UNIVERSITÀ DEGLI STUDI DI SASSARI

CORSO DI DOTTORATO DI RICERCA IN SCIENZE BIOMEDICHE

Coordinatore del Corso: Prof.ssa Maioli Margherita

CURRICULUM IN GENETICA MEDICA

Responsabile di Curriculum: Prof.ssa Maioli Margherita

XXXIV CICLO

Development of reproducible workflows to optimize data-intensive bioinformatics

Coordinatore:

Prof.ssa Maioli Margherita

Tutor:

Prof. Francesco Cucca

Prof. Andrea Angius

Tesi di dottorato di:

Dott. Vincenzo Rallo

Anno Accademico 2020-2021

Index

Summary

This study aims to test and validate different strategies based on the features of the phenotypes analyzed. The objective also involves optimizing datasets, tools, and bioinformatics and statistical approaches aimed at the most effective analysis of genetic data to identify causal and/or predisposing variants of Mendelian diseases and complex traits.

Research activity was based on the implementation of standardized and experimental procedures to perform structured genomic analysis of rare and complex diseases and related quantitative traits, using a cohort of individuals extensively characterized at the genetic and phenotypic level from the Sardinian population (SardiNIA cohort). The genetic homogeneity of the Sardinian population facilitated the analysis and allowed to highlight the presence of founder effects and/or peculiar genomic traits causing rare and complex diseases. The accuracy of the clinical data of the SardiNIA cohort and the availability of genetic data distributed over the whole genome for each individual, made possible the screening of several rare and complex phenotypes.

Concerning rare diseases, during screenings procedures related to phenotypes involved in age-related diseases, several patients affected by different rare diseases were identified according to specific clinical features. Using the whole-genome approach, we were able to describe the first causal molecular variant of Usher syndrome, its incidence and distribution in Sardinia and the specific molecular features. Collaboration with research institutes and hospitals improved to reach the molecular diagnosis of our patients. Our results reveal that this approach represents an effective and generalizable method to find causal variants in rare and/or Mendelian diseases and to set up large-scale screening programs and/or pre-symptomatic diagnosis in patients and risk groups.

In the study of complex diseases, to investigate in depth the molecular mechanisms of complex traits and their regulatory pathways, we systematically integrated association data from public Genome-Wide (GWAS) studies with whole-genome sequence data from the SardiNIA cohort using colocalization analysis. Currently,

standardization of colocalization methodologies to examine efficiently the shared association profiles between traits is still a challenge. My experimental approach involved comparing three different software programs, "*coloc*", "*gwas-pw*", and "*eCAVIAR*" to examine and evaluate the impact of each of them on results. Adopting an agnostic approach without biological assumptions, we assessed biological affinity between quantitative traits and diseases to identify novel genetic variants potentially useful in the study of several diseases. Our investigation provided a high percentage of replication of literature findings, clarified uncertain association signals, and identified novel coincident associations between quantitative traits and diseases. Concordance results between the three alternative software were more than 90% with minimal discrepancy. High concordance indicates the reliability of the used algorithms and proves that the specific software does not affect results. This study underlines the scientific contribution of colocalization analysis as a valid methodology to study phenotype-genotype relationships and to identify new susceptibility loci in complex diseases.

Our research has adopted and validated several approaches according to peculiar phenotypes starting from a robust genetic database. Biological characterization and functional annotation will make genome analyses reliable and increase our understanding of biological mechanisms of disease susceptibility.

Introduction

1. From human genome sequencing to applications

Whole human genome sequencing[1] has enabled the development of a genomics-based approach to diseases that has revolutionized diagnostic and therapeutic methods, risk assessment and therapies, especially in rare diseases as well as in complex diseases/traits.

Personalized medicine and disease molecular characterization have created genomic medicine methods that now are standardized in research, but a part of them is moving into the clinic routine and can improve the achievement of a transition from clinical research to precision medicine.

Cancer pharmacogenomics, rare disease diagnosis, monitoring infectious disease pandemics, understanding the molecular basis of autoimmune diseases, the role of the microbiome and the identification of biomarkers responsible for drug response[2] are now tangible applications.

These technological innovations and knowledge-based approaches in genomics are constantly evolving and are enabled by the continuous evolution of Next-Generation Sequencing (NGS) technologies.[3]

2. Next-Generation Sequencing massive data trend: standard processes and state-of-the-art innovative approaches

Over the past two decades, impressive advances have completely changed genome sequencing technologies, leading to radically reduced costs, increased volume of data produced at reduced timescales that have led to an increase in the number of genomes sequenced[4]. In a short time, a large amount of data has been generated, made publicly available and can be used to study different phenotypes or genomic related traits.

Advances in sequencing technology allowed genetic studies to extend the analysis to thousands of whole genome samples, enabling the identification of previously undetectable variants with possible functional consequences and accelerating translation into biological understanding. Without NGS technologies, these advances would not have been possible.

Improvements in technologies have enabled an increase in the throughput, quality, and coverage of genomic data, and making it necessary to exploit the data using approaches calibrated according to the technology applied. The next challenge will be to know how to analyze and capitalize the data generated according to the phenotype under study. It is essential to adapt existing approaches and develop new ones that can support the phenotype characteristics. Understanding functional biological effects, the pathways involved and creating the opportunity to identify genome-based drug compounds are the next challenges.

Not only technology but also several other reasons contributed to the success of genomic research: the availability of public sequences databases and detailed maps of genomic variants in different populations, together with several new bioinformatic and statistical approaches that rapidly and accurately correlate the whole genome of thousands of samples with quantitative phenotypes or diseases. These have provided an innovative tool for the scientific community to analyze biological data and generate statistically supported results[5].

3. **Omics approaches**

The integration and analysis of different datasets is essential for moving toward personalized medicine. Since the sequencing of the first whole genome to date, the data produced have undergone several improvements: higher resolution and coverage, better quality data generation, and many analysis tools that have made it possible to have more reliable data. Several technologies have been developed to support the molecular description of biological samples. High-throughput sequencing technologies allow samples to be characterized at multiple levels and with different approaches depending on the purpose for which they are to be used.

Omics approaches make possible the comprehensive study of the complex molecular interactions of cellular systems. With the advancement and development of new post-genomic technologies, omics studies are becoming more widespread and more accessible to researchers and clinicians[6].

A major challenge for systemic biology is to integrate genomics, transcriptomics, proteomics and metabolomics to achieve a more comprehensive overview of biological entities. From a genomic perspective, through the integration of epigenomic, metagenomic, single-cell analysis and immune profiling, the biological phenotypes underlying the inherited diseases will be described in a greater depth and combined with metabolic and/or functional analyses even in silico.

Omics approaches are providing to the scientific community a variety of strategies to investigate genomes more comprehensively, resulting in a more accurate genetic view of the rare and complex phenotypes. These approaches will allow us to identify diagnostic/prognostic markers of trait and disease.

4. Approaches for rare and complex disorders

Moving from genetic statistical associations to the identification of the functional genetic variants that influence disease is often a complex process. Fine mapping can select and prioritize genetic variants, but while approaches for identifying causal variants are now standardized for rare diseases, there are no established and accepted criteria for complex diseases. These gaps in the uniformity of approaches can be filled only through conducting trials to validate the pipeline towards a gold standard for the detection of causal variants underlying complex traits.

a. Rare diseases

Strategies to find variants causing rare Mendelian diseases have been now standardized mainly with the diffusion and standardization of high coverage whole-exome sequencing protocols and the creation of large databases of rare disease patients and corresponding comparison cohorts. Some protocols have already been optimized and incorporated into routine clinical diagnostics. Best practices include screening methods based on the mode of inheritance of the trait (dominant or recessive, pseudo-Mendelian), pedigree structure, trios analysis to identify *de novo* or inherited variant, minor allele frequencies (MAF), predicted biological function in silico of each variant, candidate gene and analysis of network in which the gene is involved[7]. The goal is to achieve evidences that identify a particular gene/pathway for proper diagnosis.

b. Complex traits

Concerning complex traits and common diseases, no standardized analytical methodology is currently available. Nevertheless, both advances in NGS technologies and the extensive use of the Genome-wide Associations Study (GWAS) have contributed greatly to the discovery of relationships between genomic variants and disease susceptibility[8] [9]. Considerable progress has been made in the comprehension of the genetic basis and molecular pathways of a variety of biomedically relevant common phenotypes, such as type 1 and 2 diabetes, macular degeneration, rheumatoid arthritis, coronary artery disease[10] [11] [12] and so on.

i. Genome-Wide Associations Study (GWAS)

Genome Wide Associations Studies are now systematic and robust investigations and represent a highly effective and clearly defined approach to identify genetic factors related to the causes of phenotypes and diseases. These studies are based on genetic datasets, including single nucleotide polymorphisms (SNPs) or small insertions/deletions (indels), that represent a large fraction of genome sequence variability, highly informative for specific traits.

While some approaches consider specific gene intervals, GWAS use an agnostic approach considering the entire set of available SNPs over the whole genome. This approach allows for the study and identification of genome-associated regions, including those that have a function that is not directly related to the evaluated phenotype assessed. Implementation of this approach on large and highly characterized datasets revealed more than 71,673 variant-trait associations and provided valuable information on the allelic architecture of multifactorial traits[13].

In recent years, genomic data have increased exponentially and numerous different approaches have been developed and improved to enable GWAS data investigation[14 15]. The ability to discern causal genes among the myriad of associated variants is a relevant aspect of pathological variants identification[16].

Probing entire genomic regions, both coding and non-coding ones, leads us to elucidate which genomic portions are linked to specific variants, genes and metabolic pathways whose alterations underlie the etiology of complex diseases[17 18 19 20].

Given the increasing availability of public GWAS data, their integration represents the rational next step towards a better biological interpretation of outcomes. The transition to deep-level analyses, using methodologies that review previous data in greater detail for stronger evidence, enables better interpretation of GWAS data to identify shared causal genetic variants between quantitative traits (QT) and diseases.

A huge range of gene expression measurements derived from microarray, RNA-seq or QTL studies has been widely generated and released to the scientific community[21]. A proportion of GWAS variants may have a modulatory function in gene expression. Genomic loci associated with complex phenotypes are enriched for variants regulating gene transcriptional levels (eQTLs, expression Quantitative Traits Loci) influencing the number of transcripts rather than gene sequences [22]. eQTL analysis is a powerful approach to identify target genes and to investigate the regulatory role of variants associated with complex traits and diseases[23]. The eQTLs survey potentially improves insight into gene regulation and the target genes associated. Furthermore, eQTL studies in multiple tissues provide a useful control for each new GWAS dataset, as they directly indicate the tissues and their potential candidate genes in which these effects are mediated.[24] [25]

These findings provide essential insights understanding the role of the variant impact on individual genetic risk, but further studies are needed to define the molecular mechanisms.

ii. Colocalization analyses of genetic variants of complex traits

Different statistical approaches and methodologies could exploit and integrate GWAS on omics data, providing insights on the role of traits in disease predisposition and suggestions concerning the underlying mechanism of genetic variants, such as mendelian randomization[26], fine-mapping[27] and colocalization[2829].

A statistical methodology to carry out GWAS surveys is colocalization analysis. Colocalization analysis represents a robust approach that allows associating both trait and disease directly. It is an excellent strategy to verify shared genetic variants between two independent association signals. Colocalization arises from the need to assess whether two genetic signals are indeed overlapping, indicating a shared association profile driven by the same haplotype. The main advantage is the ability to exploit summary statistics from GWAS studies to investigate and identify correlations between phenotypes or diseases, without having to access genotypic data.

The systematic application of this methodology on a large scale to publicly available data could highlight novel molecular mechanisms underlying disease risk and other traits influenced by genes previously identified in association with specific traits.

Colocalization is based on a Bayesian approach to assess genetic associations. This overcomes the limitations of the frequentist approach which underestimates essential elements which include the study's power and the number of true positive signals[30]. These statistical models use Bayes factors to test genome-wide significance by assigning credible probabilities to each of the alternative hypotheses[31]. Unlike P-values, Bayes Factors of SNPs with different frequencies and from different studies are directly comparable. Thus, the Bayesian approach allows direct SNPs comparison, facilitating probability calculation. Rather than comparing exclusively the most associated SNP for a signal, the colocalization takes into account the significance of the full set of SNPs in a region, matching association profiles and evaluating the probability of a shared causal variant/haplotype.

GWAS data analysis by colocalization has identified previously reported correlations between diseases and phenotypes, together with novel results, emphasizing the relevance of this statistical model. Several tools have been developed to perform colocalization analysis.[323334] Each of them is based on different assumptions to assess the overlap of GWAS signals. Since a standard colocalization method has not been yet defined, a methodology investigation is needed to detect an appropriate approach to identify variants influencing both traits and diseases. Due to the multitude of GWAS data available, their evaluation using colocalization can be particularly useful to investigate phenotype-genome connections.

Rationale

The main objective of genomic investigations must be the correct design of the analysis strategy. Starting from a cohort of individuals from the Sardinian population (SardiNIA cohort[35]) extensively characterized by a genomic and phenotypic perspective, my research activity was based on structured genomic analyses concerning rare and complex diseases and related quantitative traits. This study proposes to conduct, compare, and evaluate different strategies and methodologies to determine bioinformatic and statistical approaches aimed at the genetic characterization of different diseases.

The Sardinian population, settled in an isolated region with minimal admixture from outside populations for many centuries, is characterized by distinctive genetic features. Genetic homogeneity enables to analyze and highlight the presence of founder effects causing rare and complex pathologies, easing the identification of variants rare in Europeans but having higher frequency in Sardinia due to drift or selection [36] [37]. Such variants need extremely large sample sizes to be pointed out in other cosmopolitan populations. The SardiNIA project is focused on the molecular characterization of a general population cohort, deeply genetically characterized, which includes individuals affected by several disorders (autoimmune, cardiovascular, metabolic, and inflammatory diseases) that we aimed to clarify and define from an original genetic perspective[38] [39] [40] [41].

Screening a similar large cohort representing the general population and characterized with extensive genomic data, provides a unique opportunity to study rare and complex diseases, binary and quantitative traits and identify susceptibility loci.

In this context, starting from Next Generation Sequencing data (Whole Genome/Exome Sequencing and microarrays) and exploiting a dataset composed of thousands of individuals characterized for thousand phenotypes, we implemented and applied a set of standardized bioinformatics pipelines optimized according to the characteristics of the analyzed phenotypes.

The implementation of bioinformatics approaches and the use of specific algorithms that handle the querying, identification, and management of large amounts of genetic data have facilitated, accelerated, and improved the effectiveness of the analysis of complex traits and rare diseases. Our workflow involved production, collection, standardization, analysis and interpretation of genetic/biological data combined with the investigation, comparison and application of different informatics tools.

In rare diseases, a standardized analytical approach conformed to the state of the art and the best practices used internationally was adopted. Genome analysis using several tools, algorithms or custom 'scripts' has been conducted. The goal was to identify a restricted set of likely causal variants and provide a biological interpretation of them associated to phenotypes and specific clinical features.

Concerning complex diseases, we integrated association data from public GWAS with whole-genome sequence data of the SardiNIA cohort employing a colocalization analysis. Currently, no universally standardized colocalization methodologies are recognized by the scientific community and my experimental approach involved a comparison of different statistical algorithms. The strengths and weaknesses of the different colocalization approaches were examined, together with the impact of each on the obtained results. The main goal was to investigate in depth the molecular mechanisms regulating complex traits; in particular, adopting an agnostic approach without biological assumptions, the presence of a biological affinity between quantitative traits and several diseases was assessed, in order to identify novel biomarkers potentially useful in disease therapy.

Materials and Methods

1. Data Description

a. SardiNIA cohort

The SardiNIA project[35] involves over about 8000 volunteers recruited from 4 villages in Ogliastra genetically characterized with about 22M variants. All individuals are also deeply phenotypically characterized with more than 800 (anthropometric, serological, hematological, cardiovascular, ophthalmological and other measurements), with 3,757 of them immune profiled (more than 1000 measurements)[42]. Concerning this study and, more specifically aging-related ophthalmologic phenotypes, a study focused on retinal blood vessels was done by photographing the ocular fundus, which varies with age and in which automated quantitative analysis of fundus images can reveal general measures of the patient's health status.[43] From this screening, it has been possible to identify a number of eye diseases in the general population that allowed the identification of genetic syndromes related to the eye. This huge, highly detailed dataset has allowed us to study several phenotypes, both rare and complex, implementing different bioinformatics approaches according to the nature of the disorders, with the aim of improving our understanding of the molecular mechanisms underlying diseases.

i. Genetic data

SardiNIA samples were all genotyped using different Illumina microarrays (OmniExpress, ImmunoChip, Cardio-MetaboChip and ExomeChip) on 890,542 autosomal and 16,325 unique X-linked SNPs across the genome. In parallel, whole-genome shotgun sequence data from 3,514 Sardinian individuals, either from the SardiNIA cohort or participating in case-control studies from across the island, were also available[44]. Sequence data were generated with Illumina Genome Analyzer IIx, Illumina HiSeq 2000 and Illumina HiSeq 2500 instruments. Reads were aligned to the human reference genome (GRCh37 assembly) using BWA-0.5.9[45]. Variant calling and genotyping was carried out using GotCloud pipeline (see https://genome.sph.umich.edu/wiki/GotCloud). The likely functional impact of variants was annotated using CADD scores [46] and Ensembl Variant Effect Predictor [47]. To increase the genetic map resolution, the genotypes were integrated with sequencing data. More in detail, genotypes were phased using MACH software [48], using 30 iterations of the Markov haplotyping chain and 400 states per iteration. Imputation used *minimac* software [49] and a reference panel including haplotypes estimated by sequencing.

This strategy disclosed about 23 million markers available for analysis on all subjects of the SardiNIA dataset.[42,44]

2. Rare disease

a. *Recruitment of USHER patients*

Usher patients were identified during a large-scale screening on eye-related quantitative traits in the SardiNIA cohort. With a worldwide prevalence of between 4 and 17 :100000, Usher syndrome (USH) resulted to be the most common cause of combined sight and hearing loss, responsible for more than half of deaf-blindness cases.[50 51]

USH encompasses an autosomal recessive group of disorders characterized by congenital sensorineural hearing loss and retinitis pigmentosa (RP), a progressive pigmentary degeneration of rod and cone photoreceptors.

On the basis of the clinical presentation and evolution, USH has been traditionally divided into three subtypes: USH type 1 (25-44% of all USH cases), characterized by severe bilateral hearing function loss and early RP occurring during childhood; USH type 2, the most common form (over half all USH cases) with moderate hearing loss and RP presentation at the adolescent hood; USH type 3 (USH3)

(2-4% of all cases) showing progressive hearing loss and variable degree of vestibular dysfunction with RP, usually occurring in the second decade of life.[52] [53]

So far, USH has been associated with at least 19 loci, with 16 causative genes identified, encoding for a variety of proteins structurally and functionally essential for the cochlear hair cells and photoreceptors.[54]

We deeply analyzed from the ophthalmological and audio metrical point of view 3 families for a total of 22 subjects, 9 of them presenting a pathological phenotype. Five of the nine patients and their available family members were extensively studied at the Institute of Genetic and Biomedical Research of the National Research Council (CNR) in Lanusei, Italy. Our clinical hypotheses and previous phenotypic findings indicated a suspected Usher syndrome type 2.

b. *Clinical examination*

All individuals underwent a complete ophthalmic examination, including best-corrected visual acuity (BCVA) measured with Snellen charts, slit lamp biomicroscopy, funduscopy, fundus autofluorescence (FAF), and spectral domain–optical coherence tomography (SD-OCT; Spectralis; Heidelberg Engineering) and full field electroretinography.

Audiologic evaluation was performed for each patient using a clinical audiometer (Otometrics, Madsen, model Xeta 2, Denmark) according to the reference manual. Tonal audiometry in the frequency range 0.25–8 kHz and higher frequencies at 10 and 12.5 kHz were tested. The severity of hearing loss was classified as mild (20-40 dB), moderate (41-70 dB), severe (71-90 dB) or profound (>91dB).[55]

c. *Molecular screening*

Molecular diagnostics focused on finding causative variants for Usher syndrome was based on whole sequence analysis of the following selected candidate genes: MYO7A, USH1C, CDH23, PCDH15, USH1G, CIB2, USH2A, GPR98, DFNB31, CLRN1, and PDZD7 (Table1).

Syndrome subtype[1]	Locus[2]	Gene name[3]	Protein name[4]
USH1	USH1B	*MYO7A*	Myosin VIIa
	USH1C	*USH1C*	Harmonin
	USH1D	*CDH23*	Cadherin-23
	USH1F	*PCDH15*	Protocadherin-15
	USH1G	*USH1G*	Usher syndrome type-1G protein
	USH1J	*CIB2*	Calcium and integrin-binding family member 2
USH2	USH2A	*USH2A*	Usherin
	USH2C	*GPR98*	Adhesion G-protein coupled receptor V1
	USH2D	*DFNB31*	Whirlin
USH3	USH3A	*CLRN1*	Clarin-1
n/a	n/a	*PDZD7*	PDZ domain-containing protein 7

Table 1. Candidate gene list evaluated for Usher syndrome. [1]Subtype of disorder; [2]Genomic locus [2]; [3]Gene ID by *genecard.org*; [4]Protein name by *UniProtKB/Swiss-Prot*.

Using the UCSC Table Browser web software[56], a BED (Browser Extensible Data) file containing the genomic location (chromosome, initial and final genomic coordinates) of the focused regions was produced. *Bedtools intersect* software[57] was used to select variants located within the regions of interest. Ensembl's VEP (Variant Effect Predictor) software[47] was adopted to annotate the types of variants (SNPs, insertions, deletions, duplications, CNVs or structural variants) by adding for each of them information based on reference databases (dbSNP[58], gnomAD[59],1000 GENOME[60]).

Due to a large number (28,457) of variants available in the candidate genes, selection features were based on Minor Allele Frequency (MAF) < 1% and a strong potential impact on the phenotype (Stop gain, Missense, Frameshift, Splice, Stop lost, Start lost, etc.) of coding or regulatory variants that could be the cause of deleterious consequences for the protein sequence. Functional predictions for amino acid changes were conducted according to several models: SIFT, Polyphen and MutationTaster.

The disease is inherited in an autosomal recessive mode: therefore, variants were filtered for this pattern of inheritance. Sanger sequencing was used to confirm in all family members the supposed causative variant found by whole genome sequence analysis.

d. ***Haplotype reconstruction***

Linkage disequilibrium (LD) block construction and haplotype population frequency estimation was performed via *Haploview 4.2*[61]. For accurate haplotype determination, from 3514 Sardinian sequenced samples in the initial cohort, a representative subset of 1454 unrelated individuals were pooled as listed by Chiang et al.[62] using *bcftools 1.7*[63]. We selected a genomic interval of approximately 30kb, where the causal variant rs764182950 was located, through *PLINK v1.90*[64].

Block size definition was carried out using the default algorithm "confidence intervals"[65]. We set the standard parameters by including all markers independently of the MAF value and examined haplotypes exceeding frequencies above 0.9% to detect the causative mutation-related.

e. ***Molecular modelling (Homology models and Free energy calculation)***

The Iterative Threading ASSEmbly Refinement (I-TASSER) server (https://zhanglab.ccmb.med.umich.edu/) was utilized to build the models using default settings[66]. Pymol (https://pymol.org/) was used to display, analyze the built Usherin model and to build the mutated protein. To calculate the free energy, we used the following tools using the 3D structure of the usherin protein domain 3210-3402. DUET consolidates two complementary approaches mCSM and SDM in a consensus prediction, obtained by combining the results of the separate methods in an optimized predictor using Support Vector Machines (SVM)[67] [68]. Mupro developed two machine learning methods: Support Vector Machines and Neural Networks and for the calculation, we used the amino acid structure of all protein[69].

3. Complex disease

We evaluated the colocalization of several large GWAS. We combined summary statistics from the immunophenotype GWAS [42] on the SardiNIA cohort[70] and 44 publicly available case-control studies obtained from *The NHGRI-EBI GWAS Catalog*[13]. Publicly available GWAS were performed on European ancestry samples characterized with genetic and phenotypic profiles.

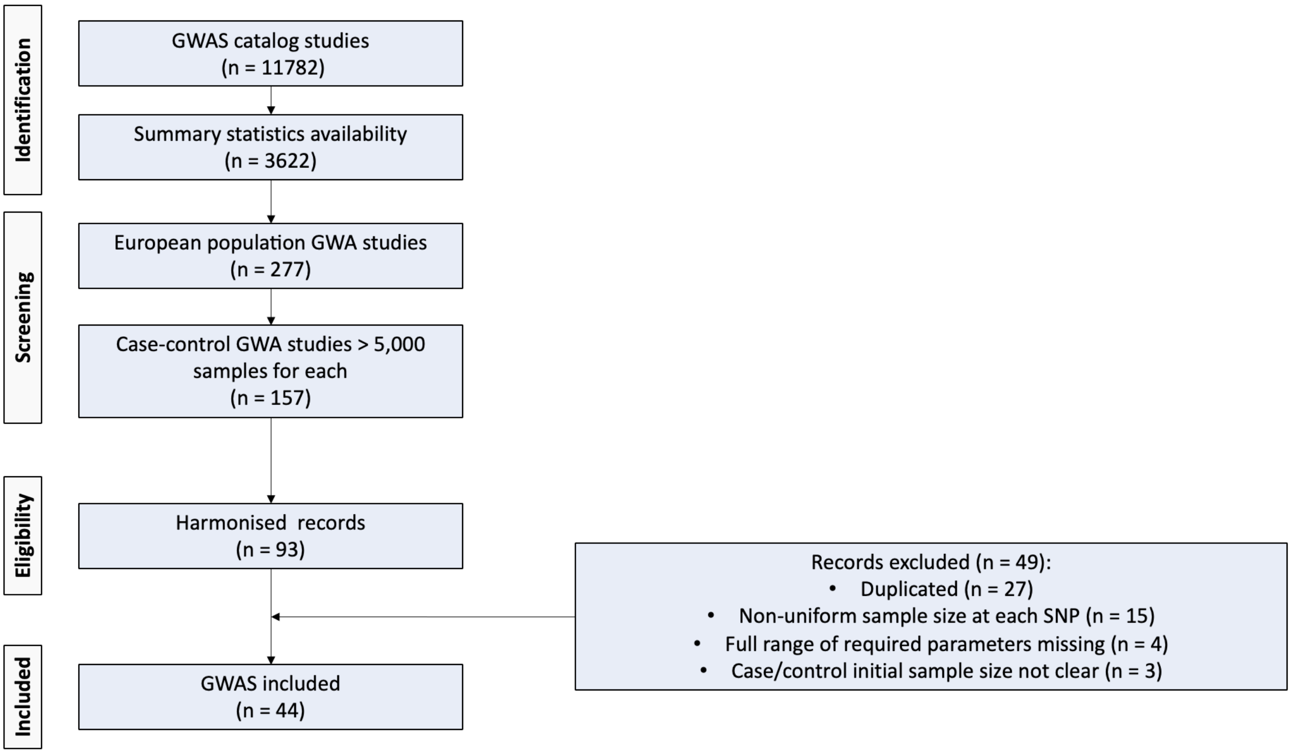
In the SardiNIA cohort, 731 immunophenotypic traits were measured on 95 types of immune cell subpopulations with the GWAS summary statistics collected. Immune traits were obtained from cellular subtypes of the major leukocyte subpopulations in peripheral blood, including monocytes, granulocytes, circulating dendritic cells (cDCs), natural killer (NK) cells, B cells, and T cells.[38 42]

We performed the analysis using an approach without biological assumptions to have higher chances to identify shared associations between specific immune traits and diseases, not immediately related to the immune phenotype. Both datasets allowed us to assess the presence of a biological affinity of quantitative immune traits with traits from case-control GWAS catalog studies and to test the involvement of immune system cells in these traits. Therefore, these datasets generate the ideal conditions to perform colocalization aimed at identifying the molecular mechanism of overlapping associations.

a. Data collection and standardization

Trait-locus signals found significantly associated with a p-value < 5x10-8 in Orrù 2020 were tested. For each of the 219 loci, there are several traits associated, because multiple traits can be associated with a locus and vice versa, for a total of 1075 unique trait-locus association signals.

A systematic approach was adopted to include studies from the GWAS catalog (see Diagram "Data selection pipeline"). To create our dataset, from all studies (11782 in March 2021) (see https://www.ebi.ac.uk/gwas/docs/file-downloads), we extracted 3622 studies from a list of published studies in which summary statistics were available (see https://www.ebi.ac.uk/gwas/downloads/summary-statistics).


Identification
GWAS catalog studies (n = 11782)
Summary statistics availability (n = 3622)
Screening
European population GWA studies (n = 277)
Case-control GWA studies > 5,000 samples for each (n = 157)
Eligibility
Harmonised records (n = 93)
Records excluded (n = 49):
• Duplicated (n = 27)
• Non-uniform sample size at each SNP (n = 15)
• Full range of required parameters missing (n = 4)
• Case/control initial sample size not clear (n = 3)
Included
GWAS included (n = 44)

Data selection pipeline

i. *European ancestry's selection*

We restricted the analysis to studies with samples from European ancestry only, excluding other populations (277 studies out of 3622 with summary statistics). European ancestry's population is the genetically closest to the Sardinian population: if samples differ by ancestral and demographic attributes, markers could be misled to disease status leading to spurious associations[71].

ii. *Case-control GWA studies*

Among the 277 studies with only Europeans, we kept 157 case-control GWA studies with more than 5,000 cases and 5,000 controls. Studies with lower sample sizes were excluded to focus only on robust colocalization signals.

Population-based case-control studies help to find common polymorphisms supposed to underlie complex traits. [72]. Discrete (case-control) phenotypes give the advantage to have a group of individuals (cases), ascertained for the phenotype of interest and presumed to have prevalence of susceptibility alleles for that trait. It can be useful to detect small-effect genes when genes with a much larger effect are present[73]. Thus, it is possible to identify the association between the samples that exhibit marker alleles and immunological traits obtained from the Sardinian population.

iii. Harmonized summary statistics data

A limitation in the use of public data is the lack of measurements standardization and the methodological approaches used in each study. The most important standardization is a clear definition of the effect/risk allele and the correspondent sign of the Beta score coefficient. To minimize the risk of getting the wrong definition we only used harmonized summary statistics. GWAS catalog provides most of the studies in a harmonized format, by means of an automated pipeline from *Open Targets*[74] that ensure that alleles are strand oriented, the effect/risk allele is consistent and data without a valid value for variant ID, chromosome, base pair position and p-value were removed. [75] [76] We focused on 93 studies whose summary statistics were harmonized.

Analogously, linkage disequilibrium must be calculated on the genetic variants of a given population. The best method of analysis by having data from a public resource, to assess the LD value between SNPs is to calculate it from the same genetic dataset that originated the summary statistics. Because often these datasets are not publicly available, an alternative option is to utilize datasets whose samples have a close genetic background, e.g. from 1000 genomes or UK10K for European populations.

iv. Duplicate studies

27 duplicate studies performed by the same author on the same or different phenotypic trait and a different number of samples were removed. If the traits were identical, GWAS with a larger sample size was preferred. Whereas when traits differed, both studies were selected. If the studies provided the summary statistics on grouped and non-grouped traits, the preference fell on GWAS with larger sample size and where the samples were not paired according to traits. If the study was performed both separately on males and females and on the whole cohort, we used ones where males and females were grouped. Where studies were carried out adjusting statistic parameters (e.g. BMI), the selection was towards the adjusted study in which confounding effects were removed. When the phenotypic trait was both self-reported and clinically diagnosed, we selected data in which the pathological trait had been clinically defined.

v. Parameter's inspection

Although some studies (GCST000392, GCST000879, GCST001255, GCST002548) featured harmonized files, they were excluded because lacked the full range of parameters required by colocalization tools such as beta value, Z Score or standard error but were limited to P-value.

vi. Sample size checking

In GWAS catalog, the sample size can be indicated in two different ways: it can be indicated for the whole set of SNPs or can be indicated for each single variant tested (e.g. in a meta-analysis some variants could be present in some cohorts but missing in others). Among the harmonized studies, several were missing the sample size at each SNP, while others (21) reported the sample size for each variant but using different formats between studies. We evaluated these cases manually.

As Pickrell 2014[77], because for some studies the sample size at each SNP was not reported, we used the overall study sample sizes as approximations of the sample size at each SNP. For datasets with a different sample size for each SNP, we approximated it to be uniform for all variants. Three studies (GCST009722, GCST003769, GCST005536), with case/control initial sample size not verifiable, were omitted. Our definitive analysis dataset includes 44 GWAS studies (Supplementary Table1).

b. Colocalization tools

A systematic review identified a list of colocalization tools: *"coloc"* v2.3-1 package"[32], *"gwas-pw*" v0.21"[33], and *"eCAVIAR"* v2.2" [34]. Software selection was supported by their published results. It has been proven the tools' ability to replicate and support previous findings.

An important feature of the 3 software is the possibility to use summary statistics to perform analyses, without need to access individual genotypes. All of them assume a Bayesian approach using different parameters to evaluate the colocalization signal. The authors evaluated the LD and its application in different manner. The colocalization score is related to each tool according to the false positive and negative rates obtained from their outcomes.

"Coloc" is a statistical method that assesses whether two association signals are consistent with a shared causal variant using two different sets of parameters. For SardiNIA quantitative traits, for each SNP we used: the beta regression coefficient, its variance (or the square of standard error) the minor allele frequency (MAF), the number of samples and we set the type of data as quantitative. For case-control datasets, we employed: P-value, number of samples, MAF, type of data ("cc") and the proportion of cases. We used the colocalization cut-off value of 0.80 as recommended by the developers. LD is not used for calculation of colocalization because *"Coloc"* assumes that samples originate from the same ethnic cohort, i.e., allele frequencies and linkage disequilibrium (LD) pattern

are similar in both populations. Another important limiting assumption of “coloc” is that the model evaluates only one causal variant inside a genomic region.

”gwas-pw” software is a method to identify pairs of traits that show evidence of a causal relationship. It uses as input the Z score, which measures the evidence of SNP association with phenotype, and similarly to “coloc” the variance of the effect size or squared standard error. The colocalization cut-off is set to 0.95. *”gwas-pw*” is based on a *“coloc”* similar model with the requirement of splitting the genome into independent, non-overlapping blocks. The authors suggest to use precomputed LD blocks from the 1000G study to ensure that the LD does not impact the results. However, also *”gwas-pw*” assumes that summary statistics derive from a population with the same genetic background and there is only one causal variant in the region under analysis.

In loci where the LD differs between the European and Sardinian populations, the outcome could be inconsistent indicating that some associations are not coincident because of the different LD patterns. To overcome this issue and to make the analysis more reliable, we used *“eCAVIAR”*, a probabilistic method that manages more than one causal variant at a given locus and can measure the likelihood of different numbers of polymorphisms[78]. This implies the knowledge of LD between different variants.

Since *“eCAVIAR”* requires LD, it can only be used when genotype data are available. In the SardiNIA cohort, we calculated the LD panel directly from whole genome genotypes. In GWAS public meta-analysis, we use genotype data from the UK10K study as proxy for European population[79]. We converted haplotype maps to *variant caller format* with *bcftools* v1.7[63] and finally using *plink* v1.90b6.6[64] we generated the LD files. In addition to LD, *“eCAVIAR”* uses the z-score and the maximum number of causal SNPs. We set *“eCAVIAR”* to evaluate 1 causal variant for each locus to have results consistent with *“coloc”* and *”gwas-pw*” because they assume that each locus contains one causal variant. We set the *“eCAVIAR”* colocalization cut-off at 0.01 according to the authors' guidelines.

c. Colocalization analysis workflow

We developed a processing workflow to automate and optimize data collection, interrogation, and analysis execution to minimize handling biases. The pipeline includes several steps.

Following the GWAS harmonized data import, we converted the genome SNPs position from assembly hg38 to hg19 using liftover[80], being the SardiNIA data produced on the hg19 reference panel. Genomic regions were considered as an arbitrary region of 50kb left and right from a signal peak identified by the top SNP at the locus. Where multiple traits have been associated to a single locus, we selected 50kb left from the leftmost top variant and 50kb right from the rightmost top variant. We then filtered variants shared in each dataset possessing the necessary summary statistics.

We standardized the input data to create different tabular files for each software using purpose-built python and UNIX shell scripts. Since each locus can be associated with more than one trait and vice versa, we performed colocalization considering 1075 locus-trait pairs found significantly associated in Orrù 2020, with shared variants found in each of the 44 studies selected from the GWAS catalog resulting in a total of 47300 colocalization analyses.

Finally, we compared the results obtained to assess the software concordance.

Results

1. Rare disease

Our goal was to uncover the molecular cause of the Usher phenotype exhibited by the patients.

Five patients affected by Usher syndrome (4 women and 1 man; aged from 38 to 61 years) and several healthy relatives from three Sardinian families living in Lanusei valley, were clinically examined (Fig.1).

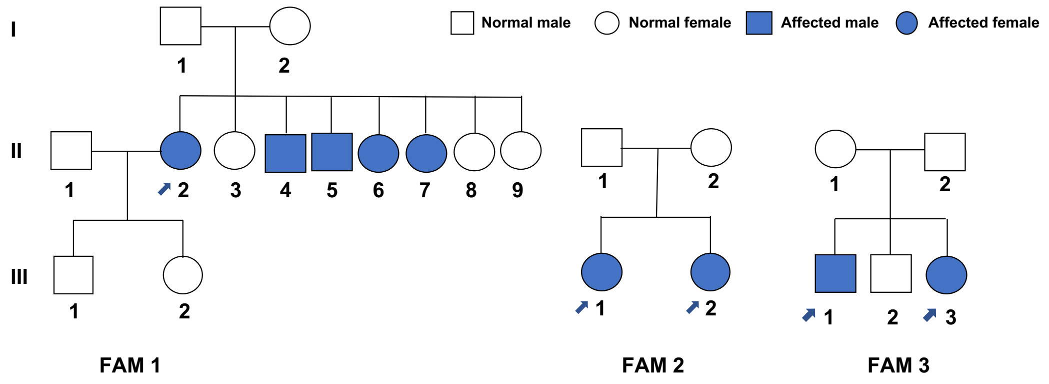


Fig. 1 | The three Sardinian pedigrees with members affected by Usher Syndrome due to USH2A mutations. In pedigree 1, we were able to find clinical/phenotypic information for three generations, but only for two generations in family 2 and 3.

All patients complained of nyctalopia as initial symptom occurred during adulthood followed by a gradual vision loss and constriction of peripheral vision, in the late stages of the disease. Best-corrected visual acuity ranged from 1/30 and 10/10 (ETDRS letters), thereby being lower in older patients.

Fundus examination showed typical RP features including pigmentary changes in the peripheral and mid-peripheral retina, attenuated arteriolar vessels and pallor of the optic disc. Fundus autofluorescence revealed a macular hyper-autofluorescent ring around the fovea and hypo-autofluorescence within and extending outward of vascular arcades.

SD-OCT scans demonstrated outer retinal atrophy with centrally preserved photoreceptor inner segment ellipsoid bands. (Fig. 2) Moreover, cystoid macular edema (CME) not responsive to topical and oral carbonic anhydrase inhibitors was detected in several eyes.

The summary of the clinical findings of all patients is shown in Table 1.

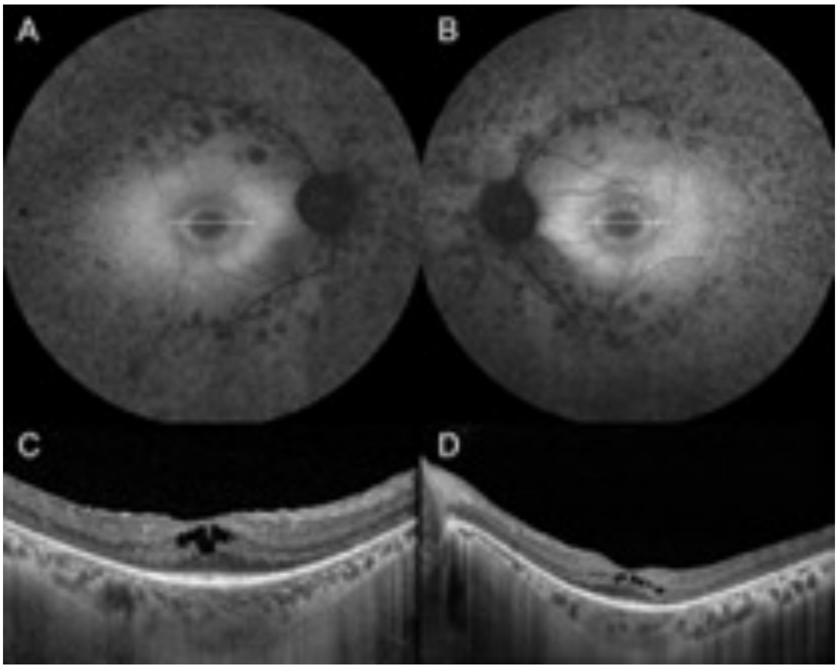


Fig. 2 | Multimodal imaging of a 38-year-old FAM III.3.3 patient affected by Usher Syndrome. (A, B) Fundus autofluorescence revealed a macular hyperautofluorescent ring around the fovea and hypoautofluorescence within and extending outward of vascular arcades, bilaterally. (C, D) In both eyes, simultaneous infrared and SD-OCT scans demonstrated outer retinal atrophy with centrally preserved photoreceptor inner segment ellipsoid bands. Note the presence of bilateral intraretinal cystoid spaces in the foveal area (see white line in A and B).

Audiologic evaluation demonstrated the presence of a bilateral profound/severe hearing loss in patients FAM-2 II.1 (Dx 91 dB/ Sx 73 dB) and FAM-2 II.2 (Dx 70dB/ Sx 56 dB) patients respectively, and moderate in the remaining patients (FAM-1 II.2 - Dx 53dB/ Sx 51 dB; FAM-3 III.3 Dx 53dB/ Sx 51 dB). Healthy relatives of these patients showed normal values (for example FAM-3 III.2 Dx 16/ Sx 18 dB) (Fig.3). We don't have enough reports to prove it, but we have noticed an exacerbation of age-related deafness.

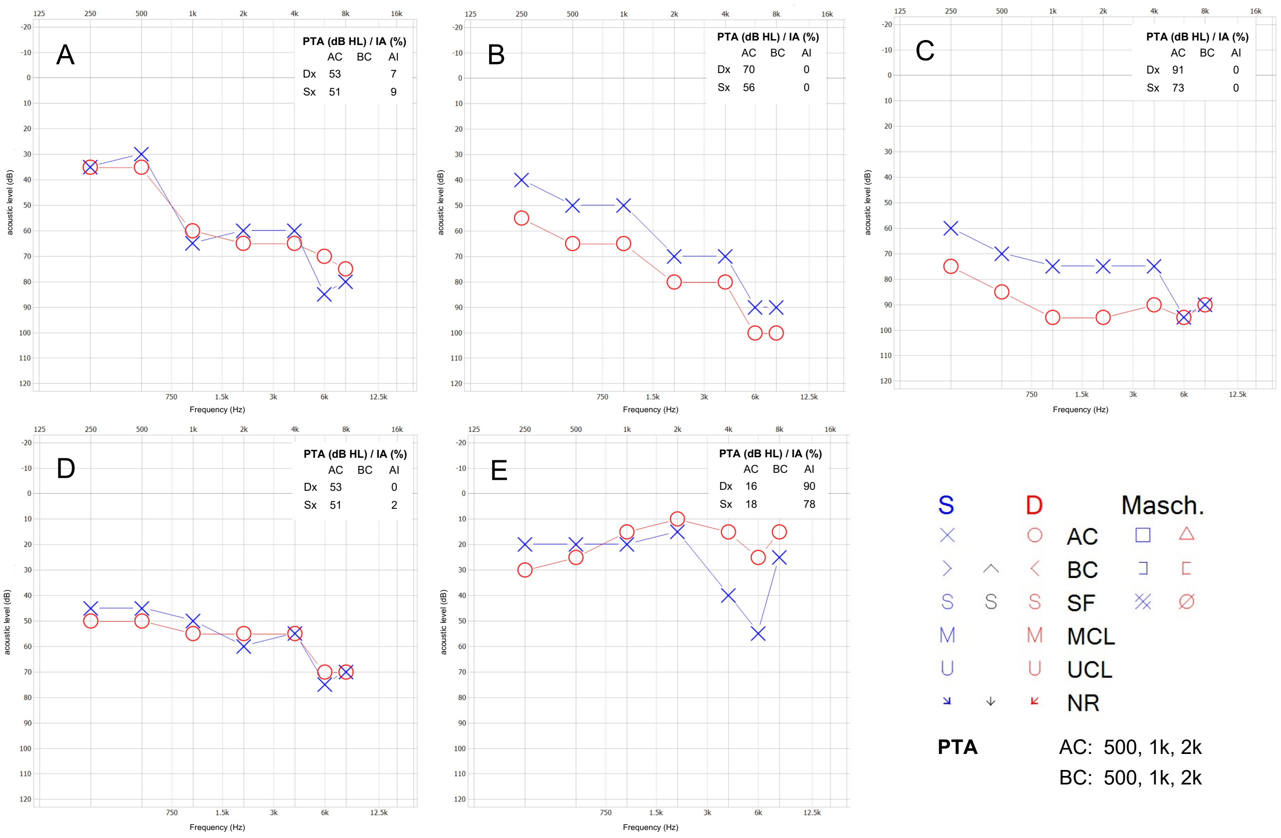


Fig. 3 | Audio profiles of Usher syndrome type 2A families. The hearing loss of patient's test (FAM-1 II.2 (A), FAM-2 III.1 (B), FAM-2 III.2 (C), FAM-3 III.3 (D)) was performed using Otometrics, Madsen Xeta ^{2} clinical audiometer. We added a healthy subject of family 3 (FAM-3 III.2 (E)) for comparison.

We selected homozygous variants for the alternative allele (both alleles mutated) and/or heterozygote compounds for two heterozygous alleles in patients versus reference homozygotes (i.e., both alleles healthy) and single heterozygotes (only one allele mutated) in healthy individuals in all coding sequences of the candidate genes we selected.

Based on autosomal recessive genetic inheritance, we do not identify candidate causative mutations except for the USH2A gene. We found several missense and/or splicing mutations in the USH2A gene (Table 2) but only two of them are predicted to be deleterious: c.9815C>T(p.Pro3272Leu) and c.1663C>G(p.Leu555Val). Based on the autosomal recessive model of inheritance, we identified a single potentially causative variant on the gene encoding for the Usherin protein the p.Pro3272Leu variant.

Chromosome Position[1]	Reference allele[2]	Alternate allele[3]	rs ID[4]	Gene region[5]	Coding Effect[6]	Protein change[7]	gnomAD Allele Frequency[8]	Polyphen[9]	SIFT[10]
1:215914826	T	C	rs35309576	exonic	missense	p.Met3868Val	0.2299	benign	tolerated
1:215916563	G	A	rs11120616	exonic	missense	p.Thr3835Ile	0.2284	benign	tolerated
1:215960167	T	G	rs10864198	exonic	missense	p.Glu3411Ala	0.5218	benign	tolerated
1:215972392	**G**	**A**	**rs764182950**	**exonic**	**missense**	**p.Pro3272Leu**	**0.00004406**	**probably_damaging**	**deleterious**
1:216172380	A	G	rs10864219	exonic	missense	p.Ile2169Thr	0.5006	benign	tolerated
1:216219781	A	G	rs6657250	exonic	missense	p.Ile2106Thr	0.6716	benign	tolerated
1:216258213	A	G	rs56222536	exonic	missense	p.Ile1665Thr	0.1417	benign	tolerated
1:216348764	C	T	rs1805049	exonic	missense	p.Arg1486Lys	0.6241	benign	tolerated
1:216371934	A	C	rs646094	splice region	splicing		0.2335	-	-
1:216465694	G	C	rs35818432	exonic	missense	p.Leu555Val	0.0017	probably_damaging	deleterious
1:216595306	C	T	rs10779261	exonic	missense	p.Ala125Thr	0.7060	benign	tolerated

Table 2. Genetic variants in *USH2A* filtered by functional consequences: Stop_gain, Missense, Frameshift, Splice, Stop lost, Start lost. The causal variant is shown in bold. [1]Genomic coordinates based on human assembly GRCh37 (hg19); [2]Reference allele; [3]Aternative allele; [4]Reference SNP number; [5]Gene location of variants; [6]Consequence of variants on the protein sequence; [7]Amino acid replacement; [8]Minor allele frequency based on gnomAD; [9-10]Transcript-specific prediction using Polyphen and SIFT present in gnomAD.

This pathogenic mutation causes base substitution in an exonic region of the gene, and the functional predictions for amino acid changes according to SIFT and Polyphen were probably damaging and deleterious, respectively. Furthermore, the frequency of this variant in the gnomAD database v2.1.1 is extremely rare (0.00004406 %) in agreement with syndrome frequencies in general population. All patients in the three families analyzed are homozygous for the c.9815C>T mutation. Based on the consideration that it is a rare syndrome, nine patients from the same geographical area affected by the same mutation assume a founder effect additionally proved by the identification of an extended haplotype.

Haplotype analysis surrounding the causal SNP, rs764182950, indicated that the haplotype "TCTCTGCACACAAC" (1,128 bp) consisting of 14 variants (rs11120635, rs184464579, rs189009821, rs909035231, rs12129576, rs12132110, rs745566865, rs764182950, rs115806383, rs11120637, rs4363405, rs74447991, rs6686574, rs12140781) represents a risk factor for Usher syndrome with a frequency of 1% in SardiNIA cohort (Fig 3).

The frequencies of p.Pro3272Leu heterozygous subjects were 2,29 % and 1,89 % in the SardiNIA cohort vs a cohort from whole island, respectively.

rs11120635	*rs184464579*	*rs189009821*	*rs909035231*	*rs12129576*	*rs12132110*	*rs745566865*	***rs764182950***	*rs115806383*	*rs11120637*	*rs4363405*	*rs74447991*	*rs6686574*	*rs12140781*	
G	C	T	C	T	G	C	G	C	A	A	A	G	C	.384
T	C	T	C	T	G	C	G	C	A	C	A	A	C	.332
T	C	T	C	T	G	C	G	C	A	C	A	G	C	.122
T	C	T	C	T	G	C	G	C	G	C	A	G	C	.099
T	C	T	C	C	A	C	G	C	G	A	T	G	A	.042
T	C	T	C	T	G	C	**A**	C	A	C	A	A	C	**.010**

Fig 3 | Haplotype reconstruction using Haploview. Vertical green column shows the reference allele, horizontal red row indicates the haplotype nucleotide sequence. The alternative allele of the causal variant is shown in red and bold. Values on the right indicate the percentages of each haplotype.

Molecular dynamics simulations of native and mutant protein-protein and protein-ligand complexes were performed. Currently, three-dimensional structure (x-ray, NMR or single particle cryo-electron microscopy) is not available, because of usherin protein's large size, membrane residence, and potential flexible conformations. To carry out our in-silico studies, we have computationally constructed a homology model of the usher protein domain from amino acid 3210 to 3402 in which the studied mutation is located (P3272L). In addition, the mutated protein was also modelled (see Fig. 4). To understand the impact of the mutation on protein structure and function, we applied in silico methodologies to study and predict the effects of SNPs on protein stability in particular the free energy change (DDG). All the tools used predicted that the amino acid change from proline to leucine (P3272L) results in a destabilization of the protein with a decrease in DDG. The tools used were DUET, mCSM SDM and CUPSAT with DDGs of -0.212, -0.377, -0.58, -2.51 Kcal/mol respectively. In addition, the Mupro tool was also used to establish DDG using only the amino acid sequence of all protein and again a destabilization was predicted with DDG =-037 using Vector Machine and -0.78 using Neural Network.

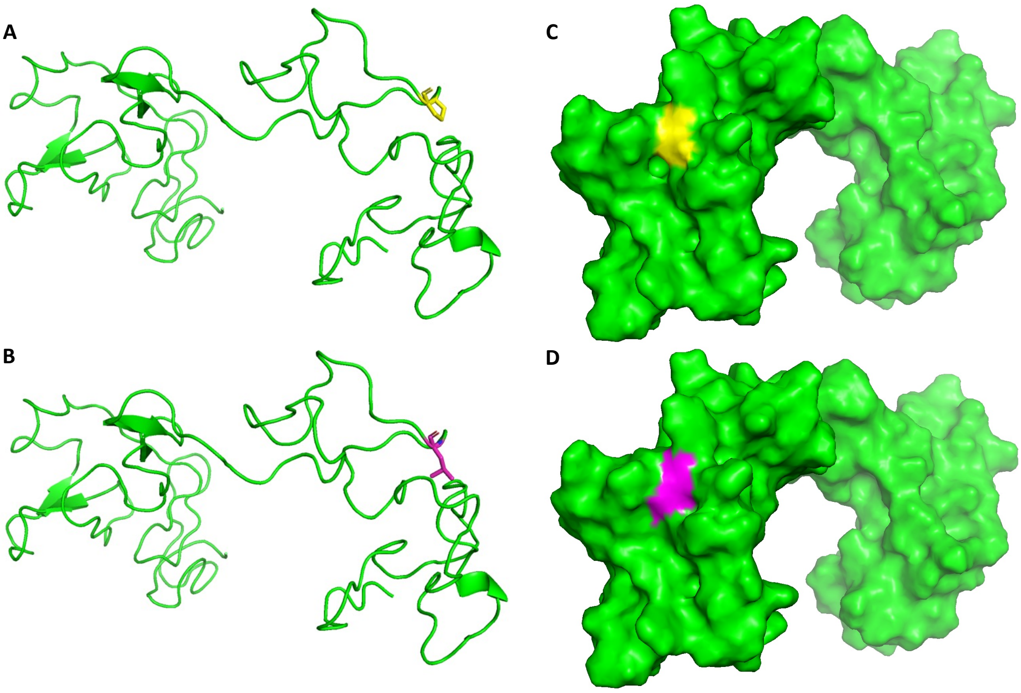


Fig. 4 | The 3-dimensional structure built through homology model of the domain 3210-3402. A: wild type protein in yellow the Proline 3272 B: mutate protein in magenta the Leucine C: surface wild type protein in yellow the Proline 3272 D: surface mutate protein in magenta the Leucine

2. Complex traits

Our goal was to compare the colocalization results obtained by the 3 software, assess the concordance between them and evaluate their impact on results.

Overall, we evaluated 47,300 trait-locus colocalizations. In 7,199 loci, there were no variants in common between the two datasets that would allow colocalization to be assessed. Using *”gwas-pw*” 35 trait-locus combinations with two studies, GCST007800 and GCST007799, failed because the software stopped in genomic regions containing SNPs with a Zscore value less or greater than 40, probably caused by numerical problems. When Zscore value reached extreme levels, the p-value becomes lower than 1e-300, and *”gwas-pw*” cannot calculate colocalization’s value.

For large pleiotropic loci with multiple associated traits (more than 1 Mb contain a large number of variants), *“eCAVIAR”*, took over 1 week determining colocalization, but were included in the analyses.

Finally, we considered 40,066 trait-locus associations colocalizing (YC) or not-colocalizing (NC) with 44 GWA studies (Fig. 5).

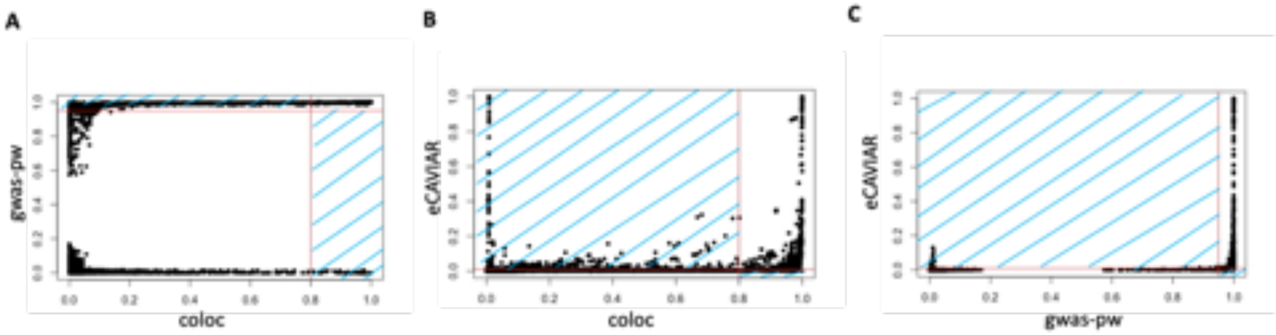


Fig. 5 | Colocalization algorithms’ outcomes

Scatter plots of colocalization outcomes were obtained by comparing three algorithms in pairs: “coloc vs gwas-pw”(A), “coloc vs eCAVIAR”(B) and ”gwas-pw vs eCAVIAR”(C). Red lines indicate the colocalization cut-off for the corresponding tool. We set the cut-off value following the authors' guidelines: 0.80 for “Coloc”, 0.95 for “gwas-pw” and 0.01 for “eCAVIAR”.

Concordant colocalization signals among the 3 tools were 964 positives and 35359 negatives. *“coloc”* identified 1255, *”gwas-pw*” 4462 and *“eCAVIAR”* 1910 positive outcomes. Negative outputs were 38811 in *“coloc”*, 35604 in *”gwas-pw*” and 38156 in *“eCAVIAR”*. Venn diagram identifying the overlapping outcomes among 3 analyzed tools (Fig. 6).

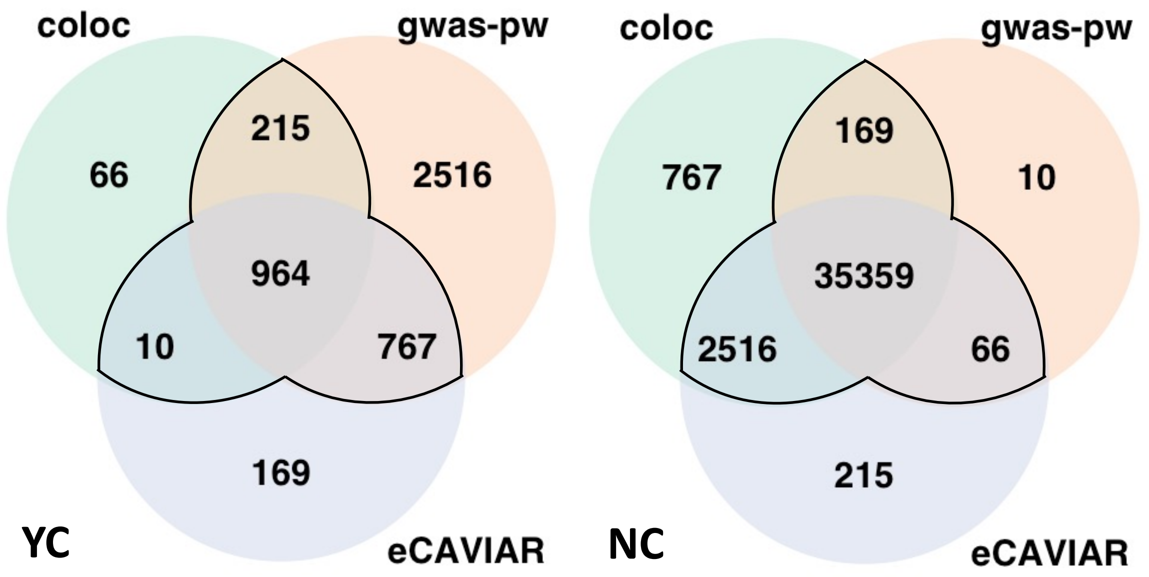


Fig 6 | Venn diagrams

Venn diagrams show trait-locus associations colocalizing (YC) or not-colocalizing (NC) with 44 GWAS catalog studies for the 3 tools. “coloc” in green,” gwas-pw” in orange and ”eCAVIAR” in blue. Black ellipses show the concordant findings in at least two out of 3 tools. We considered these as true positive trait-locus associations colocalizing (YC) or not-colocalizing (NC).

We considered as true positive the concordant findings in at least 2 out of 3 tools, i.e. the trait-locus associations colocalizing (YC) or not-colocalizing (NC) for at least 2 of the 3 tools.

To calculate the concordance, we evaluated each combination between the 3 software (See table3).

Combinations	*coloc*	*gwas-pw*	*eCAVIAR*	Count
A	< 0.8	>= 0.95	>= 0.01	767
B	>= 0.8	>= 0.95	< 0.01	215
C	>= 0.8	>= 0.95	>= 0.01	964
D	>= 0.8	< 0.95	>= 0.01	10
E	< 0.8	< 0.95	< 0.01	35359
F	< 0.8	>= 0.95	< 0.01	2516
G	>= 0.8	< 0.95	< 0.01	66
H	< 0.8	< 0.95	>= 0.01	169
Tot	-	-	-	40066

Table 3. Combinations results between software. From left to right: ID combination; coloc; gwas-pw; eCAVIAR; the number of findings.

Concerning *“coloc*” and *”gwas-pw*”, we summed up positive concordant results, B and C, and negative, E and H, dividing by the total. Likewise, we carried out the same procedure for the other tools pairs. We observed that the three tools showed comparable results: in fact, the concordance between *“Coloc”* and *”gwas-pw*” was 91%, between *“coloc”* and *“eCAVIAR”* was 96% and between *”gwas-pw*” and *“eCAVIAR”* 92%.

One limitation of employing the above strategy is that some outputs may not be consistent. One software may detect a positive result and the others miss it. Nevertheless, we are aware that if two tools based on different conceptual approaches give overlapping results, these findings are robust. Furthermore, replication across multiple datasets involving the same phenotype assumes a strong and clearly identifiable biological basis. Biological/functional testing will provide definitive evidence.

We focused on *“coloc”* and *“eCAVIAR”*, which showed the highest concordance (96%) with 974 shared positive results. Notably, the 98.9% of these results (964/974) are shared by all the 3 software used. These findings have been obtained using different disease datasets, some of which concern the same disease/phenotype.

Because of the wide range of methodologies used to determine quantitative immune traits, it is unlikely to find in the literature QT phenotypes that can be directly compared with those measured in the SardiNIA cohort.

Even when applying colocalization to the same QT phenotypes as in Orrù et al 2020, the exact overlap between results obtained here is complicated because of the different study design and the statistical analyses applied in the two studies. For these reasons, it was decided to compare the positive signals obtained here with those in Orrù et al. 2020 exclusively from a biological perspective, i.e. to assess the QTs-associated loci that colocalize with the pathological phenotypes independently of the criteria and/or dataset adopted.

QT-Locus combinations						
Binary traits (Diseases)	ID	Positive results	Replicated (+/+)	Uncertain/Confirmed (?/+)	NA/New (?/+)	Not/New (-/+)
ALLERGY		**99**	**92**	**7**	**0**	**0**
Allergic disease (asthma, hay fever or eczema)	GCST005038	48	48	0	0	0
Asthma	GCST006862	13	13	0	0	0
Asthma (adult onset)	GCST007799	7	7	0	0	0
Asthma (childhood-onset)	GCST007800	31	24	7	0	0
CD		**167**	**149**	**0**	**11**	**7**
Crohn's disease	GCST003044	105	105	0	0	0
Crohn's disease	GCST004132	62	44	0	11	7
IBD		**131**	**124**	**0**	**1**	**6**
Inflammatory bowel disease	GCST003043	101	101	0	0	0
Inflammatory bowel disease	GCST004131	30	23	0	1	6
UC		**201**	**179**	**15**	**1**	**6**
Ulcerative colitis	GCST000964	65	59	1	0	5
Ulcerative colitis	GCST003045	95	81	14	0	0
Ulcerative colitis	GCST004133	41	39	0	1	1
MS		**109**	**58**	**16**	**0**	**35**
Multiple sclerosis	GCST001198	50	30	8	0	12
Multiple sclerosis	GCST005531	59	28	8	0	23
RA		**87**	**78**	**3**	**0**	**6**
Rheumatoid arthritis	GCST000679	63	59	0	0	4
Rheumatoid arthritis	GCST005569	24	19	3	0	2
SLE		**67**	**67**	**0**	**0**	**0**
Systemic lupus erythematosus	GCST003156	67	67	0	0	0
T1D		**26**	**16**	**5**	**0**	**5**
Type 1 diabetes	GCST010681	26	16	5	0	5
TOT		**887**	**763**	**46**	**13**	**65**

Table 4. Summary table of positive results compared to Orrù et al. 2020. The columns mean (from left to right): disease phenotype reported, grouped by phenotype (in bold) and for each dataset; GWAS catalog dataset identifiers; positive results in each dataset; positive replicated results; uncertain results confirmed positive; missing results found positive; conflicting results found positive.

A list of all observed overlapping associations is provided in Table 4. Of 974 positive outcomes from *"coloc"* and *"eCAVIAR"*, 887 correspond to the same diseases treated by Orrù et al. 2020. Among them, 634 are non-redundant among diseases (Table 5).

From a biological perspective, the study replicated 524 (82,6%) previous findings, confirmed 41 (6,5%) uncertain signals and identified 69 (10,9%), including 13 (2,1%) not present and 56 (8,8%) negative results in Orrù et al. 2020. These new QT-disease associations will be characterized biologically from a functional perspective.

QT-Locus combinations							
Binary traits (Diseases)	ID	Positive results	Overall positive non-redundant	Replicated (+/+)	Uncertain/Confirmed (?/+)	NA/New (?/+)	Not/New (-/+)
ALLERGY		**99**	**65**	**58**	**7**	**0**	**0**
Allergic disease (asthma, hay fever or eczema)	GCST005038	48					
Asthma	GCST006862	13					
Asthma (adult onset)	GCST007799	7					
Asthma (childhood-onset)	GCST007800	31					
CD		**167**	**123**	**105**	**0**	**11**	**7**
Crohn's disease	GCST003044	105					
Crohn's disease	GCST004132	62					
IBD		**131**	**108**	**101**	**0**	**1**	**6**
Inflammatory bowel disease	GCST003043	101					
Inflammatory bowel disease	GCST004131	30					
UC		**201**	**103**	**81**	**15**	**1**	**6**
Ulcerative colitis	GCST000964	65					
Ulcerative colitis	GCST003045	95					
Ulcerative colitis	GCST004133	41					
MS		**109**	**70**	**32**	**11**	**0**	**27**
Multiple sclerosis	GCST001198	50					
Multiple sclerosis	GCST005531	59					
RA		**87**	**72**	**64**	**3**	**0**	**5**
Rheumatoid arthritis	GCST000679	63					
Rheumatoid arthritis	GCST005569	24					
SLE		**67**	**67**	**67**	**0**	**0**	**0**
Systemic lupus erythematosus	GCST003156	67					
T1D		**26**	**26**	**16**	**5**	**0**	**5**
Type 1 diabetes	GCST010681	26					
TOT		**887**	**634**	**524**	**41**	**13**	**56**

Table 5. Overall positive non-redundant results among disease compared to Orrù et al. 2020. The columns mean (from left to right): the disease reported, grouped by phenotype (in bold) and for each dataset; GWAS catalog dataset identifiers; positive results in each dataset; positive replicated results; uncertain results confirmed positive; missing results found positive; conflicting results found positive.

The replication rate reached 83%. If we consider the uncertain confirmed positive results and the replicated ones, the percentage rises to 89%. The discovery rate is around 11% considering both results not present or not significant in Orrù et al. 2020.

In addition, assessing other phenotypes not covered in Orrù et al. 2020, we found 87 positive outcomes, of which 63 are not redundant among diseases (Table 6).

QT-Locus combinations			
Binary traits (Diseases)	ID	Positive results	Overall positive non-redundant
Alzheimer's disease (late-onset)	GCST002245	**22**	**22**
Falling risk	GCST90012857	**1**	**1**
Insomnia complaints	GCST004695	**3**	**3**
Low hand grip strength (60 years and older) (EWGSOP)	GCST90007526	**1**	**1**
Osteoarthritis		**3**	**3**
Osteoarthritis (hip)	GCST007091	1	
Osteoarthritis (hospital diagnosed)	GCST005814	2	
Breast cancer	GCST004988	**8**	**8**
Endometrial cancer	GCST006464	**8**	**8**
Prostate cancer	GCST006085	**1**	**1**
Psoriasis	GCST005527	**8**	**8**
Stroke			**8**
Stroke	GCST006906	8	
Ischemic stroke	GCST006908	8	
Ischemic stroke (small-vessel)	GCST006909	8	
Heart failure	GCST009541	8	
TOT		**87**	**63**

Table 6. Summary positive results. The columns mean (from left to right): trait reported, grouped by phenotype (in bold) and for each dataset; GWAS catalog dataset identifiers; positive results in each dataset; overall positive non-redundant results.

In deep analysis of a trait: features of genetic regulation of CD4 and CD8 cells.

Here is an example, concerning the allergy phenotype, of uncertain results in Orrù et al 2020 that we confirmed as positive. These signals are particularly interesting because they show how important is the dataset (i.e. the genetic coverage or the sample size considered in the study) in terms of detecting signals. Using the study (GCST005038) employed in Orrù et al. 2020, which concerns a general phenotype (Allergic disease, asthma, hay fever or eczema), we replicated 48 positive results. On the other hand, when we analyzed the dataset (GCST007800) regarding a specific phenotype (Asthma childhood-onset), we found 7 additional positive signals that were characterized as uncertain in Orrù et al. 2020.

A signal led by rs1886730[C] in the *TNFRSF14* intronic region increased the level of lymphocyte T (HVEM on CD4+ cells, effect 0.2928, P = 5.643e-14) and colocalized with decreased risk for asthma (childhood-onset)[81]. Interestingly, the same signal was repeated for subclasses of CD4 and for CD8 increasing the levels of lymphocyte T (HVEM on EM CD8br, effect 0.2857, P = 8.552e-13; HVEM on CM CD4+, effect 0.2857, P = 1.907e-13; HVEM on CD45RA- CD4+, effect 0.3004, P = 2.172e-14; HVEM on CM CD8br, effect 0.3092, P = 2.104e-15;). The encoded protein functions in signal transduction pathways that activate inflammatory and inhibitory T-cell immune response[82].

The potential causal role of the upregulation of CD4 and CD8 subpopulation cells in inherited protection from asthma may clear the molecular basis of asthma and may have implications for asthma therapy.

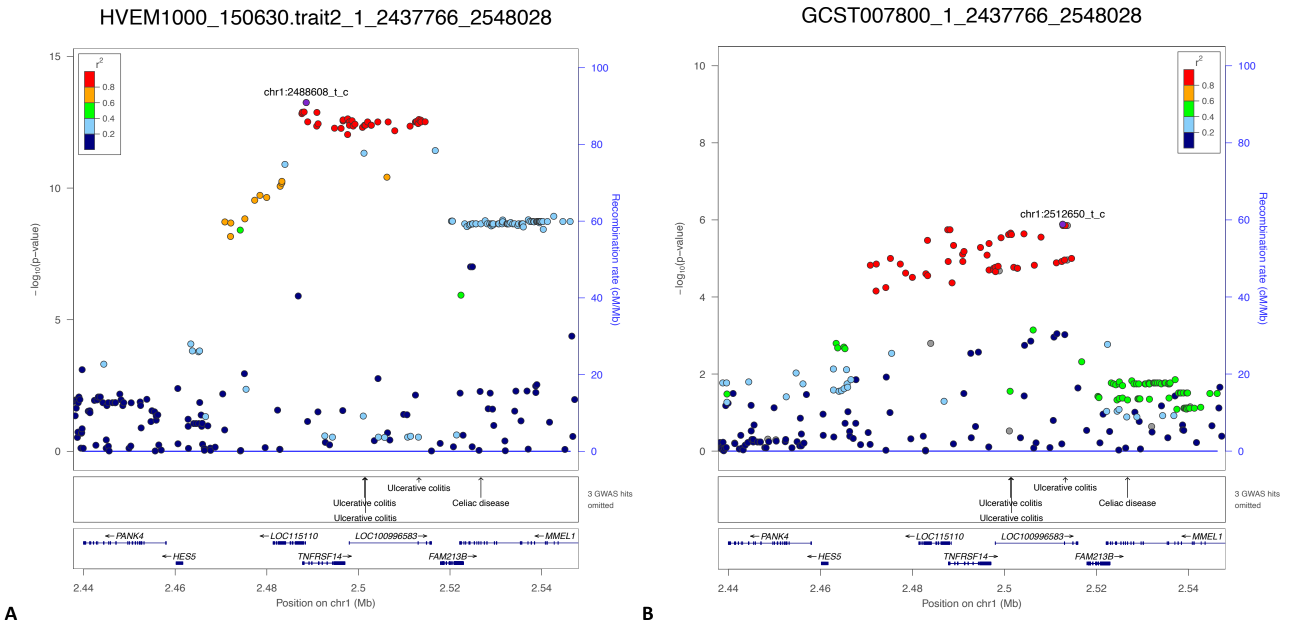


Fig. 7 | Regional association plots in the *TNFRSF14* region. The significance of the association (−log10[P value]; left y-axis) for each trait is plotted relative to the genomic positions on the hg19/GRCh37 genomic build (x-axis). SNPs are colored to reflect their LD (left from SardiNIA map, right from UK10K map) with the most significant variant (indicated with a purple dot). A, Expression of HVEM on CD4+ cells. The P values were obtained using a linear mixed association model. B, Association profiles for the autoimmune disease (asthma childhood-onset). The plots were drawn using the standalone version of LocusZoom[83].

Discussion and Conclusions

The aim of my research project was to identify datasets, tools and approaches that enable the most effective analysis of genetic data to identify causal and/or predisposing variants of Mendelian diseases and complex traits.

With this aim, I tried to maximize the availability of genetic data from public databases and to use at best the data of some projects, as the SardiNIA study, conducted at the Institute of Genetic and Biomedical Research of the CNR. At the same time, I carried out a critical evaluation of the most advanced computer instruments to optimize the approaches applied to the pathologies under study.

The data collection from public datasets is not always optimal due to differences in the generation and classification of genetic data, and the completeness of genetic and clinical information that represent a prerequisite for an appropriate analysis. In fact, a large effort is needed to harmonize the different datasets to make them comparable to include an adequate number of subjects necessary to achieve sufficient statistical power.

Compared to public datasets, the SardiNIA cohort guarantees a much more exhaustive and detailed investigation thanks to the millions of high-resolution sequenced markers and the accuracy of data cataloguing. The availability of the genotypes of each individual within the cohort has made possible the screening for multiple phenotypes.

The analytical investigations carried out achieved statistically reliable and highly reproducible results using the most suitable approaches for the analysis requirements. The systematic approach used covered numerous areas: detailed data management and processing, comparison and application of algorithms and statistical software, fine-tuning and development of sequential bioinformatic workflows and molecular characterization of morphological and functional characteristics of individuals.

When focusing on rare diseases, the collaboration with hospital centers and the involvement of specialized physicians for the screening of age-related diseases in the SardiNIA cohort, allowed the identification of several patients affected by rare diseases and their rapid and efficient molecular diagnosis. By using our approach, we were able to describe the first molecular causative variant of USH syndrome, its incidence and distribution in Sardinia. A sequential, semi-automated computational procedure was applied, leading us from sequence data to the discovery of causal variants. We verified and experimentally validated the results obtained through genomic and functional studies to give them a precise biological meaning. Our results reveal that this approach represents an effective and generalizable method for finding causal variants. Our data provide a feasible perspective for screening carrier and/or affected status and the possibility of targeted genetic counselling, in Sardinia, that can improve prevention and early treatment strategy of this inherited syndrome.

Focusing on complex traits, our investigation of tools and pipelines providing statistically reliable and highly reproducible results identified colocalization analysis as one of the most valuable and innovative methods, providing a high percentage of replicated results, confirming uncertain association signals, and identifying novel coincident quantitative trait-disease associations. The existence of different colocalization software allowed a direct comparison to verify their congruence and ascertain, and if necessary, implement the best strategy to be adopted.

One of the issues to be addressed is the need to integrate multiple databases and standardize the data to achieve a robust analysis approach. Integration of public datasets needs to take into account a number of limitations: lack of harmonization, different ethnicities, low sample size and genetic maps obtained using microarrays.

In fact, most studies included in the GWAS Catalog provided harmonized results only recently, so previous colocalization studies are based on non-harmonized data.

In addition, for some studies deriving from meta-analyses (i.e. combined GWAS on different studies), several datasets are available, some including individuals of different ethnicities, others carried out on a single population. The ethnicity of the individuals constituting the disease dataset is an important parameter to evaluate, because of the different linkage disequilibrium patterns that could affect association results. The advantage of using a multi-ethnic dataset is that, in some regions that are shared between different populations, an increase in sample size given by the grouping of individuals of different ethnicities results in better signal definition, greater statistical power and higher significance. By contrast, a dataset consisting only of individuals from a given population, even if it has a smaller sample size, may favor the identification of population-specific signals that would be lost in the multi-ethnic dataset.

Finally, certain public GWAS datasets were produced using genotyping arrays and the number of variants was insufficient to provide a solid result.

Despite these considerations, this study showed the tools' ability to identify variants shared among the association profiles. The high correlation between the results obtained here with respect to published studies indicates a high reliability of the algorithms used. Although the software used different parameters, they show high concordance outcomes among them. The results replicated by the two tools (methodological replication) and on different datasets with different sample sizes and thus different statistic power (independent replication) are evidence of consistency. The appropriate choice of the tool should be based on the dataset features (es. population ethnicity and sample size) and the data available (MAF, P-value, Z-score, Beta score, LD). Our findings confirm that colocalization is an appropriate approach to verify if different association signals overlap and, thus, share the same causal variant(s), in support of other methodologies based for example on linkage disequilibrium patterns only (fine-mapping, Mendelian randomization). We can assert the colocalization is a robust approach to investigate connections between QTs and diseases and can enhance our understanding of biological processes.

Future directions will be evaluating new findings to validate biological plausibility. Recently, a new multiple-trait-colocalization approach ("*moloc*")[84] has been developed to identify the regulatory effects of risk GWAS loci. This method can be applied to functionally relevant data to help identify disease-associated genes. The presence of multiple coincident associations, a signal coincident with that trait and the disease at multiple loci on the genome, will aid to support the role of a specific trait in a disease. Based on association results, the next step may be a systematic results assessment using Mendelian randomization to have evidence of causality and to confirm if Sardinian immune QTs are causal for the disease. A future step of the project will be a multiple trait colocalization to identify shared association profiles between QT and disease at multiple genome locations. Other goals could be to perform colocalization analyses integrating QT traits available in the GWAS catalog by prioritizing the relevant traits. The integration of data from different databases and the biological annotation from a functional perspective will make the colocalization approach increasingly robust and will advance the understanding of biological mechanisms of disease susceptibility.

References

1. Human Genome Sequencing Consortium, I. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

2. McCarthy, J. J., McLeod, H. L. & Ginsburg, G. S. Genomic medicine: a decade of successes, challenges, and opportunities. *Sci. Transl. Med.* **5**, 189sr4 (2013).

3. Kanzi, A. M. *et al.* Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. *Frontiers in Genetics* **11**, (2020).

4. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, (2016).

5. Genome-Wide Association Studies Fact Sheet - National Human Genome Research Institute (NHGRI). Available at: https://www.genome.gov/20019523/. (Accessed: 3rd September 2018)

6. Bedia, C. Experimental Approaches in Omic Sciences. in *Comprehensive Analytical Chemistry* **82**, (2018).

7. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics* **12**, (2011).

8. Feero, W. G., Guttmacher, A. E. & Manolio, T. A. *Genomic Medicine Genomewide Association Studies and Assessment of the Risk of Disease*. *N Engl J Med* **363**, (2010).

9. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).

10. Fritsche, L. G. *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.* **48**, 134–43 (2016).

11. Wellcome Trust Case Control Consortium, T. W. T. C. C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–78 (2007).

12. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).

13. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, (2019).

14. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics* **101**, (2017).

15. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics* **11**, (2020).

16. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20**, (2019).

17. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* **10**, (2009).

18. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Human Molecular Genetics* **24**, (2015).

19. Zhu, Y., Tazearslan, C. & Suh, Y. Challenges and progress in interpretation of non-coding genetic variants associated with human disease. *Experimental Biology and Medicine* **242**, (2017).

20. Miller, J. E., Veturi, Y. & Ritchie, M. D. Innovative strategies for annotating the 'relationSNP' between variants and molecular phenotypes. *BioData Mining* **12**, (2019).

21. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science (80-.).* **369**, (2020).

22. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, (2010).

23. Nica, A. C. *et al.* Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. *PLoS Genet.* **6**, e1000895 (2010).

24. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* **10**, (2009).

25. Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, (2013).

26. Smith, G. D. & Ebrahim, S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32**, (2003).

27. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Human Molecular Genetics* **24**, (2015).

28. Plagnol, V., Smyth, D. J., Todd, J. A. & Clayton, D. G. Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics* **10**, (2009).

29. Wallace, C. *et al.* Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Hum. Mol. Genet.* **21**, (2012).

30. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *J. Natl. Cancer Inst.* **96**, (2004).

31. Wakefield, J. A bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, (2007).

32. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).

33. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).

34. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

35. The SardiNIA Project - NIH. Available at: https://sardinia.nia.nih.gov/. (Accessed: 7th September 2018)

36. Lampis, R., Morelli, L., De Virgiliis, S., Congia, M. & Cucca, F. The distribution of HLA class II

haplotypes reveals that the Sardinian population is genetically differentiated from the other Caucasian populations. *Tissue Antigens* **56**, (2000).

37. Passarino, G. *et al.* Y chromosome binary markers to study the high prevalence of males in Sardinian centenarians and the genetic structure of the Sardinian population. *Hum. Hered.* **52**, (2001).

38. Orrù, V. *et al.* XGenetic variants regulating immune cell levels in health and disease. *Cell* **155**, (2013).

39. Danjou, F. *et al.* Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. *Nat. Genet.* **47**, (2015).

40. Zoledziewska, M. *et al.* Height-reducing variants and selection for short stature in Sardinia. *Nat. Genet.* **47**, (2015).

41. Pala, M. *et al.* Population- and individual-specific regulatory variation in Sardinia. *Nat. Genet.* **49**, 700–707 (2017).

42. Orrù, V. *et al.* Complex genetic signatures in immune cells underlie autoimmunity and inform therapy. *Nat. Genet.* **52**, (2020).

43. Orlov, N. V. *et al.* Age-related changes of the retinal microvasculature. *PLoS One* **14**, (2019).

44. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.* **47**, 1272–1281 (2015).

45. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).

46. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, (2014).

47. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

48. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, (2010).

49. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, (2012).

50. Kimberling, W. J. *et al.* Frequency of Usher syndrome in two pediatric populations: Implications for genetic screening of deaf and hard of hearing children. *Genet. Med.* **12**, (2010).

51. Toms, M., Pagarkar, W. & Moosajee, M. Usher syndrome: clinical features, molecular genetics and advancing therapeutics. *Ther. Adv. Ophthalmol.* **12**, (2020).

52. Smith, R. J. H. *et al.* Clinical diagnosis of the Usher syndromes. *Am. J. Med. Genet.* **50**, (1994).

53. Pakarinen, L., Karjalainen, S., Simola, K. O. J., Laippala, P. & Kaitalo, H. Usher's syndrome type 3 in finland. *Laryngoscope* **105**, (1995).

54. Hagag, A. M. *et al.* Characterisation of microvascular abnormalities using OCT angiography in patients with biallelic variants in USH2A and MYO7A. *Br. J. Ophthalmol.* **104**, (2020).

55. Leijendeckers, J. M., Pennings, R. J. E., Snik, A. F. M., Bosman, A. J. & Cremers, C. W. R. J. Audiometric characteristics of USH2a patients. *Audiol. Neurotol.* **14**, (2009).

56. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493-6 (2004).

57. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–2 (2010).

58. Sherry, S. T. *et al.* DbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

59. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

60. Clarke, L. *et al.* The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Res.* **45**, (2017).

61. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, (2005).

62. Chiang, C. W. K. *et al.* Genomic history of the Sardinian population. *Nat. Genet.* **50**, (2018).

63. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, (2009).

64. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, (2015).

65. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science (80-.).* **296**, (2002).

66. Zhang, Y. I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins Struct. Funct. Bioinforma.* **77**, (2009).

67. Pires, D. E. V., Ascher, D. B. & Blundell, T. L. DUET: A server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.* **42**, (2014).

68. Parthiban, V., Gromiha, M. M., Abhinandan, M. & Schomburg, D. Computational modeling of protein mutant stability: Analysis and optimization of statistical potentials and structural features reveal insights into prediction model development. *BMC Struct. Biol.* **7**, (2007).

69. Cheng, J., Randall, A. & Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins Struct. Funct. Genet.* **62**, (2006).

70. Pilia, G. *et al.* Heritability of Cardiovascular and Personality Traits in 6,148 Sardinians. *PLoS Genet.* **2**, e132 (2006).

71. McClellan, J. & King, M. C. Genetic heterogeneity in human disease. *Cell* **141**, (2010).

72. Zondervan, K. T. & Cardon, L. R. Designing candidate gene and genome-wide case-control association studies. *Nat. Protoc.* **2**, (2007).

73. Li, M., Boehnke, M. & Abecasis, G. R. Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *Am. J. Hum. Genet.* **78**, (2006).

74. Ghoussaini, M. *et al.* Open Targets Genetics: Systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, (2021).

75. Open Targets Genetics. Data harmonisation and aggregation process. Available at: https://genetics-

docs.opentargets.org/our-approach/data-pipeline.

76. Open Targets Genetics. Summary Statistics harmonisation. Available at: https://github.com/EBISPOT/sum-stats-formatter/tree/master/harmonisation.

77. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, (2014).

78. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, (2014).

79. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, (2015).

80. Navarro Gonzalez, J. *et al.* The UCSC genome browser database: 2021 update. *Nucleic Acids Res.* **49**, (2021).

81. Ferreira, M. A. R. *et al.* Genetic Architectures of Childhood- and Adult-Onset Asthma Are Partly Distinct. *Am. J. Hum. Genet.* **104**, (2019).

82. Stelzer, G. *et al.* The GeneCards suite: From gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinforma.* **2016**, 1.30.1-1.30.33 (2016).

83. Pruim, R. J. *et al.* LocusZoom: Regional visualization of genome-wide association scan results. in *Bioinformatics* **27**, (2011).

84. Giambartolomei, C. *et al.* A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538–2545 (2018).

Web resources

UCSC, https://genome.ucsc.edu/

Ensembl, https://www.ensembl.org/index.html

dbSNP, https://www.ncbi.nlm.nih.gov/projects/SNP/

1000 Genomes Project, http://www.internationalgenome.org/

GWAS Catalog, https://www.ebi.ac.uk/gwas/

BCFtools, https://samtools.github.io/bcftools/bcftools.html

VEP, https://grch37.ensembl.org/Homo_sapiens/Tools/VEP

Bedtools, https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html

GeneCards, https://www.genecards.org

gnomAD, https://gnomad.broadinstitute.org

Supplementary

Table 1. 44 public association studies from GWAS catalog

Study_accession	Reported_trait
GCST000679	Rheumatoid arthritis
GCST000964	Ulcerative colitis
GCST004988	Breast cancer
GCST001198	Multiple sclerosis
GCST001475	Obesity
GCST001592	Osteoarthritis
GCST002245	Alzheimer's disease (late onset)
GCST003043	Inflammatory bowel disease
GCST003044	Crohn's disease
GCST003045	Ulcerative colitis
GCST003156	Systemic lupus erythematosus
GCST004131	Inflammatory bowel disease
GCST004132	Crohn's disease
GCST004133	Ulcerative colitis
GCST004415	Invasive epithelial ovarian cancer
GCST004478	Serous invasive ovarian cancer
GCST004695	Insomnia complaints
GCST005038	Allergic disease (asthma, hay fever or eczema)
GCST005047	Type 2 diabetes
GCST005527	Psoriasis
GCST005531	Multiple sclerosis
GCST005569	Rheumatoid arthritis
GCST005647	Amyotrophic lateral sclerosis
GCST005814	Osteoarthritis (hospital diagnosed)
GCST006085	Prostate cancer
GCST006100	Strenuous sports or other exercises
GCST006414	Atrial fibrillation

GCST006464	Endometrial cancer
GCST006810	Self-reported risk-taking behaviour
GCST006862	Asthma
GCST006906	Stroke
GCST006908	Ischemic stroke
GCST006909	Ischemic stroke (small-vessel)
GCST006910	Ischemic stroke (cardioembolic)
GCST007090	Knee osteoarthritis
GCST007091	Osteoarthritis (hip)
GCST007799	Asthma (adult onset)
GCST007800	Asthma (childhood onset)
GCST009541	Heart failure
GCST009979	Major depressive disorder
GCST009982	Trauma exposure
GCST010681	Type 1 diabetes
GCST90007526	Low hand grip strength (60 years and older) (EWGSOP)
GCST90012857	Falling risk