consistent with the nature of cholesteatoma as stratified squamous epithelium. Zeng *et al.* (2024)[135] also performed pathway enrichment analysis on a global gene expression with skin as the control tissue and found epidermis development, keratinocyte differentiation and keratinization were enriched in the *down*regulated gene set. Terms related to cell adhesion and cytoskeletal function were also downregulated in their study, whereas terms related to immune function, peptidase activity and chemokine activity were upregulated.

2.4.1 Consistently dysregulated genes across papers

Inflammatory protease inhibitors: SERPINB3, SERPINB4, SERPINB7 and PI3

SERPINB3 was detected in 7 out of 9 global gene expression studies, *SERPINB4* in 5, and *SERPINB7* in 4. The SERPINs are a family of inflammatory serine protease inhibitors, of which B3 and B4 are squamous cell carcinoma markers[141]. They regulate proteases involved in the response to tissue damage, including inflammation, response to tumour cells, and wound healing. *SERPINB3* specifically has been investigated in cholesteatoma by Ho *et al.* 2012[142] and 2020[143]. In their 2012 paper, the authors found SERPIN B3 protein was localized in the epithelium of both cholesteatoma and retro-auricular skin but that three isoforms were overexpressed in cholesteatoma. In their 2020 study, the authors suggest that SERPINB3 overexpression may promote cell proliferation and prevent autophagy. Yoshikawa *et al.* (2006)[134] found that *SERPINB2* and *SERPINA8* were upregulated more strongly in cholesteatoma fibroblasts than skin fibroblasts in response to IL-1α, but the SERPINs identified in other gene expression analyses were not reported.

Another inflammatory protease inhibitor, *PI3***,** was upregulated compared to healthy skin and mucosa. Its product, elafin, is an elastase-specific inhibitor with anti-inflammatory properties, expressed as a normal part of wound healing and in inflammatory skin conditions such as psoriasis[144]. Chang *et al.* (1990)[144] found that elafin was highly expressed during the early stages of wound healing and counteracted the infiltration of polymorphonuclear cells, while it remained constantly highly expressed in chronic wounds where polymorphonuclear cell infiltration was also present. *SLPI*, another antileukoproteinase, was upregulated in cholesteatoma in 4 studies and was also investigated by Lee *et al.* (2006)[145] who detected higher expression in cholesteatoma than ordinary skin.

Expression of protease inhibitors may be part of the normal immune-regulatory response to limit the tissue damage caused by inflammatory proteases[146]. Interestingly, Jovanovic *et al.* (2020)[36] show that SERPINB3 and SLPI are *down*regulated in perimatrix compared to COM, perhaps indicating a failure to properly regulate the immune response resulting in excessive ECM breakdown. Indeed, imbalances in inflammatory elastases and their inhibitors such as PI3 and SLPI have been implicated in excessive inflammatory responses in the respiratory system[147]. Conversely, CTSC, an activator of granulocyte serine proteases[148], was *up*regulated compared to skin and mucosa across 4 papers.

S100 proteins

S100A7, S100A8 and *S100A9* were consistently upregulated in cholesteatoma compared to healthy skin, each being detected in 6 papers. The S100 proteins are a family of zinc and calcium-binding inflammatory proteins involved in recruitment of immune cells, epidermal differentiation and inflammation, and apoptosis. Pelc *et al.* (2003)[149] showed that several S100 proteins are also expressed in other cysts including epidermoid cysts and craniopharyngiomas and S100A3 is expressed in much greater quantities in cholesteatoma than other cyst types. S100A7, S100A8 and S100A9 were not measured Pelc *et al.*'s analysis, but Kim *et al.* (2008) found increased S100A7 expression in cholesteatoma[150]. Also known as psoriasin, S100A7 is an antimicrobial peptide highly expressed in psoriasis and atopic dermatitis[151].

Matrix metalloproteinases

MMP9 dysregulation was detected in 4 papers; it was upregulated compared to COM tissue and skin, but downregulated compared to tympanic membrane, the neck of cholesteatoma and middle ear mucosa. The matrix metalloproteinases (MMPs) are a family of proteases with structural collagenase and gelatinase activity, making them important degraders of the ECM with important roles in bone turnover[152]. For this reason, they have been investigated in degenerative bone diseases such as periodontal disease[153] and arthritis[154].

Due to bone-destructive nature of cholesteatoma, several studies have investigated MMPs in its pathology. Some have shown *MMP9* over expression[155–157], although others have tested *MMP9* and found no difference between cholesteatoma and control tissues[158,159]. *MMP8*, *MMP13*, and *MMP2* have also shown to be overexpressed in cholesteatoma compared to normal skin[158,160]. Another consistently upregulated protein compared to skin, LCN2, forms a heterodimer with MMP9. Unlike MMP9, LCN2 was also downregulated compared to COM according to Jovanovic *et al.* (2020)[36], who suggest that changes in the balance of MMP9/LCN2 complex may be associated with changes in signalling events modulated by their receptors.

Aside from bone loss, MMP and other protease activity may contribute to degradation of the tympanic basement membrane, facilitating invasion[161]. Interestingly, Britze *et al.* (2014)[78]

showed *MMP9* to be downregulated in cholesteatoma compared to tympanic tissue, neck of cholesteatoma and middle ear mucosa. Either these tissues typically express high levels of MMPs, perhaps contradicting a role in cholesteatoma pathology, or expression is increased in persons with cholesteatoma – Britze *et al.* took these control tissues from the same individuals as cholesteatoma tissues, so we cannot say for sure that this level of expression is normal.

CEACAM6, COCH and TNXB are consistently dysregulated but have not been individually investigated

CEACAM6 and *COCH* were detected in 5 papers, while *TNXB* was detected in 4 papers with high consistency. These genes have not been subject to individual study in cholesteatoma. Both *TNXB* and *COCH* are downregulated across tissue comparisons while *CEACAM6* is upregulated in most comparisons except cholesteatoma vs COM.

CEACAM6 is a known biomarker for several cancer types. It is also expressed in normal. epithelia, granulocytes and monocytes but is overexpressed in cancers including colorectal cancer, where it is correlated with invasiveness; it predicts poor survival and may specifically mark aggressive cancer[162].

Cochlin (COCH) is a collagen-binding protein which interacts with proteins involved in cytoskeleton remodelling and has roles in cell shape and motility in the trabecular meshwork of the eye [163]. Cochlin may also interact with cytoskeletal proteins in the ear and pathogenic variants are associated with sensorineural deafness. It has also been shown to have immune function in the inner ear[164]. Defective ciliary genes can cause defects in cochlin secretion[165], posing a potential link to our previous genetic study of cholesteatoma suggesting ciliary dysfunction[126].

TNXB is one of two gene encoding tenascin-X, a matrix glycoprotein whose deficiency causes Ehlers-Danlos syndrome; the mechanism is thought to be via impaired deposition of collagen in the extracellular matrix[166]. Furthermore, tenascin-X is thought to play a role in matrix maturation during wound repair and, alongside other tenascins, act as modulators of cell activity with anti-adhesive properties[167]. Kajitani *et al.* (2019)[168] showed that tenascin-deficient mice had increased osteoclast activity and subsequent bone loss.

2.4.2 Disrupted processes in cholesteatoma according to pathway enrichment analysis

Structural proteins and ECM dysregulation

Many enriched terms amongst DEGs from 3 or more papers, as well as in individual tissue comparisons, relate to ECM structure or degradation of the ECM. The extracellular space was enriched for up and down-regulated genes compared to both skin and mucosa; ECM proteins such as TNXB, COCH were downregulated and ECM-degrading proteases such as MMP9 and CTSC were upregulated. Some ECM protease inhibitors including PI3 and SLPI were also upregulated, indicating a complex and broad dysregulation of extracellular proteins.

ECM dysregulation may be important to cholesteatoma pathology in several ways. First, bone tissue consists mainly of mineralised ECM, so dysregulation of ECM constituents and proteases may be associated with bone loss in cholesteatoma[156,168]. Second, the ECM has important roles in coordinating cell communication, migration and cell fate[34]. Aberrant migration of epithelium is implied in retraction pocket and invasion theory, with basement membrane weakening suggested to permit invasion even if the tympanic membrane is not fully perforated[21]. ECM degradation may therefore also contribute to cholesteatoma invasiveness and hyperproliferation. ECM remodelling is also a key feature of wound-healing tissue, a tissue which shares features with cholesteatoma such as cellular proliferation, migration and differentiation, as well as inflammation[35]. Failure of the ECM to mature in chronic wound tissue may have parallels to ECM dysfunction in cholesteatoma.

Cell cycle and epidermal development

DEGs were enriched for altered cell lifecycle processes, such as adhesion, migration, proliferation, and apoptosis, overlapping with processes associated with ECM function. Differential expression of genes indicating increased epithelial proliferation is not surprising given the hyperproliferative phenotype of cholesteatoma. Increased epithelial turnover within a retraction pocket is suggested by Louw (2010)[21] to contribute to the initial formation of cholesteatoma.

Growth factor receptors and their binding proteins were not consistently detected: *FGFBP2* was downregulated compared to normal skin in 3 papers; insulin-like growth factor 2 (*IGF2*), insulin growth factor-like family member 1 (*IGFL1*) and IGF-binding protein 2 *(IGFBP2)* were upregulated in 2 papers each, although *IGFBP2* was downregulated compared to COM. The IGF2-IGFBP2 complex is associated with osteoblast activation[169], so upregulation compared to skin could represent increased bone turnover in cholesteatoma, but inadequate bone formation compared to COM. *TGFBR3* (transforming growth factor beta receptor 3) was both up- and downregulated compared to normal skin. This complex picture is difficult to interpret and is not consistent with reviews of cholesteatoma etiopathology suggesting an important role for growth factors[2,20].

Apoptotic processes were also enriched in the set of genes upregulated in cholesteatoma compared to skin tissue (*regulation of apoptotic signaling pathway* and *positive regulation of programmed cell death)*. Apoptosis has been suggested to play a role in cholesteatoma pathology: Olszewska *et al.* (2006) show increased apoptosis in cholesteatoma compared to ordinary skin and suggest this is associated with differentiation and accumulation of keratin in the middle ear[170].

Certain skin developmental terms were enriched in the upregulated DEGs compared to mucosa, but not skin. As cholesteatoma tissue consists of skin (meaning stratified keratinizing epithelium) and not mucosa, this is not surprising. However, several upregulated genes (for both skin and mucosa comparisons) are associated with hyperkeratotic skin conditions such as epidermal thickening and palmoplantar keratoderma (**SI Table 2**), consistent with excess keratin production in cholesteatoma. Additionally, some of the most consistently dysregulated genes were associated with keratinization: keratins *KRT8* and *KRT7,* and cornifin *SPRR1B*, were detected in 4 papers each. While *KRT8* and *KRT7* were downregulated compared to skin and upregulated compared to mucosa, *SPRR1B* was upregulated compared to both. Yoshikawa *et al.* (2006)[134] also found that *SPRR1B* was upregulated more in response to stimulation of IL-1α in cholesteatoma fibroblasts than skin fibroblasts. Cornifin-B is a keratinocyte envelope protein which forms a gene cluster called the epidermal differentiation complex along with profilagrin (FLG), loricrin (LORICRN) other SPRRs, and S100A genes; the complex is involved in terminal differentiation of keratinocytes[171,172]. Upregulation of *SPRR1B*, even in comparison to normal skin, supports abnormal differentiation of skin cells in cholesteatoma.

Interestingly, some DEGs in cholesteatoma are known cancer markers; prominent examples are CEACAM6 and TACSTD2[173,174], which act as adhesion molecules. Cell adhesion is of known importance in cancer via changes in signalling and migration. Loss of cell-cell adhesion can result in increased adhesion of cells to the ECM, promoting migration, proliferation and invasiveness in cancer[175]. However, cholesteatoma not malignant and cancers arising from cholesteatoma tissue are exceedingly rare[76] and likely coincidental.

Immune response

Cholesteatoma shows upregulation of inflammatory proteins: some of the most consistently reported differentially expressed genes have inflammatory roles including the SERPINs and S100A proteins. Enriched RAGE receptor binding and TLR-4 binding amongst upregulated genes compared to skin and mucosa were associated with *S100A7*, *S100A8* and *S100A9*, which may act as damage-associated activators of innate immunity[176]. Furthermore, some matrix-active peptidases including the MMPs and cathepsins (such as CTSC) act as inflammatory effectors[161,177]

Given that cholesteatoma is surrounded by an inflamed perimatrix and often preceded by chronic otitis media, upregulation of inflammatory genes compared to healthy tissues is not surprising. Sustained inflammation could be a response to ongoing tissue damage caused by the expanding cyst and may contribute to further tissue damage, degradation of the ECM and bone resorption.

Interestingly, many inflammatory proteins upregulated in cholesteatoma (including SERPINB3, S100A7 and SERPINB4) are downregulated in perimatrix compared to ordinary chronic otitis media tissue. This could indicate that these proteins do *not* have a direct role in cholesteatoma pathology but are simply expressed as part of the normal immune response to infection. However, it is also possible that *under-expression* of these proteins contributes to pathology, perhaps representing an inadequate or inappropriate immune response. Under-expression of protease inhibitors is particularly interesting as it may suggest a failure to limit the activity of ECM-active proteases in cholesteatoma.

Metal ion binding

Ion binding, specifically calcium binding, was enriched amongst upregulated DEGs compared to skin. Calcium ion binding was also identified as an enriched process in our previous whole exome study of cholesteatoma variants[126]. The role of ion binding proteins in cholesteatoma is obscure: many proteins have ion-binding function and calcium ions are involved in diverse cellular processes. Possibly, these functions are enriched only because certain families which happen to bind calcium ions (such as the S100A family) are enriched.

2.4.3 Limitations

The different approaches taken by different studies reduces the likelihood of acquiring overlapping results. First, proteomic and transcriptomic studies do not measure identical outcomes due to complex post-transcriptional regulatory systems, meaning there is not a one-to-one relationship between expressed RNAs and proteins[178]. Differences in analytical techniques will also affect which proteins/DEGs are measured depending on the number of probes within a microarray or the number of peptides in a proteome database, probably biasing results towards better-studied genes.

The papers identified in this review had generally low risk of bias in analysis and controlled for confounders using a paired design. However, cases are likely to represent most severe disease and papers generally did not report age, sex or type of cholesteatoma, so may not be representative. Also, control tissue samples taken from middle ears with cholesteatoma may not represent true healthy controls. These issues are difficult to avoid as tissue can only be taken from the middle ear during surgery.

Meta-analysis was not possible because most papers do not report raw data or the *p*-values of all genes tested, only those that are significant. Some genes may be sub-significant in individual studies but meta-analyses could reveal them to be significantly dysregulated, and vice versa. The individual studies are at high risk of type 1 and type 2 error due to their small sample sizes. By identifying the genes which appear in multiple papers, type 1 error is reduced; however, false negatives cannot be accounted for.

GO terms are not a perfect indicator of protein function; many are phylogenetically inferred and data are sometimes incomplete. For example, the MMPs have calcium and zinc ion binding function[179] but were not included in the calcium-ion binding set of genes annotated by g:Profiler; nor did some keratins appear in epidermal development or structural protein categories. GO annotations are not evenly distributed through the genome, with a large number of annotations belonging to a small subset of well-studied genes[180].

This review may have missed some global gene expression studies where this was not the main objective of the study and so was not described within the title or abstract. Most abstracts were explicit when a specific set of genes were being tested only, and I checked the full reports for cases where there was ambiguity. As reviewing was performed by me alone, there may also be some bias in selection of studies based on my interpretation of their abstracts. Although global gene expression studies are hypothesis-free and therefore unlikely to be subject to publication bias due to a negative result, it is possible for a study using a small number of probes to fail to identify any DEGs. A negative result is also possible where global gene expression study formed part of a study with a wider hypothesis, such as the expectation that a certain pathway would be over expressed.

2.5 Conclusion

This was the first systematic review of global differential gene expression studies, which aimed to identify consistently dysregulated genes in cholesteatoma and their associated pathways. 20 DEGs were reported in at least 4 of 9 studies, while 8 were present in 5 or more. *SERPIN*B4, SERPINB3, several *S100A* proteins, *TNXB, CEACAM6* and *COCH* were particularly consistently up- or downregulated and warrant further studies in cholesteatoma. ECM structural proteins and proteases were enriched in both up and downregulated gene sets compared to skin and mucosa, indicating broad ECM breakdown. Inflammatory genes were enriched amongst upregulated genes. Cholesteatoma shows upregulation of genes associated with epidermal differentiation compared to mucosa, though to a lesser extent compared to skin except for some genes in the epidermal differentiation complex.

Dysregulated ECM forms a significant aspect of cholesteatoma biology. ECM dysfunction may affect diverse processes such as cellular migration, primarily through interaction with adhesins, proliferation and differentiation. Migratory processes are relevant to cholesteatoma which may arise through impaired migration of tympanic epithelium associated with a retraction pocket or invasion through a perforation; a weakened basement membrane, possibly associated with protease action, may facilitate invasion. Furthermore, ECM degraders and downregulation of structural proteins may contribute to bone loss. Disrupted cell cycle processes such as proliferation and differentiation may be downstream of ECM dysregulation. Many inflammatory proteins have additional roles in ECM degradation and cellular development so may be central to pathology. Downregulation of certain inflammatory protease inhibitors compared to chronic otitis media tissue could indicate an overtly aggressive immune response and excess ECM and bone degradation.

3 Epidemiology of cholesteatoma in the UK Biobank

3.1 Background

While some risk factors for cholesteatoma are well established, including male sex, association with certain craniofacial dysmorphologies, and comorbid middle ear disease[1,6,7,18,22,181], others are supported by little published epidemiological data. For example, cholesteatoma is generally reported to differ in prevalence between ethnicities but as discussed in the introduction to this thesis, original epidemiological data is lacking. Reviews[1,2,13,14] may report highest incidence in white populations, low incidence in black populations, and rarity in Asian populations, although no original data is presented in the cited articles. A literature search for cholesteatoma epidemiology or prevalence only uncovered Ratnesar (1976)[15], in which it is reported that cholesteatoma in Inuit and Innu populations of Newfoundland is extremely rare despite higher levels of chronic ear disease than white populations located nearby. Meanwhile, Thornton *et al.* (2011)[16] found no difference in cholesteatoma prevalence between ethnicities in children with chronic otitis media in Nepal. This highlights another issue in cholesteatoma epidemiology: it significantly overlaps with other middle ear disease, and the nature of the relationship is uncertain. Does chronic inflammation lead to cholesteatoma development or arise from it? For example, cholesteatoma is more common in persons with orofacial cleft and had a male predominance, but the same is true for otitis media generally[19]. Exposure to cigarette smoke adversely impacts mucociliary function[182,183], and there is evidence that passive exposure raises risk of cholesteatoma[184,185], although this relationship is not always found[186]. Only one study has been performed on smoking and rates of cholesteatoma, finding that smokers had worse outcomes for ontological surgeries and higher rates of cholesteatoma[187].

Computerised databases of medical records in the forms of national registries and biobanks allow retrospective epidemiological studies of large populations. These are observational studies: a cross-sectional study looks at data from a population at a single point in time where both exposure and outcome have already occurred, while a prospective study follows both exposures and outcomes over time. Conversely, experimental studies involve the researcher applying the exposure of interest to an experimental group and comparing the outcome to a control group. As many factors differ non-randomly between cases and controls in observational studies, they constitute a low level of evidence but are relatively quick and easy to perform and make good use of large volumes of existing data[188].

The UK BioBank is a large, ongoing study including 500,000 British participants. Because both lifestyle information drawn from questionnaires and health information is available, this allows for study of demographic features and diseases associated with cholesteatoma. Careful examination of the demographic risk factors shared by and distinguishing cholesteatoma from other middle ear disease will be useful for interpretation of later genetic studies. This will also provide further evidence for risk factors which are colloquially known but have limited published data.

3.1.1 Aims and objectives

This chapter aims to use retrospective data from the UK BioBank (UKBB) to characterise lifetime prevalence and demographic factors associated with cholesteatoma. To determine which risk factors are shared between cholesteatoma and other middle ear disease and which are unique, I compare cholesteatoma demographics to a control group with other middle ear disease as well as disease-free ears. I also use ICD-10 data to identify overlapping diseases and compare the results with statistics from a Finnish biobank, FinnGen. This study follows Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines (**SI Table 3**). The results of these analyses have been published in Clinical Otolaryngology[189].

This chapter also serves as an introduction to the UKBB study population, identification of cholesteatoma cases from ICD-10 and OPC4 codes, and case-control matching as used in the remainder of this thesis. I also introduce the FinnGen biobank which provides both genetic summary statistics and demographics information and offers a comparison cohort for both epidemiological and genetic analyses.

3.2 Methods

3.2.1 Study design and setting

Participants

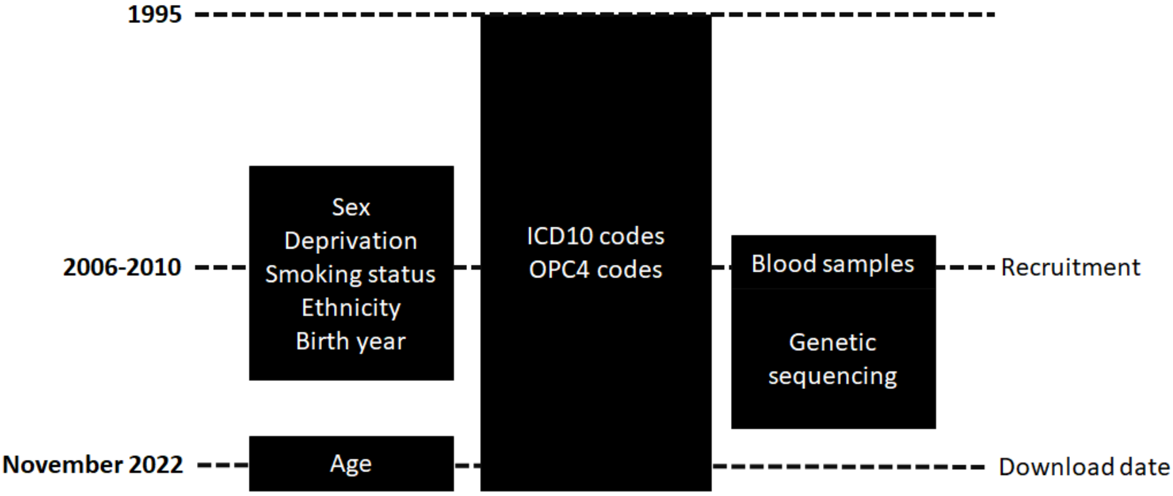
This is a retrospective case-control study using UK BioBank data under project number 61632. UKBB contains lifestyle and health data on 502,408 participants from the United Kingdom, aged 40-69 during the recruitment period 2006-2010. Its rich phenotypic data has made UKBB a valuable and widely used resource for study of genetic and non-genetic health conditions worldwide[190].

Relevant to this and later chapters are four types of data held in UKBB:

- **Medical information** in the form of ICD-10 and OPC-4 codes. The ICD-10 is an internationally standardised classification of diseases, diagnoses and medical findings. The OPC-4 is an equivalent recording operative procedures undergone by a patient. This data is used to identify cholesteatoma cases from medical records and identify overlapping diseases. This includes retrospective information (medical events prior to recruitment) and was updated with new events until the date of download in 2022 (Figure 12).
- **Questionnaire responses**, which include basic personal information such as sex and birth date, as well as demographic and lifestyle features such as economic deprivation and smoking status. These will be used for epidemiological analysis, case-control matching and/or as covariates in statistical models. Data were taken at recruitment (Figure 12).
- **Whole exome genetic data**: 450,000 participants were sequenced with dual-indexed 75 x 75 bp pared-end reads on Illumina NovaSeq 6000. Exomes were captured with the IDT xGen Exome Research panel v1.0. Sequencing was performed in two main batches: the first 50,000 samples used S2 flow cells while the remaining samples used S4 flow cells and a different IDT oligo lot to the initial batch. Whole exome sequences are available in CRAM format as well as in variant call format (VCF) called by DeepVariant[191].
- **Microarray data**: genotyping was performed using the UK BioBank Axiom Array, which directly measures ~850,000 variants. An additional ~90 million variants were imputed

using the Haplotype Reference Consortium and UK110K + 1000 Genomes reference panels[190]. This data covers the whole genome but cannot detect rare variants. The microarray data as provided has been filtered to remove markers that failed quality control, have a missingness of >5% and a MAF<0.0001[190].

Figure 12. Date of data collection for baseline, medical and genetic data.
Date of data collection for different fields. Baseline statistics such as sex, deprivation and smoking status were taken at recruitment. Medical information was taken at recruitment and extends retrospectively, although there are few records before 1995, and were continually updated until the date of download. Genetic sequencing was performed on samples taken at recruitment. The initial 50,000 WES sequences were released in 2019 and the complete set were released by 2022.



While the aim of UKBB is to provide data for the study of common diseases of later life, the large number of participants also makes it ideal for studying rarer diseases. Cholesteatoma is common enough that UKBB contains ~1,000 cases, a number which would otherwise be difficult to recruit given its low annual incidence. Furthermore, the cost of performing whole-exome sequencing for these participants would be prohibitively expensive if performed specifically for this project. Although cholesteatoma is not as common as other conditions studied by UKBB, it is still an important cause of acquired hearing loss with potentially serious complications, thus is in line with their goals of improving public health.

Ethical approval

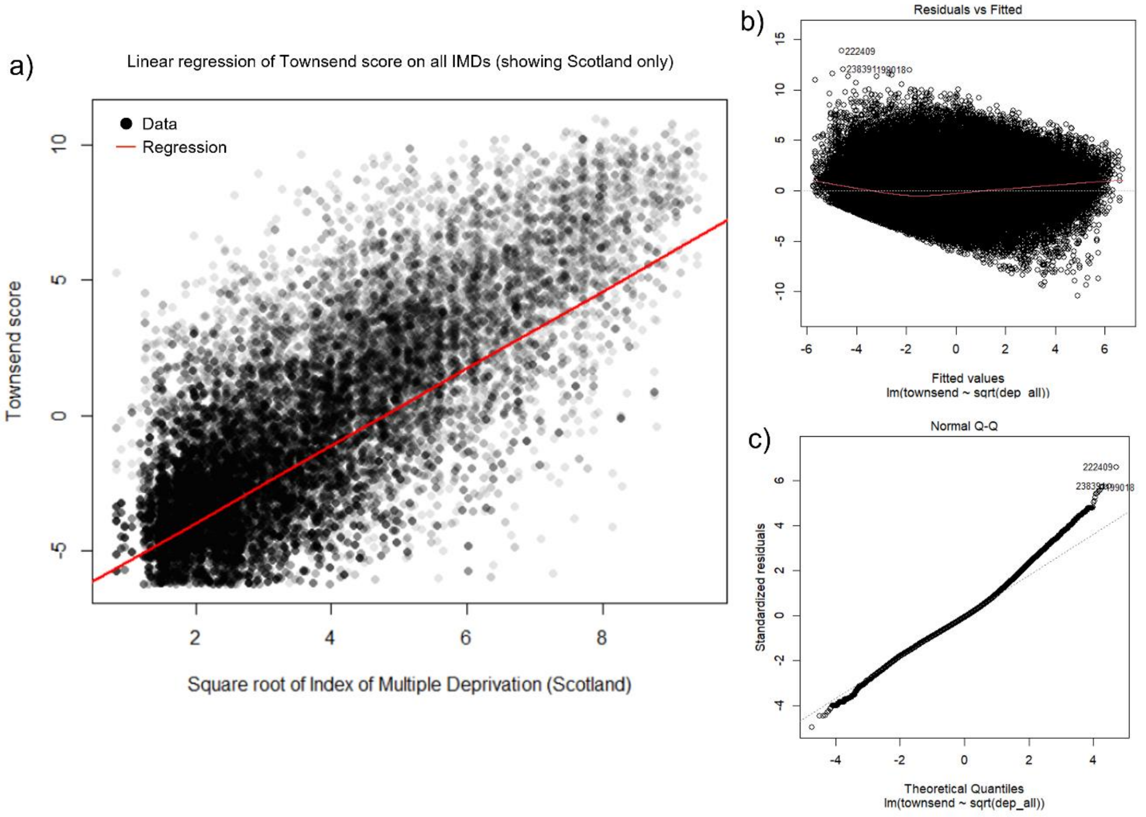
UK Biobank has obtained Research Tissue Bank (RTB) approval from its the Research Ethics Committee (approval number 16/NW/0274). Researchers can acquire UK BioBank data by registering for access: (https://www.ukbiobank.ac.uk/enable-your-research/register). Publicly available demographic data and statistical results from FinnGen release 9 were accessed via Risteys (https://r9.risteys.finngen.fi/endpoints/H8_CHOLEASTOMA) in June 2023.

Variables and processing of missing data

Key demographic features used in this analysis are as follows:

- **Sex**: There is a slight female bias in the UKBB cohort with a ratio of 1.19 females per male. One participant was missing sex data and was not included in this analysis, as all other covariates were also missing.
- **Ethnicity**: UKBB is majority white ethnicity (456,284 participants). 893 participants were missing ethnicity information. These were grouped with do not know, any other ethnicity and prefer not to say into a single 'Other/Unknown' category.
- **Age**: UKBB participants were between the ages of 51 and 85 (median 72) at the time of data download in November 2022.
- **Smoking status**: Smoking data is a composite category of questionnaire results and indicates whether a person has ever smoked, regardless of frequency or discontinuation. 298,711 (59.5%) of all participants indicated a history of smoking while data were missing for 2,885 (0.57%) participants. For subsequent analyses, except for those specifically investigating smoking as a risk factor, these were assigned to an 'unknown' category.
- **Deprivation:** deprivation data is available in the forms of Townsend deprivation index and regional indices of multiple deprivation (IMDs). Townsend deprivation index is a measure of postcode deprivation based on employment rate, car and house ownership, and household crowdedness[192]. IMD is a measure of deprivation calculated by the governments of Scotland, Wales, England and Northern Ireland combining similar metrics such as income, employment and crime[193]. As each nation calculates IMD slightly differently, I use Townsend index as the primary measure of deprivation. 624 samples were missing Townsend deprivation data. Where possible, deprivation was imputed using IMDs (available for 489,674 participants) by fitting a linear regression to the square root of IMD, $Townsend = m\sqrt{IMD} + c$, where m was found to be 1.431399 and c = -6.852356 (R2 = 0.5306). This regression was used to generate Townsend scores for those participants where only IMDs were available (**Figure 13**).

Figure 13. Imputation of missing Townsend indices from indices of multiple deprivation. *a) An example with the Scotland data showing the regression model fit to all Indices of Multiple Deprivation (IMDs). B) the residuals for all data used in the regression. C) Q-Q plot showing divergence from normality of residuals towards the upper end. Though the relationship between these indices was very noisy and did not have a perfectly linear relationship (either for IMD of square root of IMD), the small number of samples (0.12%) missing Townsend deprivation data meant that this would not affect overall distribution much and was preferable to mean imputing missing values. Only 23 participants were missing both measures of deprivation, and these were mean imputed.*



3.2.2 Case and control selection for epidemiology and genetic testing

Identification of cholesteatoma cases

UKBB data includes a list of ICD-10, ICD-9 and OPC-4 codes included in patient records. IPC-9 and -10 are two recent versions of a system for assigning codes to clinical diagnoses, while OPC-4 is a recent system for identification of surgical procedures. These codes are assigned by hospitals for billing purposes. I accessed the OPC-4.9 via the NHS website (classbrowser.nhs.uk), the ICD-10 via the WHO website (icd.who.int) and an archived copy of the ICD-9 from the CDC (cdc.gov).

Billing codes are not directly assigned by doctors, so may not be completely accurate representations of diagnoses. Some codes, such as H71 cholesteatoma, are unambiguous and very likely to correctly identify a true case. However, this may not capture all cases if records are incomplete or inaccurate. Missing or inaccurate records may arise due to misdiagnosis, inaccurate translation into ICD/OPC codes, inaccurate translation from previous versions of the ICD or during transfer from physical to electronic form. Cholesteatoma is a rare disease, so it is vital that all possible cases are identified to increase numbers.

Therefore, I expanded case criteria to include likely cases based on additional codes with guidance from clinicians Carl Philpott and Peter Prinsley. Because it is difficult to separate cholesteatoma from other middle ear conditions which often co-occur, the control group is filtered to exclude all individuals with middle ear disease. For epidemiological comparison, I also selected a cohort of non-cholesteatoma middle ear disease participants who were not part of the case group but who had any other ear disease (**Table 9**). This includes some non-middle ear disease codes including otitis externa, otalgia and effusion. Otalgia and effusion are taken as indicators of underlying inflammation or disease of the ear. While otitis externa is not a middle ear disease, inflammation of the external auditory canal is likely to affect the tympanic membrane which is a likely origin of cholesteatoma tissue. Furthermore, the boundary between otitis externa and middle ear disease becomes less clear when there is tympanic perforation. Symptoms of middle ear inflammation and otitis externa may be difficult to discern and so it possible for an otitis media case to actually have middle ear inflammation, hence these participants cannot reliably be considered middle ear disease free.

In this study, an individual is considered a cholesteatoma case if they meet the following criteria, set out in **Table 9**:

- They have one of the **confirmed** codes, which unambiguously specify cholesteatoma.
- They have one of the **suspected** OPC-4 codes, which indicate surgeries most likely performed to treat cholesteatoma, but not a **suspected exclude** code which offer plausible alternative explanations for these procedures.

- They have one of the **mastoid** ICD-10 codes but not a **mastoid exclude** code. This includes individuals with chronic mastoiditis likely to be caused by cholesteatoma with no acute explanation.

This method increased the number of cholesteatoma cases from 654 to 1,151. Data were not granular enough to distinguish between congenital and acquired cholesteatoma; no distinction was made based on age of onset on severity in this analysis.

Table 9. Case inclusion and exclusion criteria with rationales

Filter	Code	Meaning	Rationale
Confirmed	H71	Cholesteatoma of the middle ear	Unambiguous codes specify cholesteatoma
	H95.0	Recurrent cholesteatoma of postmastoidectomy cavity	
	Date H71 diagnosed		A separate column which contained some additional cases
Mastoiditis	H70.1	Chronic mastoiditis	Chronic mastoiditis most likely to result from cholesteatoma
	H70.9	Unspecified mastoiditis	
Suspected	D10.1	Radical mastoidectomy NEC	Surgeries used primarily for treatment of cholesteatoma and few other conditions (see exclude filter)
	D10.2	Modified radical mastoidectomy	
	D10.6	Revision mastoidectomy	
	D10.8	Other specified exenteration of mastoid	
	D10.9	Other unspecified exenteration of mastoid	
	D12.4	Exploration of mastoid	
	D12.1	Obliteration of mastoid	
	D12.2	Atticotomy	
	D12.7	Atticoantrostomy	
	D10.5	Excision of lesion of mastoid	Probably indicates removal of cholesteatoma due to few other lesions affecting mastoid.
Suspected exclude	D33.3	Benign neoplasm of cranial nerve	May indicate acoustic neuroma, alternative explanation for mastoidectomy.
	H93.3	Disorder of acoustic nerve	
	H70.0	Acute mastoiditis	Possible cause for mastoidectomy without cholesteatoma.
	H81.0	Meniere disease	
	D02.3	Middle ear carcinoma	Alternative explanation for excision of lesion of mastoid.
	D38.5	neoplasm of uncertain behaviour	
	C30.1	Malignant neoplasm of middle ear	
	D16.9	Benign neoplasm of bone and articular cartilage	To capture osteoma, alternative explanation for excision of lesion of mastoid.
	H65.0	Acute serous otitis media	Exclude acute cases from mastoiditis group.

Filter	Code	Meaning	Rationale
Mastoid exclude	H65.1	Other acute nonsuppurative otitis media	
	H66.0	Acute suppurative otitis media	
Other ear disease	H65-H75	Diseases of middle ear and mastoid	Include any middle ear disease
	H60	Otitis externa	Inflammatory ear disease closely related to middle ear disease
	H92	Otalgia and effusion of ear	Ear pain and discharge suggests underlying ear disease

Propensity matching

As the number of cases within UKBB is only 1,151, the case:control ratio is approximately 1:500. When uncontrolled, sample imbalance can result in biased models with large type 1 error rates[194]. Case matching to reduce this ratio was therefore performed, primarily for the later GWAS section of this thesis, but matched cases and controls were also used in epidemiological analyses. Because the same matching system was to be used for both epidemiological and genetics testing, samples were only included if they passed the following basic genetic quality controls: genetic data was available; there was no sex chromosome aneuploidy; stated sex matches measured chromosomal sex; and no close relatives were present.

I used the MatchIt[195] package for R to perform case/control matching. MatchIt first calculates propensity scores for each person which is the likelihood of being a case based on covariate values alone. Propensity can be calculated in a number of ways, but logistic regression in a common approach. A regression of outcome against covariates is performed and the predicted outcome (which for a binary variable is essentially the percent likelihood of being in the case group) is the propensity score. Matching is performed on propensity score rather than the covariates directly, which efficiently creates case and control groups of the desired ratio. The individual covariates may or may not be as well balanced as the total propensity score, which takes into account the fact that covariates contribute unequally to overall propensity.

I trialled various matching options in MatchIt and present six high-performing methods:

- **Method 1:** nearest neighbour propensity matching on data with estimated ancestry groupings (see *Appendix: Ancestry Estimation*)
- **Method 2**: the same but with exact matching for all but deprivation

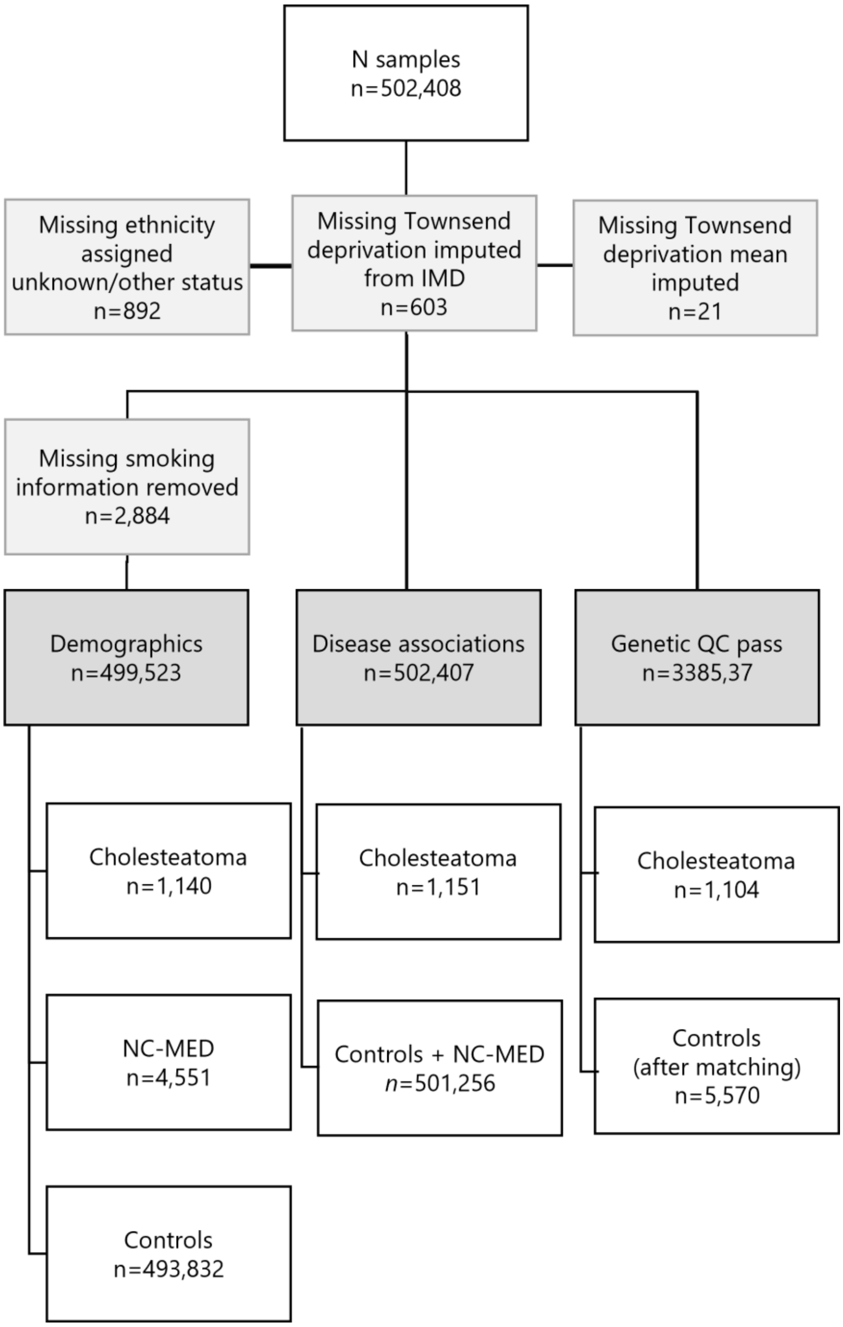
- **Method 3**: Exact matching on sex, smoking, ethnicity and age using UKBB ethnicities to the subgroup level
- **Method 4:** Exact matching on sex, smoking and ethnicity using UKBB ethnicities to the subgroup level
- **Method 5:** Exact matching on sex, ethnicity and age using UKBB ethnicities to the subgroup level
- **Method 6:** exact matching on sex and ethnicity using UKBB ethnicities to the subgroup level

For all subsequent matched analyses, I used method 6 as it was found to give the best improvement in balance with no loss of cases or controls (*see Assessment of matching performance*).

3.2.3 Final division into case, control and non-cholesteatoma middle ear disease for epidemiological analysis

For this section of the thesis, all ethnicities were retained and data were divided into cases, controls and non-cholesteatoma middle ear disease using the criteria outlined in *Case and control selection for epidemiology and genetic testing***.** For association testing of demographic risk factors, the full set of data were used. For associations with other disease codes, cholesteatoma was compared to all controls and non-cholesteatoma middle ear diseases cases. Matched data were used for some validation analyses to test for the effects of case-control imbalance (**Figure 14**). Case numbers are illustrated in, which also shows handling of missing data.

Figure 14. participant numbers and missingness information for cases and controls.
One participant was excluded for missing all covariates. Deprivation was imputed from indices of multiple deprivation (IMD) where available. Where not available, Townsend deprivation index was mean imputed. 2,884 cases with missing smoking data were excluded from demographic analysis.



3.2.4 FinnGen comparison cohort

FinnGen is a biobank containing ~500,000 samples from the Finnish population. Summary statistics for single-variant GWAS are released every six months as the number of samples sequenced increases. Phenotypes are determined by ICD-10, 9 and 8 codes and genotyping is performed using a GRCH37-aligned Thermo Fisher axiom genotype array, including ~500,000 core GWAS markers and an additional ~200,000 markers enriched in the Finnish population or of special clinical interest[196].

Concurrent with these releases, phenotype and demographic information is released via the Risteys platform (https://risteys.finregistry.fi/). This includes endpoint definition, descriptive statistics such as age distribution and sex ratio, overlapping disease endpoints and Cox Hazard regressions for other disease endpoints. Unlike UKBB, FinnGen contains individuals of all ages.

Release 9 sample size and endpoint definitions

For all analyses in this thesis, release 9 data were used. Cholesteatoma was defined as containing ICD-10 H71, ICD-19 3853 or ICD-8 38700 as a hospital discharge code or cause of death. Controls were defined by individuals containing no ICD-10, 9 or 8 codes for any condition of the middle ear or mastoid. This is very similar to my UKBB definitions but does not use OPC codes and does use older definitions of the ICD, where abscess of the middle ear was included under the same definition as cholesteatoma. The number of cases was 1447, of which the majority (1167) were defined by ICD-10 H71. The number of middle ear disease-free controls was 376,139.

Calculation of cox hazard regressions

Cox Hazard regressions for FinnGen data were calcuilated by Risteys by selecting a random sample of 10,000 individuals from the pool of cases and controls (https://r9.risteys.finngen.fi/documentation).The start of follow-up was set to 1998 due to good coverage for all registries after this date. End of follow up was date of death of 31/12/2019.

3.2.5 Statistical analyses

Logistic regressions for demographic factors

I tested for associations with demographic factors for cholesteatoma vs control and cholesteatoma vs other middle ear disease by fitting logistic regression models using the *fitglm* function with binomial error distribution and logit link function in MATLAB R2020b[197]. Cholesteatoma was assigned the binary status 1 and control/other middle ear disease was assigned a status of 0. To obtain adjusted odds ratios, I included age, sex, smoking status, deprivation, and ethnicity as covariates. 1,140 cholesteatoma cases, 4,551 other middle ear disease cases and 493,832 controls were used. 2,884 individuals with missing smoking data were excluded.

There is a large case-control imbalance, particularly for comparison of cholesteatoma to controls. Imbalance can bias regression estimates and inflate *p*-values; to test for this effect, I calculated *un*adjusted odds ratios for each covariate by testing them individually on data matched for the remaining covariates. Matching was performed using the package MatchIt (version 4.4.0)[195] in R 4.1.3[198] with 'method 6' outlined in *Propensity matching:* exact matching was used for sex and ethnicity and propensity score-based nearest neighbour matching for all remaining covariates. The regression includes no covariates; the matching process should have a similar effect to adjusting for the other covariates.

Logistic regressions for disease-disease associations

I also performed pairwise logistic regressions to test for association between ICD-10 codes. I collapsed ICD-10 codes to their parent code and removed non-relevant codes, such as those for medications or accidents (codes starting V, X, Y, Z, S, T or R). For these tests, I compared cholesteatoma to unmatched controls comprising the disease-free and other middle ear disease groups. Because I did not exclude missing smoking data for these tests, 1,151 cases and 501,256 controls were included. Each ICD-10 code was conditioned as a presence-absence binary status with cholesteatoma as a binary outcome and I again used *fitglm* to test for associations between each ICD-10 code and cholesteatoma, adjusting only for age and sex. I tested 1,312 codes, though only 751 codes had any overlap with cholesteatoma. The number of cases for each code varied between 1 and 151,022.

Cox Hazard regression for comparison to FinnGen

I performed Cox Hazard regressions on the set of codes with significant hazard ratios in FinnGen. UKBB cholesteatoma cases with time to event information (*n*=650) were compared to ear disease-free controls (*n*=496,667) with age and sex used as covariates. The start date was set to 1995 due to poor ICD-10 coverage before this date, and the end date was November 2022. Both hazards before and after cholesteatoma were calculated using the MATLAB function *coxphfit*, using age and sex as covariates.

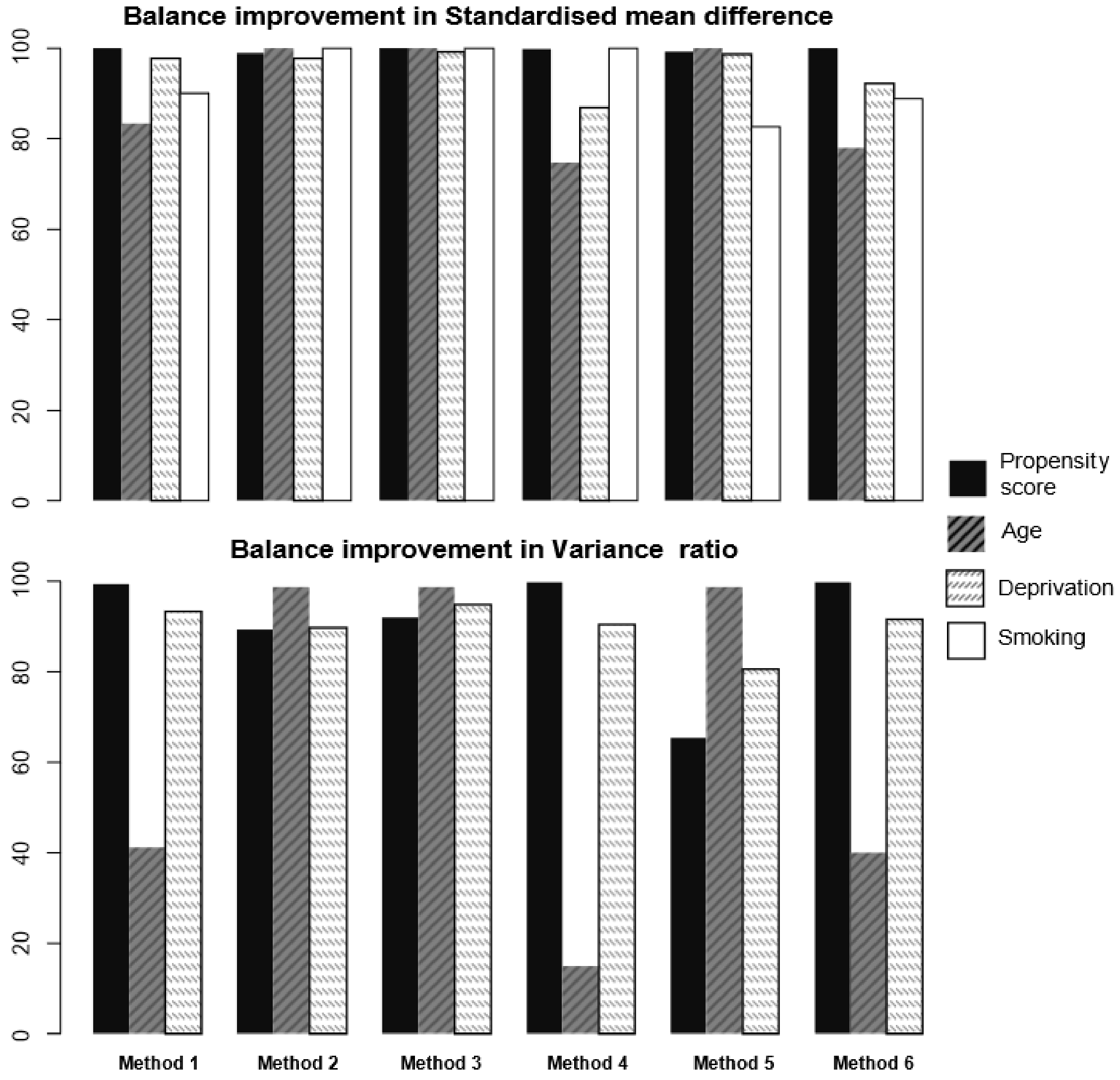
3.3 Results

3.3.1 Propensity matching

Assessment of matching performance

I assessed the performance of six matching methods, focusing on balance of the standardised mean (the mean adjusted so that covariates with different scales are comparable) and variance ratio. Good improvements for propensity score were seen for all methods, mostly varying in the degree of balance improvement for the individual covariates (**Figure 15, Table 10**).

Figure 15. Balance improvement for standardised mean difference and variance ratio for six matching methods



Exact		Age, sex, ancestry, smoking	Ethnicity, sex, age, smoking	Sex, smoking, ethnicity	Sex, Age, ethnicity	Ethnicity, sex
Nearest Neighbour	All	Deprivation	Deprivation	Age, deprivation	Smoking, deprivation	Age, smoking, deprivation

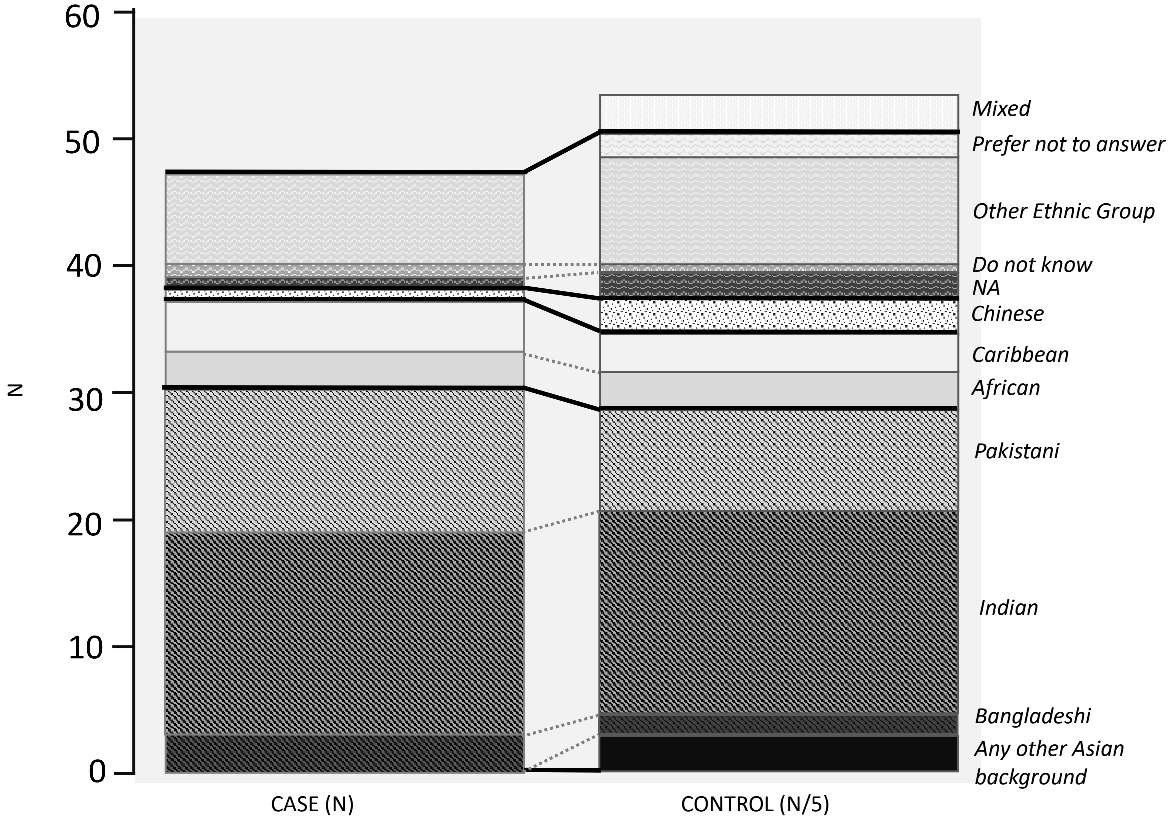
Table 10. Balance improvement for standardised mean difference (SMD) and variance ratio for matching methods

		Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
SMD	Distance	100.00	98.85	100.00	99.93	99.17	99.98
	Age	83.30	100.00	100.00	74.69	100.00	78.11
	Deprivation	97.91	97.82	99.08	87.00	98.66	92.30
	Smoking	73.04	100.00	100.00	100.00	82.61	88.85
Variance ratio	Distance	99.434	89.27	92.02	99.78	65.25	99.78
	Age	41.13	98.63	98.63	14.80	98.63	39.93
	Deprivation	93.27	89.83	95.04	90.38	80.62	91.51

Methods 1 and 2 used genetically estimated ancestry rather than ethnicity (see *Appendix: Ancestry Estimation*). Estimates were drawn from K-means clustering of the first 3 principal components. Method 2 outperformed method 1 with the better improvement in balance for most covariates, except for the overall propensity score. I then compared the original values for ethnicity, as matching was performed on ancestry estimates instead. This was to check if ethnicity was also well-balanced by this method. The matching is good but not 1:1 (**Figure 16**). I decided not to use estimated ancestry in my final case selection.

Method 3 performed very well. Propensity score SMD was improved 100% and variance ratio 92%. However, two cases could be found controls and not all cases were matched to 5 controls. 5548 controls were chosen, a ratio of 4.89 – slightly under the target of 5, meaning not all cases could be found 5 controls. Likewise, method 5 was unable to match 5 controls to each case and performed poorly for propensity score variance ratio improvement.

Figure 16. Ethnic composition of cases and controls when matched with ancestry



Methods 6 and 4 were able to match all cases to 5 controls. Method 6 outperformed method 4 and had one of the best overall balance improvement scores (99.98% for standardized mean, 99.78% for variance ratio) and balanced the individual covariates well (**Table 11**). Therefore, I chose to use method 6. After propensity matching with this method, propensity score distributions were almost identical (**Figure 17**).

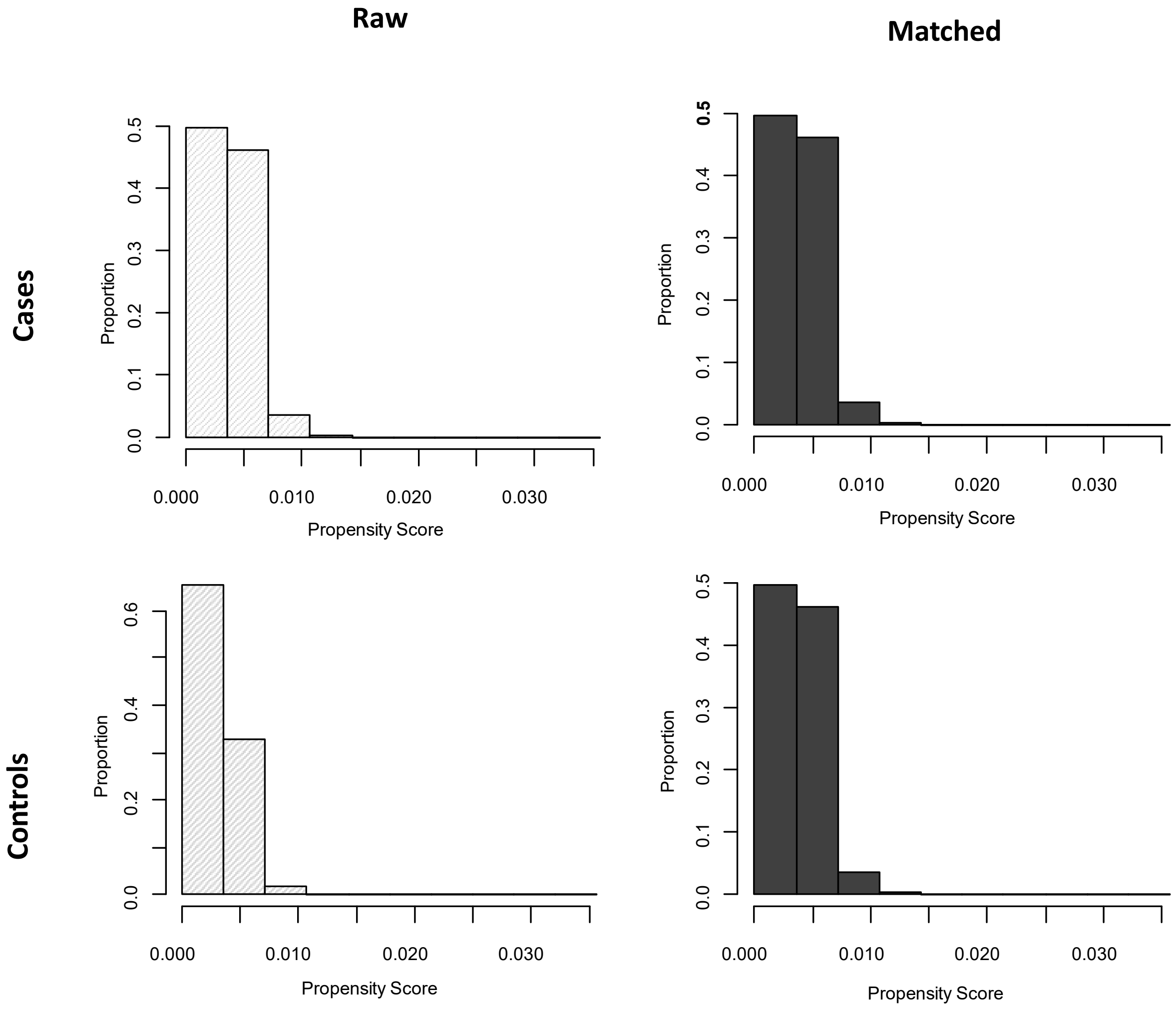
Table 11. Balance before and after matching for method 6. Variance ratio improvement cannot be calculated for categorical variables sex, ethnicity and smoking.

Balance improvement in is excellent for propensity score and most sub-categories except for age, where the initial difference was very small and the final similarity in distributions satisfactory.

	Standardised mean difference			Variance ratio		
	Before	After	% Improvement	Before	After	Improvement
Propensity score	0.370	-0.0001	100.0	1.217	0.9996	99.8
Sex	0.1526	0.0000	100			
Ethnicity	-0.0122	0.0000	100			
Smoking	0.0114	-0.00054	88.82			
Age	0.171	-0.0374	78.1	0.949	1.0319	39.9
Deprivation	0.200	0.0154	92.3	1.1441	0.9886	91.5

Figure 17. Balance improvement in propensity score distribution for method 6.

Cases and controls did not differ radically in propensity score distribution before matching (left). However, their distributions are almost identical after matching (right).



3.3.2 Demographics of cholesteatoma and other middle ear disease

Total prevalence of cholesteatoma in UKBB was 0.22%, corresponding to approximately 1 in 500 people. Prevalence was higher in males with a male:female ratio of 1:1.35. In comparison, the prevalence in FinnGen was 0.38% with a male:female ratio of 1:1.67. The median age of cholesteatoma and other middle ear disease cohorts in UKBB was 73 (IQR=12 for both), while the median age of controls was 72 (IQR=13). The median deprivation index of the cholesteatoma cohort was -1.40 (IQR=5.14); other middle ear disease -1.70 (IQR=4.82); and the controls -2.14 (IQR=4.18), where the higher scores indicate most deprivation.

Significant associations with cholesteatoma incidence were found for sex (male AOR=1.33, p<0.001), deprivation (AOR=1.08, p<0.001), age (AOR=1.02, p<0.001), Black ethnicity (AOR=0.35, p=0.0035), and other/unknown ethnicity (AOR 0.48, p=0.042) (**Table 12**). The ORs obtained in sensitivity analysis for demographic factors generally agreed with the AORs obtained from the unmatched data except for the other/unknown ethnicity, showing that imbalance did not greatly affect the results (**Table 13**).

Table 12. Descriptive and inferential statistics of cholesteatoma in the UK Biobank. *Prevalence and number of cases by demographic is shown alongside the total number of each demographic within the entire UK BioBank cohort. Adjusted odds ratios (AORs) and p-values acquired from logistic regression of demographic factors on case status compared to a middle ear disease-free control cohort and a non-cholesteatoma ear disease cohort are also shown. Bold indicates the comparison category.*

				Versus disease-free controls			Versus other ear disease		
	Prevalence (%)	N cases	N total	AOR	95% CI	p	AOR	95% CI	p
Total	0.22	1,151	502,407						
Female	**0.20**	**533**	**271,839**						
Male	0.27	607	227,684	1.33	1.179, 1.491	<0.001	1.30	1.142, 1.486	<0.001
White	**0.23**	**1,093**	**470,982**						
Mixed	0	0	2,940	0.000	0, Inf	1	0.00	0, Inf	1
Asian	0.31	30	9,769	1.237	0.857, 1.787	0.26	0.84	0.56, 1.268	0.41
Black	0.10	8	7998	0.352	0.175, 0.71	0.0015	0.60	0.28, 1.28	0.19
Chinese	0.06	1	1,569	0.287	0.04, 2.041	0.21	0.17	0.023, 1.283	0.086
Other/unknown	0.13	8	6,265	0.484	0.241, 0.973	0.042	0.45	0.212, 0.935	0.033
Non-smokers	**0.21**	**420**	**200,812**						
Smokers	0.24	720	298,711	1.06	0.934, 1.194	0.38	0.98	**0.856, 1.128**	0.80
Deprivation	--	--		**1.08**	**1.059, 1.097**	**<0.001**	**1.02**	**1.001, 1.042**	**0.040**
Age	--	--		**1.02**	**1.011, 1.026**	**<0.001**	**1**	**0.99, 1.007**	**0.69**

Comparing cholesteatoma to other middle ear disease showed a similar male bias with an AOR of 1.3 (*p*<0.001), meaning the cholesteatoma group differed from other middle ear disease about as much as it differed from the controls (AOR = 1.33; Cases vs control; **Figure 18b**). The cholesteatoma and other ear disease cohorts did not differ significantly in age or smoking status (*p*>0.05) (**Figure 18a,c**). Deprivation was significantly associated with cholesteatoma but to a lesser extent than when compared to healthy ears (AOR 1.02, *p*=0.040) and the other middle ear disease group had a more similar distribution of deprivation to the cholesteatoma group than the controls (**Figure 18e**).

Table 13. Results of matched logistic regressions testing demographic factors

Test	P value	OR
Sex	0.00028	1.267
White	0.0056	1.592
Mixed	--	1.000
Asian	0.98	0.994
Black	0.00033	0.270
Chinese	0.093	0.180
Other	0.85	10.93
Smoking	0.18	1.096
Deprivation	3.57x10-15	1.079
Age	1.02x10-5	1.019

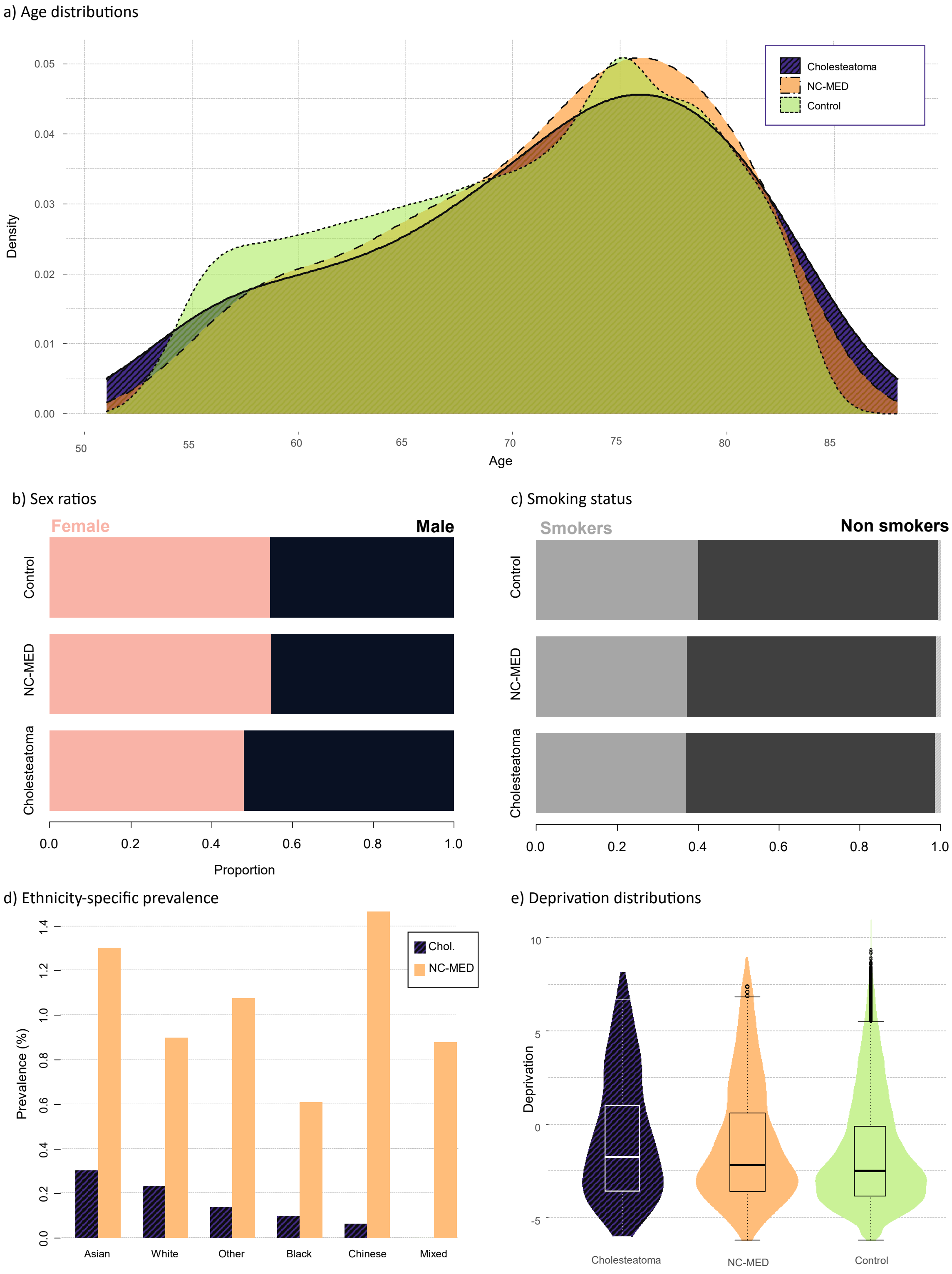
Comparison of other middle ear disease to disease-free controls

I also compared other middle ear disease to disease-free controls. There was a significant difference in age and deprivation with similar odds ratios to the cholesteatoma vs control comparison (**Table 14**). Although smoking was not significant when comparing cholesteatoma to disease-free controls, there was a significant difference for other middle ear disease vs controls with a similar odds ratio. The cholesteatoma and other middle ear disease groups had similar proportions of smokers and non-smokers (**Figure 18**), so significance may be due to the larger size of the other middle ear disease group. Meanwhile, the effect of ethnicity was difficult to quantify due to small sample sizes. The Asian and Chinese groups had higher odds of non-cholesteatoma middle ear disease compared to controls, but their odds of cholesteatoma were not significantly increased. This may also be due to the differing sample sizes of cholesteatoma and other middle ear disease.

Table 14. Logistic regression results for comparison of middle ear disease to middle ear disease-free controls

	Prevalence (%)	N	AOR	95% CI	p
Total	0.91	4589			
Female	**0.9**	**2,450**			
Male	93	2,139	1.02	0.963, 1.083	0.49
White	**0.91**	**4,283**			
Mixed	0.88	26	0.97	0.656, 1.427	0.87
Asian	1.31	130	1.46	1.222, 1.746	$3.23\text{x}10^{-5}$
Black	0.6	49	0.59	0.444, 0.791	$3.88\text{x}10^{-4}$
Chinese	1.46	23	1.714	1.133, 2.592	0.0107
Unknown	1.06	78	1.0595	0.825, 1.36	0.650
Non-smokers	**0.86**	**1,720**			
Smokers	0.94	2,831	1.07		0.038
Age			**1.02**	**1.016, 1.024**	$\mathbf{2.97\text{x}10^{-25}}$
Deprivation			**1.0557**	**1.046, 1.065**	$\mathbf{2.28\text{x}10^{-31}}$

Figure 18. Demographics of cholesteatoma and non-cholesteatoma middle ear disease



a) Age distributions

b) Sex ratios

c) Smoking status

d) Ethnicity-specific prevalence

e) Deprivation distributions

Demographics of unmatched cholesteatoma cases, non-cholesteatoma middle ear disease (NC-MED) and ear disease-free controls showing a) age distributions, b) sex ratios, c) smoking status, d) prevalence of cholesteatoma and non-cholesteatoma ear disease by ethnicity, and e) Townsend Deprivation index distribution. Plots generated in R using ggplot2 package.

3.3.3 ICD-10 associations with cholesteatoma

56 ICD-10 codes were significantly associated with cholesteatoma after Bonferroni multiple testing correction (*p*<0.05; **Table 15**). Hierarchical clustering of Jaccard distance between these codes reveals two main groups: common conditions, including *chronic obstructive pulmonary disease* (OR=2.03), *disorders of lipoprotein metabolism & other lipidaemia* (OR=1.48), and *gastro-oesophageal reflux* (OR=1.51); and diseases of the sinuses and middle ear and their complications (**Figure 19**).

The strongest associations (all p-values < 0.001) were with *other disorders of middle ear and mastoid* (OR=242.12), *other disorders of tympanic membrane* (OR=163.63), *suppurative and unspecified otitis media* (OR = 144.54), *otalgia and effusion of the ear* (OR =78.22, p<0.001) and *perforation of tympanic membrane* (OR 76.07). The overlap with cholesteatoma for each of these conditions was 100-263 persons. *Otitis externa* (OR=34.97, p<0.001) and *other diseases of inner ear* (OR=6.04, p<0.001) were also associated with cholesteatoma. Known complications of cholesteatoma were also strongly associated, including *sensorineural and conductive hearing loss* (OR=25.81), *other hearing loss* (OR=9.94), *facial nerve disorders* (OR=10.10), and *bacterial meningitis* (OR=41.78).

The most strongly associated non-ear code not known to be a cholesteatoma complication was F17, *mental and behavioural disorders due to tobacco use* with an OR of 2.34 (*p*<0.001), followed by *chronic sinusitis*, with an OR of 4.09 (*p*<0.001). Several other respiratory conditions are represented including *chronic obstructive pulmonary disease, asthma*, and chronic rhinitis with and without nasal polyps. Some congenital anomalies affecting the ear and head (Q17, Q16, Q75, Q96) were also strongly associated with cholesteatoma (OR=33.58-83.24, *p*≤0.011), although the number of overlapping cases was small (*n*=2-4).

Figure 19. Heatmap of Jaccard distance for ICD-10 codes significantly associated with cholesteatoma with hierarchical clustering.

Jaccard distance between ICD-10 codes where odds ratio adjusted p-value for cholesteatoma association < 0.05. Colour scale indicates Jaccard distance with diagonals coloured blue. Hierarchical clustering using unweighted average distance/UPGMA (left) shows two main groups of disease codes: common diseases (cyan) and sinus/middle ear infections with their rare complications (red). Table 2 contains ICD-10 code full names and statistics for all codes with adjusted p-value < 0.05. Codes marked with an asterisk (*) have child codes used in the definition of the case group.

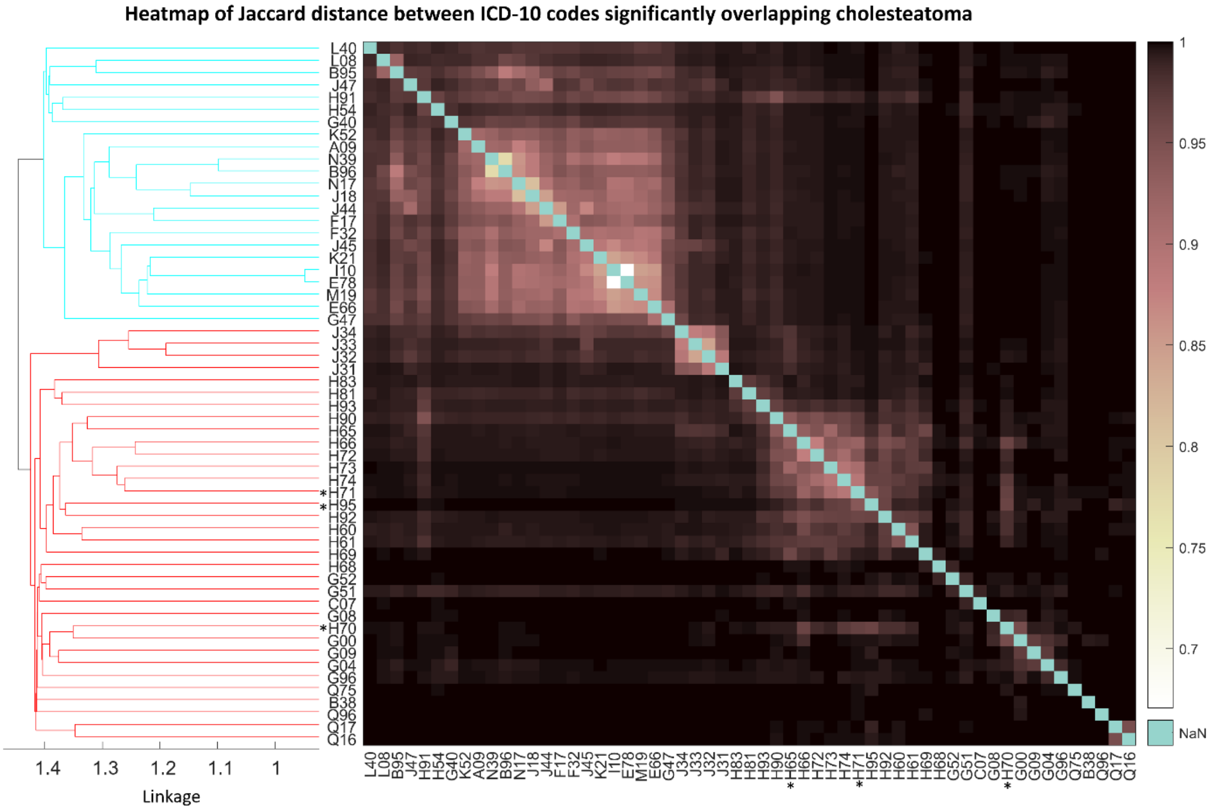


Table 15. Table of ICD-10 codes significantly associated with cholesteatoma.

Code		N	N with Cholesteatoma	Odds Ratio	95% CI	Adjusted p-value	% Overlap
H66	Suppurative and unspecified otitis media	1275	263	144.54	124.259, 168.125	<0.001	12.16
H72	Perforation of tympanic membrane	1504	193	76.07	64.549, 89.646	<0.001	7.84
H73	Other disorders of tympanic membrane	570	142	163.63	133.92 9, 199.912	<0.001	8.99
H74	Other disorders of middle ear and mastoid	695	219	242.12	203.651, 287.867	<0.001	13.46
H92	Otalgia and effusion of ear	911	126	78.22	64.143, 95.383	<0.001	6.51
H95*	Postprocedural disorders of ear and mastoid NEC (included in cholesteatoma definition)	162	117	1249.85	880.243, 1774.661	<0.001	9.78
H90	Conductive and sensorineural hearing loss	2961	152	25.81	21.65, 30.774	<0.001	3.84
H70*	Mastoiditis	181	153	2809.55	1865.538, 4231.256	<0.001	
H65	Nonsuppurative otitis media	1587	100	30.80	24.924, 38.073	<0.001	3.79
H91	Other hearing loss	10920	213	9.94	8.514, 11.608	<0.001	1.80
H60	Otitis externa	1048	75	34.97	27.429, 44.577	<0.001	3.53
H61	Other disorders of external ear	1741	88	23.73	18.972, 29.683	<0.001	3.14
H93	Other disorders of ear NEC	1874	45	10.65	7.875, 14.407	<0.001	1.51
G51	Facial Nerve disorders	1848	42	10.10	7.391, 13.793	<0.001	1.42
G00	Bacterial meningitis NEC	186	16	41.78	24.919, 70.056	<0.001	1.21
F17	Mental and behavioural disorders due to tobacco use	24886	127	2.34	1.942, 2.813	<0.001	0.49
J32	Chronic sinusitis	4446	41	4.09	2.993, 5.6	<0.001	0.74
H83	Other diseases of inner ear	1591	22	6.04	3.946, 9.235	<0.001	0.81
Q16	Congenital malformations of ear causing impairment of hearing	25	4	82.05	28.032, 240.146	<0.001	0.34
G96	Other disorders of central nervous system	562	12	9.55	5.371, 16.968	<0.001	0.71
H69	Other disorders of Eustachian tube	317	9	12.99	6.678, 25.288	<0.001	0.62
I10	Essential (primary) hypertension	151022	478	1.53	1.355, 1.737	<0.001	0.32
J44	Chronic obstructive pulmonary disease	21261	103	2.03	1.653, 2.495	<0.001	0.46
J45	Asthma	47150	172	1.71	1.454, 2.013	<0.001	0.36
Q17	Other congenital malformations of ear	42	3	36.73	11.312, 119.238	<0.001	0.25
G04	Encephalitis, myelitis and encephalomyelitis	426	8	8.06	3.991, 16.26	<0.001	0.51
B96	Sequelae of other and unspecified infectious and parasitic diseases	18715	84	1.92	1.534, 2.401	<0.001	0.42
J34	Other disorders of nose and nasal sinuses	9626	49	2.24	1.68, 2.984	<0.001	0.46
F32	Other depressive episodes	29778	110	1.74	1.425, 2.114	<0.001	0.36

Code		N	N with Cholesteatoma	Odds Ratio	95% CI	Adjusted p-value	% Overlap
Q75	Other congenital malformations of skull and face bones	17	2	59.42	13.526, 261.048	<0.001	0.17
E78	Disorders of lipoprotein metabolism and other lipidaemias	77039	262	1.48	1.28, 1.703	<0.001	0.34
J47	Bronchiectasis	5742	34	2.46	1.741, 3.465	<0.001	0.50
K21	Gastro-oesophageal reflux disease	54508	183	1.51	1.288, 1.771	<0.001	0.33
G40	Epilepsy	6849	37	2.33	1.674, 3.231	<0.001	0.46
K52	Other noninfective gastroenteritis and colitis	25678	96	1.68	1.363, 2.074	0.002	0.36
L40	Psoriasis	5499	31	2.40	1.68, 3.438	0.002	0.47
H68	Eustachian salpingitis and obstruction	28	2	33.08	7.82, 139.896	0.003	0.17
H54	Visual impairment including blindness	2813	20	2.92	1.869, 4.548	0.003	0.51
N17	Acute renal failure	22036	92	1.69	1.357, 2.094	0.003	0.40
B95	Streptococcus and staphylococcus as the cause of diseases classified to other chapters	8860	44	2.07	1.529, 2.801	0.003	0.44
G52	Disorders of other cranial nerves	158	4	10.83	4.005, 29.299	0.004	0.31
H81	Disorders of vestibular function	3088	20	2.80	1.793, 4.359	0.008	0.47
J18	Bronchopneumonia, unspecified	26445	103	1.61	1.309, 1.974	0.008	0.37
L08	Other local infections of skin and subcutaneous tissue	3198	21	2.71	1.759, 4.187	0.008	0.49
G47	Sleep disorders	11734	53	1.89	1.43, 2.491	0.009	0.41
Q96	Turner syndrome	45	2	25.18	6.078, 104.301	0.011	0.17
C07	Malignant neoplasm of parotid gland	179	4	9.50	3.519, 25.657	0.012	0.30
E66	Obesity	35634	122	1.53	1.267, 1.846	0.012	0.33
G08	Intracranial and intraspinal phlebitis and thrombophlebitis	103	3	13.29	4.205, 42.001	0.014	0.24
M19	Other arthrosis	45012	150	1.48	1.239, 1.757	0.016	0.33
A09	Other gastroenteritis and colitis of infectious and unspecified origin	19953	76	1.66	1.318, 2.102	0.025	0.36
G09	Sequelae of inflammatory diseases of central nervous system	104	3	11.97	3.788, 37.821	0.031	0.24
N39	Other disorders of urinary system	38025	128	1.49	1.24, 1.798	0.032	0.33
J31	Chronic rhinitis, nasopharyngitis and pharyngitis	1743	13	3.22	1.862, 5.582	0.038	0.45
J33	Nasal polyp	4670	26	2.28	1.545, 3.374	0.045	0.45
B38	Coccidioidomycosis	4	1	119.72	12.384, 1157.385	0.047	0.09

3.3.4 Comparison to FinnGen significant associations

Diseases occurring before cholesteatoma with significantly increased hazard ratios in both biobanks were otosclerosis and sleep apnoea (**Table 16**). Hazards of otosclerosis, epilepsy and chronic kidney disease were significantly increased after cholesteatoma in both biobanks.

Table 16. Comparison of UKBB odds ratios and hazard ratios to FinnGen

Table showing Cox Hazard ratios computed by Risteys from FinnGen data where p<0.05 and equivalent Cox Hazard Ratios and p-values computed for UKBB data. Also shown are the uncorrected p-values and odds ratios drawn from logistic regressions for the closest equivalent ICD-10 codes. Phenotype definitions vary for the following: [1]Showing H91, parent code of sudden idiopathic hearing loss (H91.2). [2]Arthrosis combines parent categories M15-M19. [3]Combination of eclampsia (O15) and pre-eclampsia (O14). No overlap with cases in UKBB. [4] Showing G47, parent code of sleep apnoea (G47.3). [5]Hybrid category of several ICD-10 codes involving pain in FinnGen, not tested in UKBB

		FinnGen		UKBB		UKBB	
		HR	P value	OR	P value (uncorrected)	HR [95% CI]	P value
Before cholesteatoma	Otosclerosis	6.80 [4.14, 11.18]	<0.001	5.18 [2.144, 12.552]	<0.001	7.98 [7.98, 31.998]	0.003
	Sudden idiopathic hearing loss	2.93 [1.62, 5.31]	<0.001	9.94 [8.514, 11.608][1]	<0.001	0 [0, 8.65x10^{171}]	0.974
	Arthrosis	0.72 [0.7,0.91]	0.0062			1.12 [1.12, 1.545]	0.974
	Pre-eclampsia or eclampsia[3]	1.80 [1.09,2.96]	0.022				
	Gonarthrosis	0.71 [0.53, 0.96]	0.025	0.98 [0.796, 1.215]	0.87	0.78 [0.78, 1.327]	0.356
	Coxarthrosis	0.59 [0.37, 0.93]	0.025	0.96 [0.739, 1.259]	0.79	0.88 [0.88, 1.711]	0.708
	Sleep apnoea	1.38 [1.04, 1.83]	0.028	1.88 [1.43, 2.491][4]	<0.001	2.1 [2.1, 4.065]	0.029
	Pain[5]	1.18 [1.01, 1.38]	0.036				
After cholesteatoma	Otosclerosis	6.06 [3.41, 10.76]	<0.001	5.18 [2.144, 12.552]	<0.001	8.7 [8.7, 34.952]	0.002
	Vascular dementia	5.02 [2.36, 10.68]	<0.001	3.66 [0.217, 2.105]	0.19	0.93 [0.93, 3.737]	0.923
	Epilepsy	1.79 [1.09, 2.93]	0.021	1.47 [1.869, 4.548]	<0.001	2.85 [2.85, 4.733]	<0.001
	Chronic kidney disease	1.77 [1.08, 2.92]	0.025	1.43 [1.119, 1.812]	0.0041	1.5 [1.5, 2.09]	0.017
	Iron deficiency anaemia	1.74 [1.07, 2.82]	0.026	1.44 [1.136, 1.829]	0.0027	1.17 [1.17, 1.746]	0.441
	Varicose veins	0.59 [0.35, 0.98]	0.041	1.061 [0.776, 1.45]	0.71	1.01 [1.01, 1.885]	0.964

3.4 Discussion

This chapter was a retrospective epidemiological study of cholesteatoma in the UK BioBank. I established demographic factors associated with cholesteatoma and compared these with factors associated with other middle ear disease. I replicated some associations found in other studies such as male sex, smoking, and deprivation. Generally, cholesteatoma and other middle ear disease were more demographically similar to each other than to disease-free controls, with cholesteatoma cases having slightly more extreme values. I also identified overlapping ICD-10 codes with cholesteatoma, which generally consisted of other inflammatory diseases of the middle ear and associated conditions. The cholesteatoma group also had higher rates of several common diseases. Associations with epilepsy and otosclerosis were repeated in the Finnish biobank FinnGen.

3.4.1 Similarity of cholesteatoma risk factors to other middle ear disease

Cholesteatoma and other middle ear disease are similar for age, deprivation and smoking

The cholesteatoma and other middle ear disease groups did not differ significantly in age, deprivation or smoking status. Both the cholesteatoma and other middle ear disease groups were median one year older than controls with the AOR for age being 1.02 for cholesteatoma. This is probably because older individuals have had longer to contract disease.

The prevalence of cholesteatoma amongst smokers was similar to the prevalence amongst other middle ear disease; while no significant effect for smoking could be detected in the cholesteatoma group, the increased rate of middle ear disease was significant for smokers, which may be due to the larger size of this group. Furthermore, a significant association with F17 *mental and behavioural disorders due to use of tobacco* was (OR 2.34) suggests that smoking does affect cholesteatoma risk. The category for 'smokers' includes people who smoke infrequently or have quit, indicating that the effect is strongest for heavy smokers. This agrees with Kaylie et. al. (2009)[187] who found a higher rate of cholesteatoma and more severe disease in smokers with chronic ear problems, but that former smokers had similar outcomes to non-smokers >5 years after quitting. Smoking raises susceptibility to middle ear disease by impaired mucociliary function, which may also explain increased rates of cholesteatoma.

Median deprivation in the cholesteatoma group was -1.40 compared to -1.70 and -2.14 for other middle ear disease and controls respectively; the odds ratios for cholesteatoma compared to other middle ear disease and controls were 1.02 and 1.08 respectively but did not differ significantly from each other (**Table 14**). Deprivation is associated with risk behaviours such as smoking[199,200], and the most disadvantaged are both more likely to require healthcare and less likely to access it[201]. These factors may contribute to greater rates of ear disease and cholesteatoma and may also explain the greater risk of other common diseases in the case group. However, it is also possible for deprivation to result from the economic impact of hearing loss and so the relationship may be reversed; deprivation was recorded at recruitment and cholesteatoma is likely to have occurred before this time. This highlights the inability to determine causality from observational data.

Prevalence of cholesteatoma and middle ear disease vary with ethnicity

Both rates of cholesteatoma and other middle ear disease varied between ethnicities in this analysis, with prevalence highest in the Asian and white groups. Black ethnicity was significantly associated with a decreased risk of both middle ear disease and cholesteatoma, but the risk of these two types of disease differed divergently for some other ethnicities. Middle ear disease was more prevalent in both the Asian and Chinese groups, but the relative rate of cholesteatoma was lower. However, most of these differences were non-significant, likely due to the small sample sizes of these groups. Rates of cholesteatoma have previously been reported to vary with ethnicity[2], although original epidemiological data is rarely presented.

Male sex differentiates cholesteatoma from other middle ear disease

Male sex had an OR of 1.33 and there was a male predominance of 1:1.3 males:females in UKBB and 1:1.67 in FinnGen, which is well-established in the literatured[6,7]. However, there was no male-predominance in the non-cholesteatoma middle ear disease group. The adjusted odds ratio for cholesteatoma was effectively the same independent of whether the controls had middle ear disease or not, and there was no significant difference between the middle ear disease group and disease-free controls. This makes male sex a major risk factor distinguishing cholesteatoma from other middle ear disease in this cohort. Either males are not at increased risk of general ear disease, *only* of cholesteatoma; or that they are at increased risk of all forms of ear disease and at increased risk of sequelae, including cholesteatoma.

3.4.2 Cholesteatoma is significantly associated with other inflammatory ear and respiratory disease

An overlap between cholesteatoma and other middle ear disease is well known: Kemppainen et. al. (1999)[6] report history of otitis media in 72.4% of cholesteatoma cases while Castle (2018)[1] reports common concurrent diseases including otic polyp and tympanosclerosis. Tympanic retraction is also a risk factor for cholesteatoma[22,23]. Meanwhile, Djurhuus *et al.* (2015)[9] found that children with repeated ear infection requiring ventilation tube insertion are at increased risk of cholesteatoma. Although there is a possibility of iatrogenic cholesteatoma following ear surgery, the authors also showed that early ventilation tube insertion reduces risk of subsequent cholesteatoma and that national rates of cholesteatoma were decreasing[7]. This suggests that children at risk of cholesteatoma also have increased likelihood of requiring ventilation tube insertion. Whether chronic OM is a cause or symptom of cholesteatoma cannot be determined from this study design and so the nature of the relationship remains unknown.

I also detected associations with certain respiratory diseases, including chronic sinusitis, asthma, and bronchiectasis. The recently introduced concept of 'unified airway disease' seeks to explain the frequent co-occurrence of upper and lower airway disease by considering the airways as a single system sharing immunological and pathophysiological features[202]. The middle ear is connected to the nasopharynx via the Eustachian tube which is responsible for middle ear drainage, pressure equalisation and protection against pathogens[19]. Therefore, it is possible that chronic inflammation of the upper airways may also contribute to susceptibility to middle ear disease by impacting on Eustachian tube function.

3.4.3 New potential association with epilepsy

Some known sequalae of cholesteatoma such as facial nerve palsy were represented amongst associated disease codes, as well as rare ones like meningitis[203]. Interestingly, epilepsy was associated with cholesteatoma in both biobanks. Epilepsy is very rarely described as a complication of cholesteatoma[204,205] and can be triggered by intracranial infection[206], which may explain this relationship. Alternatively, epilepsy itself may be a risk factor, as it is often associated with a high burden of comorbid disease[207].

Although Cox Hazards regressions in both UKBB and FinnGen report increased hazards of otosclerosis both before and after cholesteatoma, this may not be a true association. Co-occurrence of these conditions is extremely rarely reported[208] and both have similar presentations, meaning misdiagnosis may occur. Otosclerosis is a disease of the labyrinth resulting from remodelling and overgrowth of the bone at the base of the stapes[89] and while inflammation induced by cholesteatoma could feasibly contribute to risk, the small number of overlapping cases (n=5 in UKBB, 48 in FinnGen) makes it very possible that misdiagnosis is the true cause.

3.4.4 Study limitations

This study was limited by the retrospective case-control design and so shares its limitations with other studies of biobank data. Conclusions about causality cannot be drawn from observational data and there may be additional confounding variables that were not accounted for.

The UKBB population is not necessarily representative of the wider UK population: it is biased towards females, has overall lower deprivation, and participants have fewer health problems than the general population[209]. Furthermore, the study population was majority White British, making it difficult to assess prevalence in other ethnicities. Demographic information reflects participants' situations at the time of recruitment, not the time of disease diagnosis. Deprivation is based on postcode and may not reflect an individual's actual status.

Identification of cases was limited by the availability of ICD codes: ICD-10 was introduced in 1995, so records prior to this date are generally not available. The older ICD-9 code for cholesteatoma also contains the unrelated otic abscess so was not useful for assigning cases. Records may also be missing due to a failure to translate paper records into electronic format or may contain inaccurate diagnoses. As ICD-10 codes are assigned for hospital billing purposes, and not by doctors themselves, they may not accurately reflect the actual diagnosis made; additionally, cholesteatoma may be misdiagnosed as other forms of otitis media. In order to maximise the number of cases detected, I used an expanded criteria which included operative codes strongly suggestive of cholesteatoma such as mastoidectomy. Ultimately, the rate of cholesteatoma of 1 in ~500 persons was similar to what I expected, though it was less than in FinnGen. Lack of detailed information means it is not possible to determine if cholesteatoma was congenital or acquired, or uni- or bilateral. Additionally, date of diagnosis was only available for a subset of cases, meaning Cox Hazard regressions had less power. The date of diagnosis also may not represent the actual time of disease onset, only when attention was sought for the condition.

Imbalance between cases and controls could introduce bias into estimates and inflate p-values. I tested for this effect by performing matched logistic regressions for demographic factors but did not perform such tests for all ICD-10 codes which may have had more extreme imbalance. Therefore, disease codes with small numbers of cases should be interpreted with caution.

3.4.5 Future directions

Although this study provides evidence of a higher incidence in white compared to non-white ethnicities in the UK (besides South Asian ethnicities) the sample size for non-white ethnicities in the UK Biobank is too small to be conclusive and differences between populations remain poorly characterised. Large, registry-based studies reporting age- and sex-specific rates may be required to clarify this.

The demographic similarity between cholesteatoma and other middle ear disease for most risk factors suggests an overlap in pathology, whether both arise through common causes or one arises from the other. However, some risk factors were not shared between cholesteatoma and other middle ear disease. Experiments using controls with OM will acquire different results to studies using disease-free ears, as it seems that there are both overlapping and distinct risk factors. Many tissue-based studies, such as gene expression studies, necessarily compare cholesteatoma to ears with OM because the middle ear is inaccessible and tissue can only be acquired during surgery.

3.5 Conclusions

Risk factors are shared between cholesteatoma and other inflammatory middle ear disease, but male sex was a major risk factor distinguishing these groups. Cholesteatoma overlaps significantly with other inflammatory middle ear conditions, but it is difficult to disentangle the factors contributing to ear disease in general from those contributing to cholesteatoma alone. Cholesteatoma was also positively associated with epilepsy and negatively associated with arthrosis in two biobanks. These relationships are unexplored in the literature and warrant further investigation.

The results of these analyses must be considered in the following genetics studies of cholesteatoma, as the choice of control group affects the conclusions that can be drawn. The relationship between cholesteatoma and other chronic ear diseases may be due to shared environmental risk factors, shared genetic risk factors, or because one disease directly provokes the other. If cholesteatoma is a consequence of chronic middle ear disease, then using disease-free controls may essentially be studying susceptibility to ear disease in general, whereas using controls with non-cholesteatoma middle ear disease asks which factors govern cause some individuals to progress to cholesteatoma and not others. An issue with using other middle ear disease as controls is that a proportion of them may have cholesteatoma but lack specific ICD-10 codes indicating it; they may also go on to develop cholesteatoma in the future. It is better to compare cholesteatoma to disease-free controls and compare any results to what is known about the genetic risk factors for chronic ear disease in general.

4 Genome-wide association tests: Single variant, gene-based and gene set enrichment

4.1 Background

4.1.1 Rationale

In the introduction to this thesis, I discussed evidence for genetic involvement in cholesteatoma, including several observations of family clustering[54–57], family history in ~10% of cases[61], and identification of some genes containing deleterious variants amongst 10 affected families in a previous Genetic of Cholesteatoma (GoC) study[126]. In a review of cholesteatoma genetics including incidences of family clustering, Jennings *et al.* (2017)[18] concluded that there is weak evidence for an oligogenic mechanism with reduced penetrance. A total of five (2 GoC, 3 non-GoC)[65,66,68,124,126] studies have performed gene sequencing and none have detected the same set of variants; however, there has been some overlap in the function of variants identified in these studies and dysregulated processes identified by gene expression analysis[36,77,78,125,136–140] (see *Semi-systematic review of global gene expression studies*). Implicated processes include ECM organisation, immune function/inflammation, ciliary function, and calcium binding. It therefore seems likely that any genetic basis for cholesteatoma is complex, polygenic, and possibly heterogeneous.

Studies so far have had limited power due to small sample sizes and lack of control populations. Hence this study uses UK BioBank whole exome data to perform genome-wide association tests (GWAS) for cholesteatoma using 1,000 European cases and 5,000 matched controls. Whole exome data are used for consistency with previous GoC studies and to capture rare variants which may not be represented on genotyping arrays. While this sample size is relatively small for a GWAS, it will be the largest cohort of cholesteatoma cases studied outside of phenome-wide association tests (PheWAS). There are several examples of PheWAS, which apply generic GWAS methods to produce summary statistics for hundreds or thousands of phenotypes, typically generated using ICD-10 codes. This includes a PheWAS of UK BioBank data called GeneBass (https://app.genebass.org/) and FinnGen, which contains 1,447 cholesteatoma cases in its ninth release (defined by ICD-10, 9 and 8). FinnGen is notable for its large size as well as its curated control groups, excluding any participant with middle ear or mastoid disease from cholesteatoma GWAS controls. This makes the cohort useful for comparison to UK BioBank results.

4.1.2 Modern GWAS methods

A GWAS performs tests for individual variants across the genome to detect associations with an outcome. This is achieved by fitting a linear or logistic regression with the phenotype as the outcome and genotypes and covariates as independent variables. Linear regressions are used for continuous traits (e.g. height) and a logit link function is used to convert the regression to logistic for binary traits (e.g. cholesteatoma status). This was described in the thesis introduction in *Genome-wide association studies*, but I repeat the basic equation here:

$$\boldsymbol{Y} \sim \boldsymbol{W\alpha} + \boldsymbol{X\beta} + \boldsymbol{g} + \boldsymbol{e}$$

Uffelmann et al. (2021)[88]

Where $\boldsymbol{Y}$ is a continuous phenotype, $\boldsymbol{W}$ is the genotype being tested and $\boldsymbol{X}$ is a matrix of fixed covariates (such as age or sex). The terms $\boldsymbol{g}$ and $\boldsymbol{e}$ capture error due to genetic and random effects. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the effect sizes of the genotype and covariates respectively, are estimated by the model, and we are primarily interested in $\boldsymbol{\beta}$. Genotype p-values are generated by comparing the model to a null model where genotype has no effect[88]. . Numerous software packages have been developed to perform GWAS with this regression at their heart. For example, the popular toolset PLINK[210] performs GWAS with this simple regression. Other models extend or modify the regression to account for additional confounders. More sophisticated methods have recently been developed to handle non-random effects such as population stratification, sample imbalance and relatedness. These typically involve multiple steps to estimate parameters used to improve the final regression model. SAIGE[194] and REGENIE[115] are two examples of popular GWAS software packages which use multi-step methods to conduct single variant analysis and, in the case of SAIGE, downstream gene-based tests.

REGENIE

In the first step, the genome is broken up into blocks of *N* SNPs. Each block is used to perform ridge regression, which is similar to linear regression but incorporates a parameter called the shrinkage factor. The shrinkage factor is used to reduce the coefficients of the predictors towards 0 to account for correlation of SNPs in a block: if two predictors are strongly correlated, the amount of information they impart is not as much as two independent predictors and their contribution towards the model must be reduced appropriately. This is done by penalising predictors with very large coefficients, which can arise out of correlation[211]. REGENIE randomly varies the shrinkage factor in each block to reflect that the true number of predictors is not known, resulting in a set of predicted outcomes for each block. A second ridge regression combines these predictors into a single predictor representing the entire genome[115]. The predictions from this step are used as a covariates in the logistic regression (step 2).

SAIGE

The first step of SAIGE involves fitting a null logistic model to a random sample of variants against phenotype and obtaining an estimate of the random effects for each individual. The variance of test statistic scores is compared for models performed with and without the error term included. The authors show that this variance ratio is consistent for variants with minor allele count (MAC) > 20. The variance ratio is used in when fitting the logistic regression (step 2) to correct for random and non-random genetic effects, and the saddle-point approximation (SPA)[212] is used to account for imbalance. SAIGE performs correction for population structure and controls the degree of structure seen within the European population well[194].

P-value correction

Regressions usually use maximum likelihood estimates (MLE), which means they calculate how likely the given data is under a specific model and seek to configure the model to maximise this likelihood. Highly imbalanced data can bias these estimates, either because there is imbalance in the cases and controls, or certain observations are very rare. During step 2 of both SAIGE and REGENIE, additional *p*-value correction is applied to account for these effects using Firth (REGENIE) or SPA (SAIGE and REGENIE). Firth[213] regression introduces a bias term into the function used to calculate MLE[213] while SPA corrects *p*-values by substituting SPA for the normal approximation for calculating the null distribution of the test statistic[214]. Unlike the normal distribution, which only has parameters of mean and variance (the first two moments of the function), SPA uses all moments (such as skewness) so can provide better estimates when assumptions of normality are broken[194].

Gene-based tests

Analysing rare variants poses a challenge as they may require very large sample sizes for detection. Several approaches have been developed for the analysis of rare variants by aggregating them at the gene level to increase statistical power:

- A gene-based collapsing or burden approach essentially sums the number of variants within a gene. This works under the assumption that different variants in the same gene or set of genes have similar impacts on disease[111].
- A combined multivariate and collapsing test collapses variants within subgroups according to criteria such as allele frequencies, and a multivariate test is performed within subgroups[215].
- SKAT[112] is a supervised regression method to test for the joint effects of multiple variants in a region. The test aggregates weighted variant-score test statistics rather than clustering variants directly. This allows for SNP-SNP interactions and is particularly powerful when regions contain many protective, deleterious and non-causal variants. SKAT calculates a *p* value for each genome region (or gene) while adjusting for covariates such as age, sex, and population stratification.

Burden-based tests have more power to detect associations when all variants in a region have the same directional effect on the trait and most are causal; non-burden tests like SKAT do not make these assumptions and can better handle variants with non-causal or opposing effects. SKAT-O[216] is a unified burden and non-burden test, finding an optimal linear combination of SKAT and burden tests. When the burden test is more powerful, SKAT-O behaves more like a burden test and when SKAT is more powerful, it behaves more like SKAT.

Gene set analysis

The concept of gene set enrichment analysis (GSEA) was introduced in *Semi-systematic review of global gene expression studies.* Simply, GSEA identifies pathways whose members are overrepresented in a given set of genes compared to a set randomly selected from the genome. In *Semi-systematic review of global gene expression studies,* I applied this to differentially expressed genes in cholesteatoma. GSEA can also be applied to genetic variant data to identify potentially disrupted pathways and processes; much as gene-level analysis aggregates variants at the gene level to increase power, GSEA aggregates variants at the pathway level. For many diseases, several genes are involved, and these may be linked via similar functional processes or pathways[217]; even if disease is highly polygenic or heterogeneous, it is likely to arise through the same set of pathways. GSEA can provide greater biological interpretability even when no individual variants meet genome-wide significance[218].

4.1.3 Types of genetic data

There are two main approaches to genotyping for GWAS: genotyping microarray and next-generation sequencing, which may cover the protein-coding regions of the genome (whole exome sequencing; WES) or the entire genome (whole genome sequencing; WGS). These approaches were described in the introduction to this thesis in *Genotyping data*.

UK BioBank (UKBB) recently performed whole-exome sequencing of its participants, which covers all variants in protein-coding regions including rare variants. UKBB also offers genotyping data using the UK BioBank Axiom V2 array, which covers the entire genome but only measures variants included in its probe set (850,000 variants; an additional 90 million variants are imputed using the Haplotype Reference Consortium and UK10K + 1000 Genomes reference panels). Whole genome data is also available but was not released in full at the time of this study.

Whole exome data was chosen for this analysis for consistency with previous whole exome family studies performed by GoC. Compared to genotyping array, whole exome has the benefit of detecting rare variants, which are increasingly considered an important source of missing heritability in common, complex disorders as well as in rare disease due to the possibility of stronger deleterious effects[219,220]. A drawback of whole exome data is the inability to detect variants in non-coding regions, which make up most of the genome and most trait-variant associations discovered to date[98,99]. Also, most associations detected by GWAS are non-causal and arise through linkage disequilibrium with causal variants, so this does not necessarily mean most *causal* variants are non-coding. Furthermore, it has been shown that protein-coding variants can have a disproportionately high predictive power for polygenic diseases[100,221]. As they directly affect protein structure, they may be more likely to impact disease risk and their consequences may be easier to interpret.

4.1.4 Aims and objectives

Based on the expectation of genetic complexity of cholesteatoma and the relatively small sample size of this study, I aim to perform association tests at the variant, gene and pathway levels. Testing at the gene level can account for the aggregate effects of rare variants, which are difficult to detect in small sample sizes. Meanwhile, pathway analysis can be useful where disease is polygenic or heterogeneous, as disease likely to arise through common mechanisms but individual variants may not be detected by an underpowered GWAS. Identification of variants, genes or pathways associated with cholesteatoma may provide insight into disease biology. Confirmation of these results in a comparison cohort, FinnGen, would provide further support the existence of a genetic component in cholesteatoma.

In summary, the aims of this chapter are:

- To compare the UKBB WES pipeline to the pipeline used by previous GoC studies to ensure results are comparable.
- To perform association tests at the variant level and gene level for UKBB whole exome data.
- To perform gene set enrichment analysis on UKBB single variant GWAS results to identify disrupted pathways and processes.
- To perform the same pathway analysis on FinnGen summary statistics for comparison.
- To perform post-hoc sensitivity analysis to determine study power and minimum detectable effect size
- To validate results by performing GWAS using UKBB microarray data.

4.2 Methods

4.2.1 UK BioBank Whole Exome Data

In this chapter, the primary data were variant call files generated by UK BioBank from whole exome sequencing data using DeepVariant, a neural network-based variant caller[191]. Whole exome sequencing was performed on the first 50,000 participants by Regeneron and

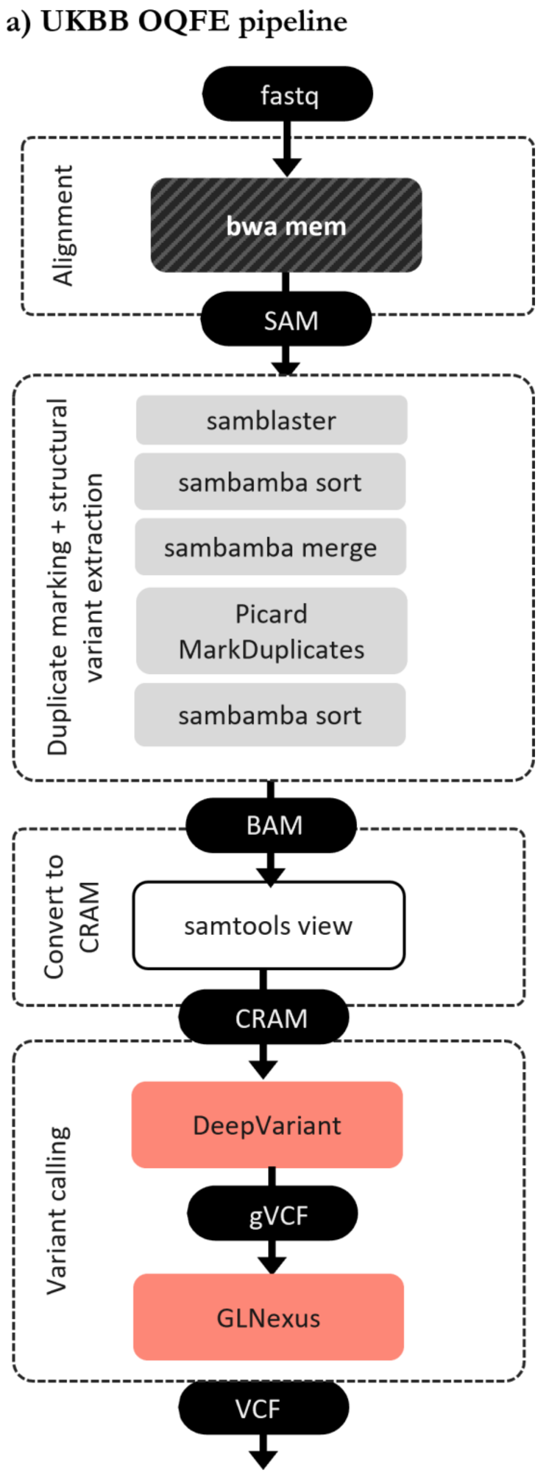
GlaxoSmithKline, with the remaining 450,000 sequenced by a consortium comprising Regeneron, AbbVie, Alnylam Pharmaceuticals, AstraZeneca, Biogen, Pfizer, Takeda and Bristol-Myers Squibb. Exomes were captured using the IDT xGen Exome Research Panel v1.0 and sequenced on the Illumina Novaseq 6000 platform using S2 (for the initial 50,000 samples) and S4 flow cells (for the remaining 450,000). The initial 50,000 samples used a different IDT oligo lot to the remaining 450,000 samples and were selected for specific enrichment in disease traits resulting in a strong batch effect. UKBB recommends the 90pc_10dp filter to control for this effect (see *Variant filtering for quality and impact*) and I also included batch as a covariate in the final GWAS. This whole exome data covers protein-coding regions, including exon sequences, intronic variants and 3′ and 5′ UTRs, but does not include intergenic regions.

Comparison of UK BioBank data generation to previous whole exome pipeline

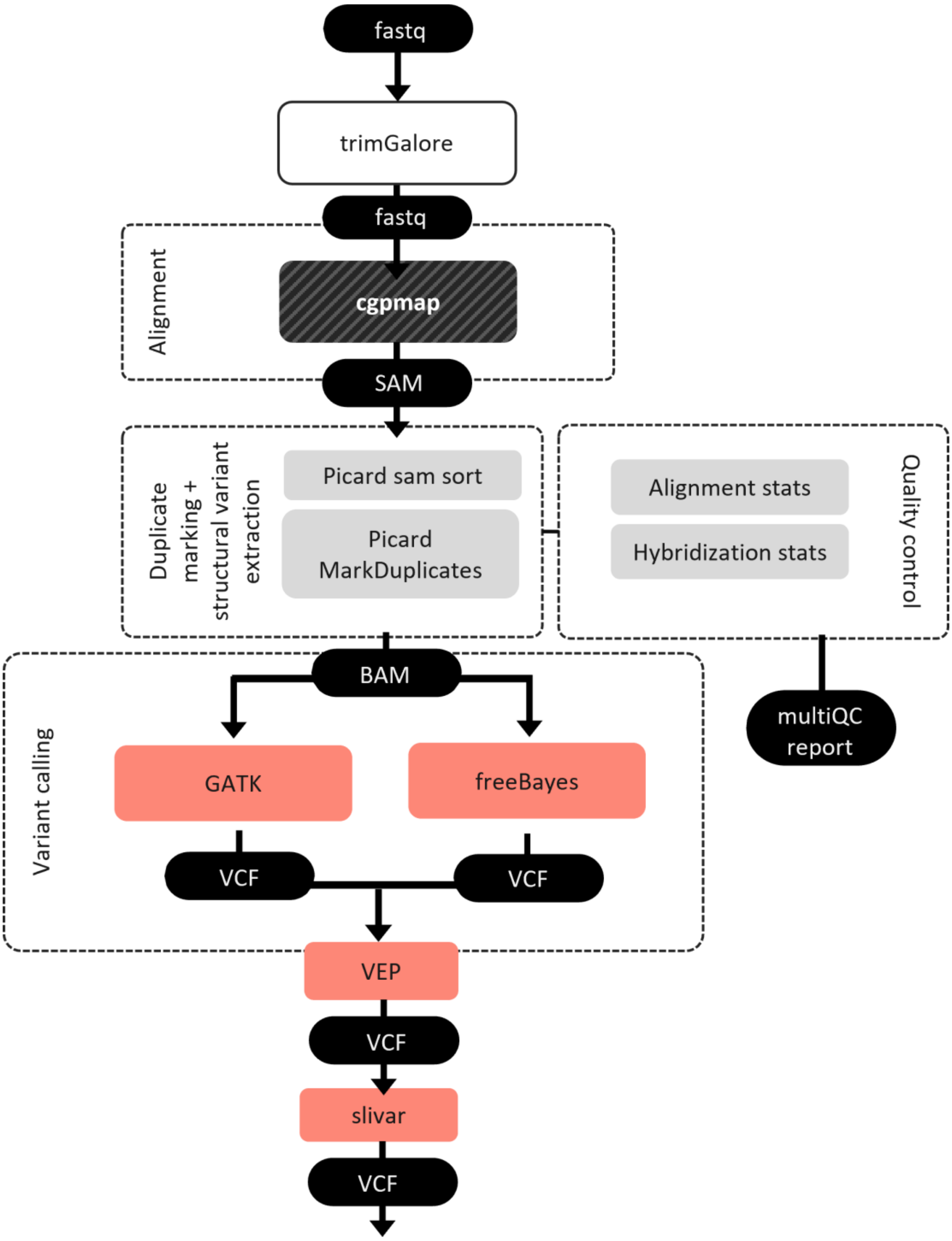
UKBB whole exome data are provided in CRAM format, a lossless file format from which the original fastq data can be reconstructed. A fastq is a text-based file containing the sequence data and quality scores per base. Several processing steps must be performed on raw sequence data to generate the CRAM file, including mapping, sorting and alignment against a reference genome. The tools used by UKBB to perform these steps are detailed in **Figure 20** and compared to those used in previous GoC cholesteatoma whole exome studies[124,126]. While many of the same tools are used (including bwa-mem[222], which is the mapping algorithm used in cgpmap[†]), differences in pipelines may lead to slight differences in the final CRAM files. I expect this effect to be small and it should not affect comparability between this study and the previous GoC WES study.

[†] Available at https://github.com/cancerit/dockstore-cgpmap

Figure 20. Comparison of UK Biobank and GoC pipelines for variant calling from whole exome data



b) MED pipeline



UKBB also provides variant call format (VCF) files containing the variable sites identified from the sequence data using DeepVariant. Different variant callers are likely to affect the comparability of data far more than differences in earlier steps.

DeepVariant performs very well for both sensitivity (proportion of true positives detected) and specificity (proportion of true negatives detected). In their 2018 paper[191], Poplin *et al.* test DeepVariant against other callers on the precisionFDA Truth Challenge data set: DeepVariant outperformed all other methods tested for both SNPs and indels, closely followed by GATK (a variant caller used in the previous GoC WES study). DeepVariant had also won the precisionFDA Truth Challenge two years earlier, and in 2020 remained one of the top performing methods on multiple platforms[223]. DeepVariant and GLNexus were also tested by Yun *et al.* (2020)[224] against WGS and WES samples from Genome in a Bottle project, Clinical Sequencing Evidence-Generating Research, and population Architecture Using Genomics and Epidemiology. They found that DeepVariant is somewhat overconfident about homozygous ALT calls (in comparison to GATK Best Practices V4.1.2.0) but is otherwise better calibrated across variant types. Precision and recall calculated separately for SNPs and indels were once again higher in DeepVariant than GATK and DeepVariant GQ (genotype quality) score distribution increases smoothly with sequence coverage whereas GATK oscillates below 99.

To determine whether to re-process the exome sequencing data using the GoC pipeline and perform variant calling consistent with previous GoC studies (**Figure 20b**) or use the DeepVariant VCFs, I reprocessed a sample of 46 randomly selected CRAMs with the GoC pipeline to compare the number and type of variants detected per person. Filtering for each call-set was performed according to the software's recommendations (**Table 17**).

Table 17. Description of filters applied to each data set.

	Filter	
GATK SNPs	QD > 2 QUAL > 30 SOR < 3 FS < 60 MQ > 40 MQRankSum > -12.5 ReadPosRankSum > -8	QD: Variant Confidence/Quality by Depth QUAL: Phred-scaled quality SOR: Symmetric Odds Ratio of 2x2 contingency table to detect strand bias FS: Phred-scaled *p*-value using Fisher's exact test to detect strand bias MQ: Root mean square mapping quality
GATK Indels	QD > 2 QUAL > 30 FS < 200 ReadPosRankSum > -20	MQRankSum: Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities ReadPosRankSum: Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias
freebayes	QUAL>5 INFO/DP SAF > 0 & SAR > 0 RPR > 1 & RPL > 1	DP: Read depth QUAL: Phred-scaled qual score RPL/RPR: Reads placed left and reads placed right – read must have neighbouring reads.
DeepVariant	DP > 6 (SNPs) DP > 10 (Indels) VAF > 0.15 (SNPs) VAF > 0.2 (Indels) QUAL > 30	DP: read depth VAF: Variant allele frequency, the proportion of reads at a site supporting the variant QUAL: Phred-scaled qual score
DeepVariant + HWE	All of the previous PLUS hwe > 1e-15 Geno < -0.1 90pc_10dp	Hwe: hardy Weinberg p value Genotype variant missingness less than 10% 90pc_10dp is a recommended filter for variant sites generated by UKBB to compensate for batch effects and difference in coverage between the initial 50k and remaining 450k exomes. A list of variants to exclude is provided. As an example, chr 2 starts with 1986436 variants, reduced to 1524976 by this filter.
DeepVariant+ Hwe+ Nonsyn	All of the previous PLUS removal of synonymous variants	Variants were annotated with SnpEff. Variant consequences were ordered by most severe to least severe. Where the most severe consequence of a variant was synonymous or less (intronic, upstream gene or downstream), the variant was removed.

4.2.2 Filtering

Sample QC

Sample quality control has largely been performed by UK BioBank. Additional metrics were provided for further filtering. Samples were removed prior to case and control selection if:

- Stated sex did not match their sex chromosomes, or there was sex chromosome aneuploidy. While these samples may belong to intersex individuals or those whose gender identity differs from their chromosomal sex, this can also indicate poor sample quality and where no distinction has been made these samples cannot be included.

- Participants with any close relatives in the same data set were removed. Cases with relatives were retained provided that relatives were not also cases, and their control relatives were removed. Controls were excluded if any relatives were present due to the large number of available controls for matching.

- After case matching, samples were filtered for missingness. Samples with > 10% site missingness were to be removed, but none failed this test.

Variant filtering for quality and impact

Variant filtering is applied to genetic data to reduce the number of variants being tested, thus reducing the dimensionality of the data and the impact of type 1 error. Filtering can be done for quality metrics as well as predicted impact, which assumes that certain types of variants are more likely to contribute to disease.

I performed filtering following Szustakowski *et al.* (2021)[225] in their paper detailing the initial release of UKBB 200k exomes (**Table 17**). Variant sites were filtered to exclude sites which violate Hardy Weinberg equilibrium with *p* values of < $1x10^{-15}$ or have genotype missingness > 0.1. Also, the 90pc_10dp filter was applied, which removes variants with less than 90% coverage at 10 depth across the population. This was performed on the whole population of ~500k participants.

Additional filters for quality at the variant level (following the method set out in Szustakowski *et al.* (2021)[225]) were:

- **Read depth > 6 (SNPs) or 10 (Indels).** Requiring a minimum number of reads (read depth, DP) at a given site is a standard quality control metric. Sites with very few reads cannot easily be distinguished from errors and should be excluded.

- **Variant allele frequency > 0.15 (SNPs) or > 0.2 (Indels):** Variant allele frequency (VAF) is a measure of the proportion of reads which support a given allele at a site and must be above a minimum threshold to strongly support a variant existing. A site may be covered by many reads but if there is no consensus on the base present at that position, their support for a variant is weak. A VAF of 0.5 suggests a heterozygote and a VAF of 1 a homozygous alt. For a SNP with four possible bases at a given site, a VAF of 0.15 may be sufficient to support a heterozygous alt at this position, but this cutoff must be higher for indels, which have more possible variations.

- **QUAL > 30**: QUAL score in DeepVariant calls is a representative of the confidence with which the neural network called the variant.

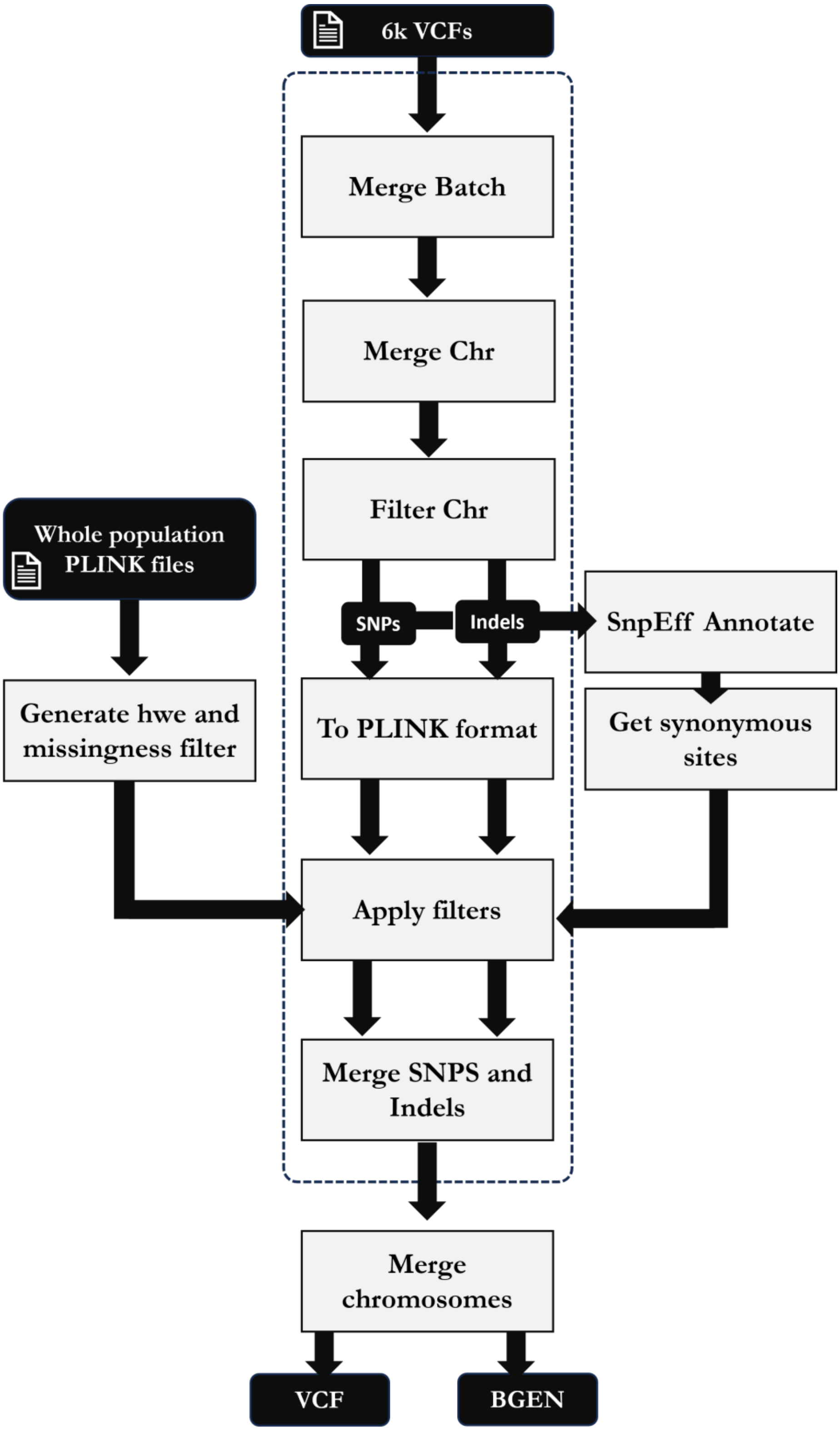
In addition, I filtered coding variants to remove synonymous, upstream and downstream variants. Synonymous variants do not change the polypeptide sequence so should not affect protein function. Although there is some evidence that synonymous variants may impact fitness in certain circumstances[226,227], these findings are generally drawn from microbial experiments and the general applicability of this has been contested[228]. Importantly, removing synonymous variants will reduce variant numbers, reducing data dimensionality, noise, and type 1 error. This is particularly important given the small sample size.

Variants were annotated with their predicted effect using SNPeff[229]. Any sites whose highest impact variant was predicted to be synonymous, upstream gene, or downstream gene was excluded. SNPeff considers upstream and downstream variants to be less impactful than synonymous variants as they should not affect protein coding or regulatory regions. In this thesis, I describe retained variants as ‘non-synonymous’ to describe the exclusion of synonymous coding variants, but certain intronic and regulatory variants such as those in splice sites and 3′ and 5′ UTRs will also be retained. No maximum or minimum MAF filter was applied.

Configuration of final pipeline

Filtering for the DeepVariant VCFs was performed on the UK BioBank Research Access Platform and accessed using the python library dxpy (**Figure 21**,**Box 1**).

Figure 21. Overview of the filtering process performed at the variant level for selected cases and controls.



Box 1. Overview of the filtering process performed at the variant level for selected cases and controls.

Merge batch: Start by merging all individuals within a batch of ~200 people, to avoid issues of merging too many files at once. Merged VCFs are split to chromosomes.

Merge chr: Each 200-person chromosome VCF is merged with the other batches to generate a single file for all participants per chromosome.

Filter chr: All sites are split to biallelic, first using `bcftools norm -N -m`. At any SNP-indel sites which also had a SNP or indel-only variant at that site will end up with repeated entries for those sites. Then phenotype likelihood (PL) is removed to prevent issues with varying numbers of entries in the PL columns for multiallelics. All sites whose alt is <*> are removed using `grep -v '<\*>` as these are non-variants. Sites are collapsed back into multiallelics, ensuring that any previous SNP-indel sites are reunited with SNP or indel-only versions of that site. Sites are split into biallelic one more time using `bcftools norm -N -m - -Oz -o.` There will now be only one copy of each variant and each row will only contain one ALT. The VCF can now be split to SNPs and indels with `bcftools view -v SNPs` and `bcftools view -V SNPs`. The first includes SNPs only, the second excludes SNPs (so will include indels, MNPs and others). DP, VAF and QUAL filters are applied.

SnpEff annotate: SnpEff annotates genes with predicted impact, starting with the most impactful consequence per gene.

Make synonymous filter: Variants whose worst impact is synonymous, upstream, downstream, or intronic are selected to generate a filter, as we don't expect these to impact protein function.

To PLINK format: The ID column for SNPs and indels is filled with the CHR:POS:REF:ALT information. The filtered SNPs and indels are converted to PLINK format. Keep-allele-order is used to prevent REF being set to the major allele. This must be done every time plink is invoked.

Apply filters: Indels and SNPs are treated separately. Plink extract is used to retain only the variants which passed HWE/missingness tests. Plink exclude is then used to remove all synonymous variants, using the text file generated in the make synonymous filter step.

Generate BGEN: The SNP and indel plink files are merged. A VCF is generated for use in the following command, which uses –a1-allele to force A1 to be the REF from the VCF (ALT is A2). The resultant plink file is used to generate a BGEN file (set to 8 bit), the BGEN is indexed ready for input to SAIGE.

REGENIE Comparison

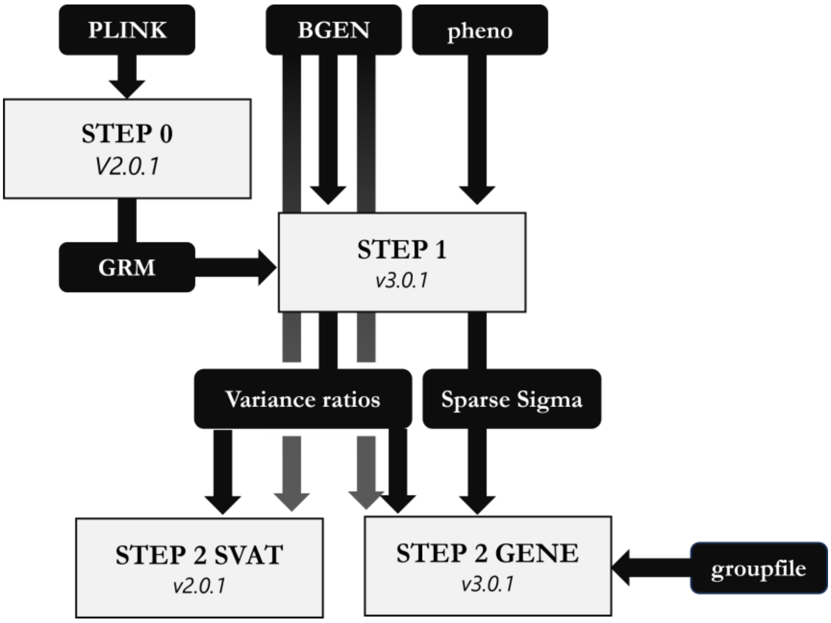
To determine which software package to use, I ran the same set of data through SAIGE and REGENIE using the same covariates. SAIGE was configured as in *Final SAIGE configuration for single variant tests*, but chromosome X was not used (due to this being part of early prototyping). REGENIE [3.1.0] was run with default settings and the following covariates: age, sex, deprivation, smoking status, and the first 10 genetic principal components.

Final SAIGE configuration for single variant tests

The final GWAS was performed using filtered whole exome data for 1,000 European cases (see Appendix: Ancestry Estimation) and 5,000 matched controls (see Propensity matching) in SAIGE. Step 0 was performed using SAIGE version 2.0.1 with the default 2,000 randomly selected markers and relatedness cutoff of 0.125. Step 1 was performed with SAIGE 3.0.1 using age, sex, deprivation, smoking, batch and the first 10 genetic principal components as covariates. Default minor allele count (MAC) categories of 1,2,3,4,5,6-10,11-20 and >20 were used for variance ratios and the minimum minor allele frequency was 0.01. Step 2 was performed with SAIGE 3.0.1 (**Figure 22**). Different versions are due to the introduction of SAIGE-GENE (3.0.1) for step 2 gene-based tests and a relevant update to step 1 allowing for generation of the sparse sigma matrix and categorical allele frequency tests. Single-variant results were filtered to MAC > 20 for subsequent analyses besides gene-based tests.

SAIGE-GENE 3.0.1 was used for gene-based analysis using the same data and output of steps 0 and 1 as the single variant association tests. The minimum MAC for this stage was 0.5 (essentially no lower cutoff – this may remove some imputed variants with very low certainty) and a maximum minor allele frequency of 0.05. These limits were chosen as the gene-based test is designed for rare variants. SAIGE-GENE performs SKAT, burden and SKAT-O tests for each gene defined by a group file listing all genes to be tested and the variants belonging to each gene. Genes were defined by Ensembl gene ID (ENSG number) using the National Center for Biotechnology Information (NCBI) human genome release 38 version p14 (GRCh38) boundaries.

Figure 22. SAIGE configuration used for final GWAS.



I generated the group file by annotating all variants output by step 2 of the single variant association tests using Ensembl Variant Effect predictor (VEP). As a single variant may affect multiple genes, variants may appear multiple times in the group file associated with different genes. Only protein-coding genes were considered in this analysis as the data is whole exome and focuses on protein-coding regions. However, VEP annotation includes many non-coding genes. These were removed by filtering for gene type 'protein coding' using the NCBI38 gene feature file[*]. Any synonymous, upstream, or downstream variants introduced by VEP annotation (possible because a missense variant affecting one gene may be upstream or downstream of another) were removed.

Annotation of variants

Variants were annotated with their reference SNP IDs (rsIDs), affected genes, consequence and predicted impact using VEP. Non-coding genes, synonymous, upstream, and downstream

[*] Downloaded from ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz

variants were removed in the same manner as generation of the group file. Repetition of removal of synonymous variants using two different variant effect predictors (VEP and SnpEff) ensures no synonymous variants are retained.

Gene Set Enrichment Analysis

I performed gene set enrichment analysis with g:Profiler[133] using the results of the single variant analysis. Variants were filtered to MAC > 20 and *p*-value < 0.05. Genes containing at least one qualifying variant were ordered by the *p*-value of their most significant SNP. A total of 2,373 genes contained at least one significant SNP, with the majority (81%) containing only 1 significant SNP. The *p*-value ranked list was supplied as an ordered query to g:Profiler using the GO molecular functions, biological process, and cellular compartment databases. I used the package gProfiler2 version 0.2.3[133] in R 4.1.3[198] to access the g:Profiler API. The version released on 13-02-2024 (reference genomes: Ensembl 111, Ensembl Genomes 57. GO release: 2024-01-17[‡]) was used.

I also performed gene-set analysis on the gene-based test results for comparison. For this analysis, g:Profiler was supplied with of all genes with *p*-value <0.05, ordered by *p*-value. While the single-variant results consider all variants with MAC>20, the gene-based results only aggregate rare variants within genes. Genes with only one or two significant variants may not themselves be significant and single variants scattered across multiple genes belonging to the same pathway will not be detected. Therefore this approach may not reflect the actual enrichment of variants affecting certain pathways. For this reason, this result is presented as secondary to the main result of gene-set enrichment analysis performed on the variant-level results.

[‡] See https://github.com/geneontology/go-announcements/issues/665 for details

In order to reduce the number of enriched GO terms and identify driver terms, I performed a term-reduction process similar to the highlighting algorithm provided by g:Profiler:

1. For all enriched terms, traverse the GO graph upwards until there are no more significant parent terms.
2. Where a term has two or more significant parents, the most significant is followed.
3. The top level of the hierarchy is noted.
4. For all terms sharing the same top parent, the most significant child term is retained.

This approach identifies the most significant term within unbroken ascending chains within the GO hierarchy. In contrast, the highlighting algorithm recursively searches subgraphs for the most significant term, eliminating its child and ancestor terms and re-running the query without any of the genes belonging to previously identified significant terms between each search. Highlighting cannot be performed for ordered queries due to the query being resubmitted.

FinnGen comparison data

FinnGen data release 9 summary statistics were downloaded from r9.finngen.fi/pheno/H8_CHOLEASTOMA. Genotyping was performed using a GRCH37-aligned Thermo Fisher axiom genotype array, including ~500,000 core GWAS markers and an additional ~200,000 markers enriched in the Finnish population or of special clinical interest[196]. GWAS was performed using REGENIE on a total of 1,447 cholesteatoma cases and 376,139 controls. Cases were selected using ICD10, IC9 and ICD8 codes. Controls were selected by excluding any ICD code indicating middle ear or mastoid disease, similarly to the procedure for UKBB.

To remove imputed variants, I downloaded probe set information from FinnGen and filtered the summary statistics to only include variants that had been directly measured. Although variants were already annotated with their nearest gene, I re-annotated the results using VEP for consistency with my analysis of UKBB data. This included assigning variant consequences, which allowed for the removal of variants considered synonymous, upstream, or downstream. I also filtered to MAC>20 for consistency and removed any entries associated with non-protein-coding genes in the same manner as for UKBB data. Gene set enrichment analysis was also applied in the same manner by filtering for genes containing at least one variant with *p*-value < 0.05 and ordering them by the *p*-value of their most significant SNP.

Comparison GWAS of microarray results

The primary analysis was of whole exome variants for inclusion of rare variants but genotyping array data were also available from UK BioBank, providing a means of internal validation for GWAS results. Quality control for this data was carried out by Affymetrix and UKBB prior to release as outlined in Bycroft *et al.* (2018)[190]; hence the only filtering I performed was to remove synonymous variants. I used the same setting for SAIGE as in the whole exome analysis to perform single-variant association tests. I did not perform gene-based burden tests for this data due to a) the presence of intergenic variants and b) the lower frequency of rare variants. Reannotation was performed using VEP in the same manner as FinnGen results. The results of this analysis are to support the whole exome results and comparison to FinnGen data, which used a similar genotyping array.

4.2.4 Post-hoc power calculation for sensitivity and ideal sample size

The sample size for this study was limited by the number of cases present in UKBB (see *Identification of cholesteatoma cases* for case definitions). The number of cases (1,000) is small for a GWAS of complex a disease, so I performed post-hoc power calculation and sensitivity analyses, to determine the minimum genetic effect, number of cases, and overall sample size necessary for achieving statistical significance at different risk allele frequencies. I used the Genetic Association Study (GAS) power calculator[§230] to calculate power and sensitivity.

Sensitivity

I used GAS to determine the minimum genetic relative risk (GRR) which a truly significant variant would need to be detected at the traditional genome-wide significance level of 5x10^{-8} with 80% power at minor allele frequencies of 0.005, 0.01, 0.05, 0.2, 0.3, 0.4 and 0.5. Sensitivity was also calculated for four hypothetical study designs (2,000, 5,000 and 10,000 cases, all 1:5 case-control ratio).

Power

I also tested three allele frequencies (0.001, 0.01, 0.2) with a range of GRRs to calculate power under the current study design and to determine the minimum sample size to give 80% power. For all tests, disease prevalence was set to 0.02 (reflecting the prevalence in UKBB as a whole) and significance level was set to 5x10^{-8}. Case-control ratio was set to 1:5 for all tests.

Genotype relative risk to odds ratio conversion

GRR is a risk ratio, which is the proportion of individuals with a given exposure who experience the outcome divided by the proportion of unexposed individuals who experience the outcome[231]. This ratio gives how many times more likely a person in the exposed group is to have the outcome. Because GRR is a risk ratio whereas the result of logistic regression is a log odds ratio, an estimate of the GRR for each variant must be calculated for comparison to GAS output. The simple formula for GRR is:

$$GRR = \frac{P(case|DD)}{P(case|Dd)} = \frac{P(case|Dd)}{P(case|dd)} \tag{1}$$

from Skol et al (2006)[232]

[§] Available at https://csg.sph.umich.edu/abecasis/gas_power_calculator/index.html where D is the effect allele and d the alternative allele. This is equal to the ratio of the probability of being in the case group when two effect alleles are present ($P(case|DD)$) versus the probability of the being in the case group when only one effect allele is present ($P(case|Dd)$). This is equivalent to the ratio of disease risk for the *Dd* genotype compared to the *dd* genotype. In terms of allele counts, this is calculated as follows:

2 $$GRR = \frac{NCase_D NTotal_d}{NCase_d NTotal_D}$$

where $NCase_D$ and $NCase_d$ are the number of D or d alleles present amongst all cases at a given site (that is 2 x homozygotes + 1 x heterozygotes), and $NTotal_D$ and $NTotal_d$ are the alleles counts amongst the entire population (both case and control). When the GRR of a variant is calculated from raw allele counts, there will be no adjustment for covariates or population structure. An estimate of risk can also be calculated from the odds ratio according to Zhang & Yu (1998)[233]:

3 $$RR = \frac{OR}{[(1 - P_0) + (P_0 \times OR)]}$$

Where P_0 is the outcome prevalence amongst the unexposed group. To estimate risk ratio for each variant with MAC > 20 and p-value < 0.05, I use the following formula:

4 $$RR = \frac{exp(\beta)}{\left[\left(1 - \frac{P(case|dd)}{s}\right) + \left(\frac{P(case|dd)}{s} \times \exp(\beta)\right)\right]}$$

Where β is the odds ratio and $P(case|dd)$ is the proportion of *dd* genotypes who have disease. s is the factor by which cases are over-sampled. With an estimated prevalence of 0.02 amongst UKBB and a case/control ratio of 1/5, the value of *s* is 8.3. Cases are oversampled by a factor of 8.3, so this correction is applied to reflect prevalence amongst *dd* individuals in the population as a whole.

I use equation 4 to calculate relative risk for each variant for comparison to the threshold risks calculated with GAS, using the absolute beta to ensure all risk ratios are in the positive direction.

4.3 Results

This GWAS used UK BioBank whole exome data from 1,000 cholesteatoma cases and 5,000 matched controls to perform variant-level and gene-level association tests. I then used the single variant results to perform gene set enrichment analysis to identify disrupted pathways and processes in cholesteatoma. Comparison data from FinnGen was used to identify enriched pathways common to both data sets. Additional validation of single variant and pathway results was performed using UKBB microarray data.

I also compared UKBB DeepVariant VCFs to variants identified by the pipeline used in our previous Genetics of Cholesteatoma (GoC) study, supporting the use of VCFs rather than re-processing raw data. Results of filtering and quality control for genetic data are also provided in this section.

4.3.1 Filtering and quality control

Comparison of callers by number of variants per person

To determine whether the UKBB OQFE pipeline with DeepVariant generated comparable call sets to the GoC pipeline using freebayes and GATK, I compared a sample of 45 UKBB whole exome CRAM files re-processed with the GoC pipeline to the DeepVariant VCFs.

The number of unfiltered variants called by freebayes and DeepVariant was of similar magnitude, though freebayes called a mean of 212,877 variants per person compared to 109,902 per person by DeepVariant (**Table 18**). GATK called far fewer variants per person (2,663). The GATK best practices pipeline applies several quality checks and filters to produce a smaller set of higher confidence variants and was configured to remove more variants during the calling process in the GoC pipeline. When equivalent filters were applied, the number of variants called by DeepVariant was less than freebayes, though again of similar magnitude (65,840 for DeepVariant vs 52,991 for freebayes; **Table 18**). The GATK callset was only reduced by 24%, possibly due to more stringent quality checks throughout the calling process. As an additional filtering step, the GoC pipeline considers only the overlap between GATK and freebayes.

Table 18. Mean variants per person in filtered and unfiltered data re-processed via the GoC pipeline compared to DeepVariant calls.

	Unfiltered			Filtered				
Variants	GATK	freebayes	DeepVariant	GATK	freebayes	DeepVariant	+hwe	nonsyn
All	2,663	212,877	109,902	2,024	52,991	65,840	23,059	7,445
SNPs	2,520	187,430	92,687	1,897	48,750	60,666	21,979	7,078
Indels	142	25,446	17,215	128	4,242	5,173	1,081	367

*'Unfiltered' refers to VCFs with no depth or quality filters applied, i.e. they are the direct output of the variant caller. 'Filtered' refers to the number of variants present after all filters from the appropriate pipeline have been applied (**Table 17**). DeepVariant counts were generated with a prototype filtering pipeline which differs slightly from the final version detailed in Configuration of final pipeline**.** The same approach is used in the final pipeline with some modification to order of steps to account for issues caused by specific variant types during VCF merging.*

Comparison of callers by overlap of variants

Most variants (99.3% of SNPs and 93.4% of indels) called by the GoC pipeline were present in the DeepVariant callset (**Table 19)**. In the GoC pipeline, variants were annotated using VEP[234] and Slivar[235] and filtered for frequency and impact. When this process was also applied to the DeepVariant callset, the discrepancy in variant numbers called increased. The number of SNPs and indels called by the GoC pipeline was reduced to a mean 344.75 and 14.5 per person respectively, while the DeepVariant SNPs and indels were reduced to 13,219 and 863 respectively. Now 100% of SNPs called by the GoC pipeline were present in DeepVariant data, but only 82.8% of indels.

Table 19. Variant overlap from UKBB and GoC pipelines

Filter	Low filter				Strict filter			
Variants	SNPS		INDELS		SNPS		INDELS	
Pipeline	GoC	UKBB	GoC	UKBB	GoC	UKBB	GoC	UKBB
Total	19,233.75	67,010.5	335.25	7778	344.75	13,219	14.5	863
Exclusive	130.25	47,996	22	7464.75	0	12,873.25	2.5	850
Overlap	19103,5		313.25		344.75		12	
Percent	99.3%	28.7%	93.4%	4%	100%	2.7%	82.8%	1.3%

Low filter: FILTER = PASS and QUAL = 5 applied to UKBB variants, as performed in GoC pipeline for freebayes variants. No post-annotation filtering was applied. GoC pipeline variants are those generated by the overlap of freebayes and GATK best practices pipeline.
*Strict filter: FILTER = PASS, QUAL=30, DP= 7 for SNPs, DP = 10 for indels (as in [225]) for UKBB variants. GoC pipeline variants are those generated by the overlap of freebayes and GATK best practices pipeline. Both are annotated with VEP and Slivar and filtered with the following criteria: impactful = TRUE, gnomad_af <0.01, variant.filter = PASS, topmed.af <0.01, variant.ALT[0]!=**

Support for use of DeepVariant VCFs

Because DeepVariant called almost all variants called by the GoC pipeline, and because DeepVariant has been shown to equal or outperform GATK in sensitivity and specificity studies[223,224] I use the DeepVariant VCFs for genetic analyses rather than reprocessing the data at considerable computing cost. Previous GoC studies benefited from stricter filtering requirements and greater reduction of candidate variants due to the small number of cases. The larger number of variants in this study is acceptable due to the larger number of cases and matched controls.

Overall reduction in variants due to population-level filters

Population-level filters for Hardy-Weinberg violation, missingness and 90pc_10dp (90% coverage at 10 depth) resulted in a reduction of ~20% from all chromosomes (**Table 20).** Most chromosomes do not lose variants to the missingness filter, probably because the 90pc_10dp filter removes them first. The reduction in variants was largely consistent across the genome, except for the X and Y chromosomes.

Table 20. Chromosome-level reduction in variants after application of population level filters

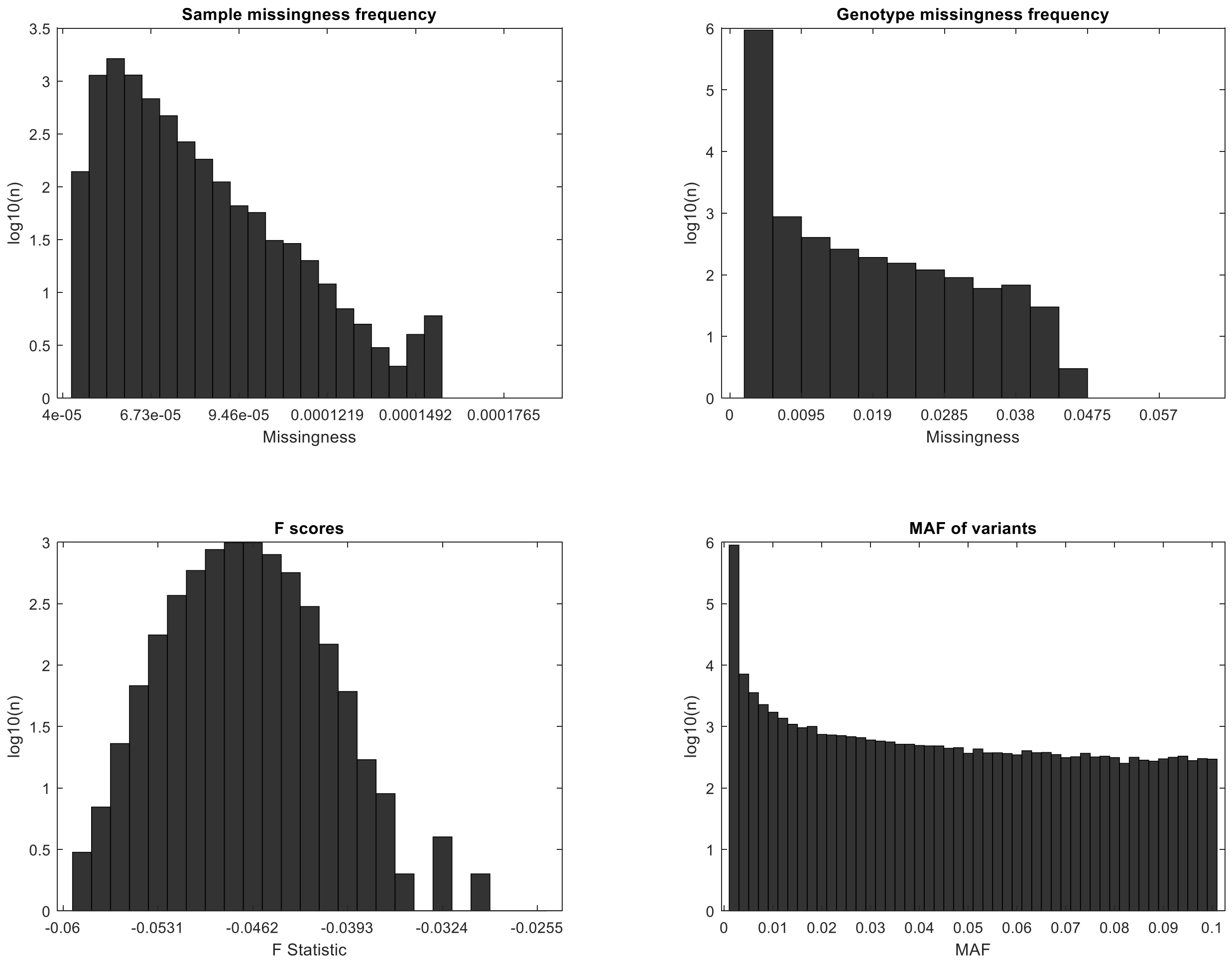
		Variants removed				
	Starting variants	90pc_10dp	hwe	Missingness	Ending variants	% reduction
1	2,687,650	562,184	14,391	0	2,111,075	21.45
2	1,986,436	461,460	8,534	0	1,516,442	23.66
3	1,572,602	356,178	6,659	0	1,209,765	23.07
4	1,088,878	279,378	4,489	0	805,011	26.07
5	1,200,708	280,428	5,342	0	914,938	23.80
6	1,343,324	304,012	6,011	0	1,033,301	23.08
7	1,290,944	283,616	6,354	0	1,000,974	22.46
8	983,410	222,187	4,231	0	756,992	23.02
9	1,160,259	243,966	5,501	0	910,792	21.50
10	1,105,522	256,916	5,228	0	843,378	23.71
11	1,589,220	297,235	7,541	0	1,284,444	19.18
12	1,435,996	324,345	6,389	0	1,105,262	23.03
13	485,358	118,804	1,874	0	364,680	24.86
14	840,031	177,011	4,264	0	658,756	21.58
15	936,831	208,664	4,616	0	723,551	22.77
16	1,300,364	228,077	6,111	0	1,066,176	18.01
17	1,565,159	280,248	7,236	0	1,277,675	18.37
18	433,149	103,014	1,837	0	328,298	24.21
19	1,791,970	262,420	10,646	0	1,518,904	15.24
20	686,915	128,316	3,052	0	555,547	19.12
21	289,748	62,683	1,451	0	225,614	22.13
22	613,853	110,413	500,209	0	500,209	18.51
Y	11,316	0	0	117	11,199	1.03
X	652,035	0	4,282	9,795	642,240	1.50

Post-filtering quality check

After all population and individual level filters are applied, the total number of variants in the population is reduced from 27,051,678 to 796,596. Most retained variants were rare, with 90% having MAF < 0.05 (***Figure 23***). Sample missingness was good with the highest missingness for any sample being 0.0015. However, a small number of variants (N=16) had missingness > 0.1 after case/control selection and filtering. This probably arose due to application of the missingness filter on a population level and subsequent sub-sampling; due to the large number of variants, it is likely for a small number to acquire higher missingness by chance. These variants were removed in subsequent analyses.

Figure 23. ***Quality control statistics for 796,596 variants.***

Sample missingness is calculated for all variants within a sample. Genotype missingness is calculated for each site across all samples. F-score is a measure of homozygosity calculated for each sample. minor allele frequency MAF shows minor allele frequencies of all variants amongst the population. Genotype missingness frequency and minor allele frequency include non-white participants as per-site information was calculated before their removal (n samples = 6,435). Sample missingness and F-scores are for the final 6,000 cases and controls.

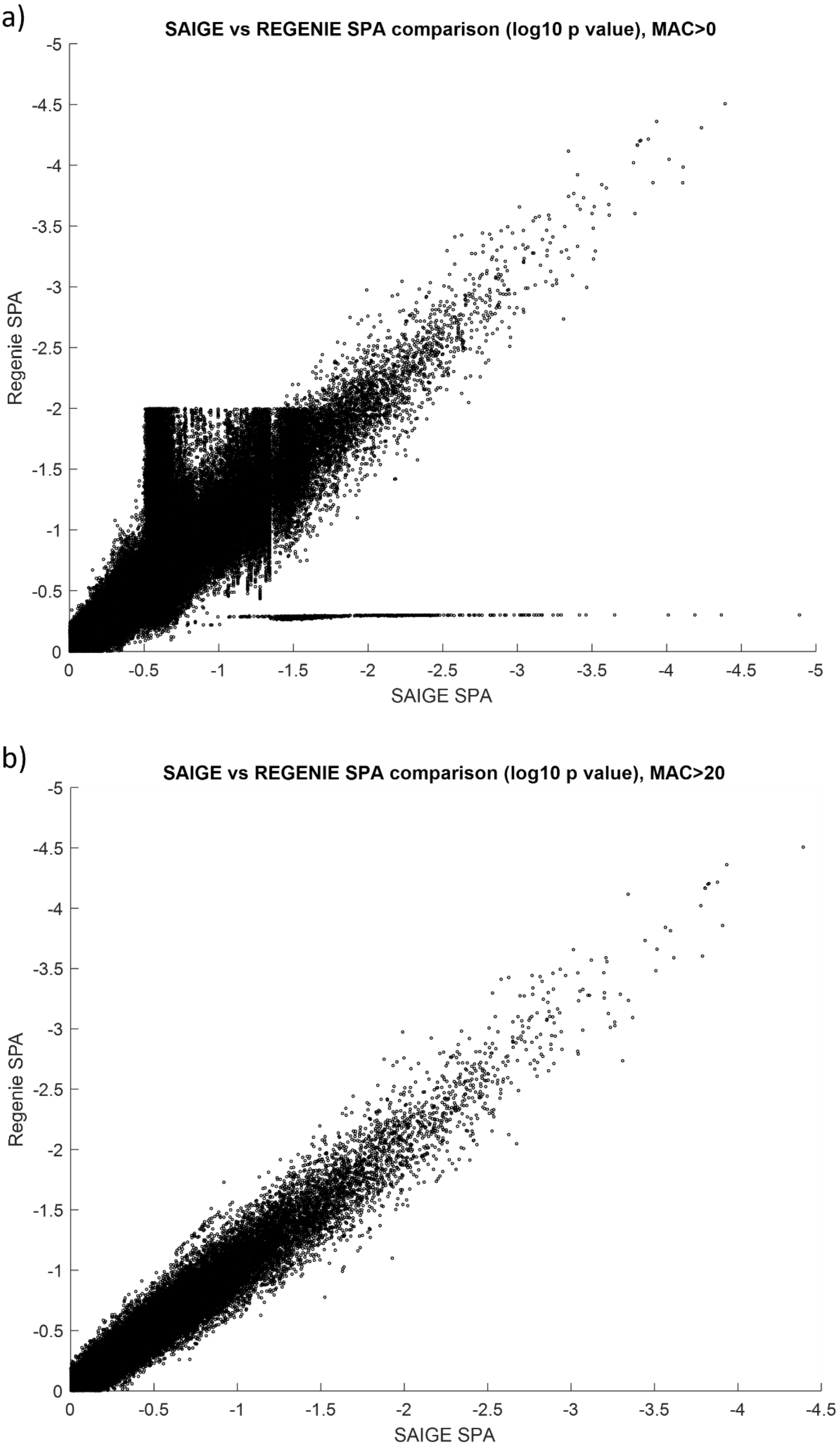


Heterozygosity was good with no samples showing *F* scores < 3 times the standard deviation below the mean (< -0.0501). *F* score indicates deviation from expected levels of homozygosity, where low homozygosity/high heterozygosity can indicate contamination. 15 samples showed *F* scores greater than 3 times above standard deviation from the population mean (>0.0534). Two samples (one case and one control) had much higher *F* scores than the rest of the population (0.2087 and 0.2838).

Comparison of GWAS methods shows good agreement between SAIGE and REGENEIE when MAC > 20

During early prototyping, I trialled REGENIE as well as SAIGE and compared the results. When no minor allele count (MAC) cutoff was applied, *p*-values generated by SAIGE and REGENIE-SPA were generally well-correlated (R^2=0.85) but two large blocks of SNPs were in disagreement (**Figure 24a**). This effect was eliminated when a MAC cutoff of 20 was applied, and R^2 increased slightly to 0.90 (**Figure 24b**). The MAC>20 cutoff was suggested by the authors of SAIGE and supported by results from this prototype, so MAC<20 variants were excluded from single-variant test results and downstream analyses, except for gene-based tests which are designed to accommodate rare variants.

Figure 24. p-value comparison for SAIGE and REGENIE using spa. a) when no MAC cutoff is applied, blocks of SNPs are in disagreement. B) filtering to MAC>20 eliminates this effect.



4.3.2 Genome-wide association test results

No single variant associations of genome-wide significance

Single variant analysis was performed for non-synonymous, coding SNPs in whole exome data. Manhattan plotting (**Figure 25)** shows no obvious signals or significant loci, and the most significant variants are scattered across the genome with no islands of linkage disequilibrium surrounding them. The traditional GWAS Bonferroni correction puts genome-wide significance at 5x10^{-8} [88]. However, whole exome studies may require different thresholds, which are not widely agreed upon. Fadista *et al.*(2016)[236] suggest 3x10^{-7} for whole exome studies where variants of MAF>=0.05. No variants met this lower threshold in our analysis. The most significant variant was a missense variant in *AMOTL2* (rsID rs139298691, *p* = 5.71x10^{-5}). Genes containing top-scoring variants (**Table 21**) were associated with various biological processes, including actin filament-based motility (*AMOTL2*), RNA binding (*RBM10*), receptor activity (*OR10A2, CMKLR1*, *PTH2R*) and calcium channel activity (*CACNA2D1, CACNA1G, ANK2*) (**Table 22**).

Figure 25. Manhattan plot of single variant results from whole exome data.
Top 20 variants are labelled with gene symbol. A cluster of 5 OR10A2 variants is labelled once.

Table 21. 20 most significant variants after single variant association tests in UKBB whole exome data

SNP	Gene	Consequence	rsID	*p*-value	BETA
chr3:134365885:C:T	*AMOTL2*	Missense	rs139298691	5.71x10^{-5}	1.776042
chr11:6869882:A:G	*OR10A2*	Missense	rs3930075	0.000111	0.194158
chrX:47186678:C:T	*RBM10*	3 prime UTR	-	0.000117	-1.66116
chr1:227816138:C:T	*PRSS38*	Missense	rs79840641	0.000123	-0.26967
chr11:6870527:A:C	*OR10A2*	Missense	rs7926083	0.000126	0.192911
chr4:113353363:G:A	*ANK2*	Missense	rs138842207	0.000127	2.273013
chr11:6870374:A:G	*OR10A2*	Missense	rs10839631	0.000128	0.192705
chr11:6869715:A:G	*OR10A2*	5 prime UTR	rs4758142	0.000133	0.19224
chr11:6869708:C:T	*OR10A2*	5 prime UTR	rs4758141	0.000136	0.191889
chr17:35764321:G:C	*C17orf50*	Missense	rs145033564	0.000139	0.56336
chr1:150308116:G:A	*MRPS21*	Missense	rs4845	0.000209	0.304099
chr7:82443499:G:C	*CACNA2D1*	5 prime UTR	rs200428602	0.000219	1.674788
chr12:108292772:A:G	*CMKLR1*	Missense	rs192034694	0.000221	1.027504
chr17:19742625:A:C	*ALDH3A1*	Missense	rs887241	0.000232	-0.18776
chr1:216421964:C:T	*USH2A*	Missense	rs10779261	0.000241	0.198952
chr2:208443511:G:T	*PTH2R*	Missense	rs144641723	0.000267	1.724242
chr4:67963391:C:T	*TMPRSS11A*	start lost	rs977728	0.000304	0.238687
chr17:50572702:G:A	*CACNA1G*	Missense	rs201875227	0.000324	2.157344
chr22:42644618:G:A	*CYB5R3*	Intron	rs578120569	0.000368	1.854875
chr9:129895300:C:G	*FNBP1*	Intron	rs17518373	0.000385283	0.642503959

Table 22. Functions of genes containing top 20 significant variants. Brief descriptions of genes functions are taken from UniProt.

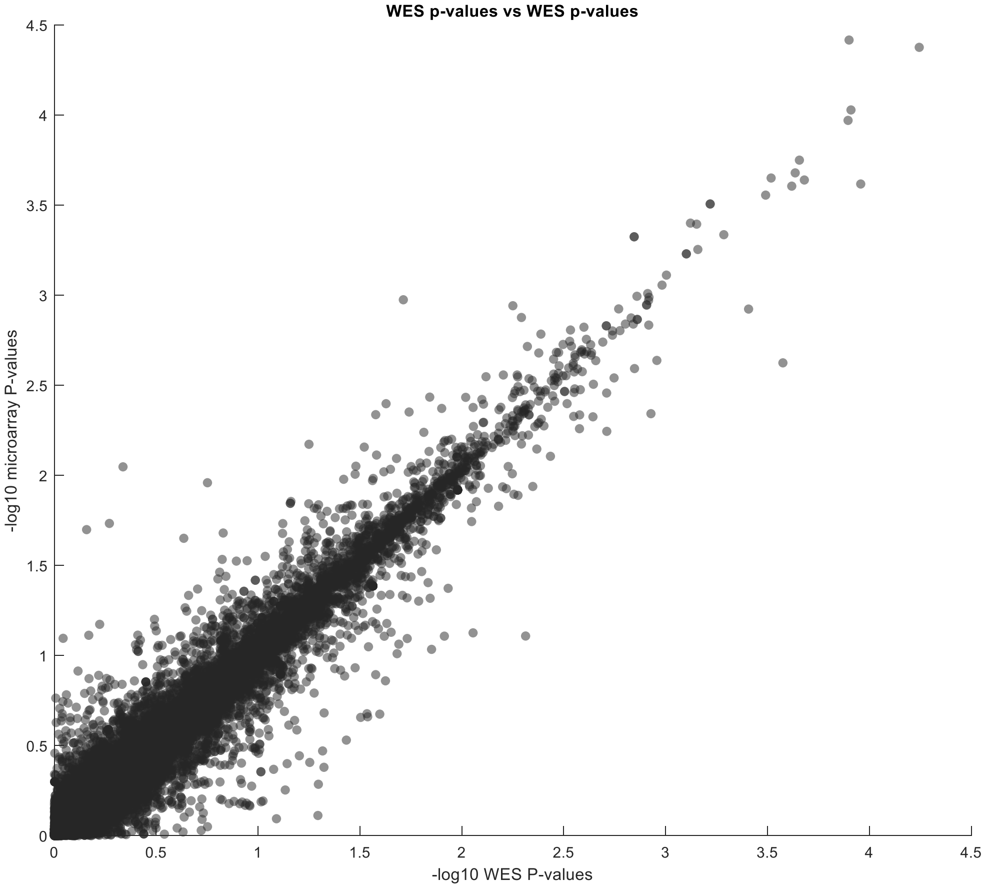
Gene	Description	Description from UniProt *(UniProt accession ID)*
AMOTL2	angiomotin like 2	Regulates translocation of phosphorylated SRC to cell matrix adhesion sites. Required for proper architecture of actin filaments, cell shape and area regulation. Inhibits Wnt/beta-catenin signaling pathway. May also be involved in endothelial migration, proliferation and polarity. *(*Q9Y2J4*)*
OR10A2	olfactory receptor family 10 subfamily A member 2	Odorant receptor. *(*Q9H208*)*
RBM10	RNA binding motif protein 10	May be involved in post-transcriptional regulation by mRNA splicing. *(*P98175*)*
PRSS38	serine protease 38	*No UniProt description (*A1L453*)*
ANK2	ankyrin 2	Essential for stabilisation of ion transporters and ion channels in various cell types, particularly cardiomyocytes and striated muscle cells. Bids dynactin to promote long-range motility of cells. Part of the ANK2/RABGAP1L complex which recycles fibronectin receptor. (Q01484)
C17orf50	chromosome 17 open reading frame 50	*No UniProt description (*Q8WW18*)*
MRPS21	mitochondrial ribosomal protein S21	*No UniProt description (*P82921)
CACNA2D1	calcium voltage-gated channel auxiliary subunit alpha2delta 1	Subunit of voltage-dependent calcium channel, regulates calcium current density *(*P54289*).*
CMKLR1	chemerin chemokine-like receptor 1	Receptor for chemoattractant adipokine chemerin and E1 molecule. Induces secondary messenger pathways such as calcium mobilisation and MAPK activation. *(*Q99788*)*
ALDH3A1	aldehyde dehydrogenase 3 family member A1	Major role in detoxification of alcohol-derived acetaldehyde *(*P30838*)*
USH2A	usherin	Part of USH2 complex involved in hearing (in growing stereocilia of cochlear hair cells) and vision (maintaining the periciliary membrane complex in photoreceptors). *(*O75445*)*
PTH2R	parathyroid hormone 2 receptor	Receptor for parathyroid hormone. May have a significant role in pancreatic function. May also function as neurotransmitter receptor. *(*P49190*)*
TMPRSS11A	transmembrane serine protease 11A	Probable serine proteinase whose overexpression inhibits cell growth and induces G1 cell cycle arrest *(*Q6ZMR5*)*
CACNA1G	calcium voltage-gated channel subunit alpha1 G	Subunit of voltage-dependent calcium channel mediating entry of calcium ions into excitable cells. Involved in a variety of calcium-dependent processes such as muscle contraction and hormone or neurotransmitter release. *(*O43497)
CYB5R3	cytochrome b5 reductase 3	Catalyses reduction of cytochrome b5 using NADH electron donor. *(*P00387*)*
FNBP1	formin binding protein 1	May act as a link between RND2 signalling and actin cytoskeleton regulation. May coordinate membrane tubulation with actin cytoskeleton reorganisation during late stage clathrin-mediated endocytosis. *(*Q96RU3*)*

Microarray validation agrees with WES where variants are overlapping

There was very little overlap between WES variants and microarray probes. With no MAC filters, the total number of variants shared was 53,199 of 796,595 variants present in the WES data and 572,358 in the microarray data. For MAC>20, the overlap was 32,011 variants. Agreement between *p*-values for MAC>20 variants is very good (Pearson correlation coefficient = 0.96, R^2 = 0.92), with some scatter due to the inclusion of a random element within the SAIGE process and possibly some measurement/sequencing errors causing differing allele counts (**Figure 26**).

Figure 26. UKBBWES vs UKBB microarray p-values for overlapping variants, sorted by ascending WES p-value.

Showing 32,011 variants present in both data sets with minor allele count>20. There is good agreement between methods with stronger scatter for high p-value (lower significance) variants.



Microarray does not identify any additional significant loci and no rare variants are included which make up the most significant results in WES.

Gene level association tests find no genes reaching genome-wide significance

Given there are approximately 20,000 protein-coding genes, genome-wide significance is set at 2.5×10^{-5} when Bonferonni correction is applied. The most significant gene was *ABCC8* (ATP binding cassette subfamily C member 8) which had a significance close to the threshold ($p=6.94 \times 10^{-5}$). No other genes met genome-wide significance. The most significant genes have diverse functions, including intracellular transport (*PLEKHA8, VPS36, BAIAP3, NXT1*), cytoskeletal organisation (*FER, TTLL12, FLNC*), cell cycle (*ESX1, TTLL12, PIMREG*), transcription regulation (*ESX1, VPS36, TTLL12, CREB5, TFEB*), and neural development or function (*BAIAP3, ADCYAP1, BCHE*) (**Table 23**).

Table 23. Gene-level GWAS results for UK BioBank whole exome data, 20 most significant genes. *A brief description of protein function taken from UniProt is given alongside the UniProt accession number. Gene-level tests do not include a beta score.*

Symbol	Name	Description of function from UniProt (accession)	*P*-value
ABCC8	ATP binding cassette subfamily C member 8	Beta-cell ATP-sensitive potassium channel subunit involved in insulin release (Q09428)	6.94×10^{-5}
PLEKHA8	pleckstrin homology domain containing A8	Cargo transport protein involved in trans-Golgi network transport, also required for cilium formation (Q96JA3)	2.60×10^{-4}
RGSL1	regulator of G protein signalling like 1	(A5PLK6)	2.61×10^{-4}
ESX1	ESX homeobox 1	Involved in cell cycle progression and spermatogenesis, arrests cell cycle at early M phase. Cleaved form ESXR1-N acts as transcriptional repressor and ESXR1-C inhibits cyclin turnover. (Q8N693)	8.95×10^{-4}
ANXA10	annexin A10	(Q9UJ72)	9.04×10^{-4}
ERLIN2	ER lipid raft associated 2	Forms complex with ERLIN2 to mediate endoplasmic reticulum-associated degradation of inositol 1,4,5-triphosphate receptors. Involved in cholesterol homeostasis. (O94905)	0.0013
ETFBKMT	electron transfer flavoprotein subunit beta lysine methyltransferase	May regulate the function of EFTB in electron transfer from Acyl-CoA dehydrogenases to the main respiratory chain (Q8IXQ9)	0.0013
NXT1	nuclear transport factor 2 like export factor 1	Nuclear export protein; stimulates export of NES-containing proteins and involved in transport of U1 snRNA, tRNA and mRNA (Q9UKK6)	0.0014
PIMREG	PICALM interacting mitotic regulator	May be involved in controlling metaphase-anaphase transition during mitosis (Q9BSJ6)	0.0014
FER	FER tyrosine kinase	Has a role in regulation of actin cytoskeleton and cell migration downstream of cell surface receptors for growth factors, including EGFR, PDGFRA and PDGRFB. Also involved in insulin receptor signalling, mast cell degranulation and leucocyte recruitment. (P16591)	0.0015

Symbol	Name	Description of function from UniProt (accession)	*P*-value
HERC5	HECT and RLD domain containing E3 ubiquitin protein ligase 5	Positively regulates innate antiviral response, also involved in bacterial clearance. (Q9UII4)	0.0015
BCHE	butyrylcholinesterase	Broad specificity esterase involved in acetylcholine inactivation. Can degrade neurotoxic organophosphate esters. (P06276)	0.0016
BAIAP3	BAI1 associated protein 3	Involved in endosome to Golgi retrograde transport. May mediate endosome fusion to trans-Golgi network via interactions with SNARE. Involved in regulation of neurotransmitter and hormone secretion (O94812)	0.0017
VPS36	vacuolar protein sorting 36 homolog	Component of the ESCRT-II complex, involved in sorting of endosomal cargo proteins into multivesicular body formation. May be involved in transcription regulation. (Q86VN1)	0.0018
TTLL12	tubulin tyrosine ligase like 12	Negatively regulates post-transcriptional modifications of tubulin. Has a role in mitosis and maintaining chromosome number stability. (Q14166)	0.0019
CREB5	cAMP responsive element binding protein 5	Activates transcription (Q02930)	0.0021
TFEB	transcription factor EB	Transcription factor, master regulator of lysosome biogenesis/exocytosis, autophagy, lipid catabolism and immune response. (P19484)	0.0021
ADCYAP1	adenylate cyclase activating polypeptide 1	Stimulates adenylate cyclase in pituitary cells. Promotes neuron projection development. Induces long-lasting increases in intracellular calcium in chromaffin cells. Involved in glucose homeostasis by inducing insulin secretion by beta cells (P18509)	0.0022
PDGFB	platelet derived growth factor subunit B	Growth factor required for normal embryonic development, cell proliferation, migration, survival and chemotaxis. Potent mitogen for mesenchymal cells. Important in wound healing. (P01127)	0.0022
FLNC	filamin C	Muscle specific filamin, important for sarcomere assembly and organisation. (Q14315)	0.0024

Thirteen of thirty-six previously reported variants[18,65,66,68,124] were detected in this GWAS. The variants were all rare with minor allele counts between 1 and 164 and generally had very small betas (median= -0.253) and non-significant *p*-values (median=0.4486). Only one *DNAH7* variant, rs115474479, had a *p*-value < 0.05. This variant had a beta of 3.1 with two heterozygotes in the case group and one heterozygote in the control group.

Table 24. Comparison of genes and variants detected in previous studies with same genes in this GWAS. For each gene, the gene-based p-value (SKAT-O) is shown.

The specific variants reported in each study are given where available. If the variant appeared within UKBB WES data, its beta and p-value are given.

		Genes	Variants		
Study	Gene	*p*-value	Variant	BETA	*p*-value
Lee *et al.* (2022)[68]	*RTN4*	0.890			
	RAB5A	0.463			
	CRYBG1	0.712			
	RGS22	0.025	rs993516236		
	APBB1IP	0.036	rs750180116		
	HEPHL1	0.808	rs756695159		
	BHLHE41	0.839	rs371168594		
	ARID3A	0.662	rs911982273		
	C5AR1	0.546	rs145736934		
	SPTLC3	0.620	rs749277943		
	CPT1B	0.417	rs745528078		
	FAM227A.	0.511			
Shaoul *et al.* (1999)[65]	*APC*	0.640			
Prinsley *et al.* (2019)[124]	*EGFL8*	0.661	rs141826798	-0.253	0.440
	BTNL9	1.000	rs367635312	-0.055	0.796
Cardenas *et al.* (2023)[126]	*NEB*	0.267	rs201548700	-1.187	0.666
			rs114089598	-0.255	0.465
			rs764064217		
	DNAH7	0.487	rs201273652		
			rs115474479	3.100	0.037
	DENND2C	0.350	rs189506550		
			rs61753528	0.224	0.469
	NBEAL1	0.119	rs199629983	-0.510	0.552
		0.119	rs180771101	0.487	0.238
	PRRC2C	0.909	rs148813704	-0.391	0.338
			rs138220849	-1.202	0.322
	SHC2	0.196	rs201010410	1.928	0.280
			rs768095487		
James *et al.* (2010)[66]	*GJB2*	0.138	rs80338939	-0.377	0.170
			rs111033196		
			rs111033222		
			rs72474224		
			rs76838169		
			rs1555046611		
			rs35887622	-0.101	0.636
			rs2274084		

4.3.3 Gene set enrichment analysis of UK BioBank data

Enriched processes in single variant analysis include adhesion, calcium transport, developmental processes and ciliary action via dyneins

150 Gene Ontology (GO) terms were enriched amongst genes with at least one $p<0.05$ variant from UKBB single variant association tests, including 58 biological processes, 67 cellular compartments and 25 molecular functions (**SI Table 4**). The large number of enriched processes probably reflects both the large number of genes containing significant SNPs (2377) and the nested nature of GO terms. These terms were collapsed to the most significant parent term in each unbroken chain of significant terms (**Table 25**). Significantly enriched pathways included:

- Cell adhesion (*cell periphery, homophilic cell adhesion via plasma membrane adhesion molecules, cell junction, cell junction assembly*, $p=1.3\text{x}10^{-21}$ - 0.0134).
- Cardiac action potential (*membrane depolarization during cardiac muscle cell action potential, voltage-gated calcium channel activity involved in cardiac muscle cell action potential, cell-cell signalling involved in cardiac conduction, regulation of heart rate by cardiac conduction, cardiac muscle cell action potential involved in contraction, cardiac muscle cell contraction*, $p=7.38\text{x}10^{-3}$ - 0.0325).
- Calcium binding and transport (*calcium ion binding, calmodulin binding, calcium channel complex, calcium ion import across plasma membrane, calcium ion transmembrane import into cytosol* $p=4.83\text{x}10^{-15}$ - 0.0385).
- Cytoskeleton organization (*cytoskeleton, cytoskeleton organization, actin filament-based process, cytoskeletal protein binding, cluster of actin-based cell projections, cell projection organization*, $p=1.34\text{x}10^{-8}$ - 0.0101)
- Ciliary activity (*minus-end-directed microtubule motor activity, dynein light chain binding, dynein intermediate chain binding, USH2 complex, stereocilia coupling link, dynein complex*, $p=5.57\text{x}10^{-7}$ - 0.043).

Table 25. Significantly enriched GO terms in UK BioBank whole exome single variant results.

Showing the most significant term in each unbroken chain of significant terms ascending the hierarchy. Results are sorted by GO database: Biological processes (GO:BP), molecular functions (GO:MF) and cellular compartments (GO:CC)

	GO term	Term ID	p	Size	N Genes
GO:BP	homophilic cell adhesion via plasma membrane adhesion molecules	GO:0007156	2.87×10^{-21}	168	61
	multicellular organismal process	GO:0032501	6.77×10^{-9}	7669	958
	cytoskeleton organization	GO:0007010	1.10×10^{-7}	1512	252
	anatomical structure development	GO:0048856	6.12×10^{-7}	5899	511
	actin filament-based process	GO:0030029	1.16×10^{-5}	805	139
	membrane depolarization during cardiac muscle cell action potential	GO:0086012	7.38×10^{-4}	21	3
	cellular glucuronidation	GO:0052695	0.0019	21	11
	sensory perception of mechanical stimulus	GO:0050954	0.0021	188	36
	detection of mechanical stimulus	GO:0050982	0.0024	55	17
	cell-cell signaling involved in cardiac conduction	GO:0086019	0.0027	32	3
	positive regulation of cellular component organization	GO:0051130	0.0053	1116	162
	regulation of heart rate by cardiac conduction	GO:0086091	0.0054	40	3
	supramolecular fiber organization	GO:0097435	0.0061	842	136
	cell projection organization	GO:0030030	0.0101	1613	239
	cardiac muscle cell action potential involved in contraction	GO:0086002	0.0108	50	3
	cell junction assembly	GO:0034329	0.0134	449	39
	calcium ion import across plasma membrane	GO:0098703	0.0155	46	3
	transport	GO:0006810	0.0167	4350	574
	cellular response to stimulus	GO:0051716	0.0181	7394	911
	trans-synaptic signaling by BDNF	GO:0099191	0.0308	5	5
	signaling	GO:0023052	0.0322	6447	781
	cardiac muscle cell contraction	GO:0086003	0.0325	72	3
	calcium ion transmembrane import into cytosol	GO:0097553	0.0385	197	4
	retina homeostasis	GO:0001895	0.0406	57	17
	striated muscle cell development	GO:0055002	0.0497	73	16
GO:CC	cell periphery	GO:0071944	1.30×10^{-21}	6228	873
	cytoplasm	GO:0005737	5.55×10^{-13}	12345	1508
	plasma membrane bounded cell projection	GO:0120025	6.34×10^{-12}	2277	303
	membrane	GO:0016020	4.71×10^{-11}	9864	1229
	endomembrane system	GO:0012505	3.09×10^{-9}	4777	613
	cytoskeleton	GO:0005856	1.34×10^{-8}	2430	360
	intracellular vesicle	GO:0097708	8.05×10^{-7}	2518	347
	cell junction	GO:0030054	1.24×10^{-4}	2230	311
	bounding membrane of organelle	GO:0098588	1.48×10^{-4}	2203	296

	GO term	Term ID	p	Size	N Genes
	cluster of actin-based cell projections	GO:0098862	5.85x10^{-4}	168	39
	cell leading edge	GO:0031252	7.68x10^{-4}	427	66
	USH2 complex	GO:1990696	0.0027	4	4
	apical part of cell	GO:0045177	0.0047	473	78
	I band	GO:0031674	0.0048	149	29
	stereocilia coupling link	GO:0002139	0.0066	8	5
	supramolecular polymer	GO:0099081	0.0069	1062	137
	calcium channel complex	GO:0034704	0.0144	84	3
	mismatch repair complex	GO:0032300	0.0263	8	3
	midbody	GO:0030496	0.0379	206	37
	dynein complex	GO:0030286	0.0430	54	16
	extracellular region	GO:0005576	0.0454	4213	428
	chiasma	GO:0005712	0.0465	2	2
	early endosome membrane	GO:0031901	0.0465	192	23
GO:MF	calcium ion binding	GO:0005509	4.83x10^{-15}	726	139
	minus-end-directed microtubule motor activity	GO:0008569	5.57x10^{-7}	17	13
	cytoskeletal protein binding	GO:0008092	1.04x10^{-4}	1002	145
	dynein intermediate chain binding	GO:0045505	1.05x10^{-4}	37	17
	calmodulin binding	GO:0005516	1.14x10^{-4}	206	48
	adenyl ribonucleotide binding	GO:0032559	3.53x10^{-4}	1560	232
	voltage-gated calcium channel activity involved in cardiac muscle cell action potential	GO:0086007	0.0024	5	2
	dynein light intermediate chain binding	GO:0051959	0.0025	28	13
	protein-containing complex binding	GO:0044877	0.0032	1752	245
	glucuronosyltransferase activity	GO:0015020	0.0056	34	13
	GTPase regulator activity	GO:0030695	0.0125	492	75
	structural constituent of muscle	GO:0008307	0.0263	42	9

Enriched pathways overlap with functions identified in previous GoC WES study

Our previous GoC study[126] used whole exome sequencing data from 21 individuals from 10 affected families to identify potentially causative genes through two methods: family overlap analysis and gene burden/TRAPD analysis. In family overlap analysis, common variants shared by cholesteatoma cases within families were identified and the set of genes carrying variants which appeared in 2 or more families was analysed used g:Profiler. For burden analysis, TRAPD[237] software was used to compare variant frequencies to frequencies recorded in public databases to identify variants which were significantly more common amongst affected families and the results subject to g:Profiler analysis. 17 significantly enriched GO terms were identified, of which 10 were also enriched in the full UKBB WES g:Profiler result (**Table 26**).

Processes with three main themes were enriched in both studies: *calcium ion binding; dynein motor activity*; and *Gtpase activity*.

Table 26. Comparison of gene set enrichment analysis results of GoC WES family study and UK BioBank.

Showing enriched GO terms only with columns showing p-value for family and burden tests from the GoC WES study as well as UKBB WES. Terms that were enriched in both the GoC WS study and UKBB WES single variant data are bolded.

Terms	Term ID	Source	Family overlap *p*-value	Burden analysis *p*-value	UKBB WES *p*-value
Cation binding	**GO:0043169**	**GO:MF**	**0.00442**	**5.48x10^{-4}**	**4.73x10-^{10}**
Calcium ion binding	**GO:0005509**	**GO:MF**	**0.00988**	**0.00383**	**4.83x10^{-15}**
Extracellular matrix structural constituent	GO:0005201	GO:MF	0.00608	3.00x10^{-4}	
Ion binding	**GO:0043167**	**GO:MF**	**3.17x10^{-4}**	**2.54x10^{-5}**	**4.12x10^{-11}**
Gtpase regulator activity	**GO:0030695**	**GO:MF**	**3.48x10^{-4}**		**0.012512**
Nucleoside-triphosphatase regulator activity	GO:0060589	GO:MF	7.24x10^{-4}		0.012512
Gtpase activator activity	GO:0005096	GO:MF	0.00608		
Guanyl-nucleotide exchange factor activity	**GO:0005085**	**GO:MF**	**0.00581**		**0.012617**
Motor activity	GO:0003774	GO:MF		1.05x10^{-5}	
Cytoskeletal protein binding	**GO:0008092**	**GO:MF**		**0.00118**	**0.000104**
Cargo receptor activity	GO:0038024	GO:MF		0.00764	
Metal ion binding	**GO:0046872**	**GO:MF**		**6.58x10^{-4}**	**4.38x10^{-10}**
Atp-dependent microtubule motor activity	GO:1990939	GO:MF		4.19x10^{-5}	
Microtubule motor activity	**GO:0003777**	**GO:MF**		**2.00x10^{-4}**	**0.00856**
Dynein intermediate chain binding	**GO:0045505**	**GO:MF**		**2.36x10^{-4}**	**0.000105**
Dynein light intermediate chain binding	**GO:0051959**	**GO:MF**		**1.08x10^{-4}**	**0.00253**

Enriched pathways from gene-based tests do not have much direct overlap with single variant gene set enrichment

Enrichment analysis of the gene-based tests generated a different, smaller set of enriched pathways- to the single variant-level enrichment tests (**Table 27**). There is some overlap in functions: the *cell junction* was an enriched compartment in both analyses and *cell junction disassembly* was enriched in gene-based test results. Different terms related to cell signalling were also enriched in both.

Table 27. Significantly enriched GO terms in UK BioBank whole gene-level variant results.

Results are divided into GO Biological Processes, GO Cellular Compartment and GO Molecular Function. Due to the small number of results, they have not been collapsed to the most significant term in each unbroken chain.

	Pathway	GO ID	*p*-value	Term size	N Genes
GO:BP	vacuolar transport	GO:0007034	0.0119	171	18
	regulation of smooth muscle cell differentiation	GO:0051150	0.0181	36	8
	nerve growth factor signaling pathway	GO:0038180	0.0302	13	2
	regulation of cAMP-dependent protein kinase activity	GO:2000479	0.0319	12	4
	ISG15-protein conjugation	GO:0032020	0.0483	6	2
GO:CC	cytoplasm	GO:0005737	2.12×10^{-06}	12345	531
	late endosome	GO:0005770	0.00406	314	9
	membrane	GO:0016020	0.00729	9864	402
	cell junction	GO:0030054	0.0224	2230	84
	endoplasmic reticulum lumen	GO:0005788	0.0479	313	23
GO:MF	protein binding	GO:0005515	0.00101	14838	602
	ISG15 transferase activity	GO:0042296	0.00647	4	2
	beta-1 adrenergic receptor binding	GO:0031697	0.0332	3	2

4.3.4 Enriched processes in FinnGen microarray data overlap with UK BioBank whole exome

Amongst FinnGen single variants with *p*-value < 0.05, 959 terms were enriched (606 GO:BP, 209 GO:CC, 144 GO:MF). This larger number is attributed to the greater number of genes containing significant SNPs due to the larger number of cases. 112 terms were enriched in both FinnGen and UKBB single variant data, meaning 75% of terms enriched in UKBB WES were also enriched in FinnGen, but only 12% of FinnGen terms were enriched in UKBB WES. Terms that were enriched in UKBB but *not* FinnGen were mostly cytoskeletal, ciliary or dynein related as well as terms related to cardiac regulation.

FinnGen enriched pathways included terms related to cell-cell adhesion, with *homophilic cell adhesion via plasma membrane adhesion molecules* being the most significantly enriched process in both FinnGen and UKBB. Calcium ion binding and transport were also enriched (under terms *localisation*, *small molecule binding*, *homeostatic process transporter activity* and *cell adhesion* in **Table 28;** *p*-value for *calcium ion binding* = 1.03x10^{-24}; *p*-value for *calcium ion transport* = 1.62x10^{-7}). Many FinnGen enriched terms were related to neuronal development and function, although these are largely collapsed into developmental process and cell projection terms (**Table 28)**. Cytoskeletal and ciliary function were also implicated in this data, including terms such as *actin-filament-based process, cytoskeletal motor activity, GTPase regulator activity,* and *minus-end-directed microtubule motor activity*. Unlike UKBB WES results, dyneins were not directly implicated.

Table 28. Enriched terms in FinnGen data collapsed to most significant term within an unbroken ascending chain of enriched terms.

P-value is corrected using g:SCS, a multiple testing correction method packaged with g:Profiler which accounts for the hierarchical nature of GO terms.

source	**Pathway**	**GO ID**	**p**	**Size**	**N genes**
GO:BP	anatomical structure development	GO:0048856	$1.53E^{-61}$	5899	2224
	multicellular organismal process	GO:0032501	$2.53E^{-51}$	7669	2733
	cell adhesion	GO:0007155	$2.81E^{-47}$	1512	648
	plasma membrane bounded cell projection organization	GO:0120036	$8.43E^{-37}$	1570	610
	Transport	GO:0006810	$6.16E^{-35}$	4350	1611
	regulation of cell communication	GO:0010646	$9.25E^{-34}$	3443	1323
	regulation of signaling	GO:0023051	$9.41x10^{-34}$	3437	1321
	cellular response to stimulus	GO:0051716	$2.78x10^{-31}$	7394	2550
	cell junction organization	GO:0034330	$2.66x10^{-28}$	759	312
	cell motility	GO:0048870	$4.21x10^{-21}$	1709	562
	actin filament-based process	GO:0030029	$7.13x10^{-20}$	805	364
	cytoskeleton organization	GO:0007010	$1.34x10^{-19}$	1512	567
	locomotion	GO:0040011	$3.54x10^{-12}$	1234	363
	phosphate-containing compound metabolic process	GO:0006796	$1.74x10^{-11}$	2571	921
	growth	GO:0040007	$7.34x10^{-11}$	939	296
	supramolecular fiber organization	GO:0097435	$2.28x10^{-10}$	842	343
	protein modification process	GO:0036211	$8.12x10^{-10}$	3031	1046
	homeostatic process	GO:0042592	$6.53x10^{-9}$	1712	614
	extracellular structure organization	GO:0043062	$1.24x10^{-8}$	325	152
	external encapsulating structure organization	GO:0045229	$1.66x10^{-8}$	326	152
	organonitrogen compound metabolic process	GO:1901564	$2.28x10^{-8}$	5986	1946
	lipid metabolic process	GO:0006629	$2.70x10^{-8}$	1388	514
	cell population proliferation	GO:0008283	$2.62x10^{-6}$	2006	703
	microtubule-based process	GO:0007017	$4.68x10^{-6}$	953	336
	cognition	GO:0050890	$9.35x10^{-5}$	322	114
	membrane organization	GO:0061024	0.000177	815	236
	cell recognition	GO:0008037	0.000316	156	55
	cell death	GO:0008219	0.000588	1988	513
	cellular glucuronidation	GO:0052695	0.000982	21	15
	negative chemotaxis	GO:0050919	0.0011	47	30
	ovulation cycle	GO:0042698	0.002504	72	24
	sensory perception of sound	GO:0007605	0.002986	165	50
	reproduction	GO:0000003	0.00358	1552	483
	AV node cell action potential	GO:0086016	0.004124	10	7
	immune response-activating signaling pathway	GO:0002757	0.004789	465	182

source	Pathway	GO ID	p	Size	N genes
	regulation of primary metabolic process	GO:0080090	0.013761	5591	1723
	protein localization to postsynaptic specialization membrane	GO:0099633	0.025999	28	11
	cell cycle process	GO:0022402	0.032969	1276	418
	cellular component maintenance	GO:0043954	0.038083	73	30
	visual perception	GO:0007601	0.038612	219	84
	regulation of nitrogen compound metabolic process	GO:0051171	0.039341	5440	1672
	Fc receptor mediated stimulatory signaling pathway	GO:0002431	0.045374	32	21
GO:CC	cytoplasm	GO:0005737	5.85x10^{-87}	12345	4083
	cell periphery	GO:0071944	1.05x10^{-78}	6228	2314
	membrane	GO:0016020	1.27x10^{-69}	9864	3341
	cell junction	GO:0030054	4.31x10^{-67}	2230	895
	cell projection	GO:0042995	3.16x10^{-66}	2389	1041
	endomembrane system	GO:0012505	5.78x10^{-43}	4777	1713
	cytoskeleton	GO:0005856	8.43x10^{-34}	2430	928
	somatodendritic compartment	GO:0036477	1.44x10^{-25}	848	330
	vesicle	GO:0031982	1.12x10^{-24}	4004	1395
	cell leading edge	GO:0031252	2.52x10^{-16}	427	204
	apical part of cell	GO:0045177	3.03x10^{-16}	473	211
	monoatomic ion channel complex	GO:0034702	5.68x10^{-13}	346	165
	cell surface	GO:0009986	7.65x10^{-13}	903	339
	cell body	GO:0044297	2.57x10^{-12}	565	199
	nucleoplasm	GO:0005654	1.00x10^{-11}	4220	1345
	organelle subcompartment	GO:0031984	1.79x10^{-10}	1521	543
	supramolecular polymer	GO:0099081	4.80x10^{-9}	1062	365
	extracellular region	GO:0005576	6.46x10^{-9}	4213	1338
	site of polarized growth	GO:0030427	2.03x10^{-8}	174	74
	basal part of cell	GO:0045178	3.52x10^{-8}	300	129
	cluster of actin-based cell projections	GO:0098862	1.89x10^{-7}	168	79
	receptor complex	GO:0043235	0.000175	523	141
	neurotransmitter receptor complex	GO:0098878	0.000859	46	25
	protein complex involved in cell adhesion	GO:0098636	0.000909	59	31
	guanyl-nucleotide exchange factor complex	GO:0032045	0.003333	24	9
	collagen trimer	GO:0005581	0.005474	91	44
	DNA repair complex	GO:1990391	0.010103	22	15
	trans-Golgi network	GO:0005802	0.038189	264	83
GO:MF	ion binding	GO:0043167	6.06x10^{-33}	6146	2247
	protein binding	GO:0005515	6.43x10^{-30}	14838	4826
	ATP binding	GO:0005524	5.61x10^{-24}	1500	632
	carbohydrate derivative binding	GO:0097367	1.25x10^{-20}	2304	911
	transporter activity	GO:0005215	8.36x10^{-15}	1239	513

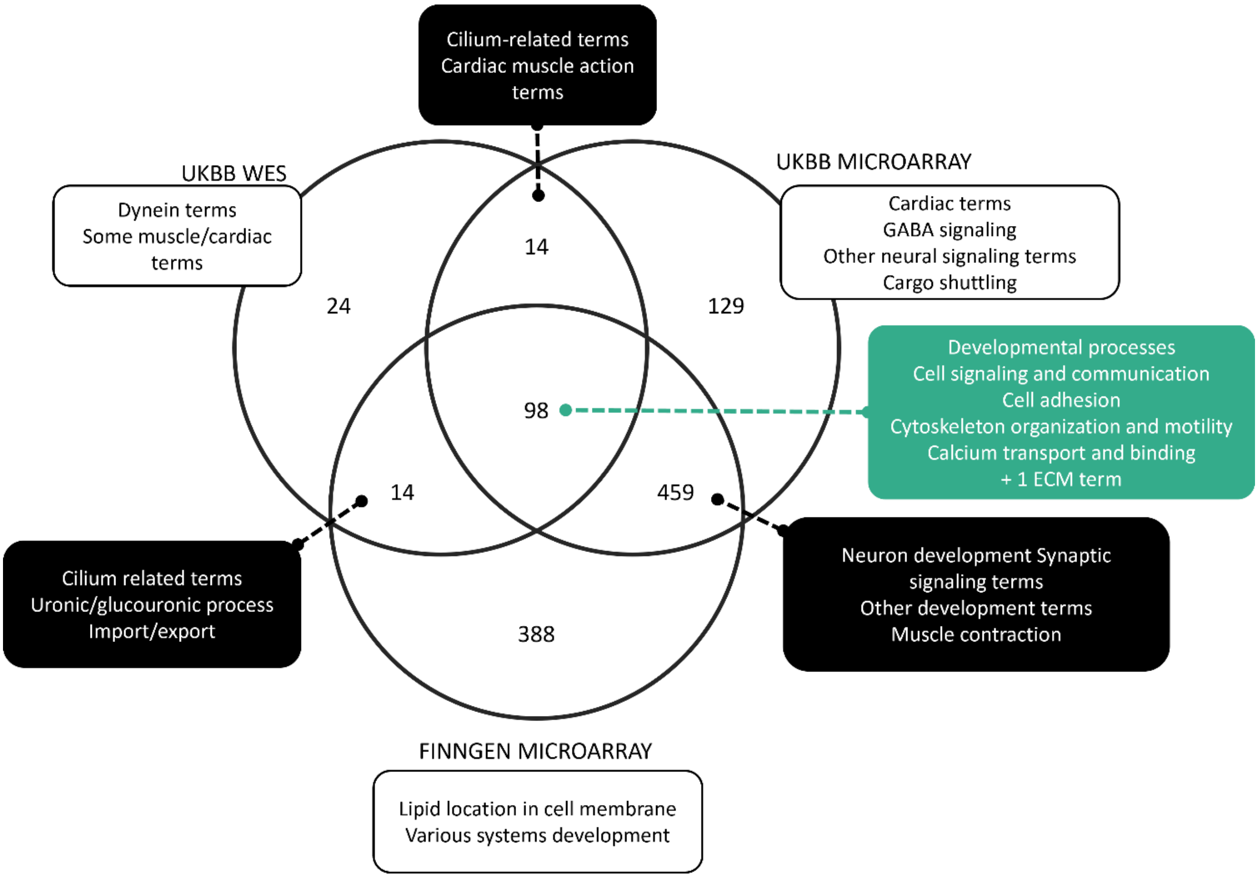
source	Pathway	GO ID	p	Size	N genes
	GTPase regulator activity	GO:0030695	1.09×10^{-11}	492	225
	phosphotransferase activity, alcohol group as acceptor	GO:0016773	1.00×10^{-9}	697	292
	ATP-dependent activity	GO:0140657	5.84×10^{-9}	580	252
	protein-containing complex binding	GO:0044877	1.49×10^{-7}	1752	648
	cytoskeletal motor activity	GO:0003774	4.70×10^{-6}	115	59
	lipid binding	GO:0008289	7.07×10^{-6}	841	321
	glutamate receptor activity	GO:0008066	7.30×10^{-6}	27	19
	protein-macromolecule adaptor activity	GO:0030674	0.000159	947	330
	extracellular matrix structural constituent	GO:0005201	0.000189	167	81
	transmembrane receptor protein tyrosine phosphatase activity	GO:0005001	0.000431	17	12
	minus-end-directed microtubule motor activity	GO:0008569	0.00057	17	14
	metallopeptidase activity	GO:0008237	0.005226	187	84
	phosphatidyl phospholipase B activity	GO:0102545	0.007775	11	7
	syntaxin-1 binding	GO:0017075	0.014796	19	12
	sulfur compound binding	GO:1901681	0.018948	268	53
	structural constituent of presynaptic active zone	GO:0098882	0.022389	5	5
	structural constituent of muscle	GO:0008307	0.032047	42	22
	postsynaptic neurotransmitter receptor activity	GO:0098960	0.033892	72	31

Results are supported by UK BioBank microarray analysis

The gene set enrichment analysis performed on UKBB microarray single variant results (**SI Table 5**) agrees with the FinnGen microarray results and includes more terms related to neural development and synapse function compared to UKBBWES. As in the FinnGen microarray results, ciliary terms were not enriched, supporting the absence of these terms being due to array-based approaches missing rare *DNAH/DNAI* variants.

In total, 93 terms were enriched across all data sets (**Figure 27**). Of the 150 enriched terms in the whole exome data, 112 were also enriched in the UKBB microarray data (75%). Meanwhile, 700 terms were enriched in the UKBB microarray data: only 16% of these terms were also enriched in the whole exome data. The UKBB microarray data shared more enriched terms with the FinnGen data (557 terms; 78% of UKBB microarray terms), suggesting that the similarity of technologies made the results more comparable. Note that slight differences in the precise processes enriched can lead to very similar processes appearing across sectors in **Figure 27**: the term *cardiac muscle contraction* is enriched in UKBB WES but not microarray, though the very similar term *cardiac muscle cell contraction* is enriched in both. For this reason, I give general descriptive terms of the types of pathways and processes enriched in different sectors of the diagram.

Figure 27. Venn diagram showing number of overlapping terms in UKBB WES, UKBB Microarray and FinnGen. *A qualitative description of common terms is given for each sector.*


Cilium-related terms
Cardiac muscle action terms
UKBB WES
Dynein terms
Some muscle/cardiac terms
UKBB MICROARRAY
Cardiac terms
GABA signaling
Other neural signaling terms
Cargo shuttling
14
24
129
Developmental processes
Cell signaling and communication
Cell adhesion
Cytoskeleton organization and motility
Calcium transport and binding
+ 1 ECM term
98
14
459
Cilium related terms
Uronic/glucouronic process
Import/export
Neuron development Synaptic signaling terms
Other development terms
Muscle contraction
388
FINNGEN MICROARRAY
Lipid location in cell membrane
Various systems development

Pathways common to all data sets included various developmental processes, cell signalling and communication processes, cell-cell adhesion, cytoskeletal processes, and calcium transport and binding. Some individual terms were also enriched in all three sets, such as *minus-end-directed microtubule motor activity* and the *extracellular matrix* cellular compartment. The terms unique to UKBB whole exome were largely dynein-related (while some cardiac muscle terms were unique to UKBB whole exome, other similar terms were enriched in other data sets).

4.3.5 Sensitivity analysis suggests study was underpowered for rare variants

The size of this study was limited by the number of cases present in UKBB. I performed post-hoc sensitivity analysis to determine power for different variant frequencies and effect sizes. I also determined the minimum effect size that would be detected under the current study design of 1,000 cases and 5,000 controls at 80% power.

Sensitivity analysis suggests that the power to detect rare variants was very low (**Table 29**), although common variants with GRR > 1.5 should be detectable with >90% sensitivity. The best improvements in predicted power with study size occur for rare variants, with considerable improvements in sensitivity going from 2,000 to 5,000 and 5,000 to 10,000 cases. However, there is little improvement in sensitivity for common variants when going from 5,000 to 10,000.

Table 29. Power under current study design and minimum sample size required for 80% power for different minor allele frequencies and genotype relative risks

	Allele frequency	Genotype relative risk*	Power with 1,000 cases	Sample size for >0.8 power
Ultra rare variants	0.001	2	0	>100,000
	0.001	5	0.002	10,000
	0.001	7	0.013	5,000
Rare variants	0.01	1.5	0	>100,000
	0.01	2	0.008	7,000
	0.01	5	0.990	700
Common variants	0.25	1.2	0.011	10,000
	0.25	1.5	0.904	900
	0.25	2	1	300

The threshold detectable genetic risk ratio (GRR) was consistently higher than the GRRs calculated from single variant association test results (**Figure 28**), which was expected given the lack of significant results. We can only confidently state that no effects exist that are above this threshold: otherwise, the effect sizes cannot be distinguished from noise. If the effect sizes observed in this study were accurate, a sample size of 5,000 would be required to detect them statistically (**Table 30**).

Figure 28. Comparison of genetic risk ratio calculated from odds ratio for different minor allele frequencies, versus threshold GRR.

The plot shows the genotype relative risks (GRRs) obtained from single variant whole exome analysis (scatter) plotted against the threshold detectable GRRs for 4 different sample sizes (line). Both the threshold GRR and calculated GRRs increase rapidly at low MAF. Rare variants have smaller effective sample sizes, making them more susceptible to larger variations in effect size estimate due to chance, leading to an increase in both the measured GRR and the threshold for significance.

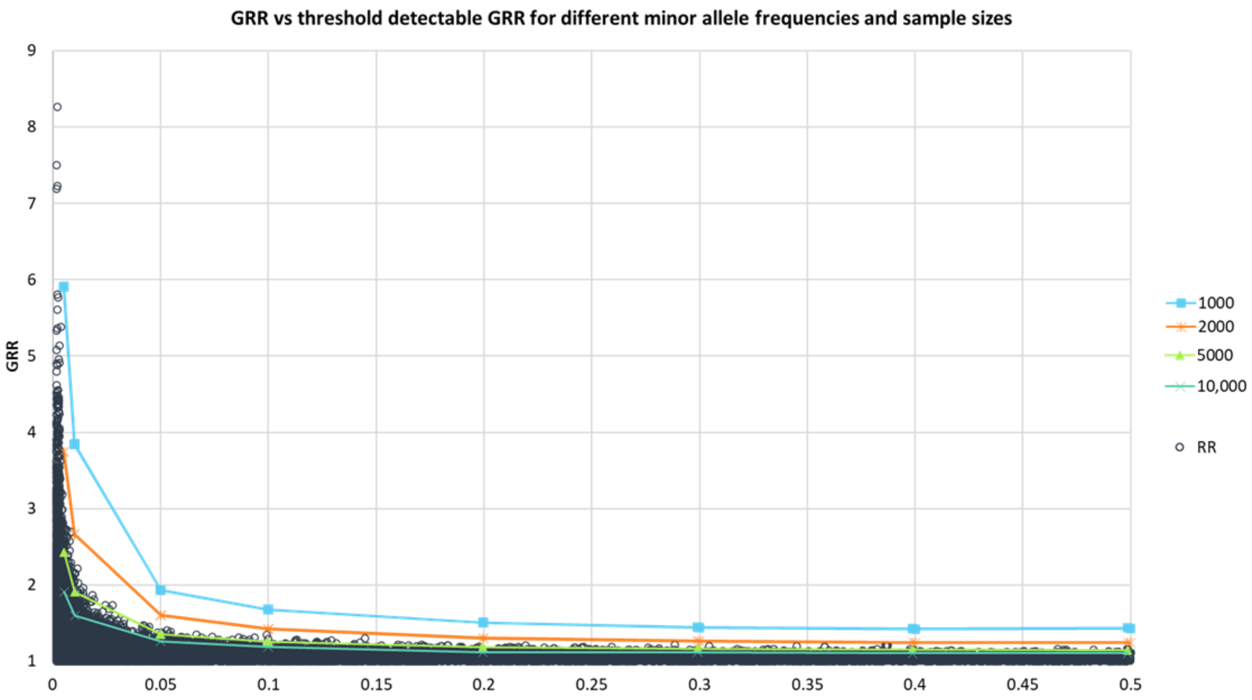


Table 30. Threshold detectable GRR for different minor allele frequencies at 80% power, with N cases and case:control ratio of 1:5

Minor Allele frequency	N cases			
	1000	2000	5000	10,000
0.005	5.9	3.74	2.42	1.9
0.01	3.84	2.66	1.9	1.59
0.05	1.93	1.6	1.35	1.25
0.1	1.67	1.42	1.25	1.18
0.2	1.5	1.3	1.18	1.11
0.3	1.44	1.26	1.16	1.11
0.4	1.42	1.24	1.15	1.1
0.5	1.43	1.24	1.14	1.1

Overall, power analysis suggests this study is underpowered to detect very rare variants unless their effect sizes are very large, though this study should be capable of detecting common variants with GRR > 1.4.

4.4 Discussion

In this chapter, I performed single-variant and gene-level genome-wide association tests with UK BioBank whole exome (UKBB WES) data. I also performed pathway level analysis using g:Profiler to detect enriched processes amongst the genes carrying variants with *p*-value < 0.05. I used FinnGen summary statistics for comparison and identified common pathways enriched across both biobanks. I also performed single-variant and gene set enrichment tests using UKBB microarray data for better comparison with FinnGen data.

In this analysis, no single variant or gene met genome-wide significance. A lack of obvious signals in the Manhattan plot of results and no significant genes suggest that no individual loci have strong enough effects to identify with this cohort of 1,000 cholesteatoma cases. The top scoring genes and variants are associated with a variety of functions, making functional interpretation difficult. Power calculations support the need for a larger number of cases as this study's power to detect rare variant associations was low. The small sample size means rare variants or those with small effect sizes cannot be detected, resulting in high risk of type 2 error.

However, gene set enrichment analysis of WES single variant results reveals several significantly enriched processes. Within UKBB WES data, this includes terms related to cell adhesion, cytoskeleton organisation, calcium binding, cardiac muscle regulation and developmental processes. Most of these were also enriched in UKBB microarray and FinnGen microarray data. The microarray results were also enriched for neural development terms, whereas dynein binding was specific to UKBB WES.

The previous GoC whole exome study[126] of twenty-one individuals from ten affected families identified an overlapping set of processes enriched for deleterious variants including calcium binding, extracellular matrix organisation and ciliary motility. The latter is particularly interesting as UKBB WES data and the GoC implicate axonemal dyneins specifically. Meanwhile, calcium binding was enriched in the UKBB whole exome single variant results, previous GoC WES paper, and in FinnGen results; calcium binding is also an enriched process amongst dysregulated genes compared to skin from several expression studies (see *Semi-systematic review of global gene expression studies*). Agreement between studies regarding enriched pathways despite a lack of individually significant genes or variants suggests a genetic effect on cholesteatoma risk exists but it may be polygenic or heterogeneous.

Our previous WES paper identified enriched deleterious variants associated with ECM degradation, which is consistent with gene expression studies showing dysregulated ECM proteins and upregulated proteases[36,77,78,125]. This study did not identify genes associated with ECM degradation in UKBB WES data. Microarray data from UKBB and FinnGen did show enrichment of some relevant terms such as *extracellular structure organization*, *extracellular matrix structural constituent* and *enrichment of the extracellular region*. However, the number of enriched terms related to these was small compared to other functions and was not repeated in the UKBB WES data.

Cholesteatoma tissue also shows aberrant expression of immune genes and inflammation has been suggested to play a role in establishment and pathology[21]. Neither this nor our previous whole exome study identified enriched deleterious variants in immune pathways. Some immune-related terms were enriched in FinnGen, but the number was very small; no terms were present in the UKBB data.

The much larger number of enriched terms in FinnGen compared to UKBB (959 vs 150) may be due to the larger sample size for FinnGen (*n* = 1,447 compared to 1000) leading to a greater number of *p*-value < 0.05 variants due to increased power. While most terms enriched in the UKBB data were also present in the FinnGen data, FinnGen contained many additional enriched processes with the large variability making them difficult to interpret.

4.4.1 Enriched processes and associated genes

Cell adhesion, actin organisation and migration

Genes related to cell-cell adhesion and various terms associated with cytoskeletal organisation were enriched across both biobanks. Cell adhesion is mediated by membrane-bound adhesion molecules which form adhesion complexes including adherens junctions, gap junctions, desmosomes, hemidesmosomes[175]. Adhesion may be between cells or with the ECM and is required for physical anchoring of tissues and cell communication. In cancer, loss of cell-cell adhesion can result in increased adhesion of cells to the ECM, promoting invasiveness[175]. On the intracellular side of adhesion complexes, the proteins interact with the cytoskeleton so can also mediate cytoskeletal processes[175]. The most significant single variant in UKBB WES data was in *AMOTL2* ($p=5.71 \times 10^{-5}$, beta=1.78), which has roles in actin filament architecture and coupling to cell junctions[238]. The AMOTL2 protein forms a complex with the adhesion molecule E-Cadherin and regulates actin filament growth and organisation to maintain cellular geometry in epithelial tissue[238].

The cytoskeleton is primarily composed of microtubules, actin filaments, and intermediate filaments and is involved in processes related to cell shape, structure, and motility[239]. In this analysis, actin organisation terms were enriched specifically. A major role of actin is in amoeboid cell motility: this involves formation of focal adhesion points, changes in cell shape, and cell contraction[239]. Actin can form dynamic protruding structures such as lamellipodia[240] and produce contractile forces through interaction with myosin to propel cells across a substrate[241]. These contractile forces also result in morphological changes and ECM remodelling necessary for cell movement[242]. Taken together, enriched genetic variants affecting adhesion molecules and cytoskeleton organisation may indicate alterations in amoeboid cell motility.

The actin cytoskeleton is also involved in exocytosis, endocytosis, and intracellular transport[239]. Interestingly, the second-best gene, *PLEKH8* ($p=2.60 \times 10^{-4}$), encodes a cargo transport protein involved in transport from the Golgi complex, synthesis of glycosphingolipids, and in primary cilium formation[243]. Other top-scoring genes according to gene-based tests were *VPS36, BAIAP3*, which are also involved in cargo sorting and transport[244,245].

Enrichment of developmental and neurodevelopmental terms

The enriched GO term *system development* may also be relevant to cholesteatoma via cranial/ear morphology promoting susceptibility to repeated infection and debris collection through poor ventilation. Conditions affecting cranial morphology are associated with higher rates of cholesteatoma, including Turner syndrome, Down syndrome, and cleft palate[18]. In fact, one study[64] showed that the siblings of children with orofacial clefts were also at modestly increased risk of cholesteatoma, despite not having a cleft themselves, which the authors suggest may be due to accumulation of subclinical muscular defects. Alternatively. The *system development* term may be relevant to cholesteatoma development via the behaviour of the cholesteatoma tissue itself, for example its increased cell turnover. However, enriched developmental processes in GSEA should be interpreted with caution as these are broad categories containing many genes, and many different processes can form part of tissue development. This does not necessarily mean they will have any influence on either the morphology of the head/ear or the behaviour of cholesteatoma tissue.

Interestingly, both *systems development* and *nervous system development* were enriched terms in the primary analysis of single-variant results, while the analysis of gene-level association tests identified significant enrichment of beta-1 adrenergic receptor binding and nerve growth factor signalling pathways. Furthermore, several FinnGen enriched pathways were related to neural development or synaptic function and nervous system development was enriched in both FinnGen and UKBB. Findings of enriched genes in neurodevelopment and neural signalling pathways is reflected by findings in the epidemiology chapter of this thesis, where odds of epilepsy were increased amongst cholesteatoma cases in both UKBB and FinnGen. This suggests that increased rates of epilepsy in these cohorts may not only be as a side effect of invasive disease but may have a genetic basis. Possibly, epilepsy and cholesteatoma may be linked via developmental defects which may not cause any obvious syndromic appearance but subtly raise risk of both. However, it is also possible that the raised epilepsy rate in these groups is coincidental and may or may not be the cause of enriched terms related to neural development. Epilepsy with cholesteatoma is very rarely reported and has only occasionally been described as a complication of cholesteatoma[204,205], making the nature of any possible association obscure.

A role for calcium ions

A role for calcium ions is supported by this study, our previous WES study, and gene expression studies reviewed in *Semi-systematic review of global gene expression studies* but interpretation of this finding is difficult due to the diverse functions of calcium in the body. In the UKBB WES data, enrichment of calcium binding was driven by variants in voltage-gated calcium channel subunits *CACNA1G, CACNA2D1* and ion channel stabiliser *ANK2* (*p*=0.00013-0.00032; beta=1.67-2.27). These were amongst the most significant single variants and also drove enrichment of terms related to cardiac regulation. Voltage-gated calcium channels are involved in calcium homeostasis, signalling in neurons and calcium influx in cardiac muscle contraction. Defects in members of the CACNA family are also associated with cardiac and neural problems, including epilepsy[246].

Most calcium in the human body is found in skeleton, where it forms an integral part of the bone matrix[247], which may be relevant to the bone loss seen in cholesteatoma. As a signalling molecule, calcium has important roles in immune function[248] and wound-healing[249]. While immune function and inflammation are implicated in cholesteatoma pathology, there is little evidence of changes to immune genes being directly responsible from this or our previous whole exome study. Cholesteatoma also resembles wound-healing tissue with ECM degradation, increased migration and proliferation, and angiogenesis[2,35]. Impaired calcium signalling could be involved in provoking a chronic wound-like response. Interestingly, calcium localisation is also thought to be important in polycystic kidney disease, where low intracellular calcium leads to upregulated cAMP, driving increased fluid secretion and activation of the MAPK pathway[250]. Thus, while there are many ways in which calcium might be involved in cholesteatoma pathology, it is difficult to determine which are relevant.

4.4.2 Ciliary dysfunction is implicated in cholesteatoma

This study identified enrichment of several GO terms related to cytoskeletal motility, specifically via action of axonemal dynein. Terms including cell projection, cytoskeleton organisation, dynein complex and dynein intermediate chain binding were enriched, mostly due to rare *DNAH* and *DNAI* variants (**SI Table 6**). Our previous study also identified deleterious *DNAH* variants in multiple affected families.

Dynein is a protein motor which moves towards the minus end of cytoskeletal microtubules. These are divided into cytoplasmic dyneins (*DYN*-) which transport vesicles along the microtubules, and axonemal dyneins (*DNA*-) which drive ciliary motility[251]. Enrichment of *DNAH* and *DNAI* variants, along with enrichment of protein products localised to the cell projection, suggests that ciliary function is important in cholesteatoma. Cilia have sensory and developmental roles, as well as physically clearing debris from the airways and middle ear.

Ciliopathies can present in a variety of ways but often result in developmental defects, with the retinal, renal and cerebral disease most common[252]. Polycystic kidney disease is an example of a ciliopathy resulting in the development of multiple fluid-filled renal cysts and is thought to arise through defects in calcium sensing via ion channels on the primary cilia[250]. Primary ciliary dyskinesia results in increased susceptibility to recurrent respiratory, ear, nose and sinus disease, and some forms are also associated with organ laterality defects or total situs inversus[253]. These arise through immobility of the primary cilia resulting in poor clearance of the ears, sinuses, and airways, as well as developmental anomalies through improper distribution of signalling molecules during embryogenesis[254]. *DNAH1* and *DNAH5* mutations are causal for some forms of primary ciliary dyskinesia[17], highlighting the importance of dyneins in resistance to infection and developmental processes. Another interesting link is between PCD, bronchiectasis and sinus infection, which were increased amongst cholesteatoma cases in this cohort. PCD with bronchiectasis, sinus infection and chronic sinusitis is known as Kartagener's syndrome[17]. The presence of these diseases in the case group further supports ciliary impairment (although probably not PCD itself as this is very rare).

Another interesting feature common to both this analysis and our previous family study is the involvement of the USH2 complex. In the family study, the USH2 complex was enriched in variants detected in family overlap analysis due to variants in *USH2A*, *ADGRV1*, and *WHRN*: one family carried an *ADGRV1* variant and an *USH2A* variant, while another family carried 2 *WHRN* variants. A variant in *USH2A*, rs10779261, was one of the most significant single variants in my whole exome analysis, leading to enrichment of the USH2 complex in my gene set analysis. The USH2 complex is involved in stereocilia development; stereocilia are non-motile cell projections required for mechanotransduction in the middle ear[255]. Defects in the USH2 complex result in Usher syndrome, a condition involving retinitis pigmentosa and deafness. In the eye, USH2A is associated with the photoreceptor ciliary complex, which consists of an outer and inner segment linked by connecting cilia. USH2A is localised to the inner segment, within a recesses surrounding the connecting cilia[72]. While stereocilia are actin-based, non-motile structures, the photoreceptor connecting cilium is an axonemal cilium[256]. Though primarily expressed in the retina and cochlea, usherin is present in the basement membranes of many tissues[257]. However, it has no established roles in the middle ear.

Ciliary function was also implicated in the FinnGen enriched terms, but dynein-binding was not implicated directly. The Finnish population is genetically distinct from other European populations due to recent bottlenecking[258], so it is possible that different causal alleles vary in prevalence between these populations. However, this may also be due to the differences between microarray and exome data. While both data sets were processed to only consider non-imputed, non-synonymous variants in protein-coding regions, whole exome can capture rare variants whereas microarray methods can only capture variants represented by a given probe set. For example, 15 of 25 significant *DNAH* family variants in UKBB had allele frequencies less than 0.001 and only 6 of these variants were present in the FinnGen probe set. Generally, differences in the regions and variants measured by microarray compared to whole exome may affect which processes appear enriched by altering the overall distribution of *p*-values between pathways.

4.4.3 Other functions of top-scoring genes and variants

The most significant gene from gene-based SKAT-O tests was *ABCC8* (p=6.94x10^{-5}), which encodes a sulfonylurea receptor involved in insulin transport hence is primarily known for its role in diabetes[259], so there is no obvious reason for an association with cholesteatoma.

PDGFB was also amongst the top-scoring genes; this encodes platelet-derived growth factor subunit B, which may be relevant to cholesteatoma as one of the chemokines involved in the interaction between fibroblasts and keratinocytes promoting hyperproliferation[20].

4.4.4 Study limitations

Limited sample size adversely affects sensitivity

A major limitation of this study was its small size. The number of cases was constrained by the number present in UKBB, though efforts were made to identify all likely cases by expanding the case definition to include codes linked to cholesteatoma and its management. Sensitivity analysis suggests that this study was underpowered for detection of rare variants and variants with small effect sizes. Based on the effect sizes calculated in this study, power calculations indicated that 5,000 cases may be needed for any of the detected variants to achieve statistical significance. This presumes that a proportion of the variants have true effects which could not be detected due to underpowering and that the effect size estimates are accurate: however, the effect size estimates have large confidence intervals due to the small sample size and effects may be detectable with <5,000 cases, whereas others may not be detectable with >5,000 cases.

Notably, recent FinnGen releases (releases 10 and 11) have identified some significant variants and show loci with strong suggestive signals with only 1,548-1,710 cases. The Finnish population has low genetic diversity, high linkage disequilibrium, and is enriched for many variants at lower frequencies in the general European population, making them an ideal population for genetic studies[258]; a British population may require different sample sizes. Significant loci in subsequent FinnGen releases are described in *Additional evidence from new FinnGen release*.

Accurate power analysis of GWAS is difficult as the power of a study is influenced by many factors, including disease prevalence, heritability, expected number of causal variants, and degree of polygenicity. As there are limited studies of cholesteatoma genetics and heritability, power calculation and estimation of ideal sample size is very difficult. Sensitivity analysis allows us to say that variants do not have a certain effect size or greater, but not whether there is or is not a true effect size smaller than the threshold genetic risk ratio (GRR). This study therefore suggests that there are no common variants with effects >1.4 (the threshold GRR for common variants).

Gene set analysis limitations

The gene-set enrichment analysis of single-variant results in this analysis did not overlap much with the gene-level tests. There are several possible reasons for this: the single variant results consider all variants with MAC>20 whereas the gene-level results *only* consider rare variants. The gene-level tests aggregate variants within a gene to determine its significance. If many genes within a pathway are individually affected by a small number of variants (such that no individual gene appear significant), this pathway is likely to appear enriched in the single-variant data but not the gene-level data.

There are some additional limitations associated with gene set analysis. Gene function databases are constantly updated with new information, so may change over time, affecting reproducibility. The cutoffs for inclusion of variants in gene set analysis may also affect the results. For example, the FinnGen data contained many more p<0.05 SNPs than UKBB WES, resulting in more genes being supplied to g:Profiler. With enough p<0.05 SNPs, all genes would eventually be included. By supplying an ordered query, the ranking of genes is considered, so the test remains valid. However, different p-value cutoffs could affect results. GO terms are also biased towards well-studied genes and pathways. Broad terms containing many genes such as 'tissue development' should be interpreted with caution. This is discussed in more detail in Limitations in functional interpretation.

4.5 Conclusion

Genes bearing variants with p-values < 0.05 in UKBB WES data were enriched for biological pathways including cell adhesion and motility, specifically via cytoskeletal and ciliary involvement with a focus on the dynein family, and calcium binding. These processes agree with the known biology of cholesteatoma and are supported by data from FinnGen as well as previous genetic studies. This is evidence for a genetic effect which may be polygenic or heterogeneous.

This study could not identify any variants or genes with genome-wide significance, nor any suggestive single variants although one gene, *ABCC8*, approached genome-wide significance. The highest-scoring genes and variants also were not detected in the previous GoC WES study nor other small genetic studies.

Several genetic factors may increase cholesteatoma risk: ciliary impairment, perhaps affecting middle ear clearance leading to excessive build-up of debris, is also supported by enriched cytoskeletal and dynein processes in this study. Cell adhesion and motility were also implicated, which may be associated with cholesteatoma invasiveness. Morphology of the ear is known to contribute to disease risk, exemplified by increased risk amongst those with craniofacial developmental anomalies which may be supported by enriched variants affecting tissue development in this study.

Immune genes and ECM degradation were not implicated in this study, suggesting that observed dysregulation in gene expression studies may be a later effect. An ear which does not properly drain may be more prone to infection without having an inadequate immune response, and inflammation may drive many subsequent changes to gene expression. Finally, calcium-binding activity may be involved, supported by this and previous WES studies and gene expression analysis. Due to calcium's many roles in the body, its relevance to cholesteatoma is unknown.

Genetics may contribute to cholesteatoma risk through the above mechanisms, though the effect is probably not due to a small number of genes with large effects due to the failure of any two genetic studies to identify the overlapping candidate genes. However, variants in the *DNAH* family are of particular interest.

5 Polygenic risk scores and machine learning classification

5.1 Background

Owing to the significant enrichment of certain pathways and processes in single variant results despite lack of individually significant variants or genes, there is a possibility for a polygenic effect in cholesteatoma. This may be detectable if the combined effect of many variants with small effect sizes is large enough that it may be detected with 1,000 cases.

A common and simple approach to aggregate effects of multiple variants is the polygenic risk score (PRS), which is a weighted sum of SNPs used to predict an elevated risk of disease compared to the general population[260]. Alternatively, machine learning (ML) classification approaches have the potential to capture complex genetics and are primarily used in multi-omics contexts (where multiple types of data such as genetic and expression data are used)[261]; recent studies have also used deep learning to enhance disease prediction risk of PRS[262,263]. A popular algorithm in bioinformatics is the Random Forest (RF)[123], which averages the output from a large number decision trees created from random re-samples of the data in order to predict an outcome, for example case or control status.

In this chapter, I calculate PRS and use RF on variant-count data to investigate polygenic risk. While the primary aim of PRS and ML is to classify disease risk based on the variants present, these approaches can also be useful tools for understanding disease biology. For example, the number of variants included in a PRS may be informative about the degree of polygenicity. The genes considered important by RF may differ from those identified through GWAS and may reveal certain combinations of genes which are particularly powerful at predicting case status. Interrogation of RF models may also allow identification of genetic subtypes or clustering of participants, so can offer insight into heterogeneity.

5.1.1 Polygenic risk score

A polygenic risk score is a weighted sum of variants present in an individual where weights are the effect sizes of those variants drawn from a GWAS in a separate population[118]. At least two populations are required: the initial base population on which GWAS is performed and a target population on which the PRS is developed, which must be of similar ancestry[260]. Because the effect size for an individual variant is its coefficient in a regression fit to the trait outcome, it reflects the difference in outcome when one, two or no copies are present (in the case of binary traits, this is the odds ratio). Therefore, by summing the effect sizes across the genome, the total gene-based risk can be calculated.

A method to reduce the number of SNPs is also usually applied and the approach to shrinkage is the main aspect which differs between PRS methods. SNP shrinkage is required as the effect sizes of neighbouring SNPs are correlated according to linkage disequilibrium – if this is not corrected, disease risk will be exaggerated. SNP clumping is a common shrinkage method where the genome is split into clumps according to the degree of correlation between SNPs and the most significant SNP per clump is selected. The clump size may be set by manually choosing a threshold SNP-SNP correlation and different cutoffs for SNP significance may be applied[118]. Also, a threshold *p*-value for variants is usually applied, such that the PRS only includes variants that meet a given threshold. Various *p*-value thresholds may be tested during PRS development.

There are several tools for performing PRS scoring, such as PRSice[264], PRSice-2[265] and LDPred2[266] and these generally only differ in the shrinkage method: PRSice uses the clumping and thresholding method, whereas LDPred2 directly models linkage disequilibrium and use this to control for correlation of SNPs.

Properties and limitations of PRS

PRS should be normally distributed amongst a population[260]. This property arises from the central limit theorem, which tells us that the summation of random effects results in a normal distribution[267]. If PRS are not normally distributed, this suggests confounding such as a difference between the base and target population. Cases should have a higher average PRS than controls, although the effect depends on which alleles are considered the effect allele. This is because the beta score used to weight the variants may be positive or negative depending on whether the effect allele was more or less common in cases than controls.

Prediction performance may be assessed via the R^2 of the model, which is the proportion of variability explained by the genetic risk according to the PRS calculated. For a continuous trait, maximum R^2 is determined by heritability so a PRS may be assessed by the proportion of heritability accounted for[268]. As sample sizes are finite and not all variants can be measured, R^2 is likely to fall short of heritability but tends towards it with increasing sample size[260]. This is not true for binary traits as their distribution is not normal and so a 'pseudo-R^2' is instead reported which cannot be directly compared to heritability[269].

PRS typically assume additive genetics and cannot consider any gene-gene or gene-environment interactions[118]. Epistatic effects, where presence of one variant can modify, mask or enhance another, will not be captured. Generally, non-additive genetic effects are not expected to have a large impact on polygenic disease risk and there is evidence that complex disorders generally follow an additive pattern[270], although it has also been shown that incorporating non-additive dominance effects can boost PRS performance[271].

While PRS are generally good at describing the risk of disease on a population level, their use for predicting disease risk in an individual or for population-level screening is controversial[272]. It has been suggested that PRS could be integrated into screening for complex diseases such as type 1 and 2 diabetes, coronary artery disease and breast cancer to identify high-risk individuals for further monitoring or interventions[273]. PRS are not yet used in any clinical settings although many commercial genetics companies offer PRS tests for common complex diseases[118].

5.1.2 Machine learning approaches

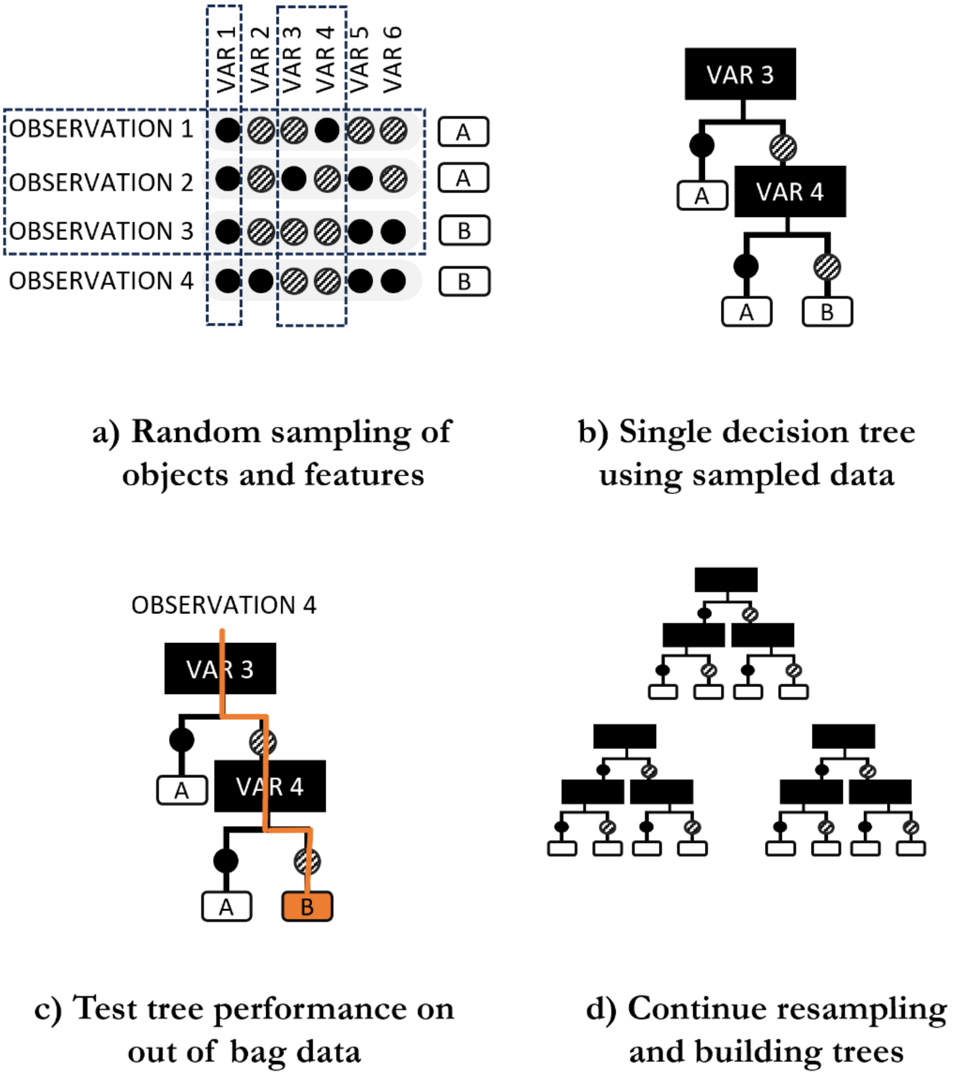
Random forests

Random Forest (RF) is a non-parametric machine learning method for regression and classification problems. An RF classifier builds several decision trees, each using a random subsample of the data, to predict an outcome. Each tree is constructed to be an optimal predictor for the subsample of data included in it. All trees vote on the outcome and the final decision is based on the consensus of all trees in the forest. This makes the method robust to overfitting and suitable for problems with high dimensional data and small sample sizes[274]. It is also particularly useful for non-linear problems with interactions between features.

First, random sampling of objects and features (**Figure 29a**) is performed: both observations (sometimes referred to as objects) and features (also called variables) are sampled randomly.

The sampled data is called the in-bag fraction. Breiman's original RF algorithm[275], and typically the default setting, is to resample a 1:1 proportion of the data with replacement, which results in about 66% of observations and features being retained. **Figure 29b** shows a single decision tree constructed from the sampled data. Trees are built by successively selecting the features which result in the greatest reduction of Gini impurity in child nodes[276]. Gini impurity is a measure of how dissimilar data are: low impurity would indicate that most of the data are of the same class, meaning each split is chosen to create the best distinction between classes until no further improvements can be made[277]. Not all sampled variables will be used: in **Figure 29**, variable 1 was sampled but not used as features 3 and 4 were sufficient to split all observations into categories A or B. The performance of the decision tree is tested using the out of bag data (**Figure 29c**) and is given as out of bag error, or the proportion of out of bag observations the tree incorrectly classifies. The process is repeated with new random resamples of the original data, thus building a 'forest' of random classification trees. To perform classification on a new sample, it is passed through the forest with each tree voting on the outcome.

Figure 29. Overview of Random Forest classification.
A) Features and objects are randomly sampled with replacement from the population, resulting in approximately one third of data being included, or 'in bag' while the remaining data is 'out of bag'. B) In bag data are used to construct a decision tree which can split the data according to its class. C) Tree performance is measured by testing on out of bag data. D) Many trees are constructed by randomly re-sampling data. Trees vote on the class of unseen data to perform classification tasks.



a) Random sampling of objects and features

b) Single decision tree using sampled data

c) Test tree performance on out of bag data

d) Continue resampling and building trees

Properties of Random Forests

RF models can be developed with the goal of predicting the class of unseen data, but also have several properties which make them useful for exploring the features of the data itself. First, features can be assigned an importance, which is a measure of how well trees including the variable perform. Variable importance can identify which features have the best predictive power. The importance of an individual predictor is reliant on the other features used to build trees: some features may be more useful when in combination with other features, for example if two variables in combination can split data well into case and control but either variable alone performs poorly. This dependence may be a better representation of underlying biological mechanisms than testing each gene individually by standard linear models as it can capture epistatic effects[278]. Conversely, variables which are highly correlated may have reduced importance as they are likely to reduce node impurity in a similar manner; once one variable has been used to split the data, the other imparts little additional information[279]. Resampling means that both variables will still be assessed, as not all variables are in-bag for a given tree. This reduces the impact of collinearity on the predictor, which is useful for classification, but harms interpretability of variable importance as correlated variables are assigned low importance regardless of possible biological relevance[279].

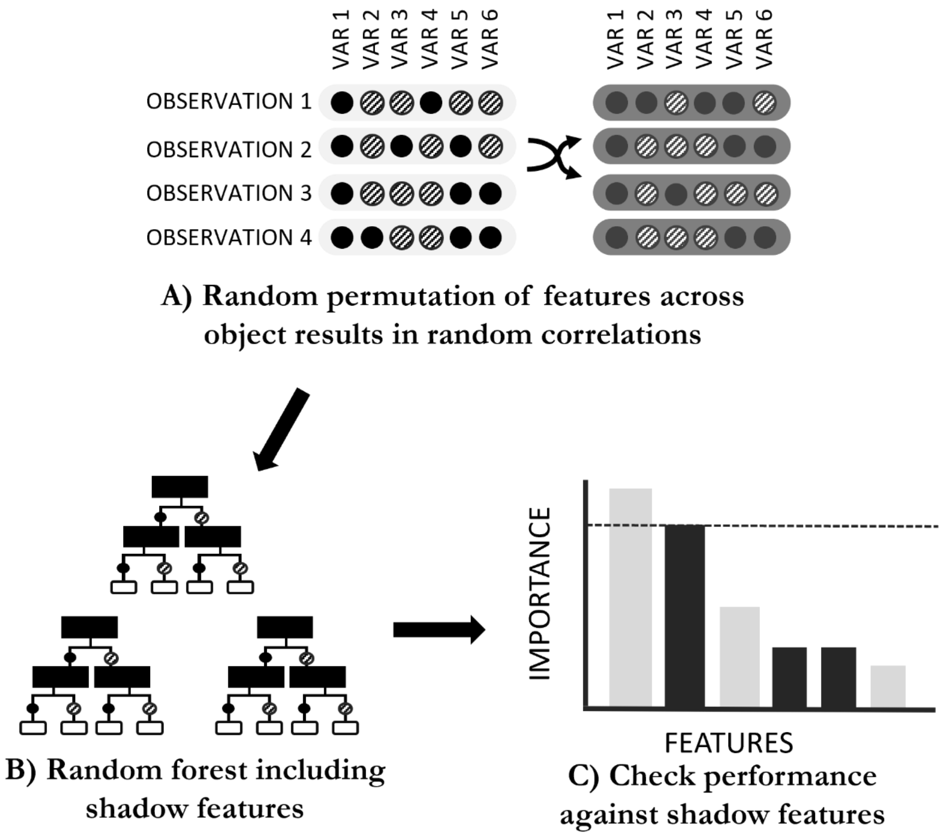
Resampling also provides robustness to overfitting, which occurs when too much information about a limited set of observations is used to train a classifier[274]. For example, if every observation had a unique combination of features, a classifier using every single variable could classify the training data perfectly. However, it would probably fail on new data as no actual pattern for discerning classes has been detected. RF avoids overfitting because the ensemble of classifiers, each trained on a random sub-sample of the data, votes on predicted class.

Another useful feature of RF models are proximity scores, which are the proportion of trees in which a given pair of observations fall on the same terminal node[123]. A decision tree classifying observations as case or control may have many terminal nodes for each class. Observations appearing together on a terminal node have the same values for the features selected to form the decision tree. If observations appear together very often, they are likely to have the same the combinations of important features. This can help identify clustering or subtypes amongst cases. This may also provide insight into the mechanisms underlying a disease, as different combinations of genes or variants which act as good predictors of disease status can be identified.

Feature selection

Feature selection may be performed to improve model performance by reducing dimensionality by excluding features which are likely to be irrelevant or redundant[280]. Boruta[281] is a feature selection method for RF which assigns ‘tentative’ or ‘confirmed’ status to variable importance by comparing performance to randomly shuffled 'shadow features' (**Figure 30**). Standard RF importance can be biased by correlations between individual trees in the forest. Random correlations between features can lead to chance associations with class, resulting in inflated importances. By creating shadow features, which are copies of the original variables with the observations randomly shuffled, the performance of actual features can be assessed by checking whether they have consistently better performance than the shadow features across many resamples. Essentially, it asks whether a feature's predictive power is better than random. This makes it useful for both improving model performance and interpreting variable importance.

Figure 30. Overview of Boruta feature selection method.
A) features are permuted by randomly shuffling the values within each feature, creating 'shadow features'. B) Random forest is performed using both the true features and shadow features. C) the performance of each feature is compared to the performance of the best shadow feature. This process is repeated many times and the frequency with which a feature outperformed all shadow features is used to determine its importance.


VAR 1
VAR 2
VAR 3
VAR 4
VAR 5
VAR 6
OBSERVATION 1
OBSERVATION 2
OBSERVATION 3
OBSERVATION 4
A) Random permutation of features across object results in random correlations
B) Random forest including shadow features
IMPORTANCE
FEATURES
C) Check performance against shadow features

5.1.3 Aims and objectives

In this chapter, I focus on the traditional PRS, calculated using PRSice-2[265], and RF with Boruta feature selection. My aim is to explore the feasibility of these approaches rather than to produce a PRS or any other model which can be used to predict disease risk in any other population. I achieve this by:

- Calculating genome-wide polygenic risk scores using FinnGen base data and UK BioBank WES variants.
- Calculating genome-wide polygenic risk scores using a 10-fold cross-validation split on UK BioBank whole exome data.
- Training RF models on 10-fold cross-validation split UK BioBank data using counts of SNPs per gene and pathway to assess model performance/predictive power.
- Using Boruta to perform feature selection to identify important genes and pathways.
- Training similar RF models on all UK BioBank data to interrogate model features including surrogate association and proximity scores.

5.2 Methods

5.2.1 Data

10-fold cross validation partitioning for PRS and RF models

For both PRS calculation and RF models, I used a *k*-fold cross-validation partition with *k*=10 using the MATLAB *cvpartition* function. This assigns the data randomly to 10 partitions without replacement. Each fold uses 9 of the partitions as train data and the remaining partition is used as test data. 10-fold cross validation was chosen as it has been shown that better predictions of error are acquired for larger values of *k*, but larger values of *k* also increase the between training sets whilst reducing the size of testing sets. *K*=10 has also been found to show a good trade-off between bias in the estimate of error and computational cost[282]. Partitions were stratified by case status to ensure an equal case:control ratio in all partitions.

For each of the ten cross-validation folds, GWAS was performed on the train set only. Settings were as in *Final SAIGE configuration for single variant* tests This was to allow train and test partitions to be used as base and target data in PRS and to prevent information leakage in RF validation.

Gene and pathway tables for Random Forests and Boruta

For machine learning analyses, tables indicating the number of significant variants per gene were constructed for all participants. Non-synonymous, coding variants from whole exome data were used and filtered to MAC>20. Results were further filtered to retain SNPs with *p*-value < 0.05, hereafter called significant SNPs (note they are not of genome-wide significance). For each person, the number of significant SNPs per gene was counted (using gene assignments from VEP as in the GWAS section). Reducing to significant SNPs only acts as a form of feature selection.

To generate equivalent tables for cross-validation folds, the same processes was applied using the relevant GWAS *p*-values for each individual fold.

To create equivalent data tables for pathways, I annotated genes with their pathways using GO definitions downloaded from the g:Profiler website for consistency with the previous gene set enrichment analysis (reference genomes: Ensembl 111, Ensembl Genomes 57. GO release: 2024-01-17). Each gene was counted for all pathways it was associated with.

5.2.2 Polygenic risk scoring

The PRS method requires base data to acquire weights, which are the effect sizes of variants according to GWAS, and target data for which PRS are calculated. The proportion of variability in phenotype explained by the PRS in the target group is the model R^2. A p-value for the PRS can be calculated by comparing the R^2 to that of a null model. PRSice-2[265] calculates the p-value empirically by shuffling phenotypes randomly to acquire the null *p*-value for comparison.

I used PRSice-2 v1.0.2[265] to perform polygenic risk score analysis with default clumping (250kb, r^2=0.1, *p*-value=1) and –beta option on (to account for reporting of beta scores rather than odds ratios in GWAS results). All covariates were as used in GWAS (age, sex, deprivation, smoking status, first 10 PCs) were used.

I ran tests using both UKBB and FinnGen summary statistics as base data and UKBB WES and microarray data genotypes as target data.

10-fold cross validation with UKBB base data

I used 10-fold cross-validation for UKBB WES data as both base and target data. I performed GWAS on each training partition (90% of the data) and used the summary statistics as base data. Target data were the test partitions for each fold (10% of the data).

FinnGen base data

I used FinnGen summary statistics from data freeze 9 as base data and UKBB WES or microarray data as target data. The FinnGen data is described in *Genome-wide association testing and downstream analyses*: *FinnGen comparison data*. All non-imputed SNPs were used with no minimum MAC cutoff.

Because the reference allele must be the same between base and target data, I identified any mismatched between base and target and flipped the reference and effect size in the FinnGen base data to match the target. Also, FinnGen uses GRCH37 whereas UKBB uses GRCH38. To avoid performing a full liftover of the FinnGen data to GRCH38, I matched SNPs on rsID (which should be stable between versions) and assigned the chromosome and position to the base data by lifting the chromosome and position of the SNP from the target data. I removed any SNPs whose reference or alternate allele did not match (**Table 31**).

Table 31. Target data parameters and number of mismatching or flipped SNPs compared to base data.

Data	Min MAC	Coding only?	overlap	Flips	Mismatches
Microarray	20	No	194926	25289	234
Microarray	0	No	200128	25293	248
Microarray	20	Yes	90422	11018	68
Microarray	0	Yes	94961	11021	79
WES	20	Yes	5711	0	18
WES	0	Yes	14935	0	278

I used coding region, non-synonymous SNPs from WES as target data. For microarray, I used all SNPs and also tested coding region, non-synonymous SNPs separately. I ran versions with MAC>20 filter and no min mac. For whole exome data, rsID was assigned using VEP while probes on the microarray data were already labelled with rsID.

5.2.3 Random forests

Feature selection using 10 cross-validation folds

I performed feature selection using the train data for each fold with the Boruta R package[283] version 8.0.0. Feature selection was performed for both genes and pathways, where data were the count of p-value < 0.05 SNPs per gene or pathway. Features are assigned confirmed or tentative status depending on whether they consistently outperform randomly generated shadow features. I ran Boruta with the default settings with class weights of 1:5 for cases to account for class imbalance. All RF models were weighted in this manner during training.

Model generation and validation using 10 cross-validation folds for gene-level data

After feature selection, new RF models were constructed using the training data from each fold (90% of the total data), with only the genes determined to be tentative or confirmed by Boruta (hereafter called 'important') for that fold. Thus, each fold used a gene table containing only important genes for that fold and only counting presence/absence of SNPs meeting the p<0.05 threshold within that fold's GWAS. I first ran 10 iterations of the model on the first fold to determine the number of trees required to attain maximum performance, judged by lowest out of bag error and used this number to construct RF models for each fold (n=500 trees). I used the treebagger function in MATLAB 2023b[132], which performs RF in the same manner as Boruta (which constructs random forests using the *ranger*[284] package) but provides a greater number of options for interrogating the results.

The model created for a given fold was then validated using that fold's test data. Test data consist of similar tables of counts of SNPs with p<0.05 per gene as the training data, but note that the *p*-value for SNPs is also drawn from the GWAS performed on the train data, not the test data itself. The data is supplied to the model which attempts to predict case-control status. I calculated sensitivity, specificity, total prediction error and mean squared error for each fold as well as the out of bag error for the model during training.

Cross-validation methods typically use the same set of features for all folds in order to evaluate the performance of a specific model configuration. Feature selection and any other pre-processing would be determined, and the validation folds used to determine variability in error. The final model configuration would then be tested on an independent data set. I did not have enough data to perform proper validation and so do not aim to build a reliable classifier. My models instead use features computed for maximum importance within each fold. Validation is not meant to assess a specific classifier, but the general approach.

Data exploration using a single RF model

The 10-fold cross-validation gives an idea of the variability in performance of the RF method by using different subsets of the data to train and test the models. I then constructed a single model using all cases and controls to explore the features of the model itself and whether any insight into genetics can be drawn from it. I constructed the model using *p*-values of SNPs drawn from the GWAS of all cases and controls (presented in the previous chapter) and any gene with confirmed importance in 2 or more of the 10 folds, or tentatively important in over half of the folds.

I also investigated the reliance of feature importance on other features by generating models which excluded each gene in turn and calculated the importance of the remaining features. While there was some reduction in OOB error from 200 to 500 trees, this was minimal. For performance reasons, I generated models using 200 trees. I repeated this 10 times and averaged the drop in performance for each predictor over the 10 runs. While this method only accounts for features pairs, it becomes computationally difficult to assess the interactions between 3 or more genes.

Pathway-level Random Forest models

To determine whether RF could return results similar to the gene-set enrichment results, I also performed tests on pathway level UKBB WES data. This was performed in the same manner as the gene-level data but included counts of SNPs per pathway rather than per gene. I performed several exploratory tests:

A) I ran Boruta on the entire dataset then constructed an RF model using the confirmed and tentative features on the entire set with no test/train split. This was to investigate pathway relationships and compare important features to GSEA results.

B) I performed Boruta on the 10-fold cross-validation sets then used the confirmed and tentative features to construct a model on the training data. I tested these models using the test data for each split. This was to test performance of this approach.

C) I performed a similar test to B but used only the pathways identified by gene set enrichment analysis instead of performing Boruta feature selection. This was for comparison to B.

All models used 50 trees and weighted cases as in previously described models.

5.3 Results

Due to the exploratory nature of this chapter, several tests were performed with different configurations (**Table 32**). PRS tests used both FinnGen base data and UKBB base data with a 10-fold cross validation test/train split. For both gene-level and pathway-level RF models, data were counts of SNPs with p<0.05 present in each gene/pathway per person. For both gene and pathway-level analyses, 10-fold test/train splits were used to calculate error and a model was also constructed using all data with no testing holdout in order to investigate the properties of the model.

Table 32. Summary of PRS and RF methods performed and their intended purpose.
For polygenic risk score (PRS), base and target data are shown. For random forest (RF) models, test and train data are shown. All UKBB data is drawn from whole exome single variant association test results, consisting of counts of p<0.05 SNPs per gene/pathway.

	Base/train Data	Target/test data	Feature selection?	Purpose
FinnGen PRS	FinnGen	UKBB WES		Create PRS using distinct target and base data. Determine if polygenic risk can be translated between populations.
		UKBB Microarray		
UKBB PRS	UKBB 10-fold cross validation testing sets	UKBB 10-fold cross validation training sets		Create PRS using same population as base and target data. Cross-validation of PRS approach. Ascertain model error on unseen data.
Gene-level RF	UKBB 10-fold cross validation testing sets	UKBB 10-fold cross validation training sets	Boruta selected features for each fold	Cross-validation of RF approach. Ascertain model error on unseen data.
	All data	N/A	Boruta selected features for each fold, features confirmed in >1 fold or tentative in >5 folds.	Investigate properties of RF model, e.g. proximity scores
Pathway-level RF	UKBB 10-fold cross validation testing sets	UKBB 10-fold cross validation training sets	Boruta selected pathways	Validation Compare to other pathway level models.
			Using pathways identified by GSEA in previous chapter	Validation Compare to other pathway level models.
	All data	N/A	Boruta selected pathways	Investigate properties of RF model to determine relationships between pathways Compare selected pathways to GSEA

5.3.1 Polygenic risk scores with FinnGen and UKBB base data

PRS using FinnGen base data

The best PRS *p*-value using FinnGen base data was achieved for whole exome data with a minimum MAC of 20 (**Table 33**). The next best performing PRS used microarray target data with no MAC cutoff and not restricted to coding only. While the WES target data required only 43 SNPs to achieve its best *p*-value of 0.0013, the microarray target data used essentially all SNPs overlapping between the base and target data and achieved a *p*-value of 0.003. The R^2 for both these methods was very small (<0.0031), indicating that most variability in disease outcome could not be explained by the genetic effect. As data are binary, R^2 is approximated using the Nagelkirke method[264]. The SNPs used in the UKBB WES PRS were not present in the UKBB microarray probe set, so these models used different variants entirely.

Table 33. PRS performance for whole exome and microarray data with different MAC cutoffs and coding/non-coding variants.

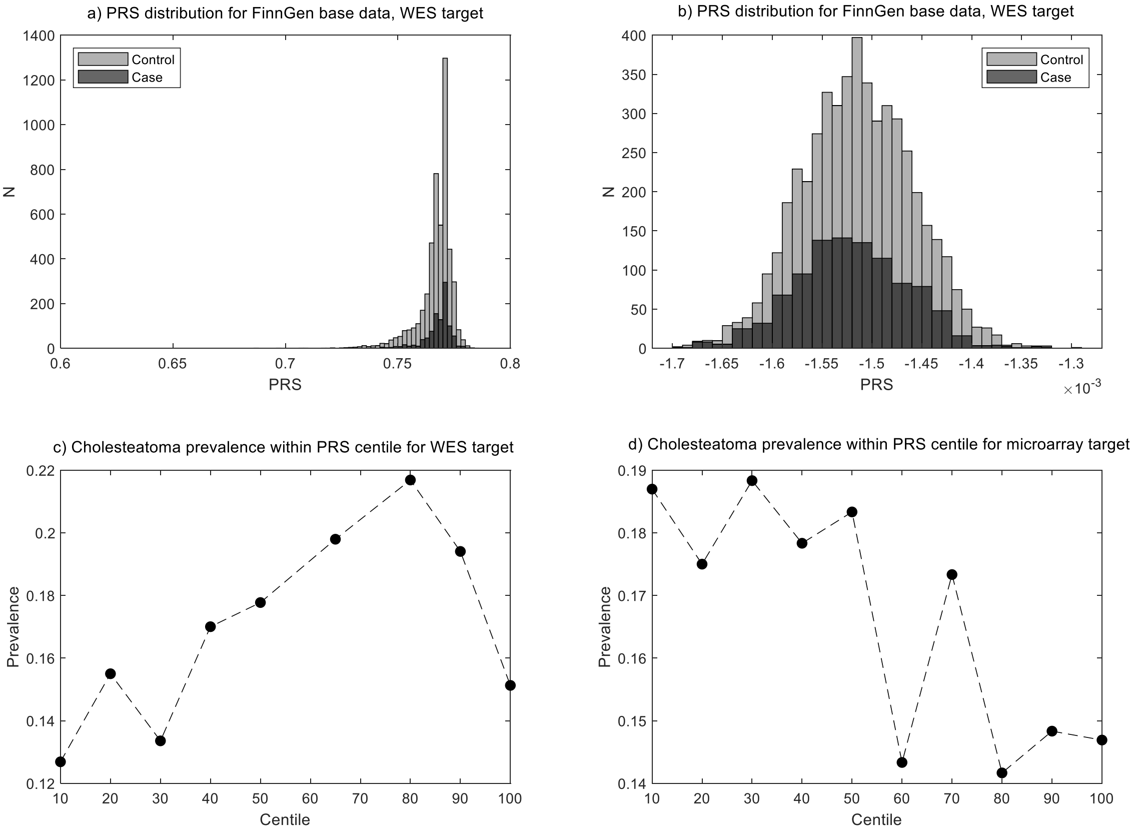
Threshold column shows the maximum p-value of variants included in the polygenic risk score (PRS). PRS R^2 is the proportion of variance in phenotype explained by the model (this is not equivalent to heritability for binary traits). Genotype refers to target data and is either UKBB whole exome (WES) or microarray.

Genotype	Set	Min MAC	Threshold *p*-value	Coefficient	PRS R^2	Standard Error	*p*-value	Num SNP
WES	Coding	20	0.0108	16.8712	0.003104	5.25	0.00130	43
Microarray	All	0	1	-1854.87	0.002463	626.61	0.00308	124659
Microarray	All	20	1	-1691.52	0.002127	614.74	0.00593	120064
WES	Coding	0	0.0108	30.8911	0.002142	11.47	0.00705	119
Microarray	Coding	20	0.0479	-129.182	0.001387	58.07	0.0261	4056
Microarray	Coding	0	0.0001	2.44164	0.001069	1.25	0.0508	4

The PRS distribution for the best WES target data had a non-normal distribution with a long-left tail (Figure 31**a**). While the difference in mean PRS for cases and controls was minimal (0.7692 and 0.7685 respectively), a comparison of PRS deciles shows a higher disease prevalence for higher PRS, although prevalence seems to decrease after the 8th decile (Figure 31**c**). This may be due to random noise or an actual decrease. The UKBB microarray PRS was normally distributed (Figure 31**c**) but negative for both cases and controls. While difference in median PRS for cases and controls did not differ (-0.0015 for both), there was a decrease in prevalence for higher deciles with prevalence decreasing from 0.1870 to 0.1469 between the first and tenth deciles. The reversed relationship is due to the negative PRS of the microarray data for both cases and controls, with cases being more extremely negative. The negative PRS indicates that both cases and controls carried more ‘protective’ alleles then causal ones. However, the direction of effect depends on which variant is considered the reference allele. In this case, I believe the wrong allele was considered the effect allele, resulting in overall negative PRS.

Figure 31. PRS distributions and prevalence per centile for FinnGen base data with best-performing UKBB target data.

Target data are UKBB WES with minor allele count (MAC) > 20 (a, c) and UKBB microarray with no min MAC and all variant types (b, d). a) WES target data show highly skewed polygenic risk score (PRS) distributions with a large tail of lower-risk individuals. B) Microarray data are more normally distributed. Note that PRS are negative. C) Cholesteatoma prevalence generally increases with PRS centile, though appears to drop towards the 90th percentile. D) Cholesteatoma prevalence generally decreases with PRS centile although the relationship is very noisy.



PRS using 10 UKBB cross-validation folds

I also performed PRS testing on UKBB test/train folds with the train fold used as base data to perform GWAS and acquire variants weights. The test fold was used as target data. The resultant PRS had *p*-values ranging from 0.0008 to 0.10 (median 0.0675; **Table 34**). Most were significant though with small R^2 values (<0.02), suggesting a polygenic effect may exist but explaining a very small proportion of overall variability. The number of SNPs used by each PRS also varied greatly from 71 to 20,215 (median 205), with better performing models generally requiring fewer SNPs with a lower *p*-value cutoff. The coefficient also varied, sometimes being negative (indicating a lower PRS for cases) and sometimes positive (indicating a higher PRS for cases).

Table 34. 10-fold PRS using UKBB WES test/train split shows highly variable results

Fold	Threshold SNP p value	PRS R^2	Coefficient	Standard Error	PRS *p*-value	Num SNPs
1	0.0844	0.0094	-276.17	150.68	0.067	3588
2	0.0734	0.0102	-274.56	144.22	0.057	3066
3	0.0027	0.0046	25.31	19.92	0.204	124
4	0.0034	0.0043	-26.36	21.23	0.214	139
5	0.2472	0.0288	-1091.41	342.50	0.001	10579
6	0.0048	0.0198	72.83	27.98	0.009	210
7	0.0002	0.0165	8.87	3.89	0.023	4
8	0.0046	0.0096	49.39	27.07	0.068	200
9	0.3709	0.0067	-730.89	473.99	0.123	15760
10	0.0025	0.0070	31.93	20.45	0.119	110

5.3.2 Boruta results for important genes and pathways

Gene level Boruta results

Feature selection was performed using Boruta to identify important variables for random forest classification. I performed feature selection on gene-level SNP counts for each training set of 10 cross-validation folds. One gene (*ESX1)* was determined as important across all folds. *AMOTL2, IL13RA2* and *RBM10* were confirmed or tentative in 9 out of 10 training folds (**Table 35**). Generally, the genes with high importance were those containing the most significantly associated SNPs such as *ESX1*, *AMOTL2*, *RBM10*, *CANA2D1*, *CACNA1G*, *PTH2R* and *ANK2*,, which were represented amongst the top 20 most significant variants (see *4.3.2 Genome-wide association test results*).

Table 35. Boruta confirmed and tentative features for all 10 folds where a feature was a confirmed important predictor in at least 2 folds or confirmed/tentative in more than 5 folds.

Confirmed features are shaded dark grey and labelled C. Tentative features are shaded light grey and labelled T. 232 genes were confirmed or tentative in at least one fold. 94 genes were confirmed or tentative in 2 or more folds. Ordered by number of times a gene was confirmed across folds.

	Fold										N		
Gene	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	C	T	Total
ESX1	C	C	C	C	C	C	T	C	C	C	9	1	10
AMOTL2	C	C	C	C	C	C		C	C	C	9	0	9
IL13RA2	C	T	C	T	C	C		C	C	C	7	2	9
RBM10	C		C	C	T	T	C	C	C	C	7	2	9
CACNA2D1	C	C	C		C			C	C	C	7	0	7
SLC25A46	C		C	T	C		C	C		C	6	1	7
ZNF41	C		C		C	C			C	C	6	0	6
BIRC2	C	C	C		C	T	T	C	T		5	3	8
CACNA1G	C		C	C	T	T	C	C			5	2	7
TRPV5	C		C		C	C	T	T		C	5	2	7
ANK2	C	C	C		T	C				C	5	1	6
PTH2R		C		T		C		C	C	C	5	1	6
TBC1D16	C				C		C	C	C		5	0	5
CYB5R3	C	C	T	C	T		T	C	T		4	4	8
RUNDC1			C	C	C	T		C	T	T	4	3	7
NIF3L1	C	C	C		T	C					4	1	5
TMEM207	C		C		C			T	C		4	1	5
IGF2R	C	C		C		C					4	0	4
TSC1		C		C			C	C			4	0	4
COL4A6	C			C		T	T	T	C	T	3	4	7
DKK1	C			T		C			C	T	3	2	5
CDKL5			C	T			T	C	C		3	2	5
UBQLNL	T		C		C			C	T		3	2	5
TXNIP			T			C	C	C	T		3	2	5

	Fold										N		
Gene	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	C	T	Total
MEGF11	C			C		T		C			3	1	4
ACOX3				C		C			C	T	3	1	4
PRDM10	C							C		C	3	0	3
ATG9B		C				C		C			3	0	3
OR6T1			C			C		C			3	0	3
MECP2	T	C					T	T		C	2	3	5
WNT10A	T	T			C	T			C		2	3	5
PPP1R26	C			T				C			2	1	3
RAPGEF2		C					T		C		2	1	3
ALLC		C						C		T	2	1	3
WFDC8			C					C		T	2	1	3
AK8			C						C	T	2	1	3
SNX13					C	C	T				2	1	3
POLN					C			T	C		2	1	3
MRPS25	C					C					2	0	2
ANGPTL4	C		C								2	0	2
SP140L	C							C			2	0	2
KBTBD13	C		C								2	0	2
GPNMB			C							C	2	0	2
IL3RA					C			C			2	0	2
NECTIN1							C		C		2	0	2
ATF6B							C	C			2	0	2
TBC1D10C									C	C	2	0	2
ZACN									C	C	2	0	2
FAM220A		T	T	T	T	T	T	T		C	1	7	8

Pathway level Boruta results

I also performed Boruta on all UKBB WES pathway-level SNP count data. I performed this both on individual training folds and on all data with no test/train split. The features identified from the individual train folds had very little overlap: out of a total 287 confirmed or tentative pathways across all 10 folds, 72 were present in 2 or more and three terms were considered important in 7 of 10 folds (*organic substance metabolic process*, *intracellular membrane-bounded organelle* and *molecular function;* **SI Table 7***)*. This may be because GO terms are hierarchical and each fold differed slightly in which level of the hierarchy was enriched; therefore, this is a suboptimal way of identifying which processes are generally important.

I also performed feature selection on all data with no test/train (**Table 36**) to compare to the results of gene set enrichment analysis (GSEA). There was little direct overlap in terms between the g:Profiler GSEA (see *4.3.3 Gene set enrichment analysis*) and Boruta results, although some general functions were similar such as cell motility and voltage-gated calcium channel activity. The results tended towards more broad/general GO terms than GSEA results, perhaps because larger terms had more genes and therefore more opportunities to split cases and controls.

Table 36. Boruta confirmed and tentative pathways when performed on all data with no cross-validation or test/train split.

The p-value of enriched terms in the full gene set enrichment analysis (GSEA) are also shown where p-value < 0.05, as g:Profiler was configured to return significant results only.

Pathway	ID	GSEA *p*-value
alcohol metabolic process	GO:0006066	-
anion binding	GO:0043168	-
Binding	GO:0005488	-
Biological process	GO:0008150	-
cardiac muscle cell membrane repolarization	GO:0099622	-
cell motility	GO:0048870	-
cellular anatomical entity	GO:0110165	-
cellular biosynthetic process	GO:0044249	-
cellular response to lipid	GO:0071396	-
cellular response to lipopolysaccharide	GO:0071222	-
Cellular component	GO:0005575	-
Cytoplasm	GO:0005737	5.55492x10-13
endoplasmic reticulum membrane	GO:0005789	-
establishment of cell polarity involved in ameboidal cell migration	GO:0003365	-
hydrolase activity, hydrolyzing N-glycosyl compounds	GO:0016799	-
intracellular anatomical structure	GO:0005622	-
intracellular membrane-bounded organelle	GO:0043231	-
intracellular organelle	GO:0043229	-
macromolecule metabolic process	GO:0043170	-
membrane-bounded organelle	GO:0043227	-
Molecular function	GO:0003674	-
neutrophil chemotaxis	GO:0030593	-

Pathway	ID	GSEA *p*-value
nitrogen compound metabolic process	GO:0006807	-
nuclear lumen	GO:0031981	-
nuclear outer membrane-endoplasmic reticulum membrane network	GO:0042175	-
Organelle	GO:0043226	-
organonitrogen compound metabolic process	GO:1901564	-
perinuclear theca	GO:0033011	-
phosphorus metabolic process	GO:0006793	-
positive regulation of biosynthetic process	GO:0009891	-
positive regulation of cellular process	GO:0048522	-
positive regulation of protein localization to cell surface	GO:2000010	-
primary metabolic process	GO:0044238	-
protein ubiquitination	GO:0016567	-
regulation of atrial cardiac muscle cell membrane depolarization	GO:0060371	-
regulation of cardiac muscle cell membrane repolarization	GO:0099623	-
regulation of cellular component organization	GO:0051128	-
regulation of developmental process	GO:0050793	-
regulation of phosphate metabolic process	GO:0019220	-
side of membrane	GO:0098552	-
Signaling	GO:0023052	0.032177767
terpenoid metabolic process	GO:0006721	-
ventricular cardiac muscle cell membrane repolarization	GO:0099625	-
voltage-gated calcium channel activity involved in cardiac muscle cell action potential	GO:0086007	0.002392465
voltage-gated calcium channel activity involved SA node cell action potential	GO:0086059	-

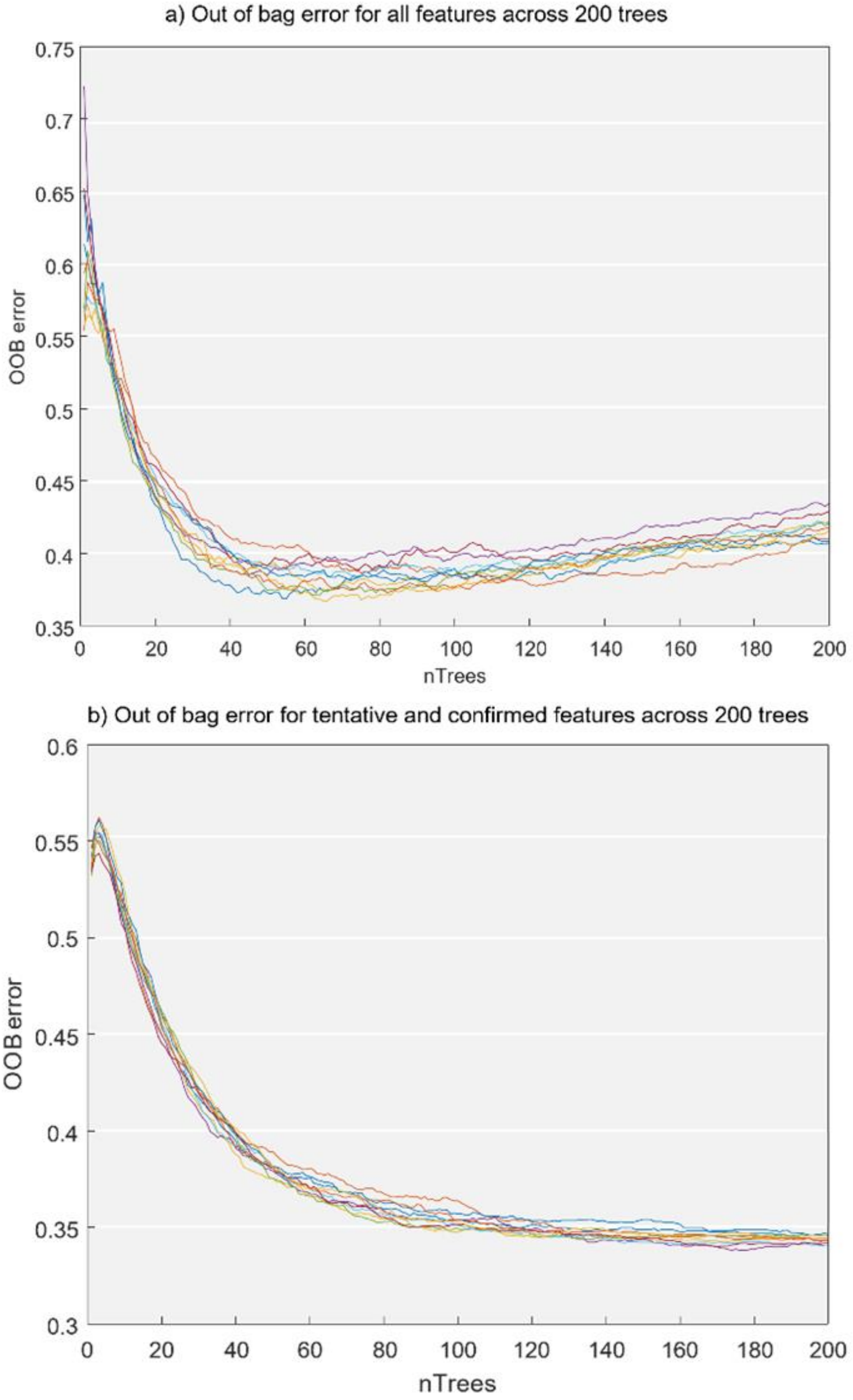
5.3.3 Performance of Random Forest models using gene-level information

Construction and validation on 10-fold data split

For individual folds, I used Boruta-important features for that fold (**SI Table 8**) and *p*-values drawn from testing on the train data with GWAS. Fold 1 is used as an example: for this fold, 36 features were confirmed important and 12 features were tentatively important. The RF model constructed using only the important features slightly outperformed a model trained on all features, with final out of bag error after 200 trees averaging 0.344 across 10 repeats (**Figure 32**). Continuing to increase the number of trees improved model accuracy up slightly with a performance of 0.339, although this is not much improvement over 200 trees.

Figure 32. Out of bag error indicates slightly improved performance when a reduced set of features are used.

Graphs show out-of bag (OOB) error for 10 models trained on the fold 1 training set. a) The best mean performance across all models when all features were included was at 67 trees (OOB error = 0.382). OOB error drops rapidly before this and slowly increases afterwards, indicating some overfitting. b) Model performance is improved by using only the confirmed and tentative features and there is no overfitting at 200 trees. The minimum mean OOB error was 0.343 at 198 trees.



a) Out of bag error for all features across 200 trees

b) Out of bag error for tentative and confirmed features across 200 trees

Cross-validation across all 10 folds shows variable performance with poor sensitivity and AUC

Random forest models trained on tentative and confirmed features were constructed for each fold using 500 trees. The models performed overall poorly with a mean out of bag error of 0.364 (**Table 37**). MSE varied between 172.44 and 178.83 (mean=175.01,). While specificity was generally high (~0.8), sensitivity was overall poor (~0.3) but occasionally reached as high as 0.79 (**Figure 33**). This reflects the model's tendency to classify test data as control. The area under the curve (AUC) of the receiver operating characteristic (ROC) was also calculated. ROC plots the true positive rate against the false positive rate. The AUC of a classifier that randomly assigns case-control status equally is 0.5. For all folds, AUC was close to 0.5 (mean 0.506). Out of bag error for individual folds was generally lower than total prediction error (mean 0.35 vs mean 0.56), showing that out of bag error is not a good indicator of actual model performance.

Figure 33. Sensitivity and specificity on testing data from 10 cross-validation folds. *Random forest classifiers were trained on 90% of UKBB WES data (counts of p-value <0.05 SNPs per gene). The remaining 10% of data was used to test each classifier. Specificity and sensitivity were calculated with the predictions of the classifier on test data.*

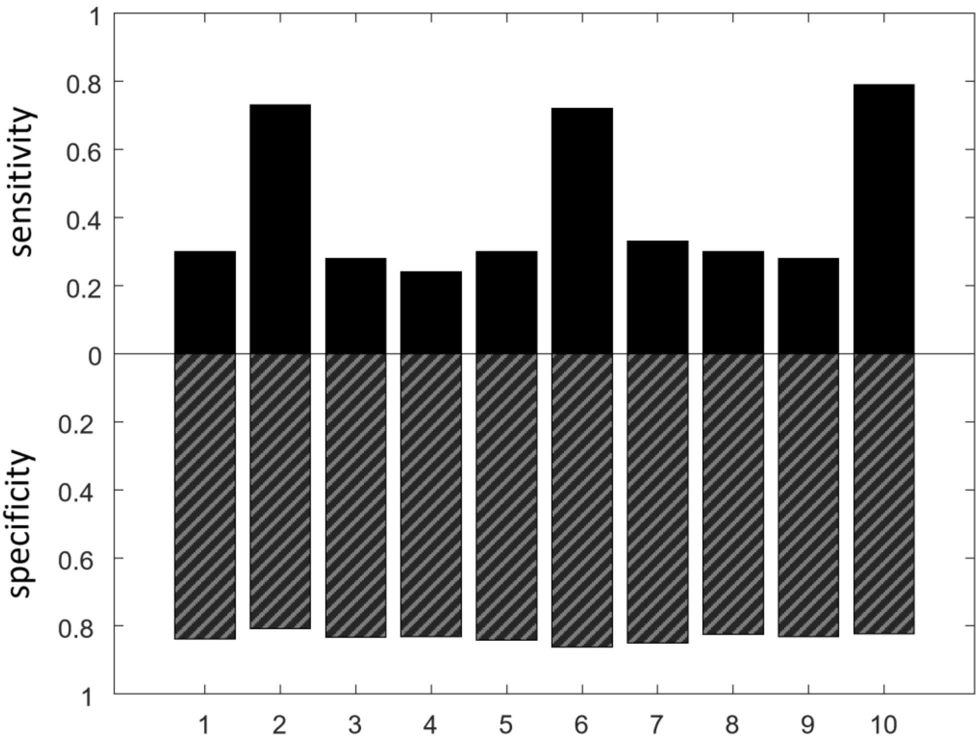


Table 37. Random forest performance across 10 folds using individually selected features for each fold.

Mean squared error, sensitivity, specificity, total prediction error and AUC calculated on the testing set are shown. Out of bag error on the training set is also shown.

Fold	Mean squared error	Sensitivity	Specificity	Total Error	AUC	Out of Bag error
1	177.4264	0.30	0.838	0.653	0.520	0.338
2	175.8483	0.73	0.809	0.312	0.523	0.355
3	172.4374	0.28	0.834	0.648	0.493	0.346
4	174.1039	0.24	0.831	0.662	0.492	0.347
5	178.8347	0.30	0.842	0.672	0.524	0.343
6	174.5274	0.72	0.863	0.413	0.463	0.329
7	172.7369	0.33	0.850	0.688	0.544	0.371
8	176.1374	0.30	0.825	0.602	0.489	0.327
9	173.4517	0.28	0.831	0.637	0.498	0.339
10	174.6137	0.79	0.824	0.295	0.510	0.364
mean	*175.0118*	*0.43*	*0.835*	*0.558*	*0.506*	*0.346*

Folds 2, 6 and 10 had higher sensitivity than other folds. This could have been because a single gene or set of genes had particularly good predictive power and performance is improved when these are well-distributed between case and control. However, these folds differed in which predictors were considered most important, suggesting this was not the case (**Table 37**).

Table 38. Most important genes for high-performing folds.
Folds 2, 6 and 10 achieved sensitivity approaching 0.8; the most important genes within these folds do not overlap. Importance score is the out of bag permuted predictor error, which reflects the reduction in performance when a feature is randomly permuted.

	Fold 2		Fold 6		10	
Rank	Importance	gene	Importance	gene	Importance	Gene
1	-3.30953	*AMOTL2*	-2.45437	*TRDMT1*	-3.82596	*ZC3H14*
2	-3.26448	*NEK10*	-2.42648	*ZKSCAN1*	-3.46661	*HERC6*
3	-3.2487	*SLC2A11*	-2.40882	*TXNIP*	-3.33528	*ANK2*
4	-3.2486	*CACNA2D1*	-2.39952	*ZSCAN32*	-3.33065	*TBC1D10C*
5	-3.24499	*IGF2R*	-2.39314	*KIAA0040*	-3.32192	*AMOTL2*
6	-3.19727	*ESX1*	-2.37577	*CUX1*	-3.31552	*KCNT1*
7	-3.17116	*CYB5R3*	-2.3717	*COL4A6*	-3.30511	*RUNDC1*
8	-3.14965	*TMEM168*	-2.34999	*NIF3L1*	-3.28167	*ESX1*
9	-3.1376	*TSC1*	-2.34764	*TBC1D10B*	-3.23144	*TBC1D10B*
10	-3.09979	*ALLC*	-2.34528	*ST6GALNAC2*	-3.18701	*RAB6A*

5.3.4 Properties of a model using confirmed and tentative features on all training data

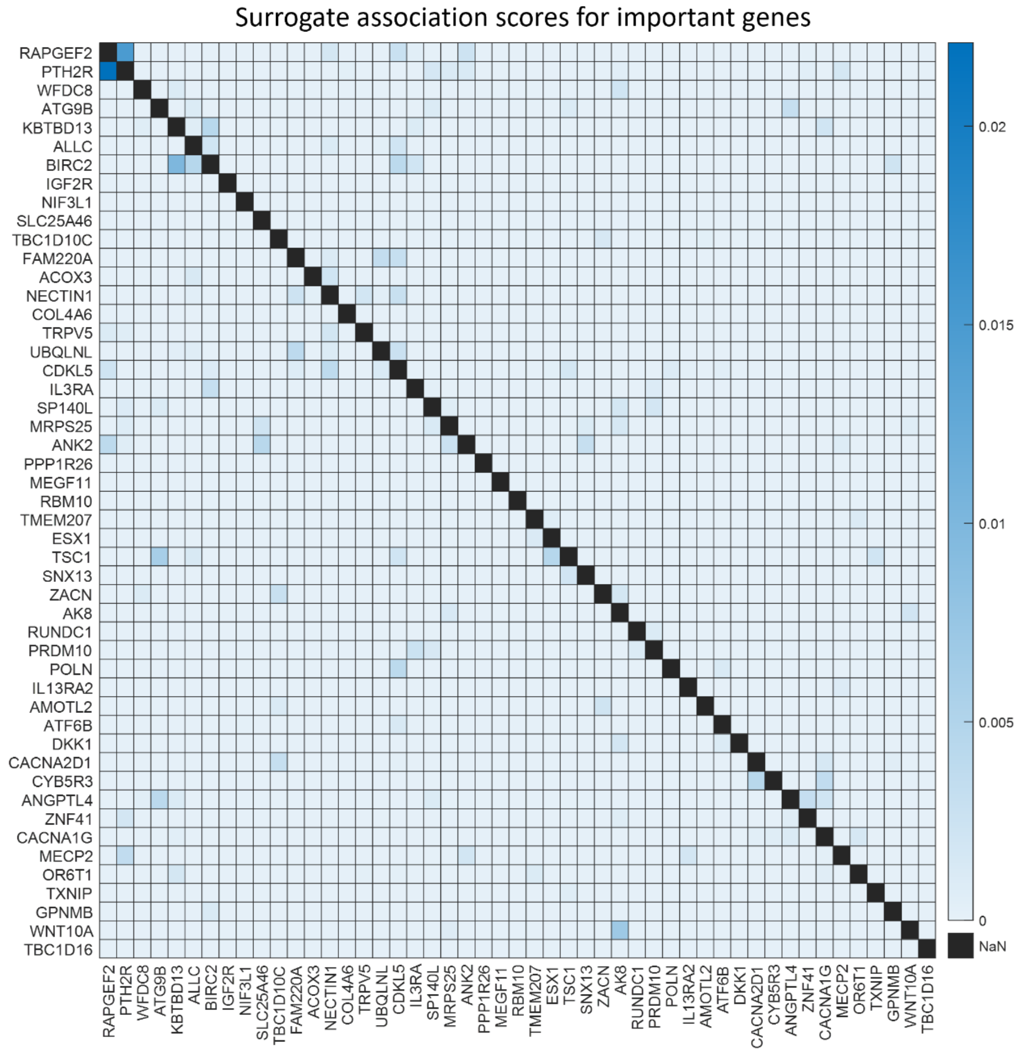
Performance is similar to individual folds

To investigate the relationships between genes in an RF classifier, I trained an additional model on all data (using GWAS summary statistics calculated on all cases and controls and Boruta-important features from across multiple folds). This performed similarly to individual test/train folds with an out of bag error of 0.35, sensitivity of 0.54 and specificity of 0.896. The sensitivity and specificity are slightly better than were achieved for most individual folds (mean sensitivity = 0.43, mean specificity = 0.835). In this case, sensitivity and specificity were calculated on train data used to fit the model. They are therefore over-estimates of model performance. The poor sensitivity and AUC from individual test/train folds indicates that caution must be taken when interpreting these results.

Surrogate association between features, proximities and case clustering

Surrogate association indicates whether features split data in a similar way: at each node in the decision tree, the best feature is chosen to split the data. For any given node, the 'next best choice' can also be determined. If features are often the next best choice for one another, they have high surrogate association. *RAPGEF2* and *PTH2R* had high surrogate association (surrogate association = 0.0148, 0.0221; **Figure 34**). The next best pairing was *BIRC2-KBTBD13* with surrogate association 0.0043-0.010 and other pairings had only slightly elevated surrogate association compared to the mean (0.00014). Scores are not symmetrical and depend on which gene was selected for use in a decision split.

Figure 34. Surrogate association scores for important genes identified by Boruta from all UKBB WES gene level data.



Proximities indicate how often observations appear on the same terminal node, indicating that they have similar combinations of important features. There was no obvious clustering of cases according to proximity scores to suggest subtypes from this data. Some small clusters of individuals classified as cases were present, but these contained both cases and controls (**SI Figure 1**). A large proportion of case individuals (as well as control individuals) carried no variants in any important genes and thus were classified alongside each other in all trees, giving them a proximity of 1.

5.3.5 Pathway level Random Forest models

I trained models using both GSEA-important (highlighted terms enriched in UKBB WES results in *Gene set enrichment analysis of UK BioBank data*) and Boruta-important pathways (identified as confirmed or tentative by Boruta in all UKBB data). Out of bag error was worse for both models trained on data at the pathway level than models trained at the gene level: when GSEA-important pathways were used, minimum out of bag error reached about 0.43 at 50 trees and began to steadily rise as trees were added (**Figure 35a**). A model of similar construction using Boruta-important features had similar performance but the increase in error after 50 trees was less steep (**Figure 35b**). I trained the final models on all data using 50 trees for both sets of features (GSEA-important and Boruta-important). Sensitivity was very poor for both models (<0.1), although slightly higher for the model trained on Boruta-important pathways. Conversely, AUC was slightly improved (GSEA-important mean AUC= 0.515, Boruta-important mean AUC = 0.541) but still poor.

The features identified by Boruta were similar to the pathways identified by GSEA although more generic (**Table 39**). Due to the very poor performance of these models, there is little information that can be reliably drawn about pathway importance or surrogate associations. While some surrogate associations exist in the pathway data for both approaches (**SI Figure 2**), these probably reflect the overlapping nature of GO terms.

Figure 35. OOB error across 10 repeats of Random Forest using pathway data.

A) Only the pathways identified by gene set enrichment analysis are used. b) The pathways identified by Boruta performed on all data are used. a) and b) show 10 RF runs on the same data used to determine the ideal number of trees for training. c) and d) show sensitivity and specificity on testing data for validation folds using the features in a) and b) when RF models were fit with 50 trees of training data for that fold. Both methods have some leakage as feature selection is based on results from all data.

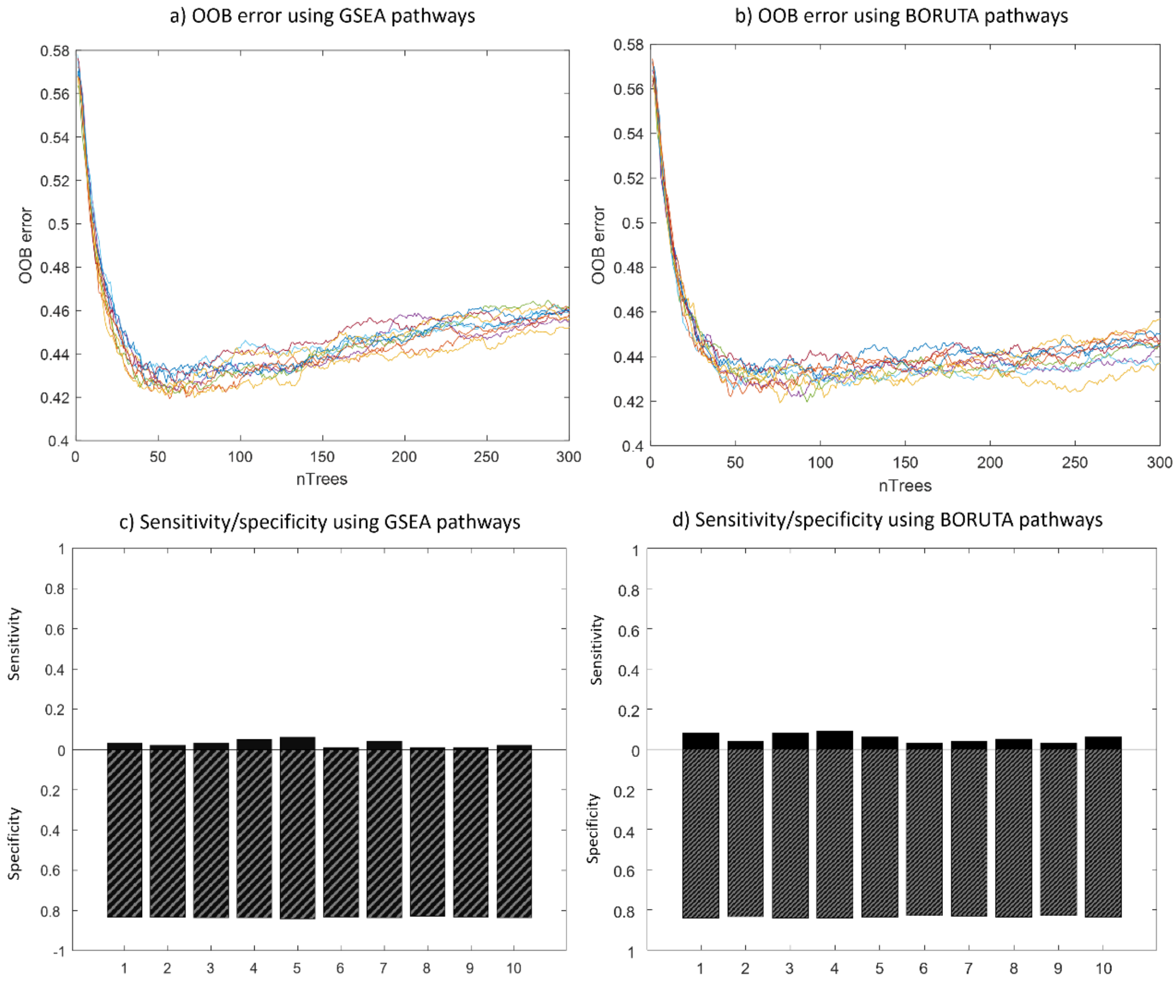


Table 39. Confirmed and tentative important pathways identified by Boruta using all UKBB WES data.

Some functions resemble enriched processes identified by GSEA such as cell motility and terms related to cardiac action potential regulation. The terms identified by Boruta are more general than those identified by enrichment analysis.

Symbol	ID	Confirmed/Tentative
establishment of cell polarity involved in ameboidal cell migration	GO:0003365	Confirmed
molecular function	GO:0003674	Confirmed
binding	GO:0005488	Confirmed
cellular component	GO:0005575	Confirmed

Symbol	ID	Confirmed/Tentative
intracellular anatomical structure	GO:0005622	Confirmed
cytoplasm	GO:0005737	Confirmed
endoplasmic reticulum membrane	GO:0005789	Confirmed
alcohol metabolic process	GO:0006066	Confirmed
terpenoid metabolic process	GO:0006721	Tentative
phosphorus metabolic process	GO:0006793	Confirmed
nitrogen compound metabolic process	GO:0006807	Confirmed
biological process	GO:0008150	Confirmed
positive regulation of biosynthetic process	GO:0009891	Confirmed
protein ubiquitination	GO:0016567	Tentative
hydrolase activity, hydrolyzing N-glycosyl compounds	GO:0016799	Tentative
regulation of phosphate metabolic process	GO:0019220	Confirmed
signaling	GO:0023052	Confirmed
neutrophil chemotaxis	GO:0030593	Confirmed
nuclear lumen	GO:0031981	Confirmed
perinuclear theca	GO:0033011	Tentative
nuclear outer membrane-endoplasmic reticulum membrane network	GO:0042175	Confirmed
anion binding	GO:0043168	Confirmed
macromolecule metabolic process	GO:0043170	Confirmed
organelle	GO:0043226	Confirmed
membrane-bounded organelle	GO:0043227	Confirmed
intracellular organelle	GO:0043229	Confirmed
intracellular membrane-bounded organelle	GO:0043231	Confirmed
primary metabolic process	GO:0044238	Confirmed
cellular biosynthetic process	GO:0044249	Confirmed
positive regulation of cellular process	GO:0048522	Confirmed
cell motility	GO:0048870	Confirmed
regulation of developmental process	GO:0050793	Confirmed
regulation of cellular component organization	GO:0051128	Confirmed
regulation of atrial cardiac muscle cell membrane depolarization	GO:0060371	Confirmed
cellular response to lipopolysaccharide	GO:0071222	Confirmed
cellular response to lipid	GO:0071396	Confirmed
voltage-gated calcium channel activity involved in cardiac muscle cell action potential	GO:0086007	Confirmed
voltage-gated calcium channel activity involved SA node cell action potential	GO:0086059	Confirmed
side of membrane	GO:0098552	Tentative
cardiac muscle cell membrane repolarization	GO:0099622	Confirmed
regulation of cardiac muscle cell membrane repolarization	GO:0099623	Confirmed
ventricular cardiac muscle cell membrane repolarization	GO:0099625	Confirmed
cellular anatomical entity	GO:0110165	Confirmed
organonitrogen compound metabolic process	GO:1901564	Confirmed
positive regulation of protein localization to cell surface	GO:2000010	Tentative

5.4 Discussion

This chapter aimed to use two methods to investigate polygenic disease mechanisms: polygenic risk scoring using PRSice-2 and machine learning, employing Random Forests with Boruta. Due to the lack of data available for properly validating models, this analysis was highly exploratory.

I performed PRS analyses with both FinnGen base data and UKBB WES data divided into test and train splits. PRS performance across all 10 UKBB test/train splits was variable with the number and *p*-value threshold for SNPs included also varying greatly. PRS models explained a small but significant amount of variance in several of the folds. A significant result was also acquired using FinnGen base data, the number of SNPs and *p*-value threshold varying depending on whether UKBB WES or microarray genotype was used as target data. Although the variable performance of these models mean they are not useful for classification of disease risk, they support the existence of a polygenic effect. The effect is likely small due to the very small R^2 of all models produced (<0.02 for UKBB WES base data; < 0.003 for FinnGen base data).

Random forest models performed using presence-absence of significant SNPs per gene did not perform well across validation folds and the best-performing folds did not agree on the most important predictors. This suggests that they performed well due to different combinations of genes being well distributed across test and train splits, which may indicate high heterogeneity. The most consistently important genes across folds according to Boruta include some where a single variant was amongst the most significant, such as *ESX1*, *AMOTL2*, *CACNA2D1* and *CACNA1G.* A large number of cases contained no variants in any important genes. Random Forest models constructed on pathway-level data had even worse performance and were essentially unable to identify cases.

The important pathways identified by Boruta when trained on all data with no test/train splits resembled the terms identified in gene set analysis, including cell motility and cardiac regulation, but were overall more generic. Identification of important pathways using Boruta may be possible but is not as specific as standard gene set enrichment analysis. Perhaps if predictive power were better, important pathways and their relationships could be better identified.

5.4.1 Significant polygenic risk scores support a polygenic effect, but results are variable due to inadequate sample size.

Significant PRS acquired using FinnGen summary statistics as base data support the existence of a polygenic effect that can be detected across populations; if the best-scoring SNPs in a Finnish population are also associated with cholesteatoma in a British population, it is likely that a proportion of these SNPs have a real effect. Performance was best with UKBB whole exome target data filtered to minor allele count (MAC)>20 and the PRS used very few variants, which turned out to be rare variants absent from the UKBB microarray set. The next best performing method used UKBB microarray genotype data with no MAC filter and not reduced to coding only, where most SNPs were used to construct the PRS. For both methods, R^2 was small (<0.003), suggesting that genetic contribution to risk is small compared to environmental effects, or poor power in the original GWAS. PRS distributions did not differ much between cases and controls, but higher disease prevalence was seen at higher PRS centiles, again supporting a small polygenic effect on risk.

However, there are several important limitations with this approach. First, the base and target populations are from a different genetic background; the Finnish population is quite distinct from other European populations and due to recent bottlenecks is enriched for some variants[258]. The WES PRS were non-normally distributed, which can occur when the base and target population differ genetically. Since these are rare variants in UKBB but are common enough to be present on the Finnish array, it is likely that the difference between base and target allele frequencies is confounding the results. This problem may also apply to the microarray data, although the PRS are normally distributed for this result.

For UKBB microarray target data, the most significant PRS occurred when all SNPs were used; this can be an indicator of inadequate power[260]. Additionally, the overlap between the SNPs on the UKBB array and FinnGen array is small compared to the total number of probes. Both biobanks used custom Axiom Array developed according to their research interests: the FinnGen array included additional probes around the major histocompatibility complex and probes for variants associated with certain diseases or known to be enriched in the Finnish population[196]. The UKBB array likewise included probes for variants associated with certain diseases and rare variants. Whole exome data includes many even rarer variants, hence the overlap is even smaller. Therefore, the set of SNPs included in PRS for each method is bound by the overlap between the data sets, rather than by which SNPs are actually the best predictors.

In the analysis using UKBB data only, each fold sampled 90% of the data to be used as 'base' data: this was used to calculate *p*-values and beta scores. PRS was fit on the remaining 10% of data. The high variability of results show that the PRS was very sensitive to the exact composition of base and target groups. For a variant to be highly predictive, it must be present in a large enough proportion of the base split to obtain a high *p*-value *and* also be present in enough of the target split to be predictive in PRS. Generally, the better performing PRS used fewer SNPs which could indicate the existence of a small set of variants with higher prediction power which must be distributed well between test and train. These PRS also suffer from the initial UKBB GWAS being underpowered, exacerbated by only using 90% of the data. Also, the target data consisted of only 10% of the data, meaning results are likely to be particularly sensitive to the coincidental presence or absence of particularly predictive SNPs. The small size of the target data probably contributes to instability of the results.

Good polygenic risk scores include sites reliably known to be associated with risk. For example, the recent release from UKBB of PRS for 53 diseases and traits uses sets of variants from meta-analysis of many GWAS[285]. PRS are not yet used in any clinical setting, with the first trials combining genetic risk with clinical predictors in the ongoing HEART study[120]; notably, this is a predictor for coronary artery disease, a very common and well-studied disease. The GWAS of cholesteatoma in this thesis was underpowered for rare variants and did not identify any significant loci, whilst background knowledge of cholesteatoma genetics is extremely limited. Therefore, I could not generate a high quality PRS, nor draw any conclusions about PRS utility in diagnostics or monitoring for this disease.

5.4.2 Poor prediction power from gene-level Random Forests

This study does not include enough observations to create a reliable classifier, and neither gene-level nor pathway-level models had good predictive power on unseen data. Cross-validation performed on 10 test-train splits of the gene-level data using features selected individually for each fold shows high variability in performance on test data despite consistent and reasonable out of bag errors of around 0.35. Selection of the best features for each fold, rather than testing the same set of features on all folds., likely contributed to this variability. Generally, sensitivity was low and AUC was very close to 0.5: overall, classification of testing data was no better than a random 1:5 assignment of cases:controls.

Good performance in some folds suggests there may be some individuals whose case/control status is more easily classified based on their genes, perhaps because a certain gene or set of genes is particularly predictive in a subset of people and performance depends on their assignment across test and train. This could indicate high polygenicity or heterogeneity. However, these genes are hard to identify and do not seem to be common to the three well-performing folds. Alternatively, this may simply indicate overfitting as participants are likely to carry distinct combinations of variants whether case or control, and so a large enough decision tree could always identify all cases. Such a model would perform badly on unseen data, as was the case in this analysis. Random forests are supposed to be robust to overfitting, but we can see in the pathway analysis that performance begins to decrease after 50 trees, suggesting overfitting can occur.

Sample proximities for a gene-level model do not reveal any obvious clustering. Small clusters containing both cases and controls probably just represent the fact that some cases and controls share variants in certain genes by chance. Also, a large proportion of both cases and controls contained no variants within any of the important genes and were always classified as controls. Overall, this seems to reflect that the best-performing genes are slightly enriched for variants in cases compared to controls, which we already knew because their *p*-values are <0.05. Not all genes containing top-ranked variants were represented, however, including *OR10A2* (olfactory receptor 10A2) which contained several high-ranking SNPs. This may be due to high polymorphism in the olfactory receptors[286]; if a gene contains many different variants, it may be more likely to contain some significant ones by chance than a less polymorphic one. Also, the presence of many non-significant SNPs in a highly polymorphic gene could introduce noise and make the gene a less useful predictor.

Two genes had stronger surrogate associations than other gene pairings: *RAPGEF2* and *PTH2R*. These are on different chromosomes (4 and 2 respectively), so this is not due to linkage disequilibrium between variants. PTH2R is a G-protein coupled receptor of parathyroid hormone with diverse functions in the central nervous system[287] and RapGEF2 is a guanine exchange factor for Rap/Ras GTPases, also with roles in the central nervous system[288]. G-protein coupled receptors are inactivated by GTPases, suggesting a possible link between these two genes which may explain their high surrogate association.

Pathway-level models failed to classify most cases and were less informative than gene set enrichment analysis.

Pathway-level models performed worse than gene-level ones, tending to classify all unseen data as control. Important pathways identified by Boruta were more generic than those identified in GSEA but some similar functions to GSEA were also detected, such as cell motility and voltage-gated calcium channel activity. The sorted query used in GSEA provides higher weight to more significantly associated SNPs whereas this analysis simply uses the number of $p<0.05$ SNPs per gene and then per pathway. As such, the effects of strongly associated SNPs may be diluted by the effects of weakly associated SNPs which are more likely to be associated by chance. As a result, the standard GSEA is probably a more powerful method for identifying important pathways for an underpowered GWAS.

5.4.3 Limitations and possible improvements

The main limitation of this analysis was a lack of data for training and validation, which affected both PRS and ML models. It is uncertain how many cases would be required to give adequate power to GWAS to construct reliable risk classification models or PRS. Little is known about cholesteatoma genetics, so its genetic architecture is difficult to estimate, and we cannot apply knowledge about established risk loci to improve prediction. Some improvements to RF models may have been possible by varying parameters, for example modifying feature selection or tree depth. Lower *p*-value cutoffs for SNP inclusion in RF models may reduce the dilution effect of non-significant SNPs being counted alongside significant SNPs within genes. However, the initially underpowered GWAS and poor performance of RF models meant I refrained from developing models further than was described.

5.5 Conclusion

When classification problems are complex, classifiers need large amounts of data. This is true for both polygenic risk scores and other machine learning methods. First, the initial data set used to generate summary statistics must be large enough to distinguish signals from noise. Next, there must be enough data to split into test and train, as testing data is crucial for internal validation of the model. Finally, there should be additional, independent data sets for external validation. This is particularly true for PRS, which need separate base and target data as well as validation data. To mitigate these issues, I used base data from FinnGen for PRS, as well as dividing UKBB data into test/train splits, which were used for both machine learning and PRS base/target data.

PRS acquired using FinnGen bas data provide some support for a polygenic role in cholesteatoma. However, use of non-British base data on British target data probably distorted the results due to enrichment of certain rare variants in the Finnish population. PRS calculated on cross-validation folds within UKBB were less significant and less consistent, varying greatly in the *p*-value and number of variants used. This was probably because GWAS was underpowered, resulting in inaccurate estimates of variant odds ratios which are particularly sensitive to the composition of the test and train groups.

Random forest classification probably failed for a similar reason; poor performance may reflect high variability in the number and composition of risk variants carried by individuals, but this is difficult to tell from inadequate sample size. Pathway-level models had even poorer performance, perhaps because assigning variants to the pathway level results in a 'diffusion' of their effects as they become grouped with other variants and spread across multiple features. As a result, little additional insight into genetic architecture could be learned.

6 Discussion

6.1 Summary of findings

In this thesis, I performed an epidemiological and genetic study of cholesteatoma in the UK BioBank using publicly available results from the Finnish biobank FinnGen for comparison and validation.

I began with a semi-systematic review of global gene expression studies to make sense of a large volume of literature examining differential expression of various proteins in cholesteatoma and determine the genes and processes that are consistently implicated across studies. Most of these studies had small sample sizes (N=2-17) and differing approaches, but some genes were consistently detected as differentially expressed including the upregulation of *SERPINB3* and *SERPINB4*, *S100A7, S100A8, S100A9,* and *CEACAM6*, and the downregulation of *TNXB* and *COCH*. Disrupted processes included tissue development, cell adhesion, extracellular matrix (ECM) constituents, metal ion binding and immune function, though many immune genes were downregulated compared to COM tissue.

In my epidemiological study, I described the identification of cases and controls from UK BioBank data as used in this and later analyses and addressed some questions about demographic risk factors by performing adjusted logistic regression. I found additional evidence for some tentative and known risk factors including male sex, deprivation and smoking. I also found that the demographics of cholesteatoma and non-cholesteatoma middle ear disease were more similar to each other than to the population with disease-free ears in UKBB, except for sex ratio and ethnicity; the male predominance is not present for non-cholesteatoma middle ear disease. Cholesteatoma prevalence was highest in White and Asian participants (in UKBB generally referring to Indian, Pakistani and Bangladeshi ethnicities) and lowest in Black participants.

The epidemiological analysis also explored some of the overlap of cholesteatoma with other inflammatory ear disease, including suppurative, nonsuppurative and unspecified otitis media, otitis externa and mastoiditis. Though these associations were generally well-known, these analyses raise useful queries about what is being investigated when we compare cholesteatoma to other middle ear disease or disease-free controls. I also detected increased rates of chronic sinusitis and respiratory disease such as asthma and bronchiectasis in cholesteatoma and identified a potentially new association with epilepsy replicated across UKBB and FinnGen. Meanwhile, an association with otosclerosis may be due to misdiagnosis due to the initial similarity of presentation and the fact that misdiagnosed ICD-10 codes are not removed from records.

In my genetic analysis, I performed genome-wide association tests (GWAS) at the variant and gene level and used gene set enrichment analysis (GSEA) to identify pathways and processes disrupted in cholesteatoma. Although GWAS identified no single variants or genes meeting genome-wide significance thresholds (traditional GWAS threshold 5x10^{-8}; WES rare variant threshold 3x10^{-7}; gene-level threshold 2.5x10^{-6}), GSEA of all whole exome variants with *p*-values < 0.05 indicated enrichment in certain processes: cell-cell adhesion, cellular motility, ciliary function via dyneins, developmental processes, and calcium binding. Notably, neither ECM nor immune function were enriched. These results were replicated in FinnGen data except for the enrichment of dynein proteins, which was due to several rare *DNAH* and *DNAI* variants in UKBB WES data. Ciliary impairment and calcium binding were also implicated by our previous whole exome study of twenty-one individuals from ten family clusters[126], with several *DNAH* genes containing rare, deleterious variants co-segregating with cholesteatoma. In the UKBB whole exome single analysis results, dynein motor processes were enriched due to both *DNAH* and *DNAI* family members.

I explored possible methods for characterising polygenic risk including polygenic risk scoring (PRS) and Random Forest (RF) machine learning approaches. PRS and RF classification were not viable with this data set due to its small size and the need for separate validation sets. I experimented with using FinnGen base data for constructing PRS and acquired significant results; however, model R^2 was always very small, and the results were very variable, showing that PRS explained a very small portion of phenotypic variance and was highly sensitive to composition of the base and target groups. PRS performed on ten cross-validation folds of UKBB data also showed highly variable results, with base/target splits varying in the number of SNPs used and significance of the PRS. RF classification was likewise poor at classifying testing data. Not much could be learned from these models due to lack of data and the probable complexity of disease, meaning no models could accurately classify cases and controls.

6.2 Cholesteatoma genetics in the context of middle ear disease

6.2.1 Factors identified in both cholesteatoma and otitis media

Possible direct causal links between cholesteatoma and OM

My epidemiological analyses support a large overlap in cholesteatoma pathology and risk factors with other middle ear diseases. A history of otitis media (OM) is common in cholesteatoma and the symptoms overlap, involving inflammation, otalgia, and otorrhea[1,6,9,23]. OM is inflammation in the middle ear, commonly in response to bacterial or viral pathogens[289,290]. Most children experience at least one episode of otitis media with effusion but disease is usually self-limiting[19]. Chronic OM may directly contribute to cholesteatoma development, for example by causing tympanic retraction. Negative pressure in the middle ear due to poor ventilation pulls the tympanic membrane inwards, resulting in a small pocket on the exterior side which may accumulate keratin debris[22]. This debris may constitute a pre-cholesteatomatous stage, which only proceeds to cholesteatoma in a small number of cases[22]. Additionally, destruction of collagen and elastin in the tympanic membrane due to chronic inflammation causes it to weaken, exacerbating the retraction[181]. Debris in the retraction pocket is thought to be the origin of primary acquired cholesteatoma according to invagination theory. Chronic inflammation may also result in perforation or provide conditions which provoke mucosal metaplasia or basal cell hyperplasia[21]. However, the symptoms of OM may also be caused by cholesteatoma, making it difficult to determine the temporal or causal relationship.

The concept of endophenotypes, drawn largely from genetic studies of psychological phenomena, may be useful for understanding this relationship. An endophenotype is an intermediate phenotype which, in combination with other endophenotypes, increases risk of another phenotype[291]. Endophenotypes may be shared between similar diseases. For example, poor ciliary function or cranial morphology may be considered endophenotypes for cholesteatoma which also act as endophenotypes for OM, hence raise the risk of both. Whether OM itself is an endophenotype contributing to risk of cholesteatoma is not clear. There may be factors governing risk of OM, risk of OM becoming chronic, and risk of proceeding to cholesteatoma, as well as distinct risk factors for cholesteatoma independent of OM. In my analysis, I excluded all ear disease from the controls, so there was no way to distinguish genetic effects contributing directly to cholesteatoma to those contributing to ear disease generally. Therefore, it is useful to compare the results of genetic analyses to the known genetics of OM.

There are no accepted OM risk loci, and study of OM is complicated by its various forms (acute, chronic, with or without effusion, suppurative or non-suppurative). However, genetic studies generally identify variants in the genes involved in the inflammatory response, mucin production and mucociliary transport, and development of the Eustachian tube[289,290]. Given that OM arises from colonisation with pathogenic bacteria, variability in host immune response is thought to play a role in susceptibility with specific variants in several interleukin genes, HLA-A, TLR4 and TNF- α having been detected in individual GWAS[289,292]. However, my results did not indicate a strong role for immune/inflammatory genes in cholesteatoma.

Mucociliary function in the middle ear is essential for preventing colonisation of pathogens through antimicrobials in the mucus, physical clearance, and recruitment of inflammatory cells[293]. Eustachian tube important for clearance of the middle ear and pressure equalisation; poorer Eustachian tube function in children compared to adults is one reason they are more susceptible to middle ear disease[293]. The evidence for a genetic involvement via morphological differences is mostly due to an association with chromosomal abnormalities[289].

Support for cilia in cholesteatoma but not variants in inflammatory response

My data do suggest a role for ciliary dysfunction in cholesteatoma. Dynein binding proteins were enriched amongst UKBB WES results, primarily due to rare (MAF < 0.001) variants in *DNAH* and *DNAI* members. Along with the *DNAL* family, these genes encode components of axonemal dyneins, which act as motor proteins responsible for powering the movement of cilia. Variants in *DNAH5, DNAH11, DNAH1, DNAI1, DNAI2, DNAL1, DNAAF1, DNAAF2, DNAAF3* are known to cause primary ciliary dyskinesia (PCD), which is associated with recurrent middle ear and respiratory infections[253]. In my epidemiological study of cholesteatoma, I found an association with chronic sinusitis and bronchiectasis which form a triad with situs inversus in Kartagener's syndrome, a form of PCD[17]. This further supports a generally poor ciliary function in persons with cholesteatoma in UKBB. A role for dyneins in cholesteatoma is also supported by the previous GoC whole exome study[126] where several rare, deleterious *DNAH* variants that co-segregated with a diagnosis of cholesteatoma were identified in families with multiple cholesteatoma cases.

6.2.2 Factors identified in cholesteatoma but not in OM

Adhesion and cytoskeleton

A role for cell adhesion is supported by gene expression studies as well as enriched terms from UKBB WES, microarray and FinnGen results. This overlaps with enrichment of cytoskeletal variants, particularly actin organization; both adhesion and reorganization of the actin cytoskeleton are required for amoeboid cell motility. Alteration of the balance between cell-cell adhesion and cell-ECM adhesion can promote altered cellular migration and invasiveness in cancer[175]. *CEACAM6* is an example of an adhesion molecule expressed by cholesteatoma which is also associated with invasiveness in cancer[162]. In some tympanic retractions, a pre-cholesteatoma in the form of micro cysts in the propria lamina can develop, but this does not usually proceed to cholesteatoma[22]. Genetic differences promoting invasiveness could be important in determining whether the cholesteatoma continues to develop or fails to establish itself in the tympanic membrane. Furthermore, altered migration may adversely impact the self-cleaning mechanism of the tympanic membrane, which involves a continuous outwards migration of cells from the centre[294]. Cell migration is also implicated by the migration theory of cholesteatoma, where aberrant migration through a perforation is considered the origin of cholesteatoma epithelium[21].

Adhesion molecules also have signalling roles, including those downstream of fibroblast growth factor receptor and epidermal growth factor receptor[175]. Excessive growth of cholesteatoma cells may also contribute to establishment under several theories of cholesteatoma formation.

Calcium binding

Calcium ions act as an important signalling molecule in many processes including muscle contraction, neural signalling, cell growth, cell migration and cell death, making its role is cholesteatoma difficult to suggest. In polycystic kidney disease, low intracellular calcium is thought to drive cyst formation through upregulated cAMP, increased fluid secretion and activation of the MAPK pathway[250]. In this case, defective calcium localisation is thought to arise through defective ciliary signalling. However, renal cysts have little in common with cholesteatoma (aside from a wider association between renal and otic disease[295,296]), as they are typically described as simple, fluid-filled sacs.

Due to its signalling roles in processes such as cell migration, cell growth, and cell death, calcium also plays an important role during wound-healing. Cholesteatoma perimatrix is a granulation tissue resembling wound tissue, in that it displays inflammation, breakdown of ECM, cellular proliferation and remodelling. Calcium is essential for proper differentiation of keratinocytes and modulation of angiogenesis during wound-healing, and deficiency in animals is associated with higher rates of chronic wound formation[249]. Calcium binding is also another large term containing many genes, although unlike the developmental process terms, there is support from gene expression studies for a role in cholesteatoma.

A major role of calcium is in the skeletal system and one report[297] found increased rates of cholesteatoma with osteoporosis, though this was not replicated in UKBB data. Meanwhile a class of drug (bisphosphonates) used to treat osteoporosis, has occasionally been noted to induce external auditory canal cholesteatoma[42]. Although bone turnover is altered in cholesteatoma, this is not until after disease is established so seems unlikely to contribute to formation.

6.2.3 Additional evidence from new FinnGen release

Following the completion of genetic analyses in this thesis which used FinnGen Release 9, Release 11 was made public. In this release, there is a single significant variant and several loci approaching significance. The only individual significant result is for a rare (case AF 0.13%) intergenic variant near to the RP11-2P2.2 pseudogene (rs766961752) (OR=107.38, *p*-value $6.4\text{x}10^{-10}$). This pseudogene was also strongly associated with cleft lip and palate in FinnGen (OR 4.73, p=$7.4\text{x}10^{-17}$), suggesting an overlap between cholesteatoma and cleft lip and palate in this cohort. This association within FinnGen was also reported by Rahimov *et al.* (2024)[298] , who attribute high incidence of cleft lip and palate in the Finnish population to intergenic variants near *IRF6*.

There was also a strong peak in chromosome 16 due to common variants around LINC02131; the most significant was rs1117410, which was less common in the case group. LINC02131 lies between *NUDT7* and *ADAMTS18* and was also strongly associated with decreased risk of diseases of middle ear and mastoid, suppurative and unspecified otitis media, chronic suppurative otitis media and acute suppurative otitis media within FinnGen. It is interesting that the variants at this locus tended to be associated with decreased disease risk; this of course depends on which allele is considered reference. The most strongly associated variants at this locus had frequencies ~50%, probably reflecting the common nature of middle ear disease.

There are several other loci where signals seem to be emerging and these may become significant in future releases; I do not discuss these now as it is difficult to determine which genes are mostly likely to be involved due to linkage disequilibrium, though interestingly most of the nearby genes at these loci are not also associated with other middle ear disease according to FinnGen PheWAS. An additional interesting finding was a variant in *DNAH7* (rs1419900187) with *p*-value 8.5x10-7 (AF 0.000408); this was 10th most significant variant in FinnGen.

Interesting, nonsuppurative otitis media, suppurative otitis media, acute suppurative otitis media, and otitis externa all have strong peaks around the MHC region in chromosome 6, indicating an immune role in these diseases. No such peak is present in cholesteatoma results, despite this peak being visible even in 'acute otitis externa noninfective', which only has 419 cases. This is interesting as it further suggests that immune function does not have a major role in cholesteatoma susceptibility, even if it is involved in pathogenesis.

6.3 Proposed genetic architectures

6.3.1 Evidence of a polygenic effect

This study did not identify any single genes or variants significantly associated with cholesteatoma, but several enriched processes were detected and supported by data from FinnGen and our previous whole exome study, suggesting a polygenic effect does exist. However, it is unclear whether individuals are enriched for variants in multiple processes, or if defects in just one pathway are sufficient to increase cholesteatoma risk.

Heritability cannot be estimated from this data, but the overall genetic effect does not appear to be large based on the small R^2 of PRS calculated using both FinnGen and UKBB base data. However, a higher better R^2 may be acquired with a better-powered GWAS for base data or better match between base and target populations. Meanwhile, power analysis of GWAS imposes a maximum possible risk ratio on single common variants of 1.4; this means a heterozygous carrier would be 40% more likely to have disease than a person with no variant. This largely discounts the possibility of a small number of variants with strong effects, although rare variants may have larger effect sizes and there is the possibility of type 2 error. Failure of previous genetic studies[65,66,68,124,126] to identify any common genes or variants supports a lack of high-penetrance causal variants but the small sizes of these studies and lack of controls may also result in type 2 error.

The omnigenic model suggests that all genes expressed in relevant tissues can be involved in disease due to complex regulatory networks. As a result, larger functional terms tend to explain more heritability than more functionally relevant terms. Boyle *et al.* (2017) suggest that genes under the omnigenic model can be sorted into 'core genes' which are directly relevant to phenotype and 'peripheral genes' affect phenotype indirectly via regulatory networks[299]. In this analysis, functions not directly related to cholesteatoma phenotype may include tissue development, neural development, and calcium binding. The omnigenic model may also explain the poor transferability of polygenic risk scores and variant effect sizes across populations as it is not only the effect of core genes that must be considered but of a large number of interacting peripheral genes which may be heterogeneous between populations[300]. Many enriched terms in genetic analysis of UKBB data had compelling links to cholesteatoma biology including ciliary function, cell adhesion, cytoskeletal organisation, and calcium binding. Some processes such as developmental processes and synaptic signalling have more obscure roles and could be considered peripheral, but further study would be required to determine this.

6.3.2 Familial and non-familial forms of disease

It is not unusual for diseases to have familial and sporadic forms. In such diseases, a portion of cases are due to a small number of highly penetrant genetic variants and run in families. The remaining portion are sporadic and have a less obvious genetic basis, seeming to be dominated more by environmental factors. However, in some diseases it has been shown that non-familial cases can be influenced by genetics in a highly polygenic manner. For example, up to 10% of cases of breast cancer are due to highly pathogenic variants on a small number of genes, primarily *BRCA1* and *BRCA2*. Meanwhile, over 100 different loci have been linked to individually small increases in breast cancer risk but can be combined in a polygenic risk score identifying up to a 3-fold increased risk due to polygenic effects[301]. Similarly, ALS has familial and sporadic forms, where the familial form (10% of cases) is typically associated with dominant inheritance of a variant affecting a single gene. Common causative genes are *C9ORF72*, *SOD1*, *TARDBP* and *FUS*, but many genes have been identified in familial ALS and are associated with different presentations of disease[302]. Genes associated with the sporadic cases are largely unknown, though a small number of individuals will also carry mutations in *C9ORF72, SOD1, TARDBP* or *FUS*. In both breast cancer and sporadic ALS, there is a large environmental risk factor.

These cases highlight the complexity of diseases where different genetic architectures can have different presentations, penetrance, and heritability. In cholesteatoma, only 10% of cases reported a family history in an online survey posted to cholesteatoma support groups[61]. Though there is likely to be a degree of bias – persons with family history may have had better awareness of cholesteatoma and more likely to seek support groups, for example – this rate is a lot higher than we might expect due to chance if lifetime risk is 1 in 500. Affected families may share particularly high polygenic risk scores, important lifestyle factors, or specific variants with high penetrance. Our previous WES study did not identify any rare, deleterious variants which co-segregated with cholesteatoma in all 10 families, but some processes were enriched in the variants identified by TRAPD analysis and in the variants common to at least 2 families. This included microtubule motor function, ECM degradation and calcium ion binding, so it is possible that a familial form of cholesteatoma is based on a smaller number of higher risk variants affecting these processes, or individual risk variants unique to each family affecting similar processes. Due to the identification of *DNAH* variants in these families and enriched dynein-binding in UKBB WES results, I suggest that rare dynein variants specifically require further investigation in affected families.

6.4 Limitations

The limitations of individual studies performed in this thesis are discussed in their relevant chapters. Here, I discuss some overarching limitations of the biobank approach.

6.4.1 Difficulties due to rare disease and sample size

A major limitation of this study was the sample size. Cholesteatoma is rare, so acquiring many cases is difficult and the number of cases was ultimately limited by the number in UKBB. In order to increase the sample size and study power, I used expanded criteria using ICD-10 and OPC codes to identify all putative cases, which almost doubled the number within UKBB. This may also have introduced some cases who were not actually cholesteatoma cases, although procedures such as mastoidectomy are unlikely to be performed for any other reason. However, ICD-10 codes are largely missing before 1995 and UKBB participants are above the age of peak incidence, meaning childhood cases may have been missed.

Despite maximising the number of cases, I have shown that the study was underpowered for rare variants unless highly penetrant. This is particularly problematic given the evidence that rare dynein variants may be important. Also, if there are familial and sporadic forms of cholesteatoma, we can expect rare variants to have a greater role in familial disease. This is because familial cholesteatoma is rare, hence we would not expect common variants to be causal. Furthermore, rare variants may have stronger deleterious effects than common ones, as negative selection prevents them from becoming common[219]. Cholesteatoma genetics are likely complex, given that no studies have yet identified the same variants and disease risk is multifactorial.

To reduce risk of type 1 error, I performed filtering to consider only non-synonymous coding variants in UKBB WES data and applied similar filters to microarray data for comparison. Alongside quality filtering, this reduces the number of variants and retains the highest quality variants most likely to be involved in disease. Generally, synonymous variants are not thought to contribute to disease[228], so it is unlikely that this will exclude any causal variants. Type 2 error is difficult to avoid, particularly given the small sample size. Type 2 error rate can be reduced by lowering *p*-value threshold, but this necessarily increases type 1 error. By conducting enrichment analysis, the effects of variants not meeting genome-wide significance can still be assessed.

Lack of data also severely restricted the polygenic risk score and machine learning analyses, as large volumes of test, train and validation data are required for accurate classification. The best PRS are built on foundations of good knowledge about genetic and environmental risk factors[121,285,303], which are not well-known in cholesteatoma. Due to these issues, most examples of PRS are for common, well-studied conditions such as breast cancer, Alzheimer's disease, coronary artery disease and type 1 diabetes[304].

6.4.2 Congenital cholesteatoma cannot be distinguished from acquired cholesteatoma in this study

Lack of granularity in the UKBB data means it was impossible to distinguish congenital and acquired cholesteatoma. The discussion in this thesis focuses on acquired cholesteatoma as the more common form (estimates for congenital cholesteatoma are between 4 and 24%)[5]. However, they have distinct features which mean factors that I have suggested to play a role in acquired cholesteatoma may not apply to congenital cholesteatoma. For example, congenital cholesteatoma is not associated with inflammation, tympanic retraction, or perforation[2]. Ciliary variants, which may increase risk of OM and prevent proper clearance of debris from the middle ear, may not be relevant. Aberrant migration may still be involved as a possible origin for congenital cholesteatoma is migration from the developing auditory canal[305]. Likewise, tissue development processes could be involved if the origin is an epithelial remnant which is inappropriately retained in the middle ear[306]. However, there is less literature regarding congenital cholesteatoma causes, making it difficult to draw conclusions.

6.4.3 Limitations in functional interpretation

I used GO term enrichment to identify disrupted processes in both gene expression data and genetic variant data. However, our knowledge of the pathways and processes that genes are involved in, or the tissues in which they are expressed, is not exhaustive and annotations change over time as databases are amended with new information. Many annotations are inferred from structural similarity via automated processes and many annotations belong to a relatively small number of well-studied genes[180]. This can cause analysis to be biased towards functions associated with well-studied genes and can also affect study replicability as GO terms are updated between studies.

Additionally, there are limitations to the interpretation of individual variants both due to incomplete knowledge of gene functions and inability to determine causal loci. This is particularly relevant to microarray approaches where it is likely that causal variants are not directly measured, particularly if they are rare. Associations with non-causal variants can arise through linkage with causal variants, making it difficult to determine which genes are actually involved in pathology. When researchers wish to identify which variant in a region of strong linkage disequilibrium is causal, they may employ fine-mapping[307]. Fine-mapping was not indicated in this study as no significant variants or strong signals were identified to begin with.

6.4.4 Wider ethical issues due to lack of diversity within biobanks

This study is only relevant to white British populations due to the small number of non-white participants in UKBB. This is part of a wider issue with biobank-based studies, as many of them contain majority European background individuals due to the large number of North American and European companies involved[308]. Very few biobanks exist outside these regions. Amongst the European and American biobanks, some have a specific aim of including participants from diverse backgrounds (e.g. Our Future Health). However, the overall effect is that knowledge of genetic disease is concentrated on white populations in Europe and America; results may not apply to other ethnicities due to varying gene frequencies. This could lead to exacerbation of existing inequalities within and between countries. This effect is amplified by the fact that many non-genetic risk factors are also associated with different ethnicities within countries, as these are often linked with socioeconomic factors.

6.5 Future study directions

6.5.1 Future studies should be guided by clinical utility

Studies should be informed by what type of information would provide the most benefit to patients. First, can genetic information be acted upon? In cholesteatoma, early identification and surgery are important to preserve hearing and reduce risk of recurrence[309]. A person at high risk could be monitored more closely, for example being checked for cholesteatoma when experiencing otalgia or effusion but there are currently no non-surgical treatments that could be given to reduce risk.

A 2022 review[53] of the current state of medical interventions for cholesteatoma lists several drugs which have been or are under trial for treatment of inflammatory disorders which may have application in cholesteatoma. Most of these are anti-inflammatory drugs, many of them targeting TLR4 and/or RAGE receptor. TLR4 and RAGE are both pattern recognition receptors involved in the immune response; TLR4 is a major activator of innate immunity whose primary target is bacterial lipopolysaccharide, while RAGE has a broader repertoire including S100 proteins. Of the suggested drugs, Ibudilast, Azeliragon, and Fenebrutinib are in phase 3 trials for other inflammatory conditions; however there have been few drugs tested on cholesteatoma.

Some mouse and human studies have been performed: Uzun *et al.* (2021)[310] investigated anti-inflammatory drugs in cell culture and concluded that tacrolimus and imiquimod are candidates for further study due to decreased expression of inflammatory cytokines, decreased cholesteatoma cell viability, and minimal known ototoxicity. Earlier studies[311,312] identified the anti-inflammatory 5-fluorouracil as a potential treatment. Vitamin A has been studied in animal models of cholesteatoma, as vitamin A deficiency is linked to chronic otitis media: Nageris *et al.* (2001)[198] report that both vitamin A and cortisporin reduced the rate of cholesteatoma formation in a gerbil model and Rao *et al.* (2009)[313] report treating 5 patients with vitamin A and completely removing cholesteatomas in 4 of them. However, Boesoirie *et al.* (2023)[314] found no difference in vitamin A or E for CSOM or CSOM with cholesteatoma in 60 patients. Overall, there are still no non-surgical cholesteatoma treatments and no drugs in development as alternatives, despite several candidates being suggested.

Any intervention based on genetic risk would currently amount to increasing awareness of disease, closer monitoring, and earlier intervention to preserve hearing. Whether genetic testing would offer any benefit over simply informing people of a familial risk would depend on the strength of the genetic effect, whether any non-surgical treatments become available, and whether any treatment course might be affected depending on which genes are involved.

Beyond genetic testing within high-risk families, the utility of PRS for predicting cholesteatoma risk is questionable. Even the most rigorously studied complex diseases do not generally have PRS in clinical use: the ongoing HEART study[120] uses a combined PRS and clinical predictor for coronary artery disease[121], one of the most common causes of death in the world, and has just passed its pilot phase. No other PRS are used for predictive purposes in a medical setting, and so are mostly offered by commercial companies offering private testing via at-home kits. PRS offered vary between companies, but include coronary artery disease, type 2 diabetes, various cancers, and Alzheimer's disease. Yet these are not clinically tested, and while PRS are generally accepted to describe population-level risk well, their utility on an individual level is not known. PRS are continuous, so a somewhat arbitrary cutoff must be placed to determine 'high risk' or which people require further screening or preventative treatment. This means that the additional benefit from including PRS in screening for disease risk may be modest[272] and will depend on the relative contribution of genetics to disease risk. Another concern is that emphasis on complex genetic risk factors draws attention away from well-known, highly impactful environmental risk factors such as smoking, obesity, and deprivation[272].

Aside from these issues, PRS generally are not constructed for rare diseases such as cholesteatoma. This is probably because sample size is often small, because rare diseases are not as widely studied as common diseases, are generally studied by methods other than GWAS, and because previous GWAS have mostly been microarray based (so have not captured rare variants). In short, we do not have reliable odds ratios for genetic variants associated with polygenic rare diseases. In order for a PRS to have any predictive power in a general population, much more about cholesteatoma biology would need to be known and there would need to be an effect size large enough to warrant investigation. PRS may be of more utility within groups already known to be a high risk, in this case when disease is in the family. However, it may be that familial cases are less polygenic; additional family studies are needed to provide insight into this matter. Future studies may also wish to compare cholesteatoma to disease-free ears, as well as to non-cholesteatoma middle ear disease.

6.5.2 Strategies for increasing power of future studies

Individual studies may struggle to acquire a large enough cohort of cholesteatoma patients for genetic testing. However, the large and increasing number of biobanks make meta-analysis an attractive possibility. Rämö *et al.* (2023)[90] recently performed such a meta-analysis for otosclerosis using results from UKBB, FinnGen and the Estonian Biobank. As more studies are performed on these biobanks and PheWAS results are generated, meta-analysis may become increasingly easy as summary statistics can be used. Biobanks such as FinnGen may release these results, or outside researchers may release them; several PheWAS browsers are available for UKBB, for example GeneBass (https://app.genebass.org/) and PheWeb (https://pheweb.sph.umich.edu/), while BioBank Japan compiles results from individual studies (https://pheweb.jp/). An additional benefit of PheWAS is the ability to check for phenotype associations for high-ranking genes which may offer more information on the reasons for a gene's significance, for example the FinnGen R11 result for cholesteatoma also being associated with orofacial cleft. Such associations may not necessarily extend beyond the individual biobank being studied but could explain the associations within that biobank.

Such studies may provide stronger evidence for a genetic basis of cholesteatoma risk and help to differentiate it from other middle ear disease. This knowledge could be used to guide family studies, as loci or pathways identified from the general public could be examined directly rather than sifting through the whole genome. In other diseases with sporadic and familial forms, there is often overlap in the genes involved in both forms[302,315–317]. A recent review of obesity genetics also identified common pathways between rare monogenic causes of severe obesity and common, polygenic risk of obesity[318], so even if identical genes are not involved, there is likely to be similarity in disease pathways. Therefore, GWAS results should provide additional insight into familial cholesteatoma even if distinct familial and sporadic forms of disease exist.

Power may also be increased if the binary outcome can be converted into a continuous trait, for example if a biomarker can be identified associated with severity[319]. Expression of the biomarker is measured, resulting in detection of expression quantitative trait loci (eQTLs)[320]. There are no cholesteatoma biomarkers at this time, though gene expression studies reveal several candidates. The matrix metalloproteins have often been studied in cholesteatoma and are thought to be involved in invasiveness and bone destruction: MMP expression has been correlated with increased destruction of bone[156,158], and they are also expressed in the tympanic membrane where they may contribute to retraction pocket formation through degradation of the lamina propria and loss of elasticity [181]. Although several different MMPs have been detected in cholesteatoma, MMP9 is amongst the most frequently described.

Khondoker *et al.* (2015)[321] increased power of an Alzheimer's disease GWAS using imaging data to measure cortical thickness and regional volume in different parts of the brain. Cholesteatoma can vary in the location and spread, extent of ossicular damage and invasion of other bony structures. They can easily be identified via MRI or CT scan[322]. This presents an alternative quantitative measurement for cholesteatoma. However, an issue with both this approach and the eQTL approach is cost effectiveness; biobanks do not typically contain large amounts of imaging data, nor are they likely to contain tissue samples from the middle ear, so participants would have to be recruited and these measurements taken. This reintroduces the issue of cholesteatoma incidence being low, making recruitment difficult as well as expensive. Another possible approach could be to use the presence of complications as biomarkers for cholesteatoma aggressiveness, for example using ICD-10 codes indicating recurrence or intracranial complications such as meningitis.

6.5.3 Perform family studies focusing on affected pathways

Family studies have the potential to better identify rare, highly penetrant variants[323]. First, other family members can be used as controls, reducing the number of candidate genes. If individual families carry distinct risk variants or combinations of risk variants, such studies may better identify them than population-level GWAS as the signal to noise ratio will be worse for the latter. A better estimate of heritability may be acquired from family studies, as although heritability can be estimated from large, unrelated populations, these methods require high power and are generally poor where rare SNPs are involved[324].

However, if candidate variants have low penetrance, family studies may lack power[323]. One issue is that many genes are likely to co-segregate with disease, meaning they must be filtered in some way to identify likely candidates. Our previous study used a combination of predicted impact filters, co-segregation analysis and TRAPD analysis, where the frequencies of variants are compared to the frequency in the general population. Filtering to genes associated with the functions identified in this GWAS may be one method to reduce multiple testing and further increase power of family studies.

Given that our family study detected rare, deleterious *DNAH* variants co-segregating with cholesteatoma in five out of ten families and the UKBB WES study detected enrichment of axonemal dyneins, I recommend further study of rare dynein variants within affected families. The known link between dynein variants and PCD, chronic ear disease and cholesteatoma is also convincing. In my study, the dynein variants were rare and therefore could not explain all cases, but they may be involved in a subset.

6.5.4 Further investigation of epilepsy link

The nature of the relationship between epilepsy and cholesteatoma in this study is unknown. Epilepsy may arise as a complication of intracranial infection or may be a risk factor itself. Another possibility is that anti-epileptic drugs may have an influence on cholesteatoma risk, as is seen with bisphosphonates given for osteoporosis. Cholesteatoma and other middle ear conditions are not amongst known side effects of anti-epileptic drugs, which mainly include nausea, headache, dizziness and cognitive effects. However, there are many anti-epileptic drugs and rarer adverse effects can involve various organs including the skin, causing conditions such as acne, rash, exfoliation and Stevens-Johnson syndrome[325]. While there is currently no established link between epilepsy or its treatments and raised risk of cholesteatoma, biobanks containing information about prescription medications could hold further insight into any potential relationship.

6.6 Impact

Little is known about what causes cholesteatoma. Although there are several convincing theories for formation, these do not explain all cases and many cases show features of multiple theories of formation. Nor is it known why these processes occur in some people and not others, despite similar conditions such as chronic inflammation and tympanic retraction. Until recent observations of family clustering, cholesteatoma has been considered non-genetic. This thesis supports a genetic role and suggests several pathways and processes which may be involved in cholesteatoma biology, which may be used to enhance future family-based studies. It also provides evidence that cholesteatoma risk is complex: it is not based on a single gene or variant, may be highly polygenic and/or heterogeneous, and the overall genetic effect within the general population is not large.

7 Conclusion

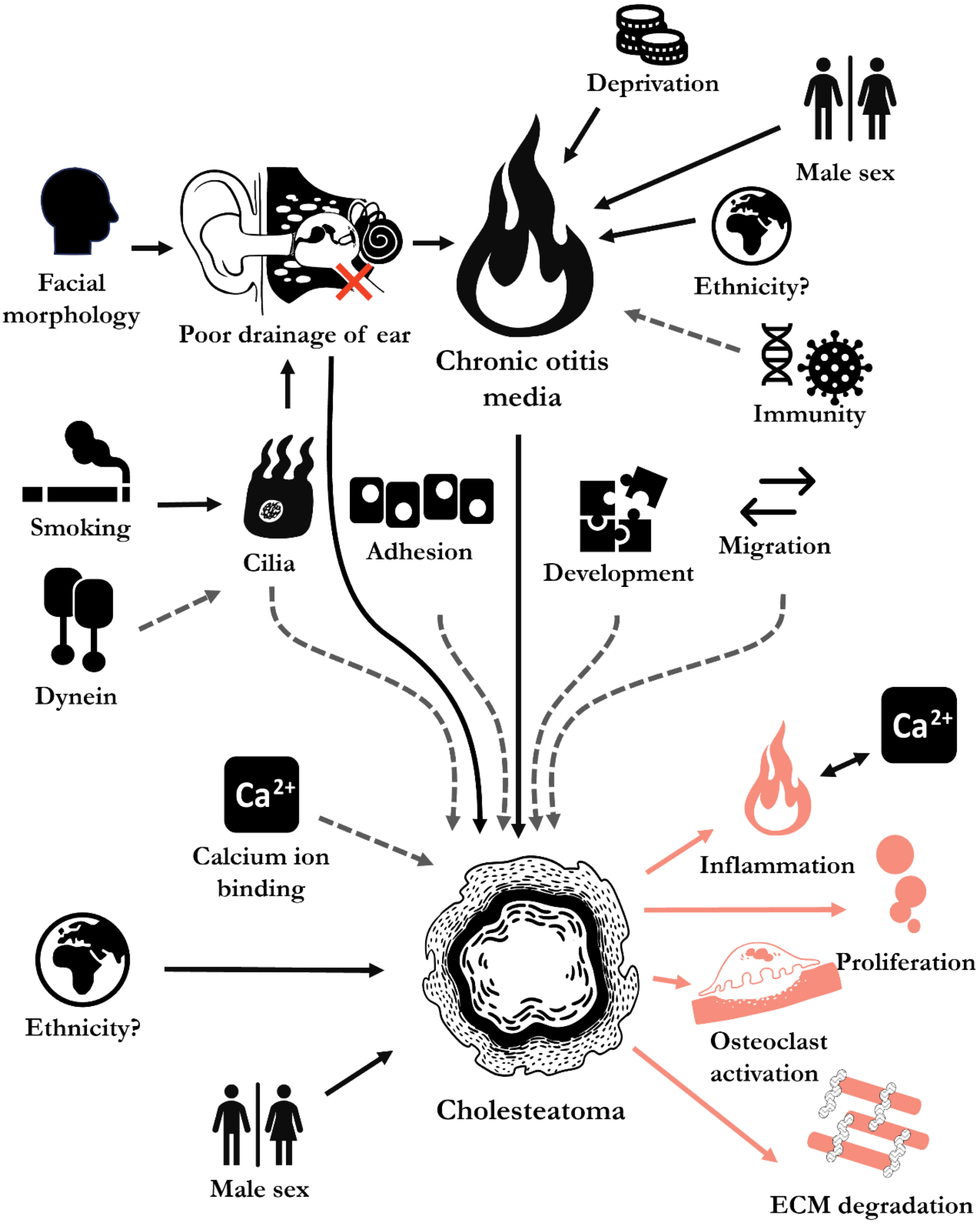
Cholesteatoma is a complex disease with heterogeneous presentation and multifactorial environmental and genetic risk. Increased susceptibility may arise through a combination of factors contributing to increased risk of otitis media, and both separate and overlapping risk factors contributing to risk of development of cholesteatoma. Genetic risk factors may be heterogeneous, possibly with both familial (high-penetrance, less polygenic) and sporadic (low-penetrance, highly polygenic, highly environmental) forms.

A major finding of his study was enrichment of pathways within UKBB whole exome single variant data with the following themes: cell adhesion, actin cytoskeletal organisation, tissue development, and calcium binding. These were supported by enrichment within the Finnish biobank FinnGen. Another important finding was enrichment of dynein binding processes within UKBB whole exome single variant data due to rare *DNAH* and *DNAI* variants; this was not replicated in the microarray-based FinnGen data nor UKBB microarray data due to rarity of these variants. As our previous whole exome study also identified rare *DNAH* variants within affected families, this supports a role for dynein function and therefore ciliopathy in cholesteatoma.

Variants in these pathways may contribute to cholesteatoma risk directly or indirectly via risk of otitis media (**Figure 36**). Ciliary impairment may raise susceptibility to OM and prevent proper clearage of keratin debris from the middle ear. Altered cell motility due to adhesion and cytoskeletal variants may also contribute to dysfunctional epithelial turnover on the tympanic membrane or promote invasiveness.

Poor Eustachian tube function has a known role in risk of chronic ear disease and chromosomal abnormalities have previously been shown to raise risk of otitis media and cholesteatoma, but it is unclear if enriched tissue development processes discovered in this analysis are associated with differences in Eustachian tube morphology. A role for calcium in cholesteatoma seems well supported by genetic and gene expression studies. While immune genes are implicated in OM susceptibility, no immune involvement in cholesteatoma was detected in this thesis.

Figure 36. Diagram showing cholesteatoma risk factors (black) and pathological features (pink).



Likewise, extracellular matrix dysfunction is implicated in cholesteatoma from gene expression studies. My semi-systematic review of global gene expression studies identified upregulated proteases and broad dysregulation of extracellular matrix structural components including consistent downregulated *TNXB* and *COCH*. However, my genetic analysis did not implicate ECM function directly.

Overall, this study supports the existence of a complex genetic component to cholesteatoma disease risk. Whether individuals have some combination of the identified genetic pathways or defects in just one is not known. Although very few genetic studies of cholesteatoma have been performed, evidence from this and our previous whole exome study support the dynein family of ciliary motor proteins as targets for future research.

8 Appendix: Ancestry Estimation

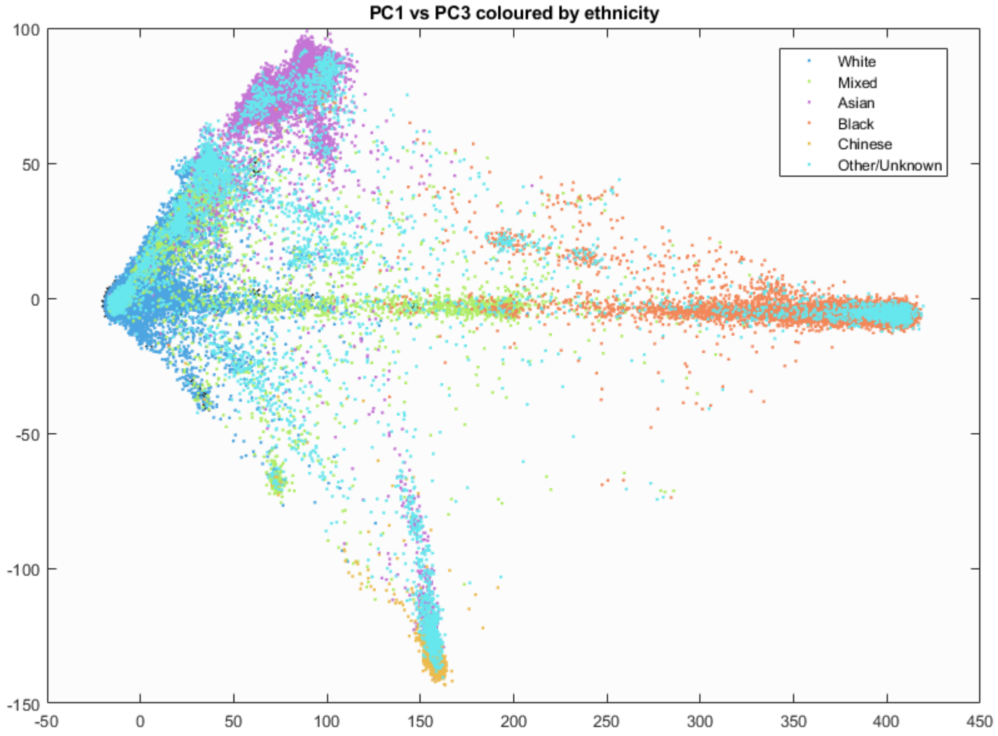
To avoid confounding due to population structure, GWAS usually use participants of a homogenous genetic background. While it is possible to perform multi-ancestry GWAS, large numbers of each ancestry are required. The UKBB cohort includes individuals of a wide array of backgrounds but is majority White British. The metadata contains three methods for identification of genetic background: ethnicity, as stated by participants at recruitment; genetic principal components, from which genetic structure can be determined; and an indicator for genetically determined White British participants, determined by Bycroft *et al.* (2018)c using a combination of both.

Use of genetically-determined ancestry may be more appropriate for genetic study, as the intent is to control for genetic factors. However, this may not be appropriate for epidemiological study where sociological factors are likely to be important. I investigated ancestry based on genetic principal components and compared it to ethnicity to determine whether ancestry or ethnicity provided better case-control matching for epidemiological and genetic study, and whether ethnicity would be a suitable proxy for genetic ancestry in genetic study.

Ancestry estimation from ethnicity and k-means clustering

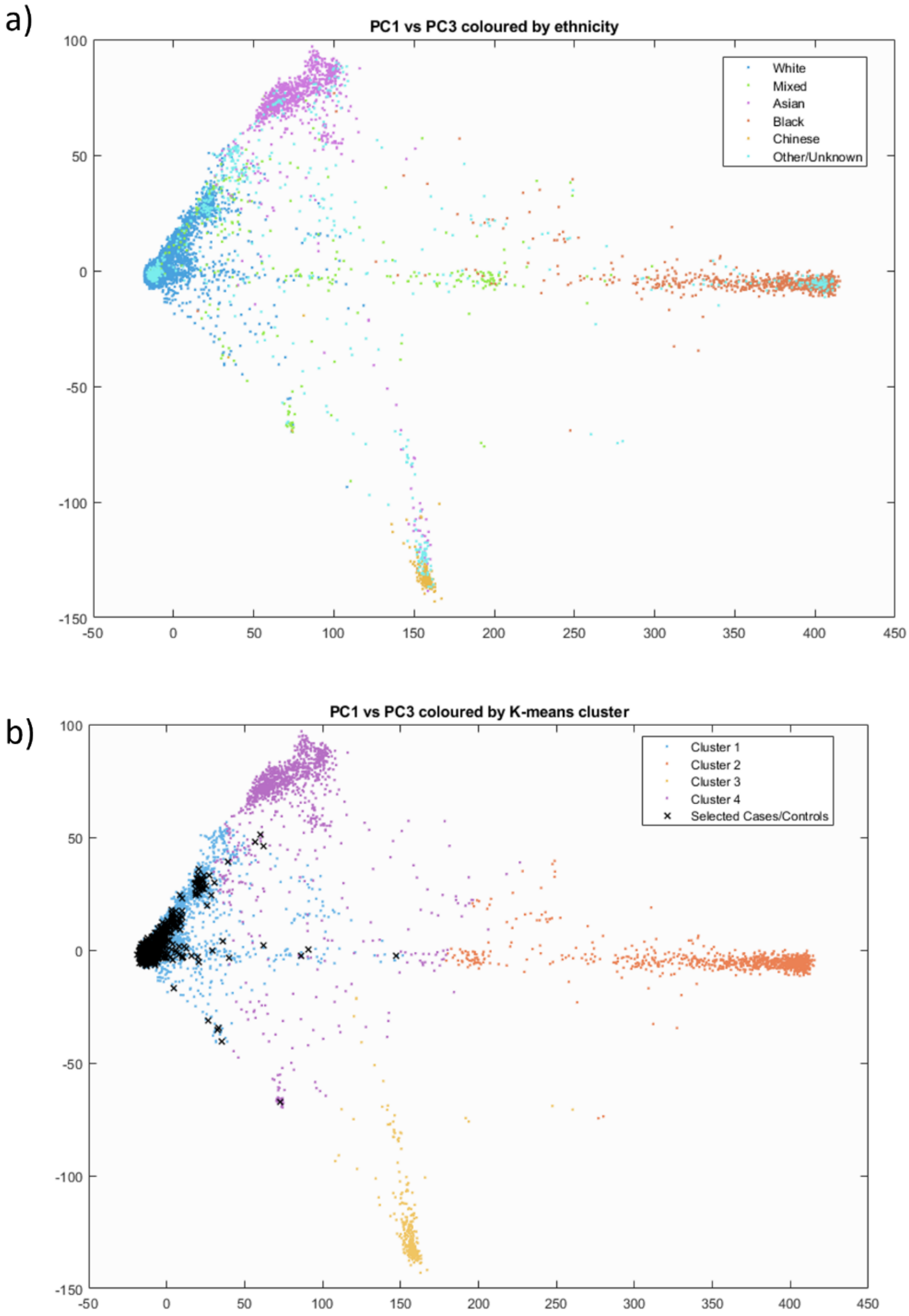
K-means is a method for sorting multi-dimensional data into groups by calculating the position of the centroid of each cluster. The centroid is the coordinate at the centre of any given cluster. Individuals are clustered as to minimise the average distance to the centroid. This was performed on the genetic principal components supplied by UKBB. Genetics of different populations do not form self-contained, discrete clusters as populations are rarely entirely isolated. They are similar to neighbouring populations and may have significant admixture with other populations leading to the PCs forming continuous gradients between more densely clustered regions (**Figure 37**). The clustering algorithm must create artificial cutoffs in order to assign individuals to a given cluster.

Figure 37. Genetic principal components coloured by ethnicity given in UKBB questionnaire



For this reason, K-means clustering fails when all individuals are supplied at once. I randomly sampled 5000 individuals at a time to assign them to 4 clusters with k-means using the first 10 PCs. I chose 4 clusters to represent the 4 major continental ethnicities present In UKBB – European, African, South Asian and East Asian. Mixed and other ethnicities were assigned if the individual was above a threshold distance from their centroid (**Figure 38b**).

Figure 38. Genetic principal components of a) a random sample of individuals, coloured by ethnicity and b) the same individuals coloured by assigned cluster, showing selected cases and controls in black.



Use of ethnicity over ancestry in genetic studies

In the GWAS section of this thesis, I use white ethnicity as a proxy for European ancestry. To check that this was a suitable approximation, I compared the selected cases and controls to the assigned ancestry by k-means. The points fell within the broader European cluster, though one or two could also have been placed in 'other', I decided this was a suitable approximation (**Figure 38**). This was beneficial to using Bycroft *et al.*'s definition of White British[190] as it maintained a larger number of cases and allowed direct comparison to epidemiological analysis where ethnicity was a more appropriate variate. Also, matching performed better when ethnicity was used rather than ancestry (see *Assessment of matching performance*). Finally, approximating ancestry based on genetic principal components also involves assigning somewhat arbitrary cutoffs as the components form gradients between more densely clustered regions.

Glossary

Abbreviations

ALS	Amyotrophic lateral sclerosis
AOR	Adjusted odds ratio
CF	Cystic fibrosis
COM	Chronic otitis media
DEG	Differentially expressed gene
ECM	Extracellular matrix
ENSG	Ensembl gene ID
GO	Gene ontology (GO:BP biological process; GO CC cellular compartment; GO MF molecular function)
GoC	Genetics of cholesteatoma
GRR	Genetic risk ratio
GSEA	Gene set enrichment analysis
GWAS	Genome-wide association study
HPV	Human papillomavirus
HR	Hazard ratio
MAC	Minor allele count
MAF	Minor allele frequency
MLE	Maximum likelihood estimate
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
OM	Otitis media
OR	Odds ratio
PCD	Primary ciliary dyskinesia
PheWAS	Phenome-wide association study.
PRS	Polygenic risk score
rsID	Reference SNP cluster ID
SNP	Single nucleotide polymorphism
SPA	Saddle point approximation
UKBB	UK BioBank
VCF	Variant call file
WES	Whole exome sequencing

WGS	Whole genome sequencing

Definitions

Autosomal: (A variant carried on) non-sex chromosomes

Connective tissue: A type of tissue with mostly structural function, consisting mostly of elastic and collagen fibres.

Coverage: the percent of target bases represented by a minimum read depth

Decision tree: A structured series of conditions used to split data into classes, consisting of nodes which represent the condition being tested and branches representing the outcome. Terminal nodes represent the class of data.

Dominant: Genetic mechanism where one copy of a variant gene is sufficient to display phenotype

Epithelial tissue / epithelium: A type of tissue made up of a thin layer of cells. Makes up the external and internal surfaces of the body. One of the four types of animal tissue.

Genetic architecture: The number, location and effect size of variants contributing to a disease, as well as the genetic heterogeneity

Genome-wide association study: A type of study where association tests are performed on variants across the genome.

Genotype: The combination of genetic variants present in an individual

Genotyping array: A microarray used to detect specific genetic variants using a number of 'probes' (sequences complementary to the target sequences to be identified).

Haplotype: A stretch of DNA or set of variants on a single (haploid) chromosome which tends to be the same amongst members of a population because it is inherited together from a parent.

Indel: a type of genetic variant where a single or short stretch of bases have been inserted or deleted.

Labyrinth: The bony inner ear which includes the vestibule, semicircular canals and the cochlea. Involved in sensorineural hearing.

Microarray: A chip that assays a large number of biological entities such as DNA or RNA fragments. See genotyping array.

Minor allele count: The number of the less common of a pair of possible alleles at a site.

Minor allele frequency: The frequency of the less common of a pair of possible alleles at a site.

Mucosa: A type of epithelial tissue consisting of simple cuboidal cells.

Mutation: Any change in DNA from the wild type. While effectively synonymous with a variant, this word tends to be reserved for rare variants.

Ossicular chain: The set of three small, delicate bones of the middle ear which transmit vibrations from the ear drum to the inner ear, involved in conductive hearing. The ossicular chain includes the incus, malleus and stapes.

Penetrance: The proportion of individuals with a trait-associated genotype who have the trait

Phenome-wide association study: a type of study where variants are searched for associations with a wide array of phenotypes.

Phenotype: The observed traits of an individual

Polymorphism: A naturally occurring variant which is common within a population. A polymorphic site is one which has multiple possible variants.

Read: A small fragment of sequenced DNA usually <100 base pairs long. When DNA is sequenced, it is done so in numerous overlapping reads.

Read depth: the number of overlapping reads representing a particular base/position in the gene sequence.

Recessive: Genetic mechanism where two copies of a variant gene must be carried to display phenotype

Sex-linked: Where a causal gene for a trait is carried on a sex chromosome

Single nucleotide polymorphism: A type of variant where a single base is substituted for another. The terms single nucleotide polymorphism (SNP) and single nucleotide variant (SNV) refer to the same type of variation, though it is more appropriate to call the variant a polymorphism if variation at that site is common and a variant if it is rare.

Single nucleotide polymorphism: A type of variant where only one base pair is changed.

Stratified squamous epithelium: A type of epithelial tissue consisting of several layers of epithelial cells arranged on a basal membrane. Keratinising stratified squamous epithelium produces keratin and makes up the skin.

Tympanic membrane: The ear drum, a tight membrane continuous with the wall of the ear canal separating it from the middle ear. It is responsible for conducting vibrations from sound waves to the ossicular chain. The lateral surface consists of keratinising stratified squamous epithelium, supported by a connective tissue layer. The interior surface is mucosa.

Variant: Any change in a genetic sequence compared to a reference genome. This may be in the form of a change to a single base pair or the insertion or deletion of many bases. A rare variant may be called a mutation while a common variant may be considered a polymorphism.

References

1. Castle JT. Cholesteatoma Pearls: Practical Points and Update. *Head and Neck Pathol*. 2018;12(3):419-429. doi:10.1007/s12105-018-0915-5

2. Olszewska E, Wagner M, Bernal-Sprekelsen M, et al. Etiopathogenesis of cholesteatoma. *European Archives of Oto-Rhino-Laryngology*. 2004;261(1):6-24. doi:10.1007/s00405-003-0623-x

3. Reddy CE, Goodyear P, Ghosh S, Lesser T. Intratympanic membrane cholesteatoma: a rare incidental finding. *Eur Arch Otorhinolaryngol*. 2006;263(12):1061-1064. doi:10.1007/s00405-006-0156-1

4. Persaud RAP, Hajioff D, Thevasagayam MS, Wareing MJ, Wright A. Keratosis obturans and external ear canal cholesteatoma: how and why we should distinguish between these conditions. *Clin Otolaryngol*. 2004;29(6):577-581. doi:10.1111/j.1365-2273.2004.00898.x

5. Kazahaya K, Potsic WP. Congenital cholesteatoma: *Current Opinion in Otolaryngology & Head and Neck Surgery*. 2004;12(5):398-403. doi:10.1097/01.moo.0000136875.41630.d6

6. Kemppainen HO, Puhakka HJ, Laippala PJ, Sipila MM, Manninen MP, Karma PH. Epidemiology and aetiology of middle ear cholesteatoma. *Acta Otolaryngol*. 1999;119(5):568-572. doi:10.1080/00016489950180801

7. Djurhuus BD, Faber CE, Skytthe A. Decreasing incidence rate for surgically treated middle ear cholesteatoma in Denmark 1977-2007. *Danish medical bulletin*. 2010;57:A4186.

8. Im GJ, Do Han K, Park KH, et al. Rate of chronic otitis media operations and cholesteatoma surgeries in South Korea: a nationwide population-based study (2006–2018). *Sci Rep*. 2020;10(1):11356. doi:10.1038/s41598-020-67799-5

9. Djurhuus BD, Christensen K, Skytthe A, Faber CE. The impact of ventilation tubes in otitis media on the risk of cholesteatoma on a national level. *International Journal of Pediatric Otorhinolaryngology*. 2015;79(4):605-609. doi:10.1016/j.ijporl.2015.02.005

10. Shibata S, Murakami K, Umeno Y, Komune S. Epidemiological study of cholesteatoma in Fukuoka City. *Journal of Laryngology and Otology*. 2015;129(S2):S6-S11. doi:10.1017/S002221511400231X

11. Britze A, Moller ML, Ovesen T. Incidence, 10-year recidivism rate and prognostic factors for cholesteatoma. *Journal of Laryngology and Otology*. 2017;131(4):319-328. doi:10.1017/S0022215117000299

12. Padgham N, Mills R, Christmas H. Has the increasing use of grommets influenced the frequency of surgery for cholesteatoma? *The Journal of Laryngology & Otology*. 1989;103(11):1034-1035. doi:10.1017/S0022215100110898

13. Nevoux J, Lenoir M, Roger G, Denoyelle F, Ducou Le Pointe H, Garabédian EN. Childhood cholesteatoma. *European Annals of Otorhinolaryngology, Head and Neck Diseases*. 2010;127(4):143-150. doi:10.1016/j.anorl.2010.07.001

14. Kuo CL, Shiao AS, Yung M, et al. Updates and knowledge gaps in cholesteatoma research. *Biomed Res Int*. 2015;2015:854024. doi:10.1155/2015/854024

15. Ratnesar P. Chronic ear disease along the coasts of Labrador and Northern Newfoundland. *J Otolaryngol*. 1976;5(2):122-130.

16. Thornton D, Martin TPC, Amin P, Haque S, Wilson S, Smith MCF. Chronic suppurative otitis media in Nepal: ethnicity does not determine whether disease is associated with cholesteatoma or not. *J Laryngol Otol*. 2011;125(1):22-26. doi:10.1017/S0022215110001878

17. Kapania EM, Stern BM, Sharma G. Ciliary Dysfunction. In: *StatPearls*. ; 2021.

18. Jennings BA, Prinsley P, Philpott C, Willis G, Bhutta MF. The genetics of cholesteatoma. A systematic review using narrative synthesis. *Clinical Otolaryngology*. 2018;43(1):55-67. doi:10.1111/coa.12900

19. Schilder AGM, Chonmaitree T, Cripps AW, et al. Otitis media. *Nat Rev Dis Primers*. 2016;2(1):16063. doi:10.1038/nrdp.2016.63

20. Kuo CL. Etiopathogenesis of acquired cholesteatoma: prominent theories and recent advances in biomolecular research. *Laryngoscope*. 2015;125(1):234-240. doi:10.1002/lary.24890

21. Louw L. Acquired cholesteatoma pathogenesis: stepwise explanations. *J Laryngol Otol*. 2010;124(6):587-593. doi:10.1017/S0022215109992763

22. Bayoumy AB, Veugen CCAFM, Rijssen LB, Yung M, Bok JWM. The Natural Course of Tympanic Membrane Retractions in the Posterosuperior Quadrant of Pars Tensa: A Watchful Waiting Policy. *Otology & Neurotology*. 2021;42(1):e50-e59. doi:10.1097/MAO.0000000000002834

23. Cutajar J, Nowghani M, Tulsidas-Mahtani B, Hamilton J. The Natural History of Asymptomatic Deep Pars Tensa Retraction. *Int Adv Otol*. Published online May 17, 2018:10-14. doi:10.5152/iao.2018.5234

24. Clark MP, Pretorius PM, Beaumont D, Milford CA. Congenital cholesteatoma of occipital bone or intradiploic epidermoid cyst? One and the same disease. *J Laryngol Otol*. 2009;123(6):673-675. doi:10.1017/S0022215108003083

25. Yamamoto-Fukuda T, Takahashi H, Koji T. Animal Models of Middle Ear Cholesteatoma. *Journal of Biomedicine and Biotechnology*. 2011;2011:1-11. doi:10.1155/2011/394241

26. Sudhoff H, Tos M. Pathogenesis of attic cholesteatoma: clinical and immunohistochemical support for combination of retraction theory and proliferation theory. *Am J Otol*. 2000;21(6):786-792.

27. Sasaki T, Iino Y, Masuda M, Ito M. Intratympanic membrane cholesteatoma: A case report. *Acta Oto-Laryngologica Case Reports*. 2016;1(1):75-78. doi:10.1080/23772484.2016.1241662

28. Hoang VT, Trinh CT, Nguyen CH, Chansomphou V, Chansomphou V, Tran TTT. Overview of epidermoid cyst. *Eur J Radiol Open*. 2019;6:291-301. doi:10.1016/j.ejro.2019.08.003

29. Zhang X, Yin M, Zhang LJ. Keratin 6, 16 and 17-Critical Barrier Alarmin Molecules in Skin Wounds and Psoriasis. *Cells*. 2019;8(8). doi:10.3390/cells8080807

30. Preciado DA. Biology of cholesteatoma: Special considerations in pediatric patients. *International Journal of Pediatric Otorhinolaryngology*. 2012;76(3):319-321. doi:10.1016/j.ijporl.2011.12.014

31. Fontes Lima A, Carvalho Moreira F, Sousa Menezes A, et al. Is pediatric cholesteatoma more aggressive in children than in adults? A comparative study using the EAONO/JOS classification. *International Journal of Pediatric Otorhinolaryngology*. 2020;138:110170. doi:10.1016/j.ijporl.2020.110170

32. De Carvalho Dornelles C, Da Costa SS, Meurer L, Rosito LPS, Da Silva AR, Alves SL. Comparison of acquired cholesteatoma between pediatric and adult patients. *Eur Arch Otorhinolaryngol*. 2009;266(10):1553-1561. doi:10.1007/s00405-009-0957-0

33. Lynrah ZA, Bakshi J, Panda NK, Khandelwal NK. Aggressiveness of Pediatric Cholesteatoma. Do We Have an Evidence? *Indian J Otolaryngol Head Neck Surg*. 2013;65(3):264-268. doi:10.1007/s12070-012-0548-z

34. Frantz C, Stewart KM, Weaver VM. The extracellular matrix at a glance. *J Cell Sci*. 2010;123(Pt 24):4195-4200. doi:10.1242/jcs.023820

35. Alhajj M, Bansal P, Goyal A. Physiology, Granulation Tissue. In: *StatPearls*. StatPearls Publishing; 2021.

36. Jovanovic I, Zivkovic M, Djuric T, Stojkovic L, Jesic S, Stankovic A. Perimatrix of middle ear cholesteatoma: A granulation tissue with a specific transcriptomic signature. *Laryngoscope*. 2020;130(4):E220-E227. doi:10.1002/lary.28084

37. Tang P, Xiong Q, Ge W, Zhang L. The role of MicroRNAs in osteoclasts and osteoporosis. *RNA Biology*. 2014;11(11):1355-1363. doi:10.1080/15476286.2014.996462

38. Boyle WJ, Simonet WS, Lacey DL. Osteoclast differentiation and activation. *Nature*. 2003;423:337-342.

39. Imai R, Sato T, Iwamoto Y, et al. Osteoclasts Modulate Bone Erosion in Cholesteatoma via RANKL Signaling. *Journal of the Association for Research in Otolaryngology*. 2019;20(5):449-459. doi:10.1007/s10162-019-00727-1

40. Cinamon U, Kronenberg J, Benayahu D. Structural changes and protein expression in the mastoid bone adjacent to cholesteatoma. *Laryngoscope*. 2000;110(7):1198-1203. doi:10.1097/00005537-200007000-00025

41. Koizumi H, Suzuki H, Kawaguchi R, et al. Presence of osteoclasts in middle ear cholesteatoma: a study of undecalcified bone sections. *Acta Oto-Laryngologica*. 2017;137(2):127-130. doi:10.1080/00016489.2016.1222549

42. Thorsteinsson AL, Vestergaard P, Eiken P. External auditory canal and middle ear cholesteatoma and osteonecrosis in bisphosphonate-treated osteoporosis patients: A Danish national register-based cohort study and literature review. *Osteoporosis International*. 2014;25(7):1937-1944. doi:10.1007/s00198-014-2684-7

43. Minami SB, Mutai H, Suzuki T, et al. Microbiomes of the normal middle ear and ears with chronic otitis media. *Laryngoscope*. 2017;127(10):E371-E377. doi:10.1002/lary.26579

44. Samarrai R, Frank S, Lum A, Woodis K, Weinstock G, Roberts D. Defining the microbiome of the head and neck: A contemporary review. *American Journal of Otolaryngology - Head and Neck Medicine and Surgery*. 2022;43(1). doi:10.1016/j.amjoto.2021.103224

45. Weiss JP, Antonelli PJ, Dirain CO. Microbiome Analysis of Cholesteatoma by Gene Sequencing. *Otology and Neurotology*. 2019;40(9):1186-1193. doi:10.1097/MAO.0000000000002355

46. Dunne EF, Park IU. HPV and HPV-associated diseases. *Infectious Disease Clinics of North America*. 2013;27(4):765-778. doi:10.1016/j.idc.2013.09.001

47. Mittal S, Ravichandran J, Shanmugasundaram S, Kanagarajan A. Verrucous Cyst with Cutaneous Horn. *Indian Dermatol Online J*. 2024;15(3):518-520. doi:10.4103/idoj.idoj_314_23

48. Chao WY, Chang SJ, Jin YT. Detection of human papillomavirus in cholesteatomas. *European Archives of Oto-Rhino-Laryngology*. 2000;257(3):0120. doi:10.1007/s004050050206

49. Franz P, Teschendorf M, Wohlschläger J, Fischer M. Prevalence of human papillomavirus DNA in cholesteatomas. *Orl*. 2007;69(4):251-255. doi:10.1159/000101547

50. Viana RMM, Souza JP, Jorge DMM, et al. Detection of respiratory viruses in primary cholesteatoma tissues. *Journal of Medical Virology*. 2021;93(11):6132-6139. doi:10.1002/jmv.27107

51. Skoulakis A, Lachanas V, Florou Z, et al. High-Prevalence of Various High-Risk Sub-types of Human Papilloma Virus in Patients with Acquired Cholesteatoma in Greece. *Acta Scientific Microbiology*. 2018;1(4):3-5. doi:10.31080/asmi.2018.01.0031

52. Rydzewski B, Goździcka-Józefiak A, Sokalski J, Matusiak M, Durzyński Ł. Identification of human papilloma viruses (HPV) in inflammatory states and ear neoplasms. *Otolaryngologia Polska*. 2007;61(2):137-141. doi:10.1016/s0030-6657(07)70401-8

53. Schürmann M, Goon P, Sudhoff H. Review of potential medical treatments for middle ear cholesteatoma. *Cell Commun Signal*. 2022;20(1):148. doi:10.1186/s12964-022-00953-w

54. Prinsley P. Familial cholesteatoma in East Anglia, UK. *Journal of Laryngology and Otology*. 2009;123(3):294-297. doi:10.1017/S0022215108002673

55. Landegger LD, Cohen MS. Congenital cholesteatoma in siblings. *J Laryngol Otol*. 2013;127(11):1143-1144. doi:10.1017/S0022215113002284

56. Homoe P, Rosborg J. Family cluster of cholesteatoma. *J Laryngol Otol*. 2007;121(1):65-67. doi:10.1017/S0022215106004117

57. Lipkin AF, Coker NJ, Jenkins HA. Hereditary Congenital Cholesteatoma: A Variant of Branchio-oto Dysplasia. *Archives of Otolaryngology - Head and Neck Surgery*. 1986;112(10):1097-1100. doi:10.1001/archotol.1986.03780100085014

58. Al Balushi T, Naik JZ, Al Khabori M. Congenital cholesteatoma in identical twins. *J Laryngol Otol*. 2013;127(1):67-69. doi:10.1017/S0022215112002757

59. Pinzas L, Glaun M, Liu YCC. Congenital cholesteatoma in identical twins. *International Journal of Pediatric Otorhinolaryngology*. 2022;162:111330. doi:10.1016/j.ijporl.2022.111330

60. Brar S, Wolf DM, Faoury M, Barwell J, Saggar A, Daya H. Monozygotic twins and cholesteatomas: nature or nuture? *Eur Arch Otorhinolaryngol*. 2023;280(12):5649-5654. doi:10.1007/s00405-023-08239-8

61. Collins R, Ta NH, Jennings BA, et al. Cholesteatoma and family history: An international survey. *Clinical Otolaryngology*. 2020;45(4):500-505. doi:10.1111/coa.13544

62. Podoshin L, Margalit A, Fradis M, Tamir A, Ben-David Y, Epstein L. Cholesteatoma: An Epidemiologic Study among Members of Kibbutzim in Northern Israel. *Ann Otol Rhinol Laryngol*. 1986;95(4):365-368. doi:10.1177/000348948609500408

63. Bonnard Å, Engmér Berglin C, Wincent J, et al. The Risk of Cholesteatoma in Individuals With First-degree Relatives Surgically Treated for the Disease. *JAMA Otolaryngol Head Neck Surg*. 2023;149(5):390. doi:10.1001/jamaoto.2023.0048

64. Djurhuus BD, Skytthe A, Faber CE, Christensen K. Cholesteatoma risk in 8,593 orofacial cleft cases and 6,989 siblings: A nationwide study. *The Laryngoscope*. 2015;125(5):1225-1229. doi:10.1002/lary.25022

65. Shaoul R, Papsin B, Cutz E, Durie P. Congenital cholesteatoma in a child carrying a gene mutation for adenomatous polyposis coli. *J Pediatr Gastroenterol Nutr*. 1999;28(1):100-103. doi:10.1097/00005176-199901000-00023

66. James AL, Chadha NK, Papsin BC, Stockley TL. Pediatric cholesteatoma and variants in the gene encoding connexin 26. *Laryngoscope*. 2010;120(1):183-187. doi:10.1002/lary.20649

67. Smith RJH, Jones MK. Nonsyndromic Hearing Loss and Deafness, DFNB1. In: *GeneReviews [Internet]*. ; 1998:1-14.

68. Lee NK, Cass SP, Gubbels SP, et al. Novel candidate genes for cholesteatoma in chronic otitis media. *Front Genet*. 2022;13:1033965. doi:10.3389/fgene.2022.1033965

69. Chen X, Qin Z. Post-Transcriptional Regulation by MicroRNA-21 and *let-7a* MicroRNA in Paediatric Cholesteatoma. *J Int Med Res*. 2011;39(6):2110-2118. doi:10.1177/147323001103900607

70. Zang J, Hui L, Yang N, Yang B, Jiang X. Downregulation of MiR-203a Disinhibits Bmi1 and Promotes Growth and Proliferation of Keratinocytes in Cholesteatoma. *Int J Med Sci*. 2018;15(5):447-455. doi:10.7150/ijms.22410

71. Xie S, Liu X, Pan Z, et al. Microarray Analysis of Differentially-expressed MicroRNAs in Acquired Middle Ear Cholesteatoma. *Int J Med Sci*. 2018;15(13):1547-1554. doi:10.7150/ijms.26329

72. Liu X, Bulgakov OV, Darrow KN, et al. Usherin is required for maintenance of retinal photoreceptors and normal development of cochlear hair cells. *Proc Natl Acad Sci USA*. 2007;104(11):4413-4418. doi:10.1073/pnas.0610950104

73. Gao J, Tang Q, Zhu X, et al. Long noncoding RNAs show differential expression profiles and display ceRNA potential in cholesteatoma pathogenesis. *Oncol Rep*. Published online March 16, 2018. doi:10.3892/or.2018.6320

74. Gao J, Tang Q, Xue R, et al. Comprehensive circular RNA expression profiling with associated ceRNA network reveals their therapeutic potential in cholesteatoma. *Oncol Rep*. Published online February 12, 2020. doi:10.3892/or.2020.7501

75. Jovanovic I, Zivkovic M, Jesic S, Stankovic A. Non-coding RNA and cholesteatoma. *Laryngoscope Investig Oto*. 2022;7(1):60-66. doi:10.1002/lio2.728

76. Yanez-Siller JC, Wentland C, Bowers K, Litofsky NS, Rivera AL. Squamous Cell Carcinoma of the Temporal Bone Arising from Cholesteatoma: A Case Report and Review of the Literature. *J Neurol Surg Rep*. 2022;83(01):e13-e18. doi:10.1055/s-0041-1741069

77. Klenke C, Janowski S, Borck D, et al. Identification of Novel Cholesteatoma-Related Gene Expression Signatures Using Full-Genome Microarrays. *PLoS One*. 2012;7(12). doi:ARTN e52718 10.1371/journal.pone.0052718

78. Britze A, Birkler RI, Gregersen N, Ovesen T, Palmfeldt J. Large-scale proteomics differentiates cholesteatoma from surrounding tissues and identifies novel proteins related to the pathogenesis. *PLoS One*. 2014;9(8):e104103. doi:10.1371/journal.pone.0104103

79. Albino AP, Kimmelman CP, Parisier SC. Cholesteatoma: a molecular and cellular puzzle. *Am J Otol*. 1998;19(1):7-19.

80. Satoh C, Yoshiura K ichiro, Mishima H, Yoshida H, Takahashi H, Kumai Y. Proto-oncogene mutations in middle ear cholesteatoma contribute to its pathogenesis. *BMC Med Genomics*. 2023;16(1):288. doi:10.1186/s12920-023-01640-6

81. Hill WG. Understanding and using quantitative genetic variation. *Phil Trans R Soc B*. 2010;365(1537):73-85. doi:10.1098/rstb.2009.0203

82. Yu E, Sharma S. Cystic Fibrosis. In: *StatPearls*. StatPearls Publishing; 2023. Accessed November 15, 2023. http://www.ncbi.nlm.nih.gov/books/NBK493206/

83. Walker FO. Huntingdon's disease. 2007;369(9557):218-228.

84. Pereira SVN, Ribeiro JD, Ribeiro AF, Bertuzzo CS, Marson FAL. Novel, rare and common pathogenic variants in the CFTR gene screened by high-throughput sequencing technology and predicted by in silico tools. *Sci Rep*. 2019;9(1):6234. doi:10.1038/s41598-019-42404-6

85. Sosnay PR, Siklosi KR, Van Goor F, et al. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat Genet*. 2013;45(10):1160-1167. doi:10.1038/ng.2745

86. David N Cooper, Michael Krawczak, Constantin Polychronakos, Chris Tyler-Smith, Hildegard Kehrer-Sawatzki. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human genetics*. 2013;132(10):1077-1130.

87. Shawky RM. Reduced penetrance in human inherited disease. *Egyptian Journal of Medical Human Genetics*. 2014;15(2):103-111. doi:10.1016/j.ejmhg.2014.01.003

88. Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, et al. Genome-wide association studies. *Nature Reviews Methods Primers*. 2021;1.

89. Ealy M, Smith RJH. The Genetics of otosclerosis. *Hearing Research*. 2010;266(1-2):70-74. doi:10.1016/j.heares.2009.07.002

90. Rämö JT, Kiiskinen T, Seist R, et al. Genome-wide screen of otosclerosis in population biobanks: 27 loci and shared associations with skeletal structure. *Nat Commun*. 2023;14(1):157. doi:10.1038/s41467-022-32936-3

91. Tcheandjieu C, Zhu X, Hilliard AT, et al. Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nat Med*. 2022;28(8):1679-1692. doi:10.1038/s41591-022-01891-3

92. Lander ES. Initial impact of the sequencing of the human genome. *Nature*. 2011;470(7333):187-197. doi:10.1038/nature09792

93. Verlouw JAM, Clemens E, De Vries JH, et al. A comparison of genotyping arrays. *Eur J Hum Genet*. 2021;29(11):1611-1624. doi:10.1038/s41431-021-00917-7

94. Freeman JL, Perry GH, Feuk L, et al. Copy number variation: New insights in genome diversity. *Genome Res*. 2006;16(8):949-961. doi:10.1101/gr.3677206

95. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. *Nat Rev Genet*. 2015;16(3):172-183. doi:10.1038/nrg3871

96. The Wellcome Trust Case Control Consortium, Management Committee, Burton PR, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661-678. doi:10.1038/nature05911

97. Venter JC, Adams MD, Myers EW, et al. The Sequence of the Human Genome. *Science*. 2001;291(5507):1304-1351. doi:10.1126/science.1058040

98. Farh KK, Marson A, Zhu J, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. 2015;518(7539):337-343. doi:10.1038/nature13835

99. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337(6099):1190-1195. doi:10.1126/science.1222794

100. Yong SY, Raben TG, Lello L, Hsu SDH. Genetic architecture of complex traits and disease risk predictors. *Sci Rep*. 2020;10(1):12055. doi:10.1038/s41598-020-68881-8

101. Frequently Asked Questions. Genome Reference Consortium. Accessed June 17, 2024. https://www.ncbi.nlm.nih.gov/grc/help/faq/#human-reference-genome-individuals

102. Rautiainen M, Nurk S, Walenz BP, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol*. 2023;41(10):1474-1482. doi:10.1038/s41587-023-01662-6

103. Wang T, Antonacci-Fulton L, Howe K, et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature*. 2022;604(7906):437-446. doi:10.1038/s41586-022-04601-8

104. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. Published online 2012. doi:10.48550/ARXIV.1207.3907

105. Auwera G van der, O'Connor BD. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. First edition. O'Reilly Media; 2020.

106. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet*. 2018;19(5):253-268. doi:10.1038/nrg.2017.116

107. Deltas C. Digenic inheritance and genetic modifiers. *Clinical Genetics*. 2018;93(3):429-438. doi:10.1111/cge.13150

108. Dixon P, Keeney E, Taylor JC, Wordsworth S, Martin RM. Can polygenic risk scores contribute to cost-effective cancer screening? A systematic review. *Genetics in Medicine*. 2022;24(8):1604-1617. doi:10.1016/j.gim.2022.04.020

109. Khoury MJ, Janssens ACJW, Ransohoff DF. How can polygenic inheritance be used in population screening for common diseases? *Genetics in Medicine*. 2013;15(6):437-443. doi:10.1038/gim.2012.182

110. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005;308(5720):385-389. doi:10.1126/science.1109557

111. Guo MH, Dauber A, Lippincott MF, Chan YM, Salem RM, Hirschhorn JN. Determinants of Power in Gene-Based Burden Testing for Monogenic Disorders. *Am J Hum Genet*. 2016;99(3):527-539. doi:10.1016/j.ajhg.2016.06.031

112. Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, Xihong Lin. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics*. 2011;89(1):82-93.

113. Jacklyn Hellwege, Jacob Keaton, Ayush Giri, Xiaoyi Gao, Digna R. Velez Edwards, Todd L. Edwards. Population Stratification in Genetic Association Studies. *Current Protocols in Human Genetics*. 2018;95(1):1.22-1.22.23.

114. Alkes L. Price, Noah A. Zaitlen, David Reich, Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature reviews genetics*. 2011;11(7):459-463.

115. Mbatchou J, Barnard L, Backman J, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*. 2021;53(7):1097-1103. doi:10.1038/s41588-021-00870-7

116. Zhou W, Bi W, Zhao Z, et al. SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. *Nat Genet*. 2022;54(10):1466-1469. doi:10.1038/s41588-022-01178-w

117. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50(9):1219-1224. doi:10.1038/s41588-018-0183-z

118. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med*. 2020;12(1):44. doi:10.1186/s13073-020-00742-5

119. Genomics Plc announces successful world-first pilot using improved genomic risk assessment in cardiovascular disease prevention in the NHS. Genomics Plc. July 11, 2022. Accessed August 21, 2024. https://www.genomicsplc.com/news/successful-world-first-pilot-using-improved-genomic-risk-assessment-in-cardiovascular-disease-prevention-in-the-nhs/#:~:text=The%20HEART%20study%20%28Healthcare%20Evaluation%20of%20Absolute%20Risk,GPs%20to%20help%20prevent%20and%20manage%20CVD%20%28QRISK%C2%A92%29.

120. The HEART Study and version 1.0. NHS Health Research Authority. August 17, 2021. Accessed August 21, 2024. https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/the-heart-study-and-version-10/

121. Riveros-Mckay F, Weale ME, Moore R, et al. Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. *Circ: Genomic and Precision Medicine*. 2021;14(2):e003304. doi:10.1161/CIRCGEN.120.003304

122. Kokol P, Kokol M, Zagoranski S. Machine learning on small size samples: A synthetic knowledge synthesis. *Science Progress*. 2022;105(1):00368504211029777. doi:10.1177/00368504211029777

123. Touw WG, Bayjanov JR, Overmars L, et al. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics*. 2013;14(3):315-326. doi:10.1093/bib/bbs034

124. Prinsley P, Jennings BA, Bhutta M, Swan D, Willis G, Philpott C. The genetics of cholesteatoma study. Loss-of-function variants in an affected family. *Clinical Otolaryngology*. 2019;44(5):826-830. doi:10.1111/coa.13365

125. Macias JD, Gerkin RD, Locke D, Macias MMP. Differential gene expression in cholesteatoma by DNA chip analysis. *Laryngoscope*. 2013;123(SUPPL.5):S1-21. doi:10.1002/lary.24176

126. Cardenas R, Prinsley P, Philpott C, et al. Whole exome sequencing study identifies candidate loss of function variants and locus heterogeneity in familial cholesteatoma. Monsanto RDC, ed. *PLoS ONE*. 2023;18(3):e0272174. doi:10.1371/journal.pone.0272174

127. Steiling K, Christenson S. Tools for genetics and genomics: Gene expression profiling. UpToDate. 2023. https://www.uptodate.com/contents/tools-for-genetics-and-genomics-gene-expression-profiling#H3544264

128. Schönermark M, Mester B, Kempf HG, Bläser J, Tschesche H, Lenarz T. Expression of Matrix-Metalloproteinases and their Inhibitors in Human Cholesteatomas. *Acta Oto-Laryngologica*. 1996;116(3):451-456. doi:10.3109/00016489609137872

129. Jeong J, Park C, Tae K, et al. Expression of RANKL and OPG in Middle Ear Cholesteatoma Tissue. *The Laryngoscope*. 2006;116(7):1180-1184. doi:10.1097/01.mlg.0000224345.59291.da

130. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. Published online March 29, 2021:n71. doi:10.1136/bmj.n71

131. Wells G, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa Hospital Research Institute. 2021. https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

132. The MathWorks Inc. MATLAB. Published online 2023. https://www.mathworks.com

133. Reimand J, Kull M, Peterson H, Hansen J, Vilo J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*. 2007;35(suppl_2):W193-W200. doi:10.1093/nar/gkm226

134. Yoshikawa M, Kojima H, Wada K, et al. Identification of Specific Gene Expression Profiles in Fibroblasts Derived From Middle Ear Cholesteatoma. *Arch Otolaryngol Head Neck Surg*. 2006;132(7):734. doi:10.1001/archotol.132.7.734

135. Zeng L, Xie L, Hu J, et al. Osteopontin-driven partial epithelial-mesenchymal transition governs the development of middle ear cholesteatoma. *Cell Cycle*. 2024;23(5):537-554. doi:10.1080/15384101.2024.2345481

136. Tokuriki M, Noda I, Saito T, et al. Gene expression analysis of human middle ear cholesteatoma using complementary DNA arrays. *Laryngoscope*. 2003;113(5):808-814. doi:Doi 10.1097/00005537-200305000-00008

137. Randall DR, Park PS, Chau JK. Identification of altered protein abundances in cholesteatoma matrix via mass spectrometry-based proteomic analysis. *J Otolaryngol Head Neck Surg*. 2015;44:50. doi:10.1186/s40463-015-0104-4

138. Gao M, Xiao H, Liang Y, et al. The Hyperproliferation Mechanism of Cholesteatoma Based on Proteomics: SNCA Promotes Autophagy-Mediated Cell Proliferation Through the PI3K/AKT/CyclinD1 Signaling Pathway. *Molecular & Cellular Proteomics*. 2023;22(9):100628. doi:10.1016/j.mcpro.2023.100628

139. Shimizu K, Kikuta J, Ohta Y, et al. Single-cell transcriptomics of human cholesteatoma identifies an activin A-producing osteoclastogenic fibroblast subset inducing bone destruction. *Nat Commun*. 2023;14(1):4417. doi:10.1038/s41467-023-40094-3

140. Baschal EE, Larson ED, Bootpetch Roberts TC, et al. Identification of Novel Genes and Biological Pathways That Overlap in Infectious and Nonallergic Diseases of the Upper and Lower Airways Using Network Analyses. *Front Genet*. 2019;10:1352. doi:10.3389/fgene.2019.01352

141. Gettins PG. Serpin structure, mechanism, and function. *Chemical Reviews*. 2002;102(12):4751-4804. doi:10.1021/cr010170+

142. Ho K, Huang HH, Hung K, et al. Cholesteatoma growth and proliferation: Relevance with serpin B3. *The Laryngoscope*. 2012;122(12):2818-2823. doi:10.1002/lary.23547

143. Ho KY, Huang CJ, Hung CC, et al. Autophagy Is Deficient and May be Negatively Regulated by SERPINB3 in Middle Ear Cholesteatoma. *Otology & Neurotology*. 2020;41(7):e881-e888. doi:10.1097/MAO.0000000000002690

144. Chang A, Schalkwijk J, Happle R, Van de Kerkhof PCM. Elastase-inhibiting activity in scaling skin disorders. *Acta Dermato-Venereologica*. 1990;70(2):147-151. doi:102340/0001555570147151

145. Lee JK, Chae SW, Cho JG, Lee HM, Hwang SJ, Jung HH. Expression of secretory leukocyte protease inhibitor in middle ear cholesteatoma. *Eur Arch Otorhinolaryngol*. 2006;263(12):1077-1081. doi:10.1007/s00405-006-0126-7

146. Kelly-Robinson GA, Reihill JA, Lundy FT, et al. The Serpin Superfamily and Their Role in the Regulation and Dysfunction of Serine Protease Activity in COPD and Other Chronic Lung Diseases. *IJMS*. 2021;22(12):6351. doi:10.3390/ijms22126351

147. Wang Z, Chen F, Zhai R, et al. Plasma neutrophil elastase and elafin imbalance is associated with Acute Respiratory Distress Syndrome (ARDS) development. *PLoS ONE*. 2009;4(2):e4380. doi:10.1371/journal.pone.0004380

148. Brix K, Dunkhorst A, Mayer K, Jordans S. Cysteine cathepsins: Cellular roadmap to different functions. *Biochimie*. 2008;90(2):194-207. doi:10.1016/j.biochi.2007.07.024

149. Pelc P, Vanmuylder N, Lefranc F, et al. Differential expression of S100 calcium-binding proteins in epidermoid cysts, branchial cysts, craniopharyngiomas and cholesteatomas. *Histopathology*. 2003;42(4):387-394. doi:10.1046/j.1365-2559.2003.01588.x

150. Kim KH, Cho JG, Song JJ, et al. Psoriasin (S100A7), an antimicrobial peptide, is increased in human middle ear cholesteatoma. *Acta Oto-Laryngologica*. 2009;129(10):1067-1071. doi:10.1080/00016480802455291

151. Eui Dong Son, Hyoung-June Kim, Kyu Han Kim, et al. S100A7 (psoriasin) inhibits human epidermal differentiation by enhanced IL-6 secretion through IκB/NF-κB signalling. *Experimental Dermatology*. 2016;25(8):636-641.

152. Hardy E, Fernandez-Patron C. Destroy to Rebuild: The Connection Between Bone Tissue Remodeling and Matrix Metalloproteinases. *Front Physiol*. 2020;11:47. doi:10.3389/fphys.2020.00047

153. Khuda F, Najmi Mohamad Anuar N, Baharin B, et al. A mini review on the associations of matrix metalloproteinases (MMPs) -1, -8, -13 with periodontal disease. *AIMS Molecular Science*. 2021;8(1):13-31. doi:10.3934/molsci.2021002

154. Mehana ESE, Khafaga AF, El-Blehi SS. The role of matrix metalloproteinases in osteoarthritis pathogenesis: An updated review. *Life Sciences*. 2019;234:116786. doi:10.1016/j.lfs.2019.116786

155. Palkó E, Póliska S, Sziklai I, Penyige A. Analysis of KRT1, KRT10, KRT19, TP53 and MMP9 expression in pediatric and adult cholesteatoma. Ahmad A, ed. *PLoS ONE*. 2018;13(7):e0200840. doi:10.1371/journal.pone.0200840

156. Yulius S, Asroel HA, Aboet A, Zaluchu F. Correlation of Matrix Metalloproteinase-9 (MMP-9) expression and bone destruction in Chronic Suppurative Otitis Media (CSOM) patients with cholesteatoma at Adam Malik General Hospital Medan-Indonesia. *Bali Med J*. 2018;7(1):195. doi:10.15562/bmj.v7i1.751

157. Olszewska E, Matulka M, Mroczko B, et al. Diagnostic value of matrix metalloproteinase 9 and tissue inhibitor of matrix metalloproteinases 1 in cholesteatoma. *Histology and Histopathology*. 2016;(31):307-315. doi:10.14670/HH-11-677

158. Rezende CEB, Souto RPD, Rapoport PB, Campos LD, Generato MB. Avaliação da expressão gênica de metaloproteinases de matriz e seus inibidores em colesteatomas por amplificação de ácidos nucleicos. *Braz j otorhinolaryngol*. 2012;78(3):116-121. doi:10.1590/s1808-86942012000300019

159. Dambergs K, Sumeraga G, Pilmane M. Complex Evaluation of Tissue Factors in Pediatric Cholesteatoma. *Children*. 2021;8(10):926. doi:10.3390/children8100926

160. Mehta D, Daudia >A., Birchall JP, Banerjee AR. The localization of matrix metalloproteinases-8 and -13 in cholesteatoma, deep-meatal and post-auricular skin: a comparative analysis. *Acta Oto-Laryngologica*. 2007;127(2):138-142. doi:10.1080/00016480600781807

161. Cabral-Pacheco GA, Garza-Veloz I, Castruita-De La Rosa C, et al. The Roles of Matrix Metalloproteinases and Their Inhibitors in Human Diseases. *IJMS*. 2020;21(24):9739. doi:10.3390/ijms21249739

162. Beauchemin N, Arabzadeh A. Carcinoembryonic antigen-related cell adhesion molecules (CEACAMs) in cancer progression and metastasis. *Cancer and Metastasis Reviews*. 2013;32(3-4):643-671. doi:10.1007/s10555-013-9444-6

163. Goel M, Sienkiewicz AE, Picciani R, Lee RK, Bhattacharya SK. Cochlin induced TREK-1 co-expression and annexin A2 secretion: Role in trabecular meshwork cell elongation and motility. *PLoS ONE*. 2011;6(8):e23070. doi:10.1371/journal.pone.0023070

164. Youn JY, Dunham WH, Hong SJ, et al. High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Molecular Cell*. 2018;69(3):517-532.e11. doi:10.1016/j.molcel.2017.12.020

165. Leventea E, Zhu Z, Fang X, et al. Ciliopathy genes are required for apical secretion of Cochlin, an otolith crystallization factor. *Proceedings of the National Academy of Sciences of the United States of America*. 2021;118(28). doi:10.1073/pnas.2102562118

166. Bristow J, Carey W, Egging D, Schalkwijk J. Tenascin-X, collagen, elastin, and the Ehlers–Danlos syndrome. *American J of Med Genetics Pt C*. 2005;139C(1):24-30. doi:10.1002/ajmg.c.30071

167. Valcourt U, Alcaraz LB, Exposito JY, Lethias C, Bartholin L. Tenascin-X: beyond the architectural function. *Cell Adhesion & Migration*. 2015;9(1-2):154-165. doi:10.4161/19336918.2014.994893

168. Kajitani N, Yamada T, Kawakami K, Matsumoto K ichi. TNX deficiency results in bone loss due to an increase in multinucleated osteoclasts. *Biochemical and Biophysical Research Communications*. 2019;512(4):659-664. doi:10.1016/j.bbrc.2019.03.134

169. Lund J, Søndergaard MT, Conover CA, Overgaard MT. Heparin-binding mechanism of the IGF2/IGF-binding protein 2 complex. *Journal of Molecular Endocrinology*. 2014;52(3):345-355. doi:10.1530/JME-13-0184

170. Olszewska E, Chodynicki S, Chyczewski L. Apoptosis in the pathogenesis of cholesteatoma in adults. *Eur Arch Otorhinolaryngol*. 2006;263(5):409-413. doi:10.1007/s00405-005-1026-y

171. Abhishek S, Palamadai Krishnan S. Epidermal Differentiation Complex: A Review on Its Epigenetic Regulation and Potential Drug Targets. *Cell J*. 2016;18(1). doi:10.22074/cellj.2016.3980

172. Mischke D, Korge BP, Marenholz I, Volz A, Ziegler A. Genes Encoding Structural Proteins of Epidermal Cornification and S100 Calcium-Binding Proteins Form a Gene Complex ("Epidermal Differentiation Complex") on Human Chromosome 1q21. *Journal of Investigative Dermatology*. 1996;106(5):989-992. doi:10.1111/1523-1747.ep12338501

173. Wu G, Wang D, Xiong F, et al. The emerging roles of CEACAM6 in human cancer (Review). *Int J Oncol*. 2024;64(3):27. doi:10.3892/ijo.2024.5615

174. Pak MG, Shin DH, Lee CH, Lee MK. Significance of EpCAM and TROP2 expression in non-small cell lung cancer. *World J Surg Onc*. 2012;10(1):53. doi:10.1186/1477-7819-10-53

175. Janiszewska M, Primi MC, Izard T. Cell adhesion in cancer: Beyond the migration of single cells. *Journal of Biological Chemistry*. 2020;295(8):2495-2505. doi:10.1074/jbc.REV119.007759

176. Schiopu A, Cotoi OS. S100A8 and S100A9: DAMPs at the Crossroads between Innate Immunity, Traditional Risk Factors, and Cardiovascular Disease. *Mediators of Inflammation*. 2013;2013:1-10. doi:10.1155/2013/828354

177. Patel S, Homaei A, El-Seedi HR, Akhtar N. Cathepsins: Proteases that are vital for survival but can also be fatal. *Biomedicine & Pharmacotherapy*. 2018;105:526-532. doi:10.1016/j.biopha.2018.05.148

178. Patel A, McGrosso D, Hefner Y, et al. Proteome allocation is linked to transcriptional regulation through a modularized transcriptome. *Nat Commun*. 2024;15(1):5234. doi:10.1038/s41467-024-49231-y

179. Laronha H, Caldeira J. Structure and Function of Human Matrix Metalloproteinases. *Cells*. 2020;9(5):1076. doi:10.3390/cells9051076

180. Tomczak A, Mortensen JM, Winnenburg R, et al. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Sci Rep*. 2018;8(1):5115. doi:10.1038/s41598-018-23395-2

181. Urík M, Tedla M, Hurník P. Pathogenesis of Retraction Pocket of the Tympanic Membrane—A Narrative Review. *Medicina*. 2021;57(5):425. doi:10.3390/medicina57050425

182. Xavier RF, Ramos D, Ito JT, et al. Effects of Cigarette Smoking Intensity on the Mucociliary Clearance of Active Smokers. *Respiration*. 2013;86(6):479-485. doi:10.1159/000348398

183. Habesoglu M, Demir K, Yumusakhuylu AC, Sahin Yilmaz A, Oysu C. Does Passive Smoking Have an Effect on Nasal Mucociliary Clearance? *Otolaryngol--head neck surg*. 2012;147(1):152-156. doi:10.1177/0194599812439004

184. Hammarén-Malmi S, Saxen H, Tarkkanen J, Mattila PS. Passive smoking after tympanostomy and risk of recurrent acute otitis media. *International Journal of Pediatric Otorhinolaryngology*. 2007;71(8):1305-1310. doi:10.1016/j.ijporl.2007.05.010

185. Csákányi Z, Czinner A, Spangler J, Rogers T, Katona G. Relationship of environmental tobacco smoke to otitis media (OM) in children. *International Journal of Pediatric Otorhinolaryngology*. 2012;76(7):989-993. doi:10.1016/j.ijporl.2012.03.017

186. Rowe-Jones JM, Brockbank MJ. Parental smoking and persistent otitis media with effusion in children. *International Journal of Pediatric Otorhinolaryngology*. 1992;24(1):19-24. doi:10.1016/0165-5876(92)90062-T

187. Kaylie DM, Bennett ML, Davis B, Jackson CG. Effects of smoking on otologic surgery outcomes: Smoking and Otologic Surgery Outcomes. *The Laryngoscope*. 2009;119(7):1384-1390. doi:10.1002/lary.20256

188. Rezigalla AA. Observational Study Designs: Synopsis for Selecting an Appropriate Study Design. *Cureus*. Published online January 17, 2020. doi:10.7759/cureus.6692

189. Wilson E, Jennings BA, Khondoker M, Philpott CM, Prinsley P, Brewer DS. Epidemiology of Cholesteatoma in the UK Biobank. *Clinical Otolaryngology*. 2025;50(2):316-329. doi:10.1111/coa.14257

190. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z

191. Poplin R, Chang PC, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;36(10):983-987. doi:10.1038/nbt.4235

192. Townsend P, Phillimore P, Beattie A. *Health and Deprivation: Inequality and the North*. Croom Helm; 1988.

193. Consumer Data Research Centre. Index of Multiple Deprivation (IMD). Consumer Data Research Centre.

194. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet*. 2018;50(9):1335-1341. doi:10.1038/s41588-018-0184-y

195. Ho DE, Imai K, King G, Stuart EA. **MatchIt**: Nonparametric Preprocessing for Parametric Causal Inference. *J Stat Soft*. 2011;42(8). doi:10.18637/jss.v042.i08

196. Genotyping. FinnGen. https://www.finngen.fi/en/researchers/genotyping

197. The MathWorks Inc. MATLAB. Published online 2022. https://www.mathworks.com

198. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2022. https://www.R-project.or

199. Bellis MA, Hughes K, Nicholls J, Sheron N, Gilmore I, Jones L. The alcohol harm paradox: using a national survey to explore how alcohol may disproportionately impact health in deprived individuals. *BMC Public Health*. 2016;16(1):111. doi:10.1186/s12889-016-2766-x

200. Foster HME, Celis-Morales CA, Nicholl BI, et al. The effect of socioeconomic deprivation on the association between an extended measurement of unhealthy lifestyle factors and health outcomes: a prospective analysis of the UK Biobank cohort. *The Lancet Public Health*. 2018;3(12):e576-e585. doi:10.1016/S2468-2667(18)30200-7

201. Butler DC, Petterson S, Phillips RL, Bazemore AW. Measures of Social Deprivation That Predict Health Care Access and Need within a Rational Area of Primary Care Service Delivery. *Health Serv Res*. 2013;48(2pt1):539-559. doi:10.1111/j.1475-6773.2012.01449.x

202. Bachert C, Luong AU, Gevaert P, et al. The Unified Airway Hypothesis: Evidence From Specific Intervention With Anti–IL-5 Biologic Therapy. *The Journal of Allergy and Clinical Immunology: In Practice*. 2023;11(9):2630-2641. doi:10.1016/j.jaip.2023.05.011

203. Bhutta MF, Williamson IG, Sudhoff HH. Cholesteatoma. *BMJ*. 2011;342(mar03 1):d1088-d1088. doi:10.1136/bmj.d1088

204. Migirov L, Duvdevani S, Kronenberg J. Otogenic intracranial complications: A review of 28 cases. *Acta Oto-Laryngologica*. 2005;125(8):819-822. doi:10.1080/00016480510038590

205. Deng Y, Ren Y, Wang C, Zhang Y. Clinical features and prognosis of cerebellopontine angle cholesteatoma in 54 patient. *Int J Clin Exp Med*. 2020;13(6):4426-4433.

206. Vezzani A, Fujinami RS, White HS, et al. Infections, inflammation and epilepsy. *Acta Neuropathol*. 2016;131(2):211-234. doi:10.1007/s00401-015-1481-5

207. Keezer MR, Sisodiya SM, Sander JW. Comorbidities of epilepsy: current concepts and future perspectives. *The Lancet Neurology*. 2016;15(1):106-115. doi:10.1016/S1474-4422(15)00225-2

208. Lyoubi M, Douimi L, El Krimi Z, et al. Management of the association of otosclerosis and cholesteatoma: Which pathology to treat first? *International Journal of Surgery Case Reports*. 2022;96:107281. doi:10.1016/j.ijscr.2022.107281

209. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology*. 2017;186(9):1026-1034. doi:10.1093/aje/kwx246

210. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaSci*. 2015;4(1):7. doi:10.1186/s13742-015-0047-8

211. Murel J, Kavlakoglu E. What is ridge regression? IBM. 2023. https://www.ibm.com/topics/ridge-regression

212. Daniels HE. Saddlepoint Approximations in Statistics. *Ann Math Statist*. 1954;25(4):631-650. doi:10.1214/aoms/1177728652

213. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27-38. doi:10.1093/biomet/80.1.27

214. Dey R, Schmidt EM, Abecasis GR, Lee S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am J Hum Genet*. 2017;101(1):37-49. doi:10.1016/j.ajhg.2017.05.014

215. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 2008;83(3):311-321. doi:10.1016/j.ajhg.2008.06.024

216. Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*. 2012;91(2):224-237. doi:10.1016/j.ajhg.2012.06.007

217. Li Y, Agarwal P. A Pathway-Based View of Human Diseases and Disease Relationships. Hide W, ed. *PLoS ONE*. 2009;4(2):e4346. doi:10.1371/journal.pone.0004346

218. Zhu X, Stephens M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat Commun*. 2018;9(1):4361. doi:10.1038/s41467-018-06805-x

219. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev*. 2009;19(3):212-219. doi:10.1016/j.gde.2009.04.010

220. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics*. 2014;95(1):5-23. doi:10.1016/j.ajhg.2014.06.009

221. Gaynor SM, Joseph T, Bai X, et al. Yield of genetic association signals from genomes, exomes and imputation in the UK Biobank. *Nat Genet*. Published online September 25, 2024. doi:10.1038/s41588-024-01930-4

222. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Published online 2013. doi:10.48550/ARXIV.1303.3997

223. Truth Challenge V2: Calling Variants from Short and Long Reads in Difficult-to-Map Regions. precisionFDA. 2020. Accessed June 17, 2024. https://precision.fda.gov/challenges/10/results

224. Yun T, Li H, Chang PC, Lin MF, Carroll A, McLean CY. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Robinson P, ed. *Bioinformatics*. 2021;36(24):5582-5589. doi:10.1093/bioinformatics/btaa1081

225. Szustakowski JD, Balasubramanian S, Kvikstad E, et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nature Genetics*. 2021;53(7):942-948. doi:10.1038/s41588-021-00885-0

226. Bailey SF, Alonso Morales LA, Kassen R. Effects of Synonymous Mutations beyond Codon Bias: The Evidence for Adaptive Synonymous Substitutions from Microbial Evolution

Experiments. Bedhomme S, ed. *Genome Biology and Evolution*. 2021;13(9):evab141. doi:10.1093/gbe/evab141

227. Shen X, Song S, Li C, Zhang J. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature*. 2022;606(7915):725-731. doi:10.1038/s41586-022-04823-w

228. Kruglyak L, Beyer A, Bloom JS, et al. Insufficient evidence for non-neutrality of synonymous mutations. *Nature*. 2023;616(7957):E8-E9. doi:10.1038/s41586-023-05865-4

229. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;6(2):80-92. doi:10.4161/fly.19695

230. Johnson JL, Abecasis GR. GAS Power Calculator: web-based power calculator for genetic association studies. Published online July 17, 2017. doi:10.1101/164343

231. Newland MC. The Proper Calculation of Risk Ratios: How and Why. *Perspect Behav Sci*. Published online October 7, 2024. doi:10.1007/s40614-024-00423-3

232. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet*. 2006;38(2):209-213. doi:10.1038/ng1706

233. Zhang J, Yu KF. What's the Relative Risk?: A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. *JAMA*. 1998;280(19):1690. doi:10.1001/jama.280.19.1690

234. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122. doi:10.1186/s13059-016-0974-4

235. Pedersen BS, Brown JM, Dashnow H, et al. Effective variant filtering and expected candidate variant yield in studies of rare human disease. *npj Genom Med*. 2021;6(1):60. doi:10.1038/s41525-021-00227-3

236. Fadista J, Manning AK, Florez JC, Groop L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur J Hum Genet*. 2016;24(8):1202-1205. doi:10.1038/ejhg.2015.269

237. Guo MH, Plummer L, Chan YM, Hirschhorn JN, Lippincott MF. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. *The American Journal of Human Genetics*. 2018;103(4):522-534. doi:10.1016/j.ajhg.2018.08.016

238. Hildebrand S, Hultin S, Subramani A, et al. The E-cadherin/AmotL2 complex organizes actin filaments required for epithelial hexagonal packing and blastocyst hatching. *Sci Rep*. 2017;7(1):9540. doi:10.1038/s41598-017-10102-w

239. Hohmann T, Dehghani F. The Cytoskeleton—A Complex Interacting Meshwork. *Cells*. 2019;8(4):362. doi:10.3390/cells8040362

240. Safiejko-Mroczka B, Bell PB. Reorganization of the actin cytoskeleton in the protruding lamellae of human fibroblasts. *Cell Motil Cytoskeleton*. 2001;50(1):13-32. doi:10.1002/cm.1038

241. Svitkina T. The Actin Cytoskeleton and Actin-Based Motility. *Cold Spring Harb Perspect Biol*. 2018;10(1):a018267. doi:10.1101/cshperspect.a018267

242. Olson MF, Sahai E. The actin cytoskeleton in cancer cell motility. *Clin Exp Metastasis*. 2009;26(4):273. doi:10.1007/s10585-008-9174-2

243. D'Angelo G, Rega LR, De Matteis MA. Connecting vesicular transport with lipid synthesis: FAPP2. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*. 2012;1821(8):1089-1095. doi:10.1016/j.bbalip.2012.01.003

244. Lockridge O. Review of human butyrylcholinesterase structure, function, genetic variants, history of use in the clinic, and potential therapeutic uses. *Pharmacology & Therapeutics*. 2015;148:34-46. doi:10.1016/j.pharmthera.2014.11.011

245. Wang T, Hong W. RILP interacts with VPS22 and VPS36 of ESCRT-II and regulates their membrane recruitment. *Biochemical and Biophysical Research Communications*. 2006;350(2):413-423. doi:10.1016/j.bbrc.2006.09.064

246. Szymanowicz O, Drużdż A, Słowikowski B, et al. A Review of the CACNA Gene Family: Its Role in Neurological Disorders. *Diseases*. 2024;12(5):90. doi:10.3390/diseases12050090

247. Ross AC, Taylor CL, Yaktine AL, Del Valle HB, eds. Overview of Calcium. In: *Dietary Reference Intakes for Calcium and Vitamin D*. National Academies Press; 2011.

248. Vig M, Kinet JP. Calcium signaling in immune cells. *Nat Immunol*. 2009;10(1):21-27. doi:10.1038/ni.f.220

249. Subramaniam T, Fauzi MB, Lokanathan Y, Law JX. The Role of Calcium in Wound Healing. *Int J Mol Sci*. 2021;22(12):6486. doi:10.3390/ijms22126486

250. Cornec-Le Gall E, Alam A, Perrone RD. Autosomal dominant polycystic kidney disease. *The Lancet*. 2019;393(10174):919-935. doi:10.1016/S0140-6736(18)32782-X

251. Milisav I. Dynein and dynein-related genes. *Cell Motil Cytoskeleton*. 1998;39(4):261-272. doi:10.1002/(SICI)1097-0169(1998)39:4<261::AID-CM2>3.0.CO;2-6

252. McConnachie DJ, Stow JL, Mallett AJ. Ciliopathies and the Kidney: A Review. *American Journal of Kidney Diseases*. 2021;77(3):410-419. doi:10.1053/j.ajkd.2020.08.012

253. Knowles MR, Daniels LA, Davis SD, Zariwala MA, Leigh MW. Primary Ciliary Dyskinesia. Recent Advances in Diagnostics, Genetics, and Characterization of Clinical Disease. *Am J Respir Crit Care Med*. 2013;188(8):913-922. doi:10.1164/rccm.201301-0059CI

254. Smith DJ, Montenegro-Johnson TD, Lopes SS. Symmetry-Breaking Cilia-Driven Flow in Embryogenesis. *Annu Rev Fluid Mech*. 2019;51(1):105-128. doi:10.1146/annurev-fluid-010518-040231

255. Wang H, Du H, Ren R, et al. Temporal and spatial assembly of inner ear hair cell ankle link condensate through phase separation. *Nat Commun*. 2023;14(1):1657. doi:10.1038/s41467-023-37267-5

256. Ran J, Zhou J. Targeting the photoreceptor cilium for the treatment of retinal diseases. *Acta Pharmacol Sin*. 2020;41(11):1410-1415. doi:10.1038/s41401-020-0486-3

257. Pearsall N, Bhattacharya G, Wisecarver J, Adams J, Cosgrove D, Kimberling W. Usherin expression is highly conserved in mouse and human tissues. *Hearing Research*. 2002;174(1-2):55-63. doi:10.1016/S0378-5955(02)00635-4

258. Palo JU, Ulmanen I, Lukka M, Ellonen P, Sajantila A. Genetic markers and population history: Finland revisited. *Eur J Hum Genet*. 2009;17(10):1336-1346. doi:10.1038/ejhg.2009.53

259. Haghverdizadeh P, Sadat Haerian M, Haghverdizadeh P, Sadat Haerian B. ABCC8 genetic variants and risk of diabetes mellitus. *Gene*. 2014;545(2):198-204. doi:10.1016/j.gene.2014.04.040

260. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*. 2020;15(9):2759-2772. doi:10.1038/s41596-020-0353-1

261. Monaco A, Pantaleo E, Amoroso N, et al. A primer on machine learning techniques for genomic applications. *Computational and Structural Biotechnology Journal*. 2021;19:4345-4359. doi:10.1016/j.csbj.2021.07.021

262. Gao XR, Chiariglione M, Qin K, et al. Explainable machine learning aggregates polygenic risk scores and electronic health records for Alzheimer's disease prediction. *Sci Rep*. 2023;13(1):450. doi:10.1038/s41598-023-27551-1

263. Badré A, Zhang L, Muchero W, Reynolds JC, Pan C. Deep neural network improves the estimation of polygenic risk scores for breast cancer. *J Hum Genet*. 2021;66(4):359-369. doi:10.1038/s10038-020-00832-7

264. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. *Bioinformatics*. 2015;31(9):1466-1468. doi:10.1093/bioinformatics/btu848

265. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience*. 2019;8(7):giz082. doi:10.1093/gigascience/giz082

266. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. Schwartz R, ed. *Bioinformatics*. 2021;36(22-23):5424-5431. doi:10.1093/bioinformatics/btaa1029

267. Dodge Y. Central Limit Theorem. In: *The Concise Encyclopedia of Statistics*. Springer New York; 2008:66-68. doi:10.1007/978-0-387-32833-1_50

268. Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. Wray NR, ed. *PLoS Genet*. 2013;9(3):e1003348. doi:10.1371/journal.pgen.1003348

269. Lee SH, Goddard ME, Wray NR, Visscher PM. A Better Coefficient of Determination for Genetic Profile Analysis. *Genetic Epidemiology*. 2012;36(3):214-224. doi:10.1002/gepi.21614

270. Pozarickij A, Williams C, Guggenheim JA. Non-additive (dominance) effects of genetic variants associated with refractive error and myopia. *Mol Genet Genomics*. 2020;295(4):843-853. doi:10.1007/s00438-020-01666-w

271. Ohta R, Tanigawa Y, Suzuki Y, Kellis M, Morishita S. A polygenic score method boosted by non-additive models. *Nature Communications*. 2024;15(443).

272. Sud A, Horton RH, Hingorani AD, et al. Realistic expectations are key to realising the benefits of polygenic scores. *BMJ*. Published online February 28, 2023:e073149. doi:10.1136/bmj-2022-073149

273. Polygenic Risk Score Task Force of the International Common Disease Alliance, Adeyemo A, Balaconis MK, et al. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat Med*. 2021;27(11):1876-1884. doi:10.1038/s41591-021-01549-6

274. Biau G, Scornet E. A random forest guided tour. *TEST*. 2016;25(2):197-227. doi:10.1007/s11749-016-0481-7

275. Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

276. fitctree. MathWorks. https://uk.mathworks.com/help/stats/fitctree.html

277. Krzywinski M, Altman N. Classification and regression trees. *Nat Methods*. 2017;14(8):757-758. doi:10.1038/nmeth.4370

278. Stephan J, Stegle O, Beyer A. A random forest approach to capture genetic effects in the presence of population structure. *Nat Commun*. 2015;6(1):7432. doi:10.1038/ncomms8432

279. Toloşi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*. 2011;27(14):1986-1994. doi:10.1093/bioinformatics/btr300

280. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. *Neurocomputing*. 2018;300:70-79. doi:10.1016/j.neucom.2017.11.077

281. Kursa MB, Jankowski A, Rudnicki WR. Boruta – A System for Feature Selection. *Fundamenta Informaticae*. 2010;101(4):271-285. doi:10.3233/FI-2010-288

282. Fushiki T. Estimation of prediction error by using K-fold cross-validation. *Stat Comput*. 2011;21(2):137-146. doi:10.1007/s11222-009-9153-8

283. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. *J Stat Soft*. 2010;36(11). doi:10.18637/jss.v036.i11

284. Wright MN, Ziegler A. **ranger**: A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. *J Stat Soft*. 2017;77(1). doi:10.18637/jss.v077.i01

285. Thompson DJ, Wells D, Selzam S, et al. UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. Published online June 16, 2022. doi:10.1101/2022.06.16.22276246

286. Olender T, Waszak SM, Viavant M, et al. Personal receptor repertoires: olfaction as a model. *BMC Genomics*. 2012;13(1):414. doi:10.1186/1471-2164-13-414

287. Dettori C, Ronca F, Scalese M, Saponaro F. Parathyroid Hormone (PTH)-Related Peptides Family: An Intriguing Role in the Central Nervous System. *JPM*. 2023;13(5):714. doi:10.3390/jpm13050714

288. Xu W, Dahlke SP, Emery AC, et al. Cyclic AMP-dependent activation of ERK via GLP-1 receptor signalling requires the neuroendocrine cell-specific guanine nucleotide exchanger NCS-RapGEF2. *J Neuroendocrinology*. 2021;33(7):e12974. doi:10.1111/jne.12974

289. Giese APJ, Ali S, Isaiah A, Aziz I, Riazuddin S, Ahmed ZM. Genomics of Otitis Media (OM): Molecular Genetics Approaches to Characterize Disease Pathophysiology. *Front Genet*. 2020;11:313. doi:10.3389/fgene.2020.00313

290. Geng R, Wang Q, Chen E, Zheng QY. Current Understanding of Host Genetics of Otitis Media. *Front Genet*. 2020;10:1395. doi:10.3389/fgene.2019.01395

291. Iacono WG, Malone SM, Vrieze SI. Endophenotype best practices. *International Journal of Psychophysiology*. 2017;111:115-144. doi:10.1016/j.ijpsycho.2016.07.516

292. Bhutta. Evolution and Otitis Media: A Review, and a Model to Explain High Prevalence in Indigenous Populations. *Human Biology*. 2015;87(2):92. doi:10.13110/humanbiology.87.2.0092

293. De Ru JA, Grote JJ. Otitis media with effusion: disease or defense? *International Journal of Pediatric Otorhinolaryngology*. 2004;68(3):331-339. doi:10.1016/j.ijporl.2003.11.003

294. Tang IP, Prepageran N, Raman R, Sharizhal T. Epithelial migration in the atelectatic tympanic membrane. *J Laryngol Otol*. 2009;123(12):1321-1324. doi:10.1017/S0022215109990806

295. Phelan PJ, Rheault MN. Hearing loss and renal syndromes. *Pediatr Nephrol*. 2018;33(10):1671-1683. doi:10.1007/s00467-017-3835-9

296. Greenberg D, Rosenblum ND, Tonelli M. The multifaceted links between hearing loss and chronic kidney disease. *Nat Rev Nephrol*. 2024;20(5):295-312. doi:10.1038/s41581-024-00808-2

297. Wang TC, Lin CC, Lin C Der, et al. Increased acquired cholesteatoma risk in patients with osteoporosis: A retrospective cohort study. *PLoS ONE*. 2015;10(7):e0132447. doi:10.1371/journal.pone.0132447

298. Rahimov F, Nieminen P, Kumari P, et al. High incidence and geographic distribution of cleft palate in Finland are associated with the IRF6 gene. *Nat Commun*. 2024;15(1):9568. doi:10.1038/s41467-024-53634-2

299. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*. 2017;169(7):1177-1186. doi:10.1016/j.cell.2017.05.038

300. Mathieson I. The omnigenic model and polygenic prediction of complex traits. *The American Journal of Human Genetics*. 2021;108(9):1558-1563. doi:10.1016/j.ajhg.2021.07.003

301. Mars N, Widén E, Kerminen S, et al. The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nat Commun*. 2020;11(1):6383. doi:10.1038/s41467-020-19966-5

302. Suzuki N, Nishiyama A, Warita H, Aoki M. Genetics of amyotrophic lateral sclerosis: seeking therapeutic targets in the era of gene therapy. *J Hum Genet*. 2023;68(3):131-152. doi:10.1038/s10038-022-01055-8

303. Wang Y, Namba S, Lopera E, et al. Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. *Cell Genomics*. 2023;3(1):100241. doi:10.1016/j.xgen.2022.100241

304. Polygenic Scores (PGS). PGS Catalog. https://www.pgscatalog.org/browse/scores/

305. Aimi K. Role of the tympanic ring in the pathogenesis of congenital cholesteatoma. *The Laryngoscope*. 1983;93(9):1140-1146. doi:10.1288/00005537-198309000-00005

306. Michaels L. Origin of congenital cholestecetoma from a normally occurring epidermoid rest in the developing middle ear. *International Journal of Pediatric Otorhinolaryngology*. 1988;15(1):51-65. doi:10.1016/0165-5876(88)90050-X

307. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet*. 2018;19(8):491-504. doi:10.1038/s41576-018-0016-z

308. Fitipaldi H, Franks PW. Ethnic, gender and other sociodemographic biases in genome-wide association studies for the most burdensome non-communicable diseases: 2005–2022. *Human Molecular Genetics*. 2023;32(3):520-532. doi:10.1093/hmg/ddac245

309. Song IS, Han WG, Lim KH, et al. Clinical Characteristics and Treatment Outcomes of Congenital Cholesteatoma. *Int Adv Otol*. 2019;15(3):386-390. doi:10.5152/iao.2019.6279

310. Uzun T, Çaklı H, Coşan DT, et al. In vitro study on immune response modifiers as novel medical treatment options for cholesteatoma. *International Journal of Pediatric Otorhinolaryngology*. 2021;145:110743. doi:10.1016/j.ijporl.2021.110743

311. Yamamoto-Fukuda T, Terakado M, Hishikawa Y, Koji T, Takahashi H. Topical application of 5-fluorouracil on attic cholesteatoma results in downregulation of keratinocyte growth factor and reduction of proliferative activity. *Eur Arch Otorhinolaryngol*. 2008;265(10):1173-1178. doi:10.1007/s00405-008-0597-9

312. Wright CG, Bird LL, Meyerhoff WL. Effect of 5-fluorouracil in cholesteatoma development in an animal model. *American Journal of Otolaryngology*. 1991;12(3):133-138. doi:10.1016/0196-0709(91)90142-3

313. Rao USV, Srinivas DR, Humbarwadi RS, Malhotra BK. Role of vitamin A in the evolution of cholesteatoma. *Indian J Otolaryngol Head Neck Surg*. 2009;61(2):150-152. doi:10.1007/s12070-009-0056-y

314. Boesoirie S, Hasansulama W, Lasminingrum L, et al. The Role of Vitamins A and E Level in Chronic Suppurative Otitis Media with and without Cholesteatoma. *JMDH*. 2023;Volume 16:3435-3442. doi:10.2147/JMDH.S414115

315. Andrade-Guerrero J, Santiago-Balmaseda A, Jeronimo-Aguilar P, et al. Alzheimer's Disease: An Updated Overview of Its Genetics. *IJMS*. 2023;24(4):3754. doi:10.3390/ijms24043754

316. Ye H, Robak LA, Yu M, Cykowski M, Shulman JM. Genetics and Pathogenesis of Parkinson's Syndrome. *Annu Rev Pathol Mech Dis*. 2023;18(1):95-121. doi:10.1146/annurev-pathmechdis-031521-034145

317. Lanoiselée HM, Nicolas G, Wallon D, et al. APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: A genetic screening study of familial and sporadic cases. Miller BL, ed. *PLoS Med*. 2017;14(3):e1002270. doi:10.1371/journal.pmed.1002270

318. Loos RJF, Yeo GSH. The genetics of obesity: from discovery to biology. *Nat Rev Genet*. 2022;23(2):120-133. doi:10.1038/s41576-021-00414-z

319. Potkin SG, Turner JA, Guffanti G, et al. Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: Methodological considerations. *Cognitive Neuropsychiatry*. 2009;14(4-5):391-418. doi:10.1080/13546800903059829

320. Yaguchi H, Togawa K, Moritani M, Itakura M. Identification of candidate genes in the type 2 diabetes modifier locus using expression QTL. *Genomics*. 2005;85(5):591-599. doi:10.1016/j.ygeno.2005.01.006

321. Khondoker M, Newhouse S, Westman E, et al. Linking Genetics of Brain Changes to Alzheimer's Disease: Sparse Whole Genome Association Scan of Regional MRI Volumes in the ADNI and AddNeuroMed Cohorts. *JAD*. 2015;45(3):851-864. doi:10.3233/JAD-142214

322. Vaid S, Kamble Y, Vaid N, et al. Role of Magnetic Resonance Imaging in Cholesteatoma: The Indian Experience. *Indian J Otolaryngol Head Neck Surg*. 2013;65(S3):485-492. doi:10.1007/s12070-011-0360-1

323. Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med*. 2015;7(1):16. doi:10.1186/s13073-015-0138-2

324. Barry CJS, Walker VM, Cheesman R, Davey Smith G, Morris TT, Davies NM. How to estimate heritability: a guide for genetic epidemiologists. *International Journal of Epidemiology*. 2023;52(2):624-632. doi:10.1093/ije/dyac224

325. Brodie MJ, Dichter MA. Antiepileptic Drugs. Wood AJJ, ed. *N Engl J Med*. 1996;334(3):168-175. doi:10.1056/NEJM199601183340308